Teaching Introductory Students How to Evaluate Evidence

Kari Lock Morgan
Department of Statistics
Penn State University

USCOTS
May 18th, 2019

Teaching Introductory Students

+ stat ed community
(YOU!!)

How to Evaluate ^Causal Evidence

+ causal inference community

Everything an expert should know about evaluating evidence

What an intro student should know about evaluating evidence

Evaluating Evidence

- Suppose we are comparing A vs B
- In our sample, the A group has better outcomes than the B group
- Possible explanations?
  1) A causes better outcomes than B
  2) the groups differed at baseline
  3) just random chance

Evaluating evidence for (1) requires evaluating evidence *against* (2) and (3)

Yeah, yeah…


"Yeah, yeah… obviously I already cover confounding and inference in intro stat."

GOALS National Post-Test Data

Related to confounding:

Learning Objective	% Correct
Able to reason about the purpose of random assignment	26.9%
Able to reason about how correlation does not imply causation	22.5%

Significantly worse than random guessing

Thanks to Bob DelMas for the GOALS data!

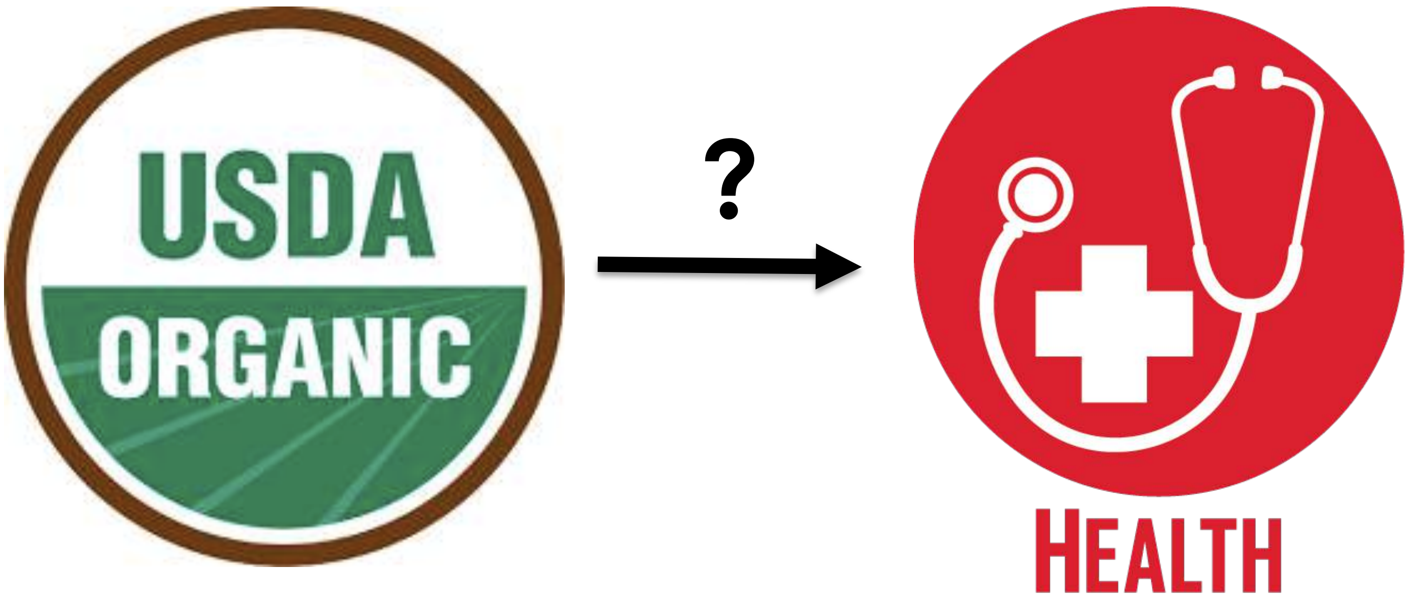
GOALS reference: Sabbag, A. G. & Zieffler, A. (2015). "Assessing Learning Outcomes: An Analysis of the GOALS-2 Instrument," *Statistics Education Research Journal (SERJ)*, **14**(2), 92-116.

GOALS National Post-Test Data

Related to p-values:

Learning Objective	**% Correct**
Able to reason that a smaller *p*-value provides stronger evidence against the null hypothesis than a larger *p*-value.	45.2%
~~Able to reason about a conclusion based on a statistically significant *p*-value in the context of a research study that compares two groups~~	~~58.3%~~
Able to reason about an incorrect interpretation of a *p*-value (probability of a treatment being more effective).	50%

Question of the Day



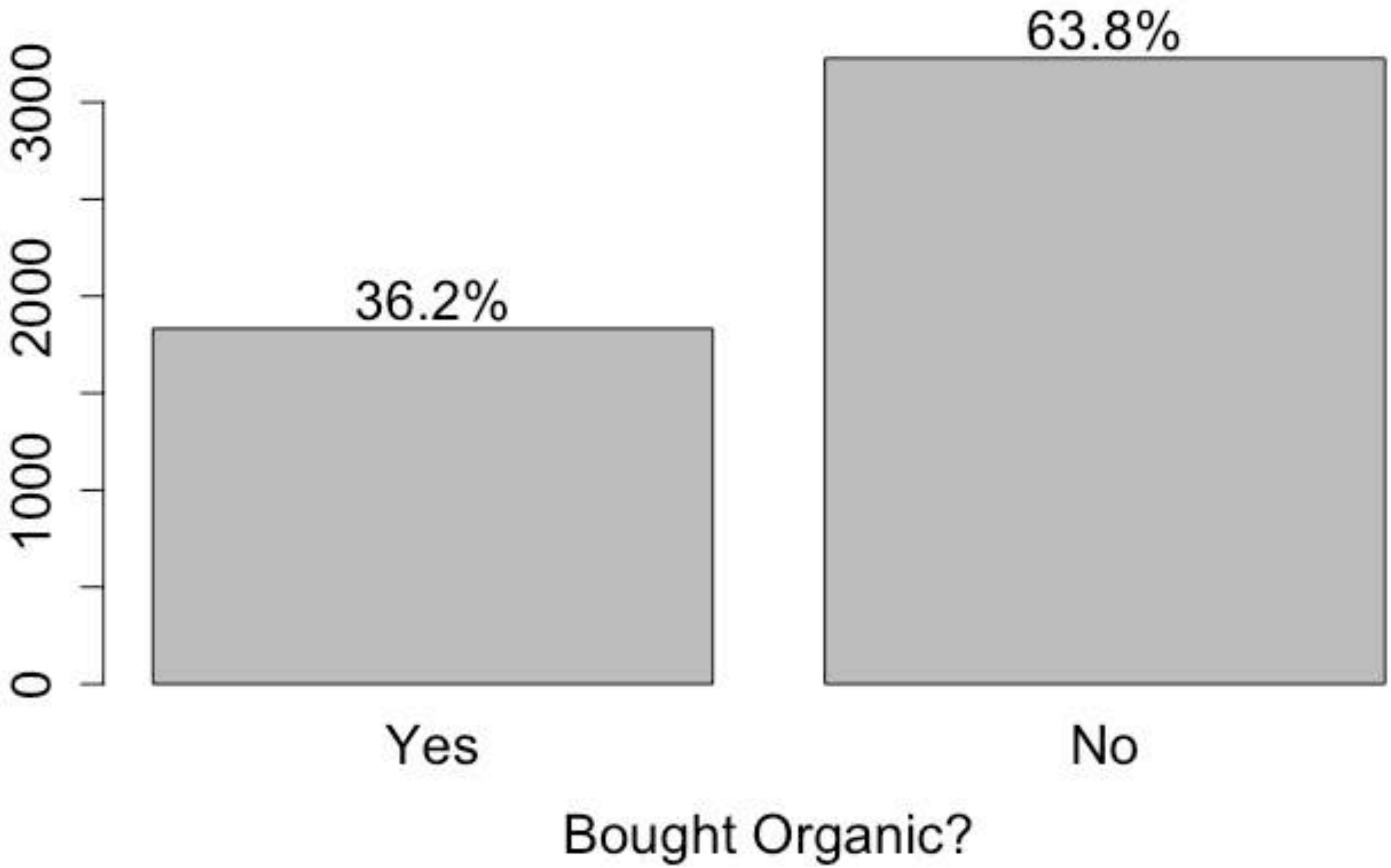
Does eating organic improve your health?

Let's evaluate the evidence!

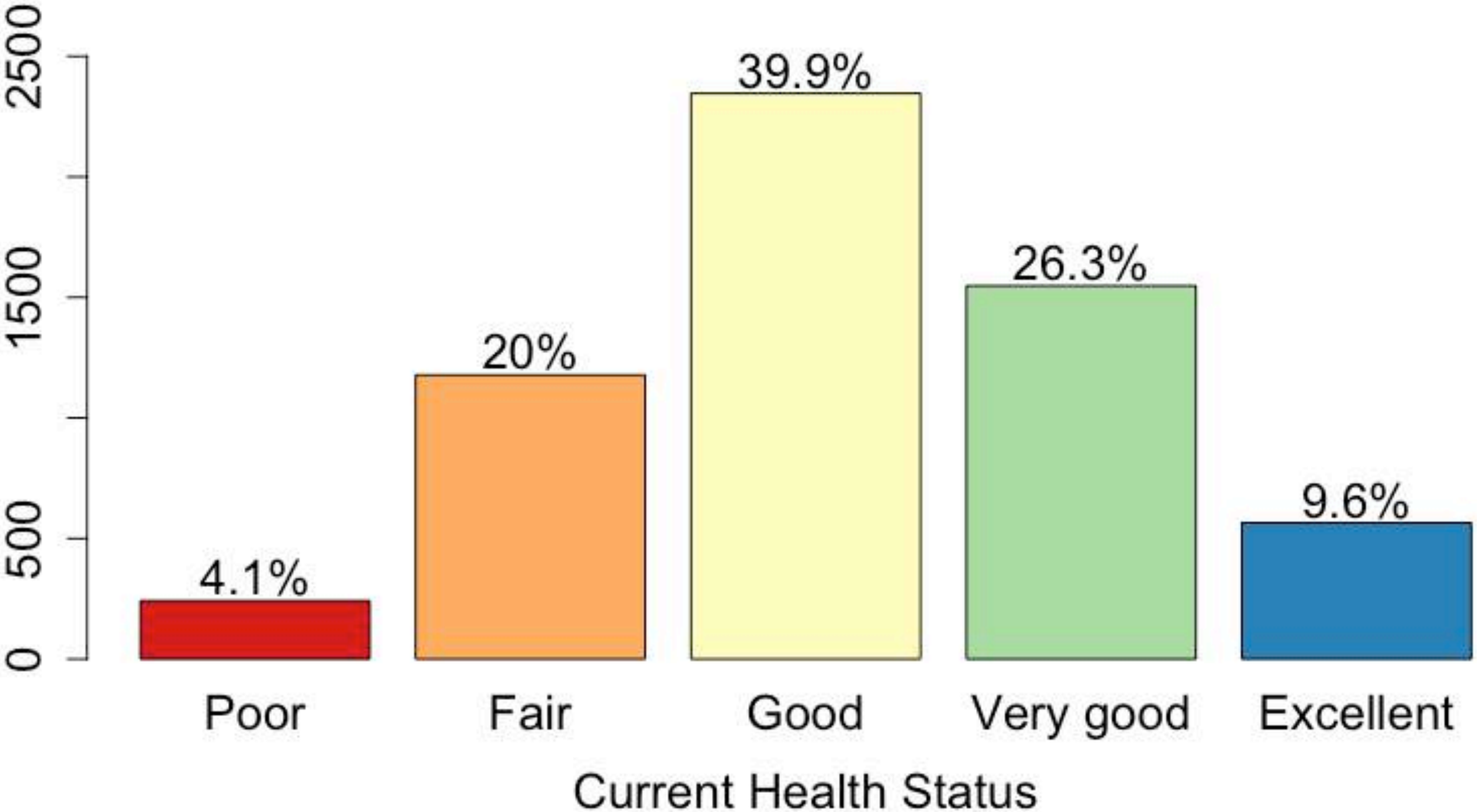
Dataset #1: NHANES

- NHANES: National Health and Nutrition Examination Survey
- Large national random sample
- 2009 – 2010 data
- n = 5060

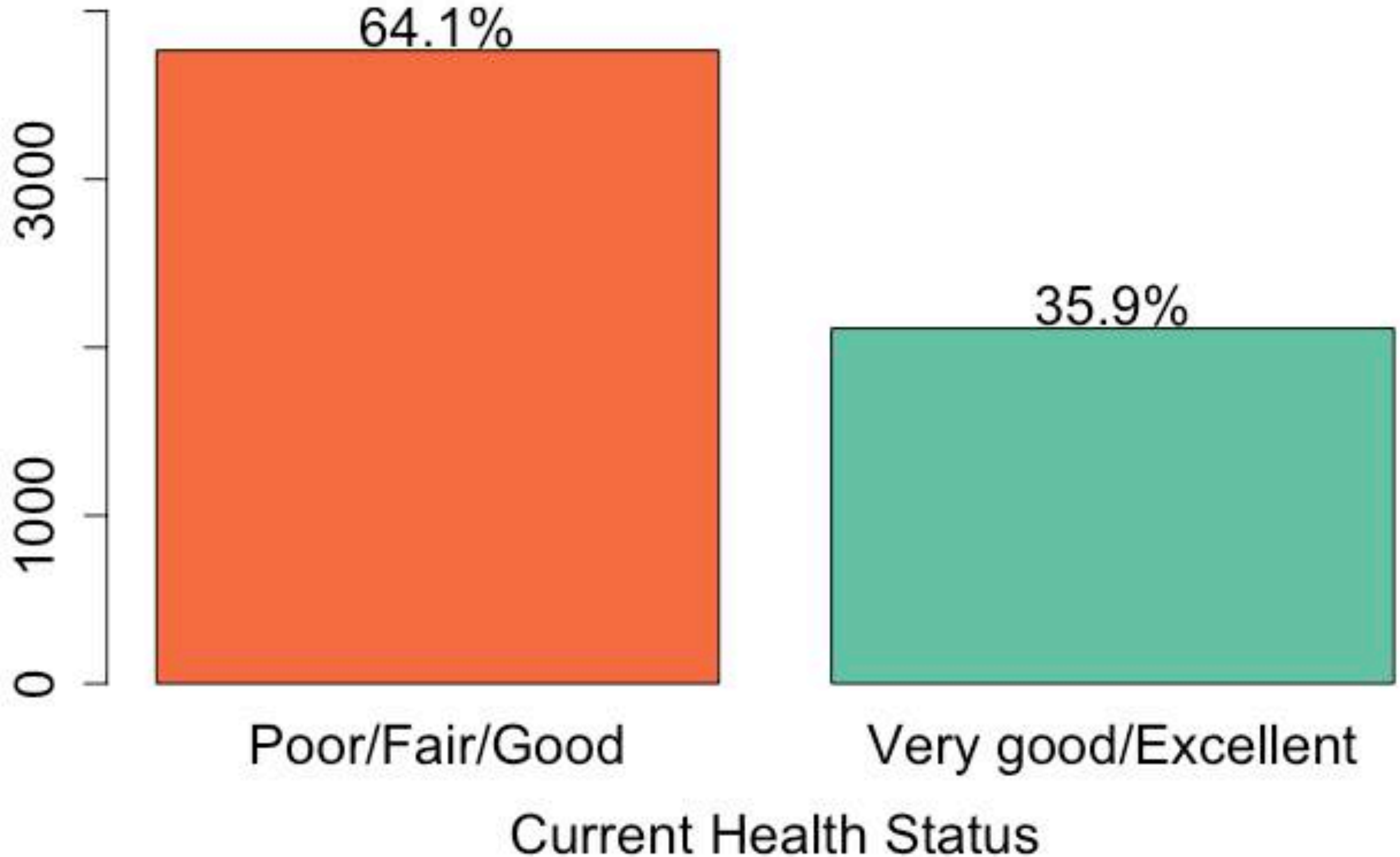
"In the past 30 days, did you buy any food that had the word 'organic' on the package?"



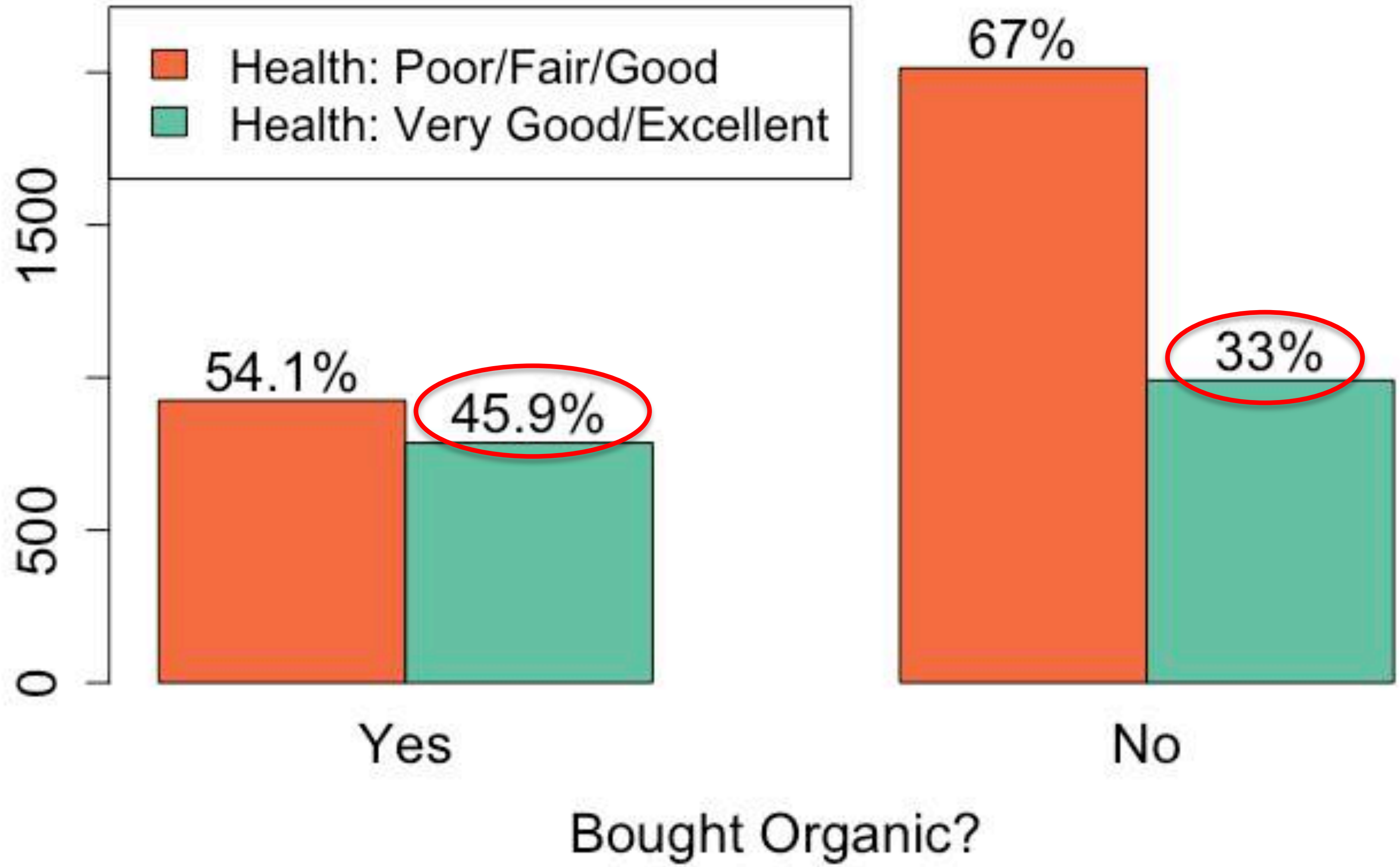
"Would you say your health in general is Excellent, Very good, Good, Fair, or Poor?"



Current Health Status



Health by Organic



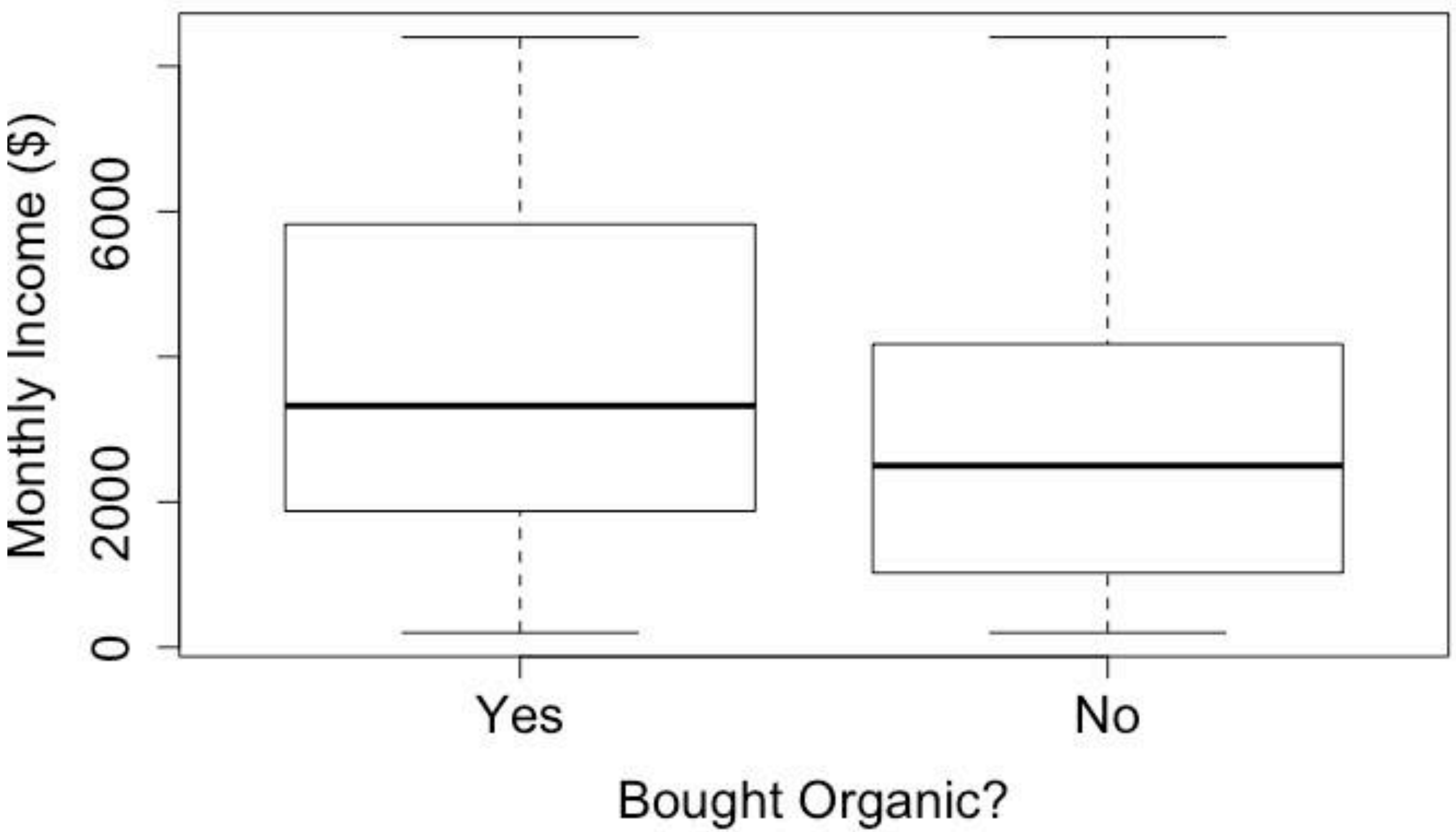
$$\hat{p}_{organic} - \hat{p}_{not\ organic} = 0.459 - 0.330 = 0.129$$

Evaluating Evidence

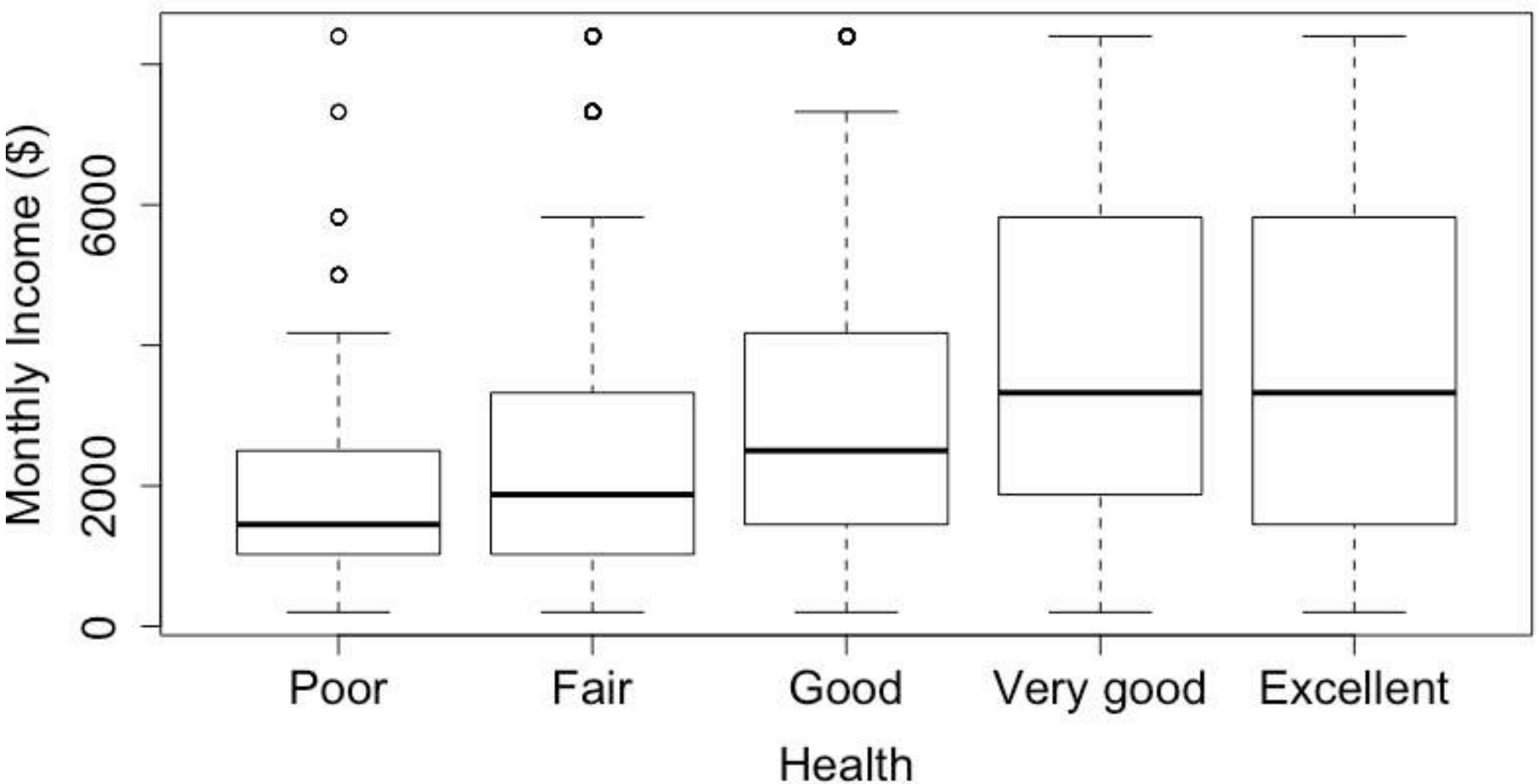
- In our sample, people who bought organic are healthier
- Possible explanations?
  1) eating organic improves health
  2) the groups differed at baseline ???
  3) ~~just random chance~~

p-value < 0.0000000000000002
(lowest p-value R will give)

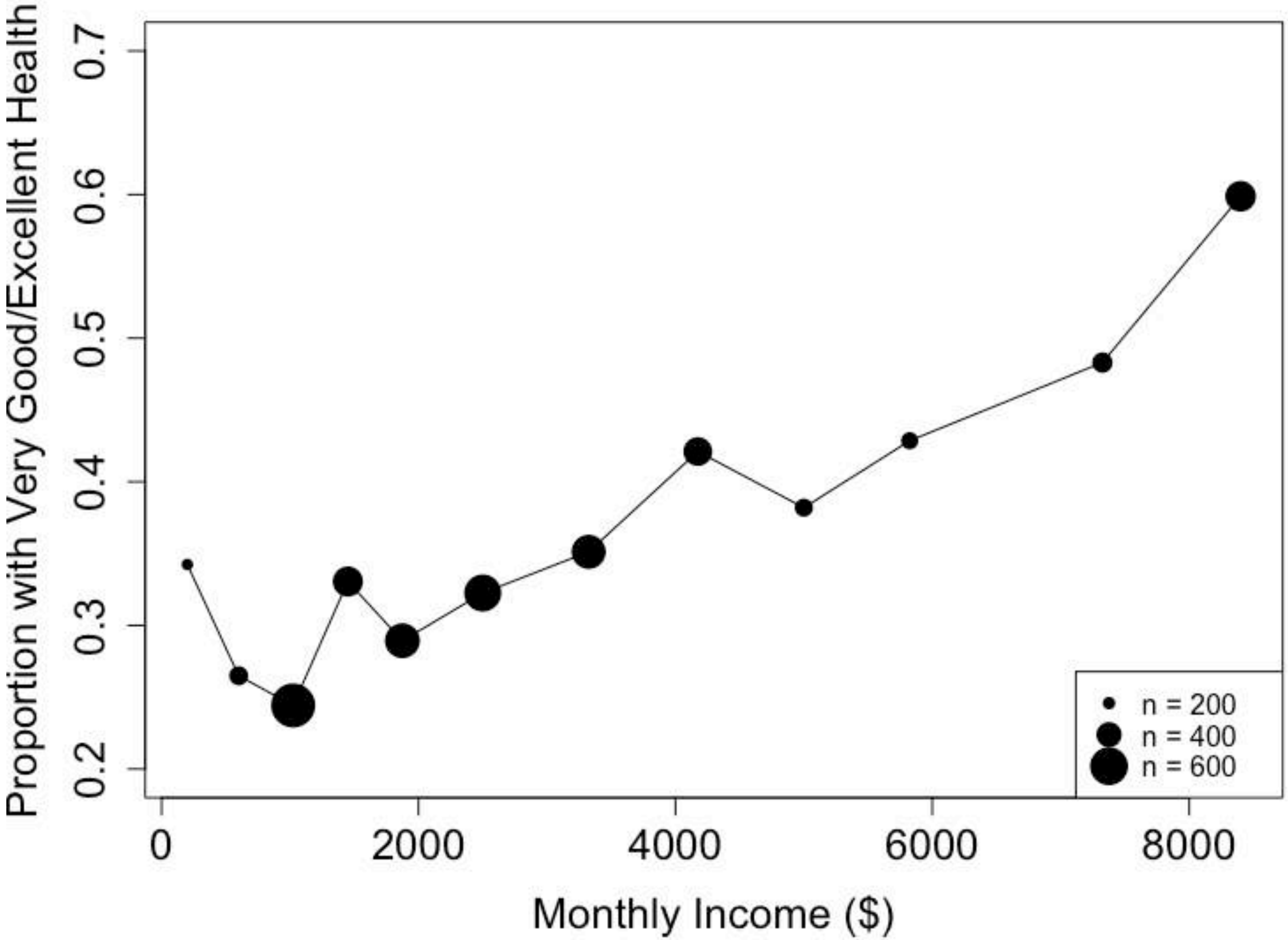
People who buy organic are richer

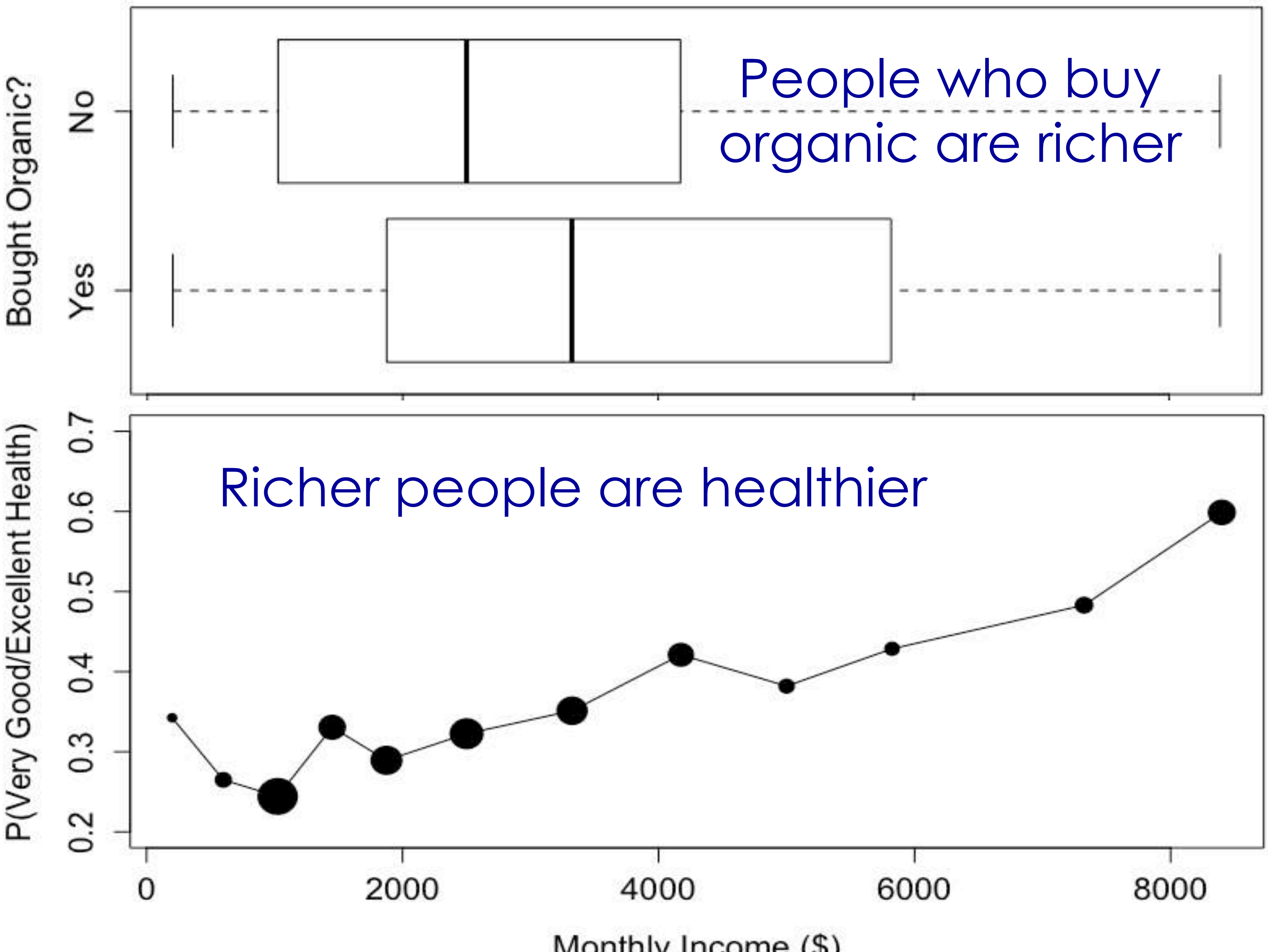


Richer people are healthier



Richer people are healthier





People who buy organic are richer
Richer people are healthier
Bought Organic?
No
Yes
P(Very Good/Excellent Health)
0.2
0.3
0.4
0.5
0.6
0.7
0
2000
4000
6000
8000
Monthly Income ($)

Three "-ations"

VISUALIZATION

"Pictures speak louder than words"

Multivariable thinking!

Evaluating Evidence

- In our sample, people who bought organic are healthier
- Possible explanations?
  1) eating organic improves health
  2) the groups differed at baseline
  3) ~~just random chance~~

With more than one possible explanation, we cannot determine causal evidence!

But we also can't rule it out!

Non-comparable groups

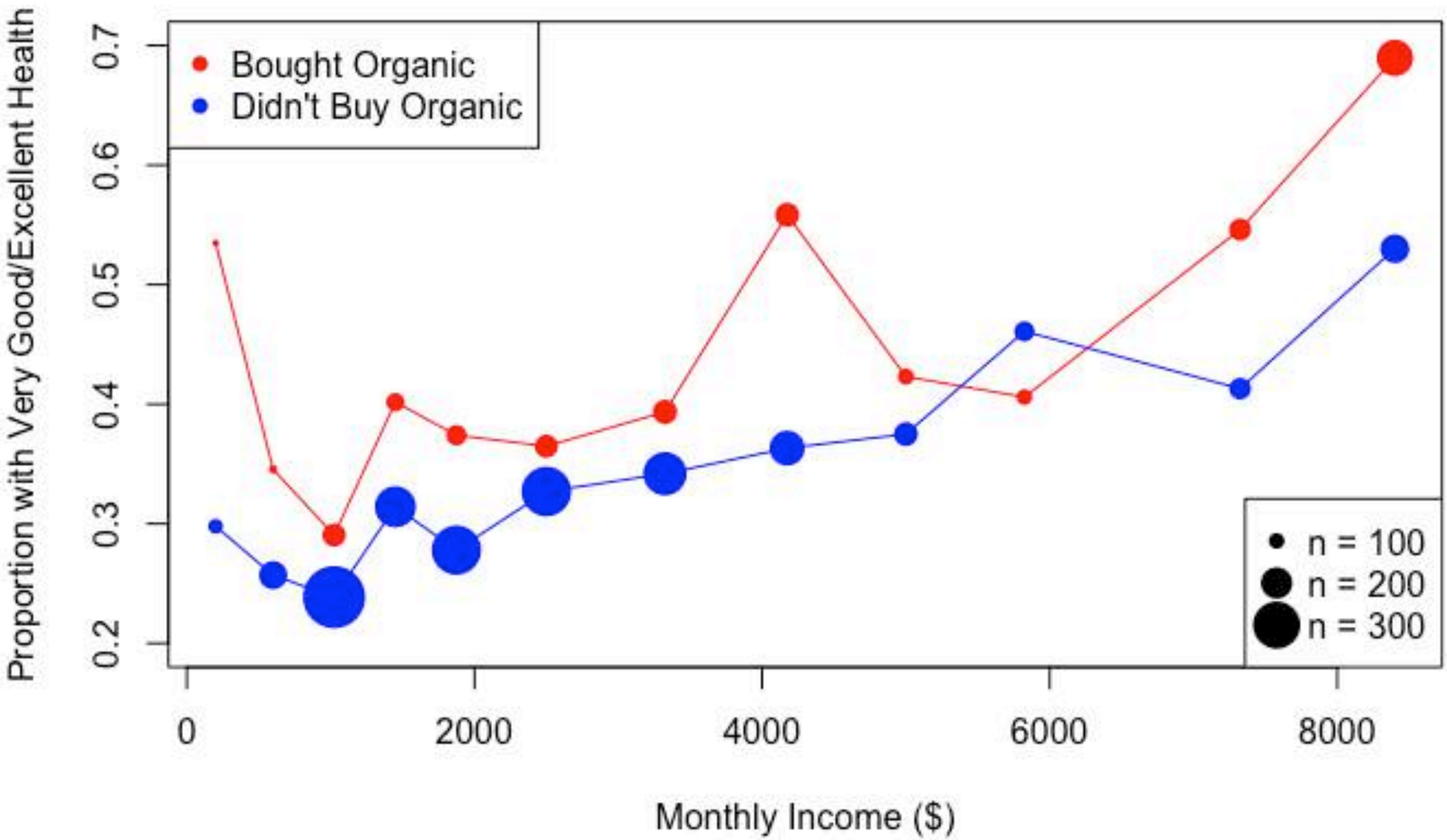
Directly comparing groups that are not comparable (groups differ at baseline) cannot yield causal evidence!

(and can be very misleading)

What can we do if groups differ at baseline?

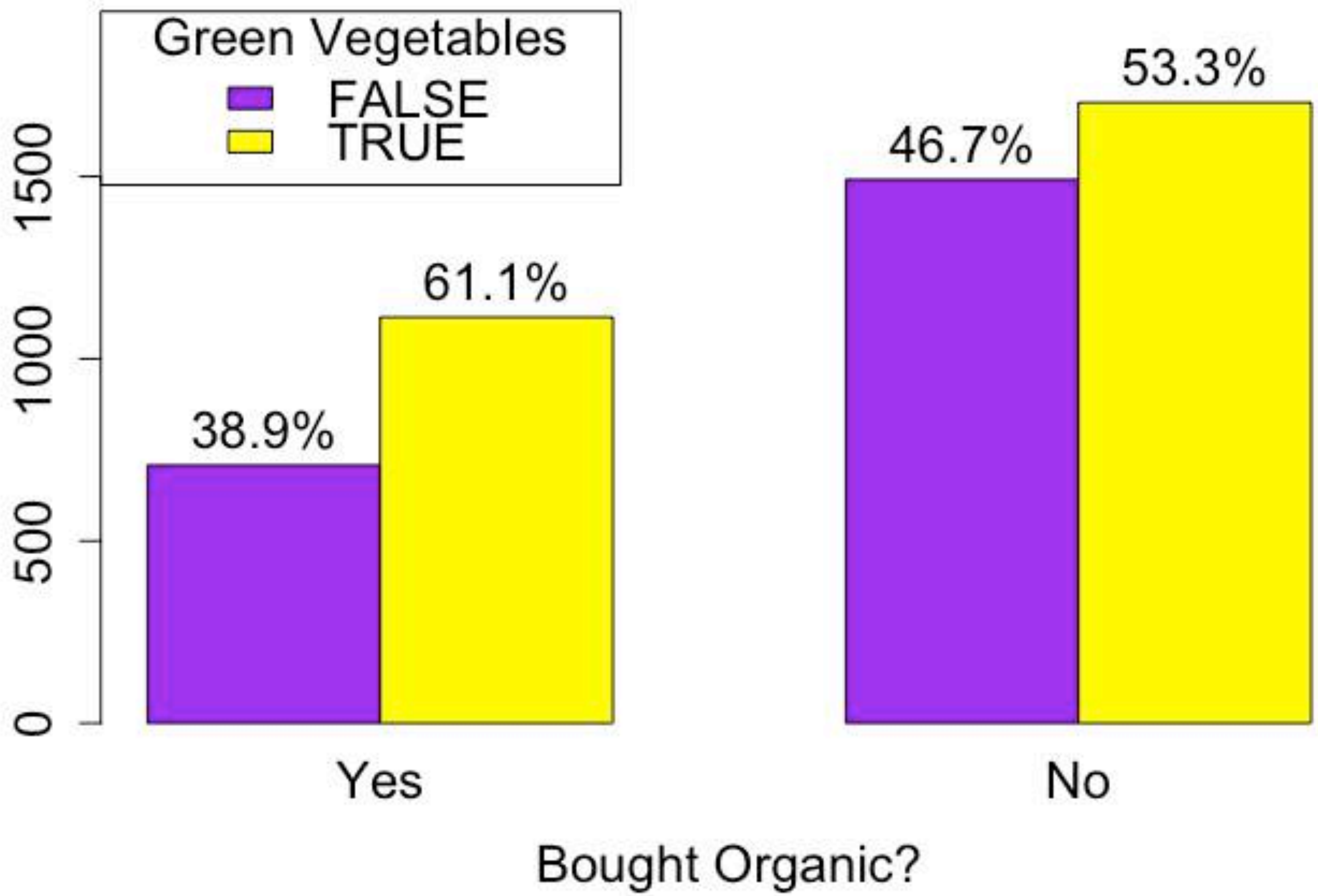
Look within similar groups

Health by Organic, by Income

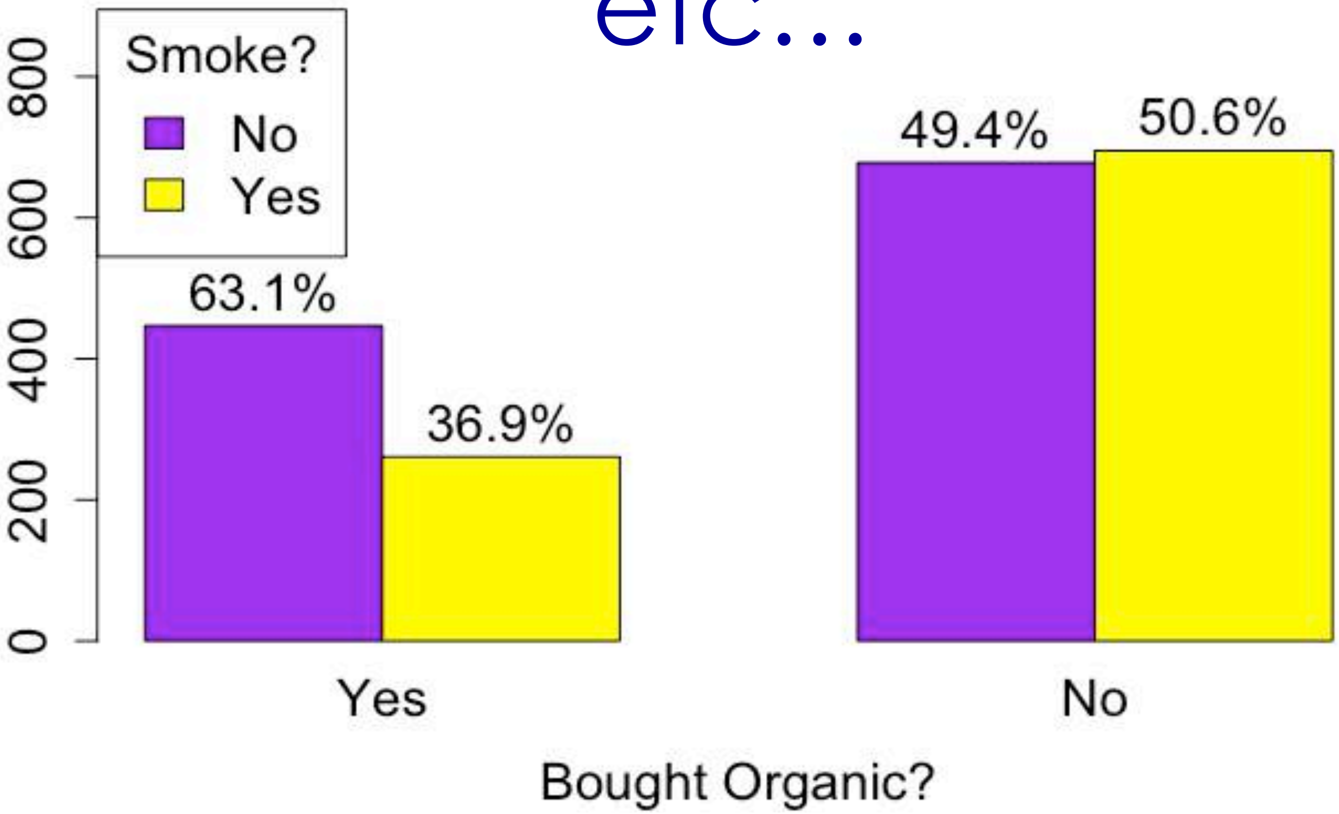


Evidence that eating organic makes you healthier?

People who buy organic are more likely to have green vegetables



People who buy organic are less likely to smoke etc…



Differences on many measured variables…

… ***and countless unmeasured variables!***

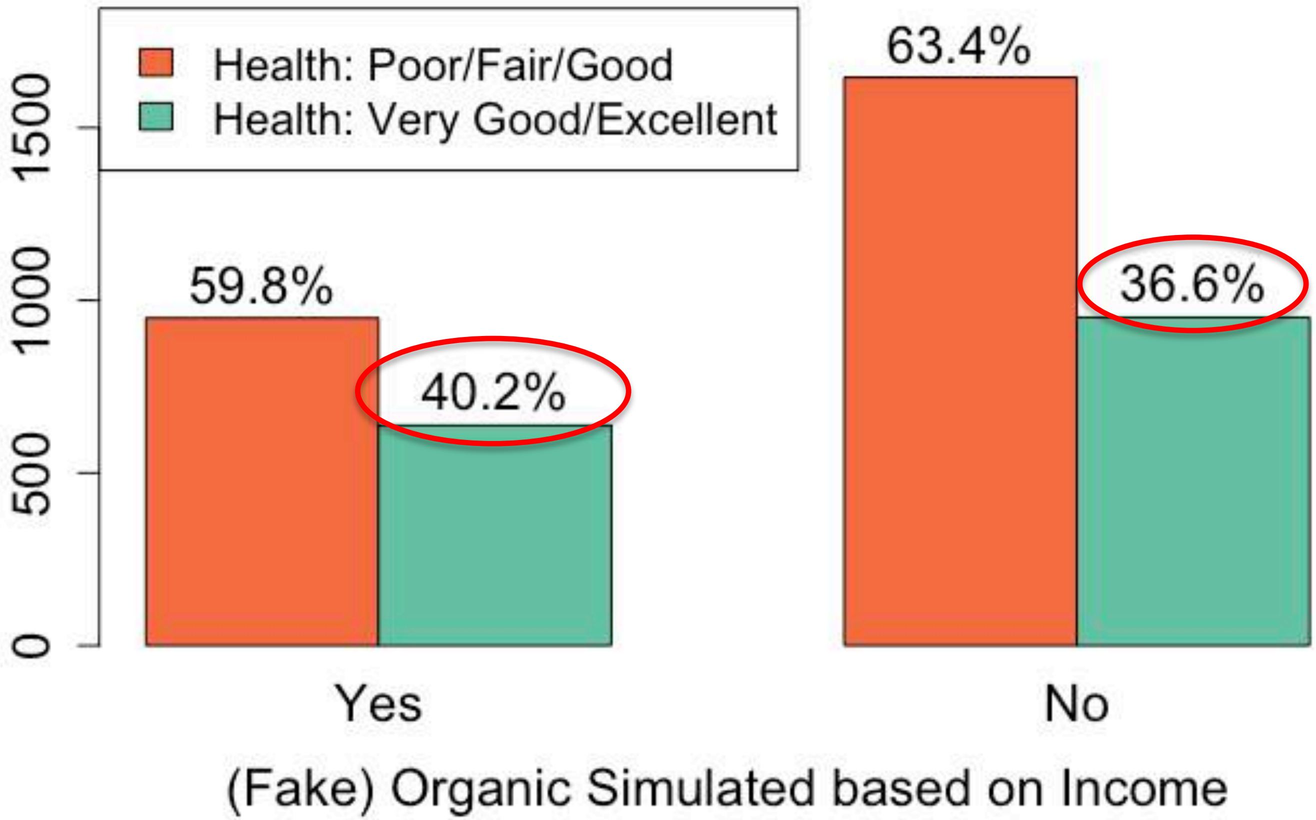
Multiple Examples

Ideally, students will see multiple examples illustrating the dangers of comparing non-comparable groups, including…

- Common sense baseline difference(s)
- Baseline difference shifts effect
- Baseline difference reverses effect
- Baseline difference masks true effect
- Baseline difference creates false effect…

Simulate "organic" based ***only*** on income

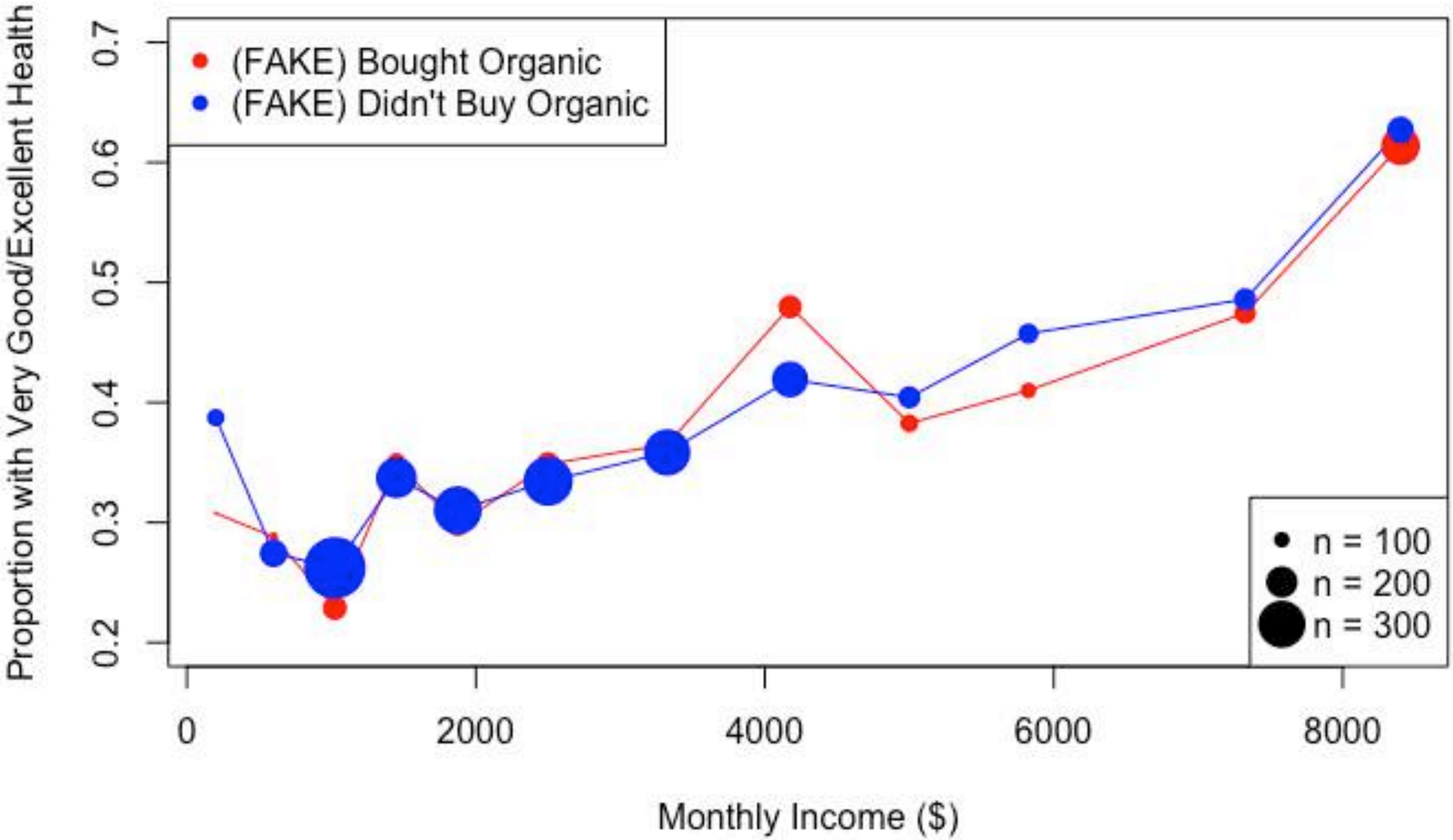
*… so it has **no real causal effect** on health!*



p-value = 0.02

(solely due to baseline difference in income!)

Fake Organic by Health by Income



Randomization

- Simply observing data as it is will almost always result in baseline differences
- How to alleviate this problem?

RANDOMIZE treatment assignment!

$\Rightarrow$ Baseline differences should be minimal (just due to random chance)

$\Rightarrow$ Allows for causal evidence!!!

Three "-ations"

VISUALIZATION

"Pictures speak louder than words"

Multivariable thinking!

RANDOMIZATION

Allows for causal evidence!

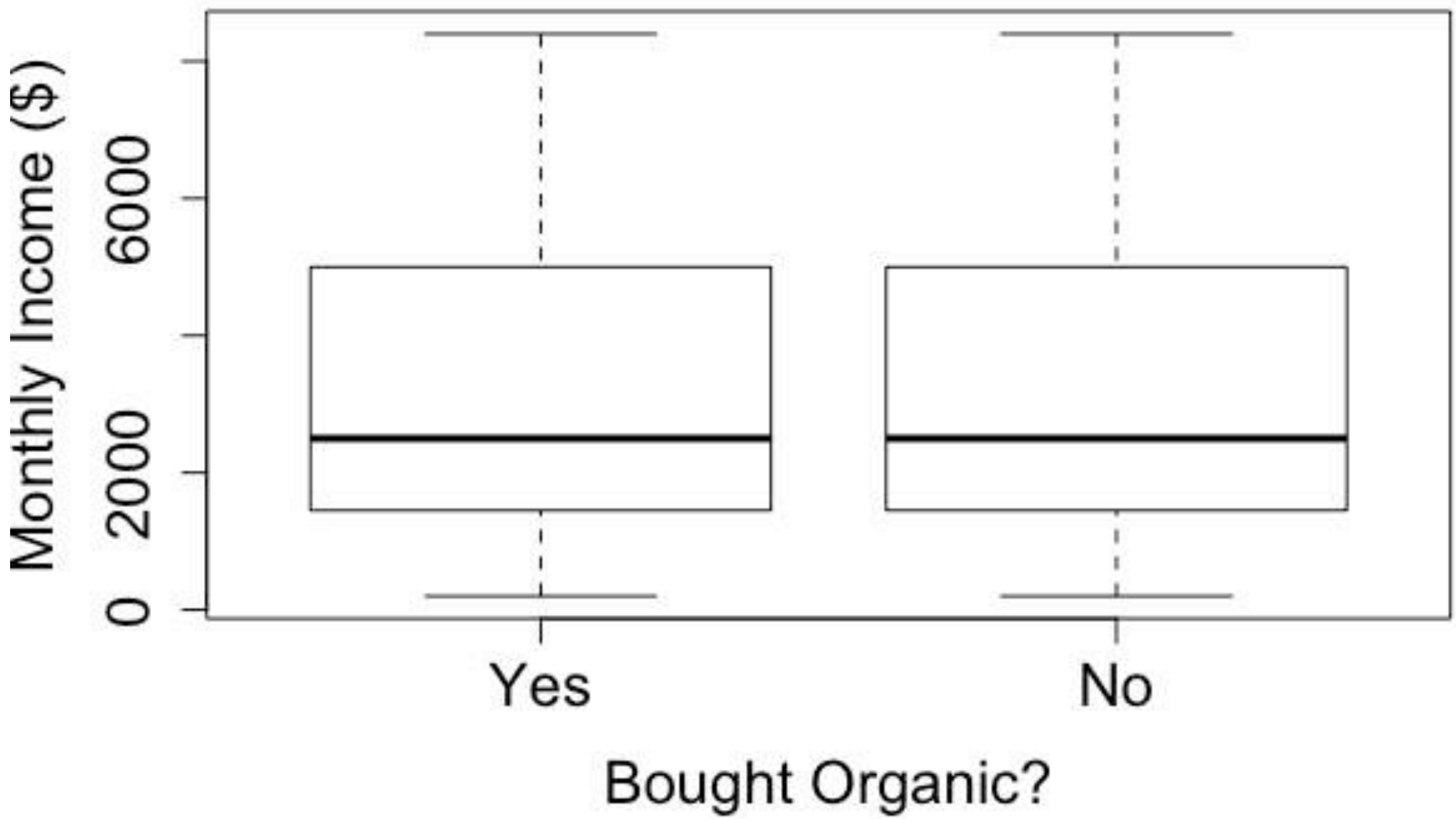
Foundation for inference

Simulate a randomization

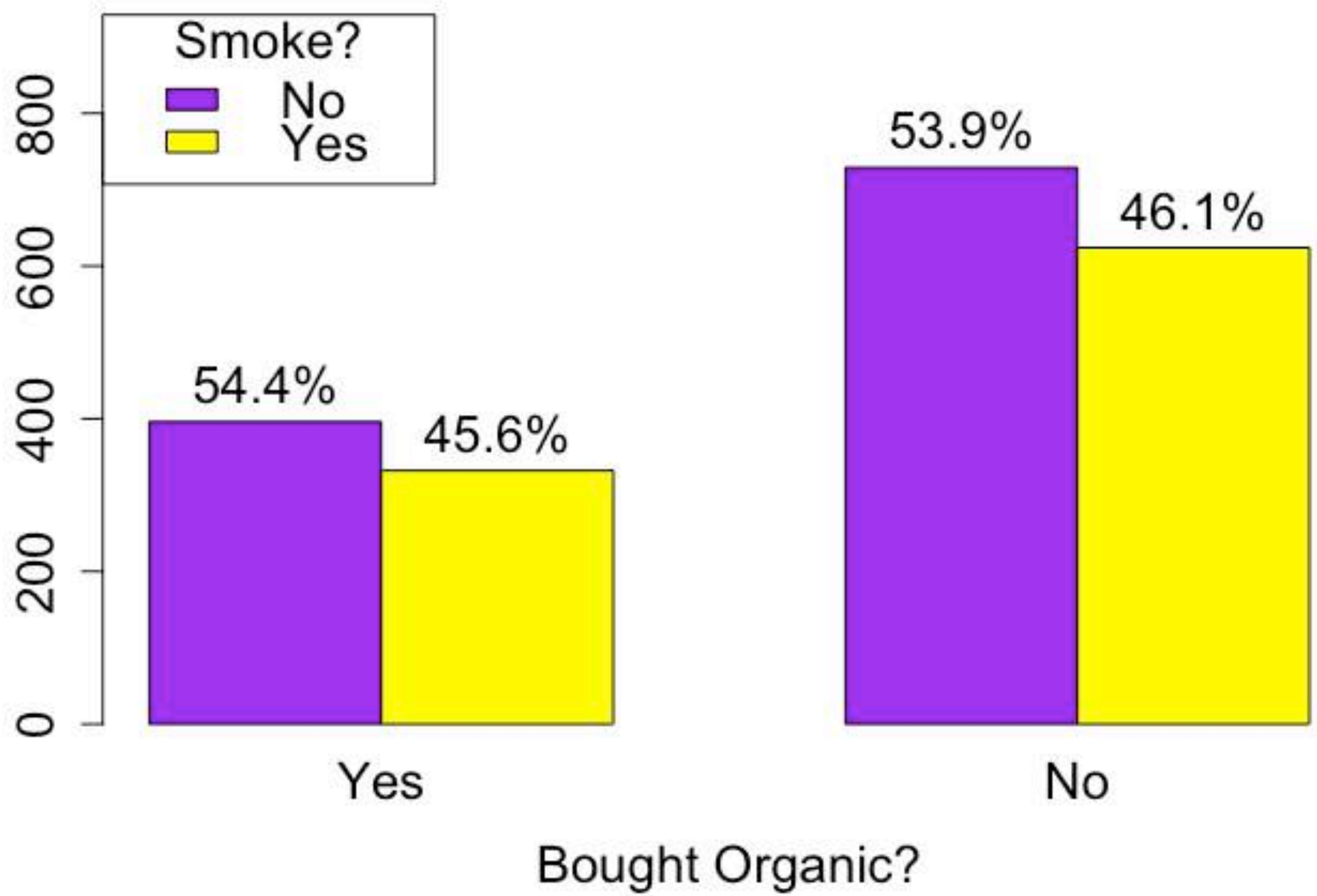
- Simulate a "randomly assigned" version of your treatment (permute it)

	green.veggies	smoke	income	health	organic	organic.sim
34	Most of the time	No	3324.5	Good	No	No
60	Always	Yes	1024.0	Fair	No	No
49	Always	Yes	2500.0	Good	Yes	No
80	Most of the time	No	1450.0	Excellent	No	No
80	Sometimes	No	1450.0	Good	No	Yes
17	Sometimes	NA	5824.0	Good	No	No
42	Sometimes	NA	3324.5	Poor	Yes	No
45	Always	NA	5000.0	Very good	No	No
28	Always	No	600.0	Fair	No	Yes
19	Always	NA	1450.0	Poor	Yes	No

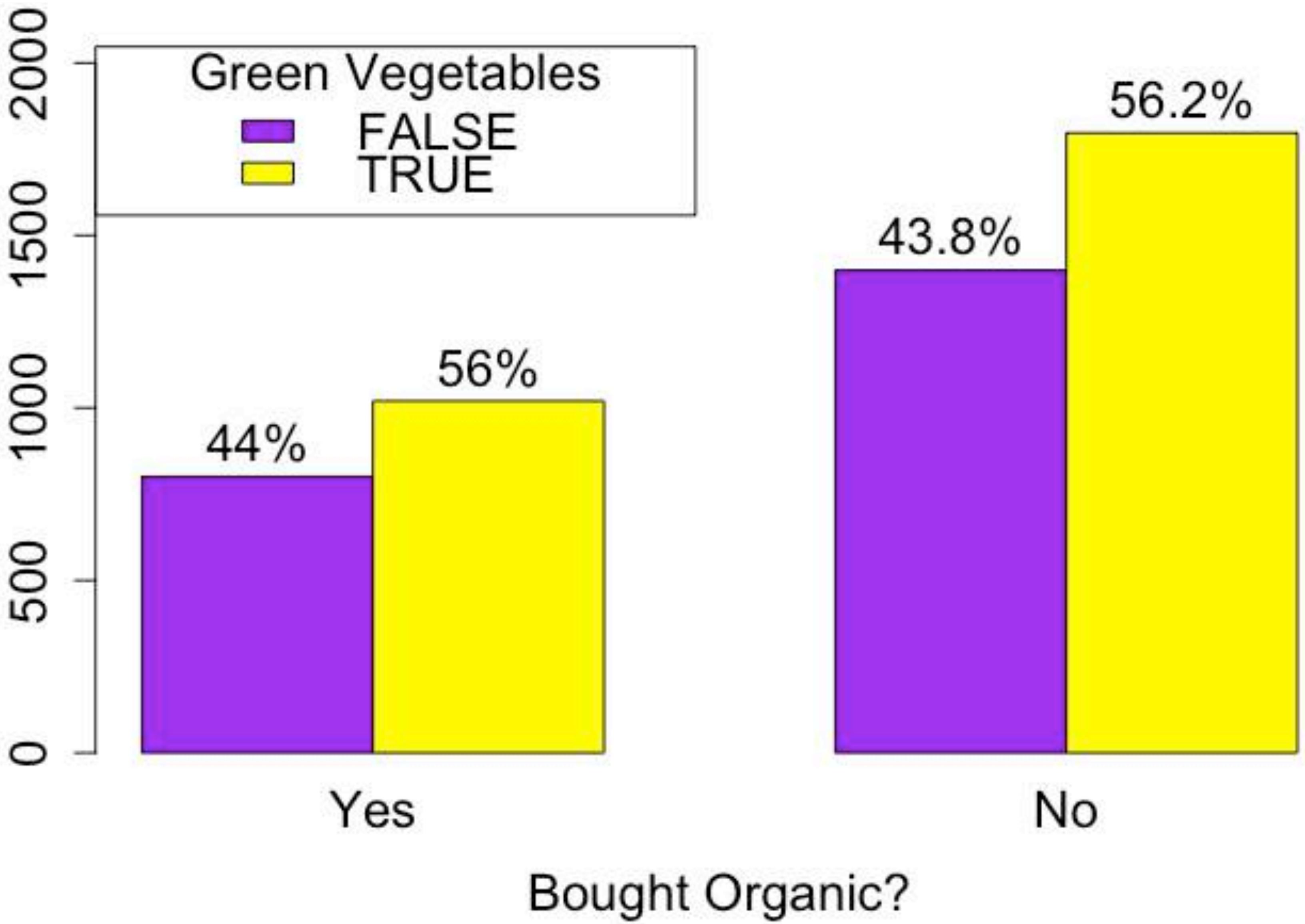
Baseline Differences?



Baseline Differences?



Baseline Differences?



Randomization!!

- <u>Without randomization</u>…
  … groups will differ at baseline
  … so very hard to find causal evidence

- <u>With randomization</u>…
  … groups should look similar at baseline
  … so can find causal evidence!

- *Can't check for all baseline differences…*

*… **but CAN check for random assignment!***

Evaluating Evidence

- Suppose A has better outcomes than B
- Possible explanations?
  1) A causes better outcomes than B
  2) the groups differed at baseline ???
  3) random chance

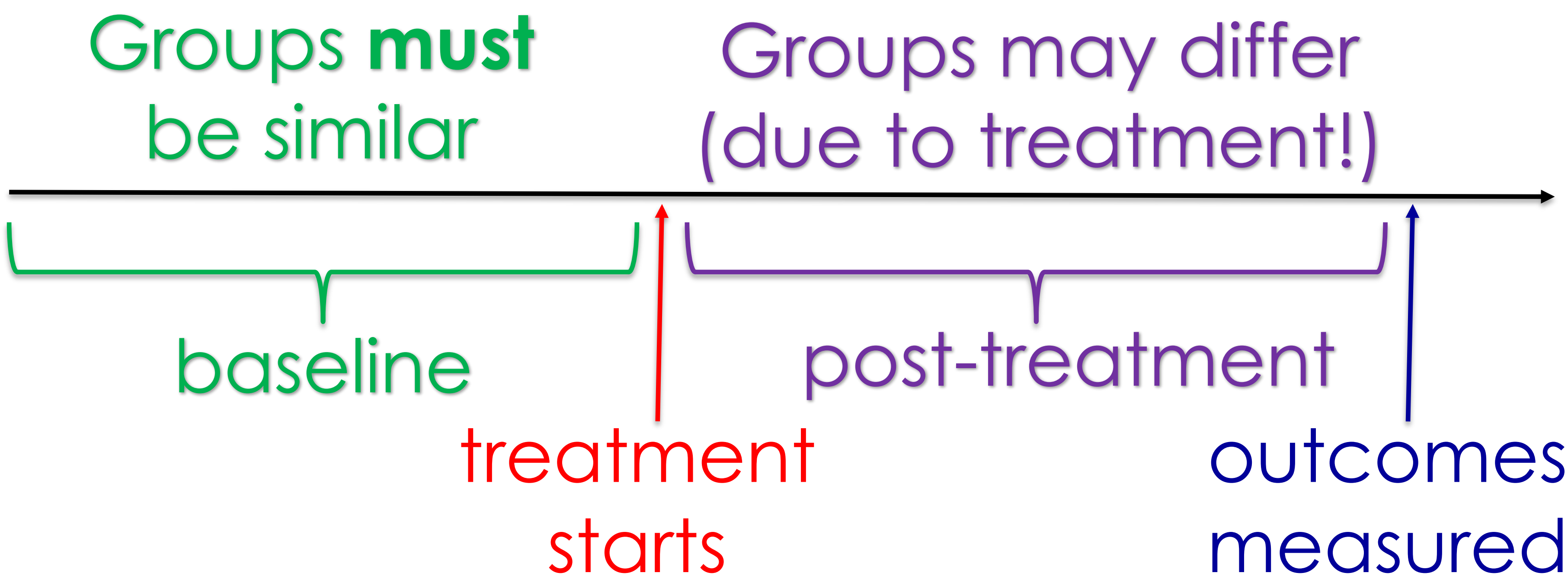
The best evidence against groups differing at baseline is the use of random assignment to treatment groups.

Randomizing Within Similar Groups

Cal and Axel Lock Morgan

Why "groups differ at baseline"?

- Why "baseline"?



- Why groups?

Teaching Confounding

- Requires multivariable thinking!
  - Help students reason with a third variable
  - Use data, don't just rely on intuition
  - (Thoughtfully) visualize the confounding
  - (Show) data broken down by confounder
- Random assignment is important!
- Not just about study design!
- Simulation can help understanding!
  - simulate treatment based on confounder
  - simulate random assignment; no differences!

Three "-ations"

VISUALIZATION

"Pictures speak louder than words"

Multivariable thinking!

RANDOMIZATION

Allows for causal evidence!

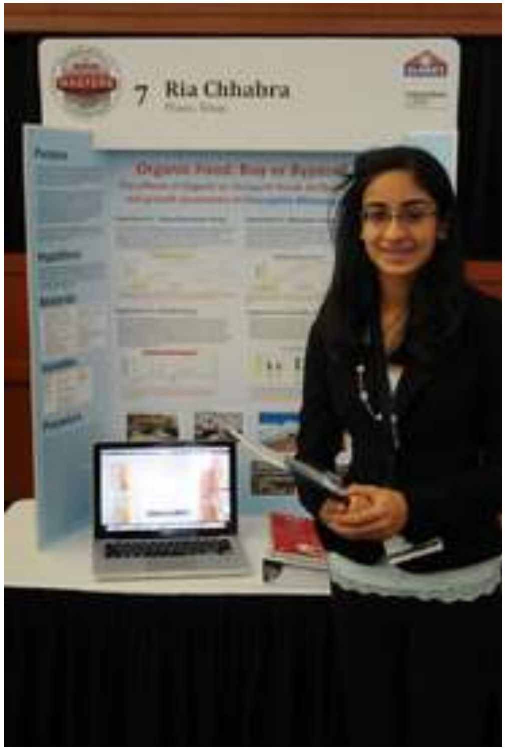
Foundation for inference

SIMULATION

Makes the abstract concrete

Dataset #2: Fruit Flies

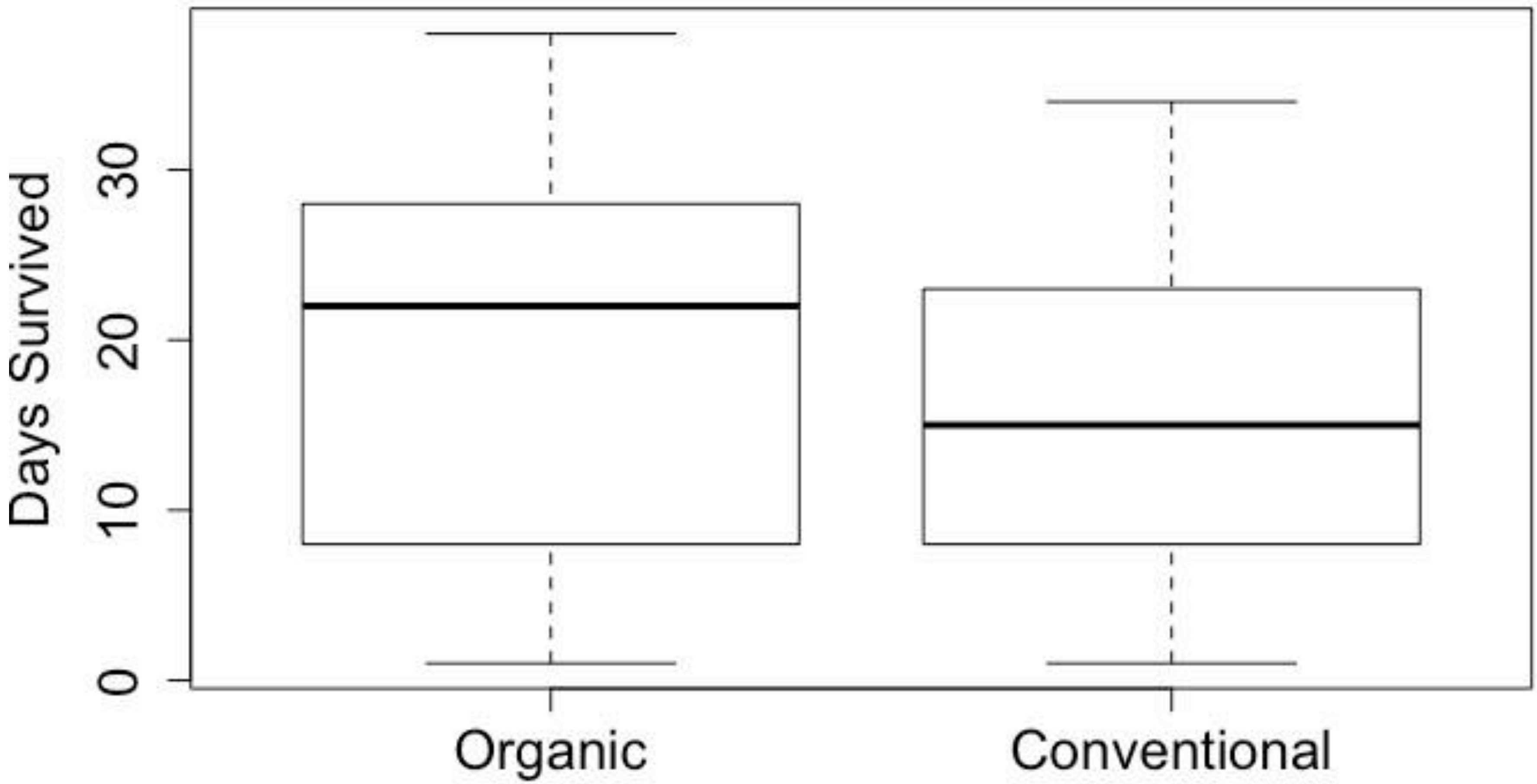
- Fruit flies **randomly** divided into two groups of 500 each
- One group fed organic food, the other conventional food
- Measured longevity, fertility, stress resistance, activity
- Study done by a 16-year-old girl for a science project!



The New York Times

Chhabra R, Kolli S, Bauer JH (2013) Organically Grown Food Provides Health Benefits to Drosophila melanogaster. PLoS ONE 8(1): e52988.

Longevity by Organic



$$\bar{Y}_O - \bar{Y}_C = 18.9 - 15.5 = 3.4 \text{ days}$$

*Data approximated from figure in paper

Evaluating Evidence

- In this sample, the fruit flies who ate organic lived longer
- Possible explanations?
  1) Eating organic increases longevity
  2) ~~The groups differed at baseline~~
  3) Just random chance ???

What kinds of results would we see, just by random chance, if there were no difference?

We can simulate to find out!!!

Simulating Random Chance

Days	Group
31	T
26	T
27	T
18	T
⋮	⋮
10	C
27	C
10	C
27	C

1. Assume no difference (days survived the same regardless of organic)
2. Mimic random chance: Re-randomize into groups
3. Compute the statistic (difference in means)

$$\bar{Y}_O - \bar{Y}_C = -0.684$$

4. Do this thousands of times!

StatKey

Evaluating Evidence

- In our sample, the fruit flies who ate organic lived longer
- Possible explanations?
  1) Eating organic increases longevity
  2) ~~The groups differed at baseline~~
  3) ~~Just random chance~~

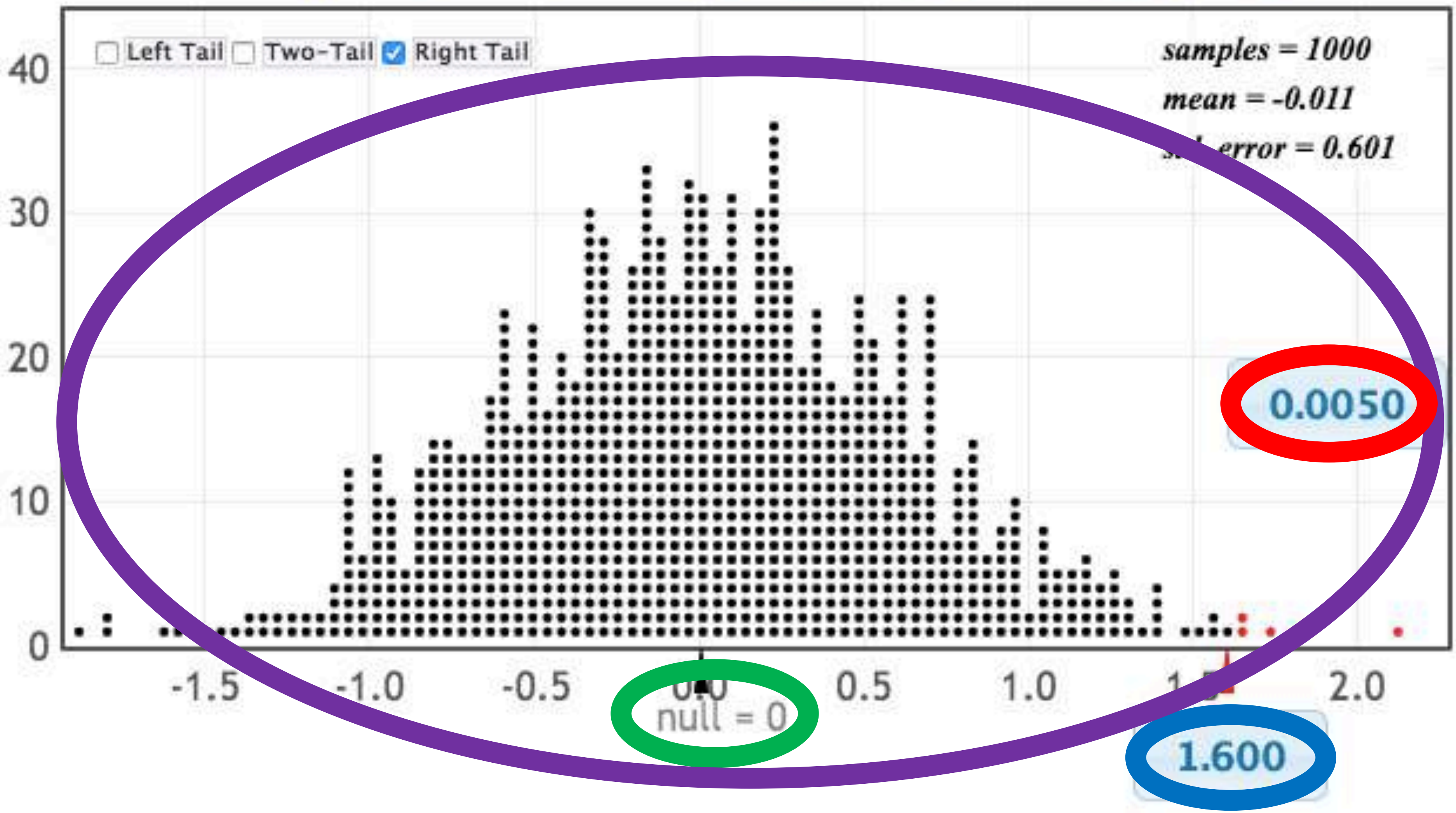
EAT ORGANIC!!!

(if you're a fruit fly)

What about a p-value???

- Students need to see, *and understand*, the concept of a p-value
- But, maybe start with extreme examples where an exact calculation isn't needed
- (and where an exact threshold isn't needed!)
- Get students comfortable with "would I expect a result this extreme just by chance?"
- THEN, p-value is a natural quantification…

p-value: The chance of obtaining a statistic as extreme as that observed, just by random chance, if the null hypothesis is true



Three "-ations"

VISUALIZATION

RANDOMIZATION

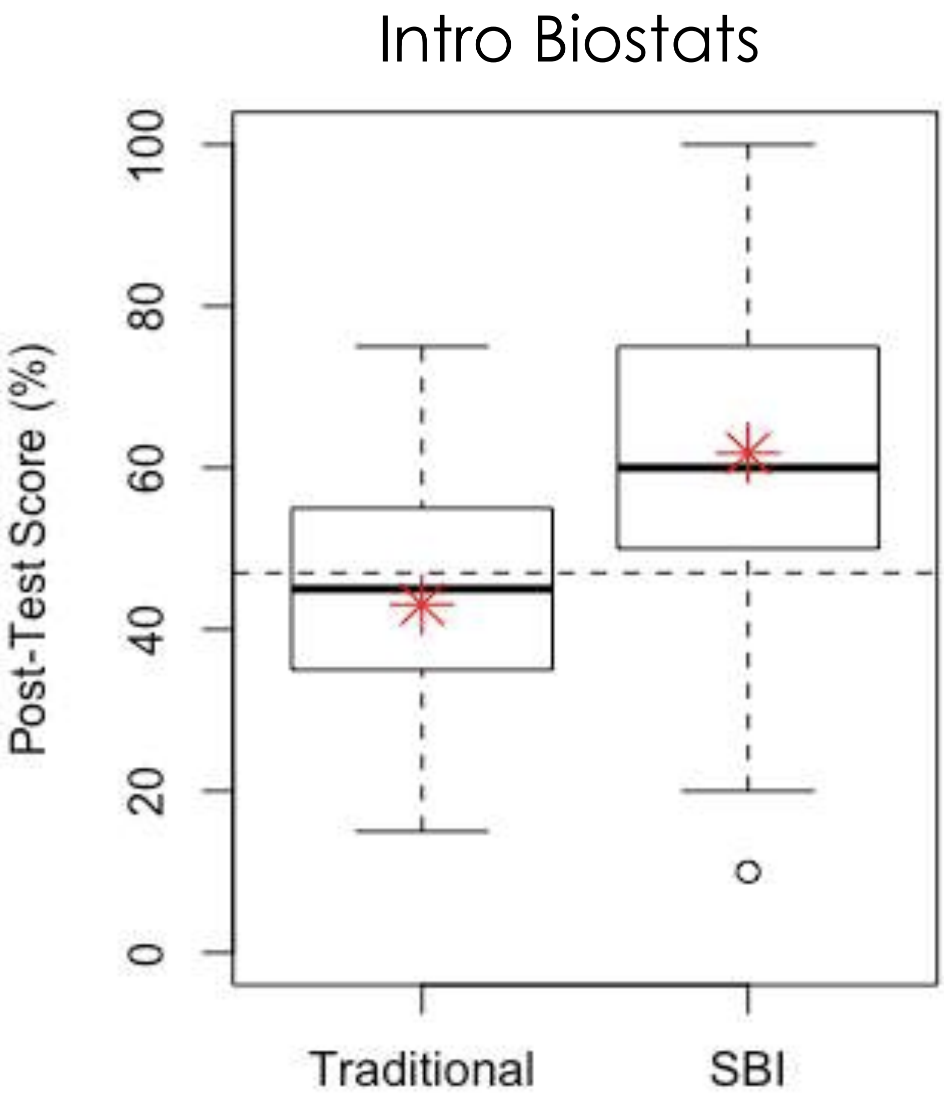
SIMULATION

Simulation-Based Inference

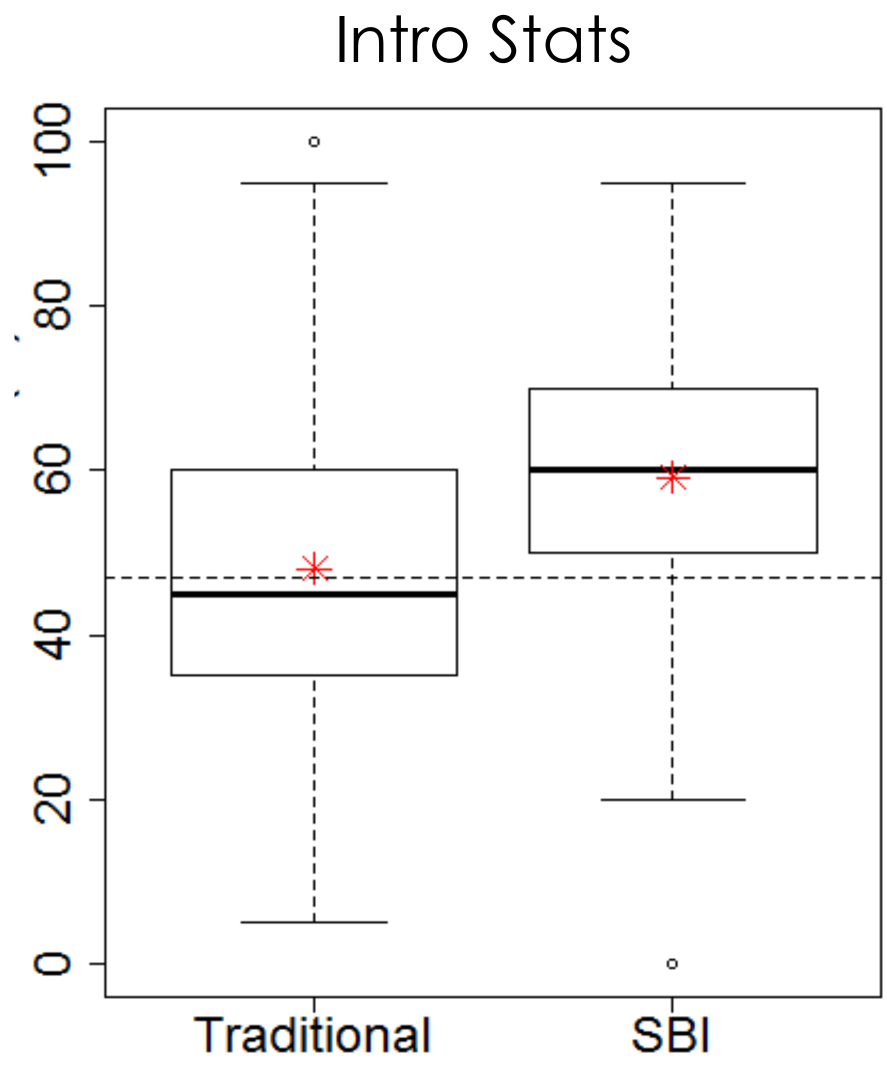
- Directly connected to key concepts!!!
- Same process for many statistics
- More flexible
- Fewer conditions
- Better connection with data collection
- Less reliance on prerequisite knowledge
- Promotes better understanding???

Let's evaluate the evidence!

GOALS Post-Test: Penn State



$\bar{Y}_{SBI} - \bar{Y}_{trad} = 18.7$

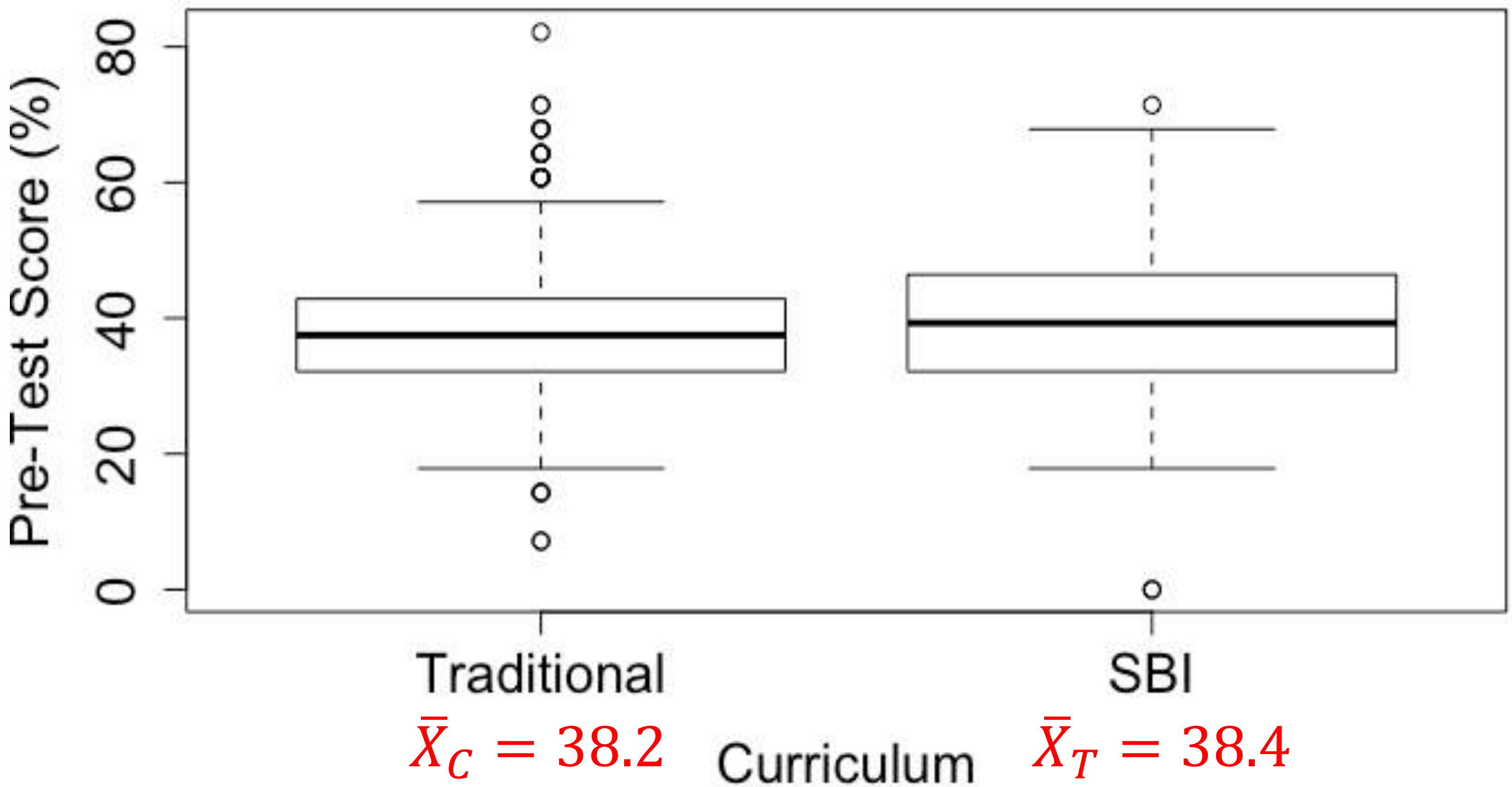


$\bar{Y}_{SBI} - \bar{Y}_{trad} = 11.4$

Evaluating Evidence

- In our sample, the students in the SBI classes had higher GOALS scores
- Possible explanations?
  1) SBI better for conceptual understanding
  2) The groups differed at baseline ???
  3) ~~Just random chance~~ p-value $< 10^{-16}$

Baseline Differences?



$\bar{X}_C = 38.2$ $\bar{X}_T = 38.4$

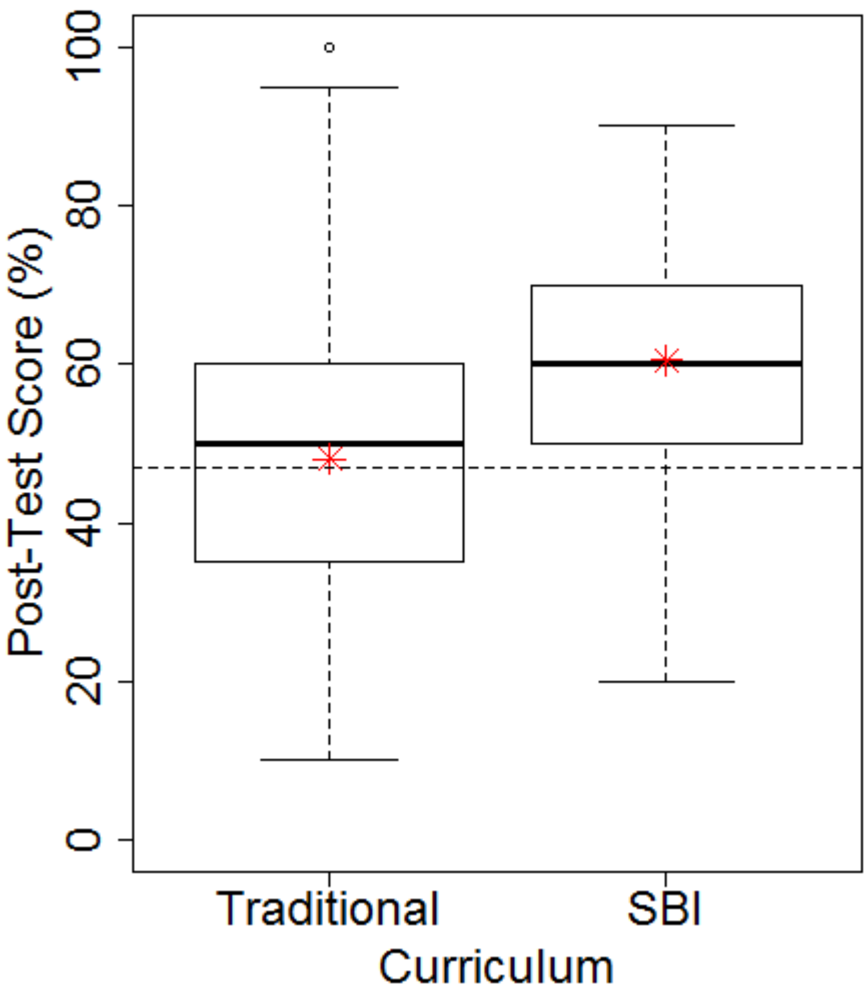
Propensity score matching to create groups similar with respect to all measured baseline variables:

Post-test difference: $\bar{Y}_{SBI} - \bar{Y}_{trad} = 10.9$

Within-Instructor Comparisons

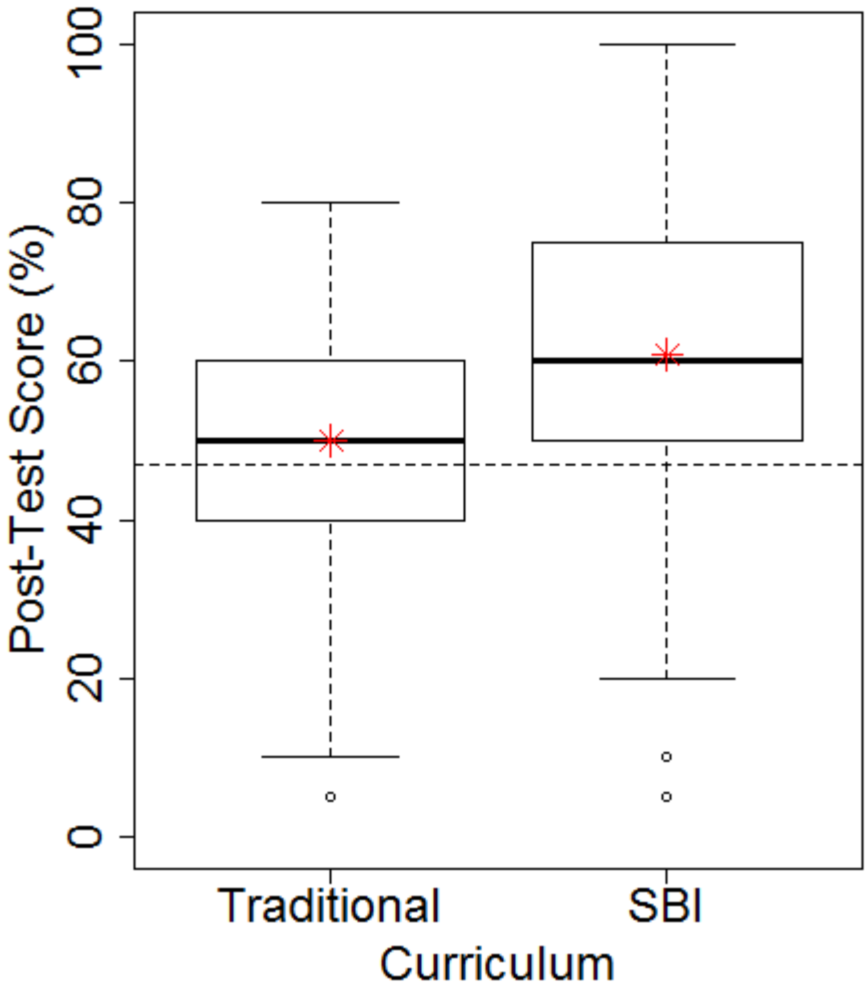
Same instructor A, S17/F17

$\bar{Y}_{SBI} - \bar{Y}_{trad} = 12.5$



Same instructor B, S17/F17

$\bar{Y}_{SBI} - \bar{Y}_{trad} = 10.8$



What else? (not exhaustive!!!)

- Meaningful effect size?
  - *Intro Biostat:*
    - Difference in means: 18.7 percentage points
    - 95% CI: (11.6, 25.9)
    - Standardized effect size = 1.04
  - *Intro Stat:*
    - Difference in means: 11.4 percentage points
    - 95% CI: (9.0, 13.8)
    - Standardized effect size = 0.76

- Missing data?
  - Penn state intro stat SBI data: Post-test missing for…
    - 36% of control students
    - 8% of treatment students

Replication

- Maurer, K. & Lock, D. (2016). "Comparison on Learning Outcomes for Simulation-based and Traditional Inference Curricula in a Designed Educational Experiment," *TISE*, **9**(1). **[random assignment!!]**
- Chance, B., Mendoza, S., Tintle, N. (2018). "Student Gains in Conceptual Understanding in Introductory Statistic With and Without a Curriculum Focused on Simulation-Based Inference," *ICOTS 10*.
- Tintle, N., Clark, J., Fisher, K., Chance, B., Cobb, G. Roy, S. (2018). "Assessing the Association Between Precourse Metrics of Student Preparation and Student Performance in Introductory Statistics: Results from Early Data on Simulation-Based Inference vs. Nonsimulation-Based Inference," *JSE*, **26**(2).
- Chance, B., Wong, J., & Tintle, N. (2016). "Student Performance in Curricula Centered on Simulation-Based Inference: A Preliminary Report," *JSE*, **24**(3).

All find better conceptual understanding with SBI!

Evaluating Evidence

- Suppose, in our sample, group A has better outcomes than group B
- Possible explanations?
  1) A causes better outcomes than B
  2) the groups differed at baseline
  3) just random chance

Evaluating evidence for (1) requires evaluating evidence *against* (2) and (3)

Three “-ations”

VISUALIZATION

RANDOMIZATION

SIMULATION

klm47@psu.edu

Want to continue the conversation?

Join us for a collaborative discussion in Room 208!