Causal Inference

Midterm

Beer and Hot-wings

A student in Minneapolis studied the relationship between beer and hot-wing consumption. At a bar she recorded ounces of beer a person drank, number of hot-wing he/she consumed, and their gender. Then she fit a linear regression model.

```
##                 Estimate Std. Error
## (Intercept)          0.3        5.7
## Hotwings             2.3        0.6
## Male                 2.0        8.9
## Hotwings:Male       -0.4        0.7
```

5. Based on the above result, on average how many ounces of beer do you expect a person to consume if the person eats 10 hot-wings and is a male.

   a) 19.6 b) 21.3 c) 21.5 d) 25.4

6. At $\alpha = 0.05$ could we conclude that males consume more hot-wings than females on average?

   a) Yes b) No

7. Based on the above the student concluded that "eating more hot-wings makes people drink more beer controlling for gender". Do you think this is a valid conclusion from the study?

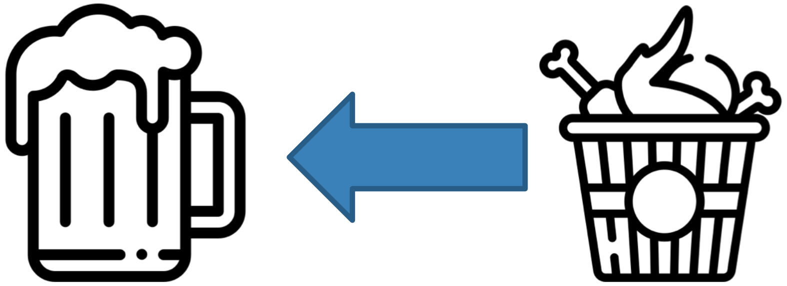
   a) Yes b) No

Let’s focus here

- “eating more hot-wings makes people drink more beer controlling for gender.”

```
##                  Estimate Std. Error
## (Intercept)           0.3        5.7
## Hotwings              2.3        0.6
## Male                  2.0        8.9
## Hotwings:Male        -0.4        0.7
```

- The outcome is the expected number of beer.
- Coefficient for Hotwings is positive and “significant”
- So, yeah!?



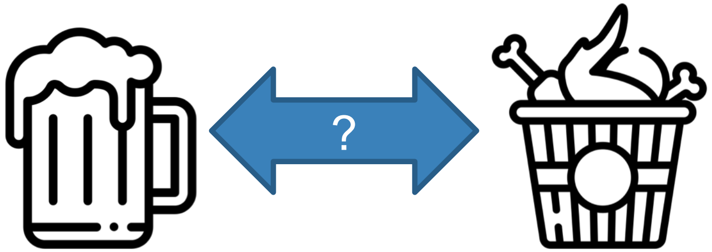
What if we flip the predictor and outcome

- Wait… are the wings making you drink more beer or are the beers making you eat more wings?

```
> lmhw<-lm(Beer~Hotwings*Male,data=bw)
>
> round(summary(lmhw)$coef[,1:2],1)
              Estimate Std. Error
(Intercept)        0.3        5.7
Hotwings           2.3        0.6
Male               2.0        8.9
Hotwings:Male     -0.4        0.7
```

```
> lmhw2<-lm(Hotwings~Beer*Male,data=bw)
>
> round(summary(lmhw2)$coef[,1:2],1)
            Estimate Std. Error
(Intercept)      3.2        1.7
Beer             0.3        0.1
Male             3.2        2.6
Beer:Male        0.0        0.1
```

- I'm confused..



Causality

- Many of you lost points in the midterm for stating that eating hot wings will cause you to drink more beer.

Effects of alcohol on food and energy intake in human subjects: evidence for passive and active over-consumption of energy.

Yeomans MR[1].

Author information

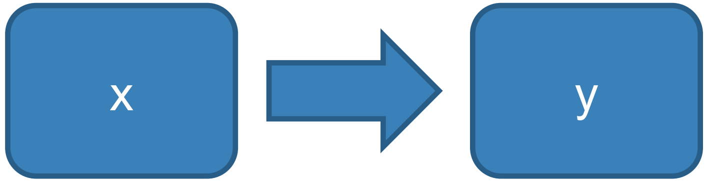
Abstract

The effects of alcohol on food and energy intake in human subjects have been the subject of a number of controlled studies recently. Unlike the evidence for other macronutrients, there is minimal evidence for any compensatory reduction in food intake in response to energy ingested as alcohol. In contrast, all studies testing intake within 1 h of preload ingestion report a higher intake of food following alcohol relative to energy-matched controls, although this short-term stimulatory effect is not evident if the test meal is delayed beyond 1 h. This time-course suggests that short-term stimulation of appetite may be mediated by the pharmacological action of alcohol on the appetite control system, either through enhanced orosensory reward or impaired satiety. In the long term, energy ingested as alcohol is additive to energy from other sources, suggesting that moderate alcohol consumption results in long-term passive over-consumption alongside short-term active over-consumption of energy through appetite stimulation. Despite the consistency of enhanced energy intake after moderate alcohol, evidence of an association between alcohol in the diet and obesity remains contentious, although the most recent results suggest that alcohol intake correlates with BMI. Future research needs to address this issue and clarify the mechanisms underlying appetite stimulation by alcohol.

Regression effect does NOT mean causality

- Just because coefficient is “significant” does not ensure the direction of cause.

```
x<-runif(100,0,10)
y<- 1+ 10*x+rnorm(100,0,3)
```



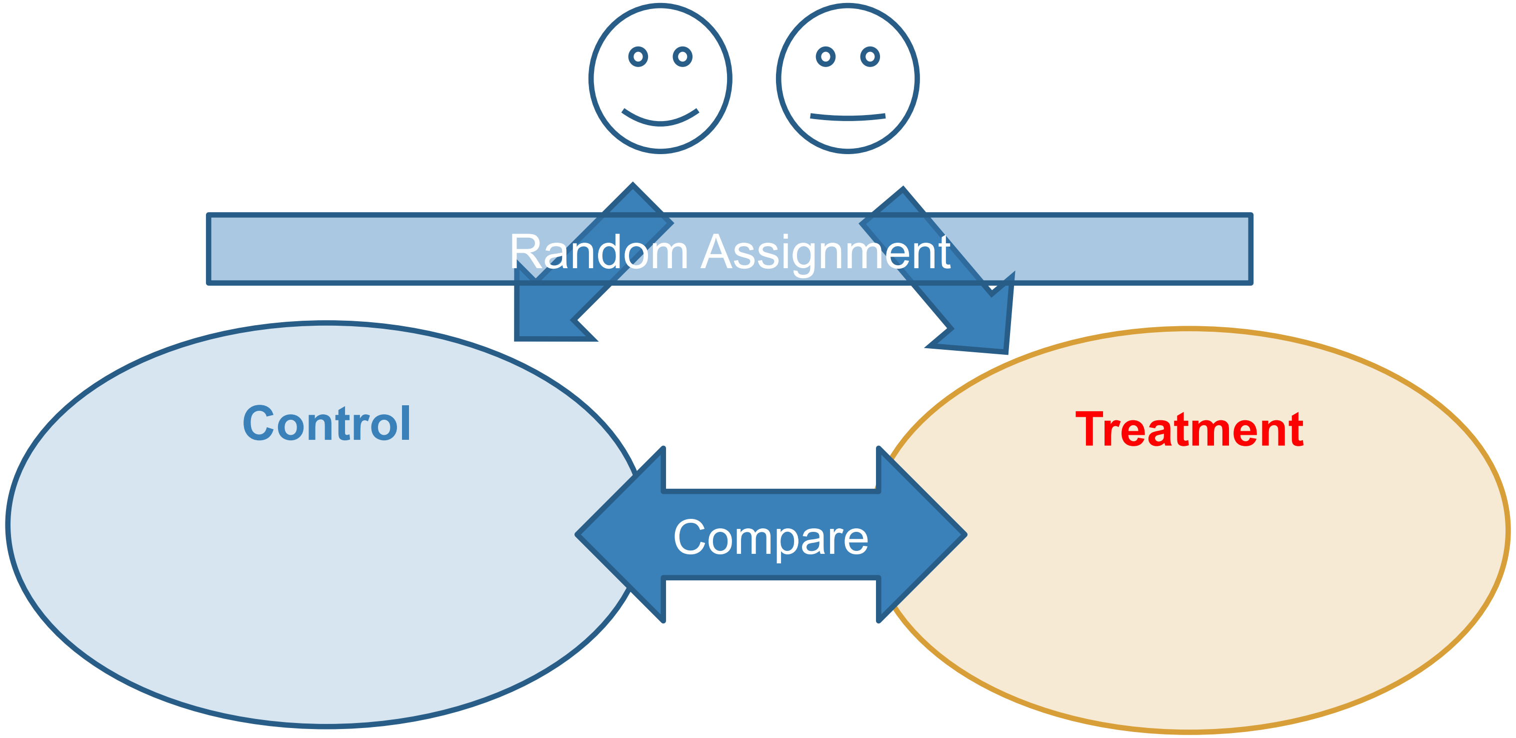

```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.37704    0.60412   2.279   0.0248 *
x           10.03114    0.09743 102.955   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 2.943 on 98 degrees of freedom
Multiple R-squared:  0.9908, Adjusted R-squared:  0.9907
F-statistic: 1.06e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = x ~ y)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0864138  0.0608937  -1.419    0.159
y            0.0987763  0.0009594 102.955   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.292 on 98 degrees of freedom
Multiple R-squared:  0.9908, Adjusted R-squared:  0.9907
F-statistic: 1.06e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

Randomized Trials

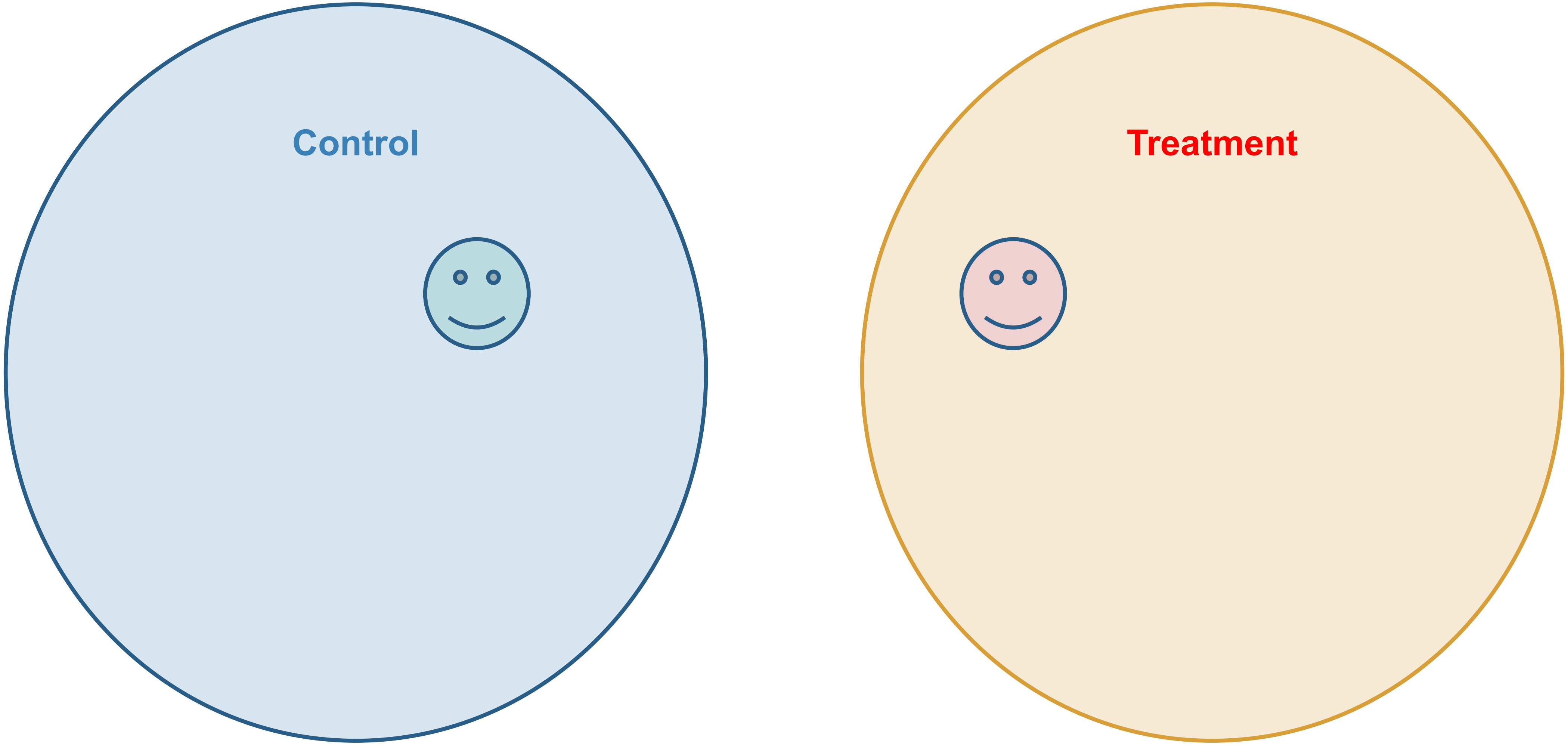
Randomized Trials

- The golden standard for clinical trials.

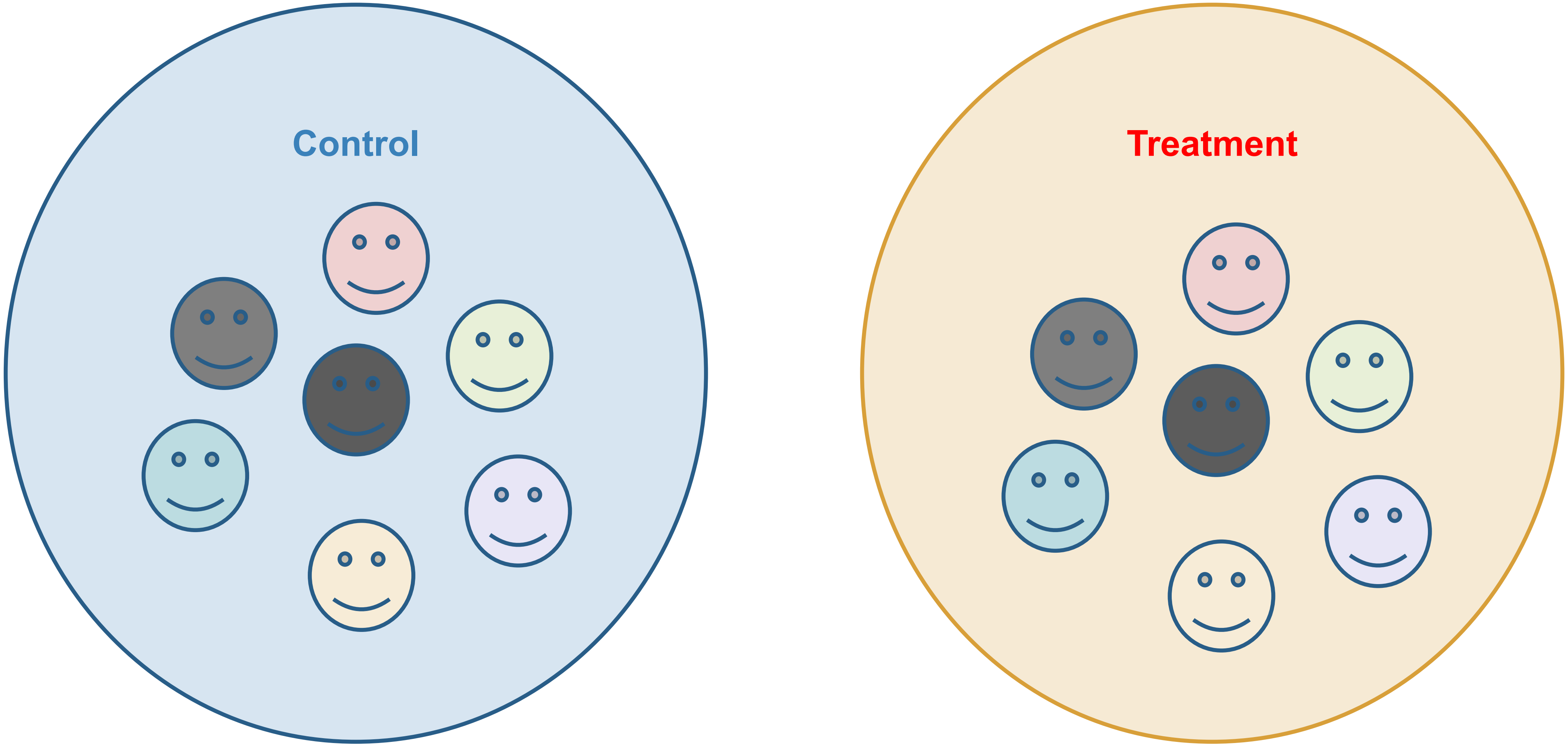


- But why? Can it infer causality?

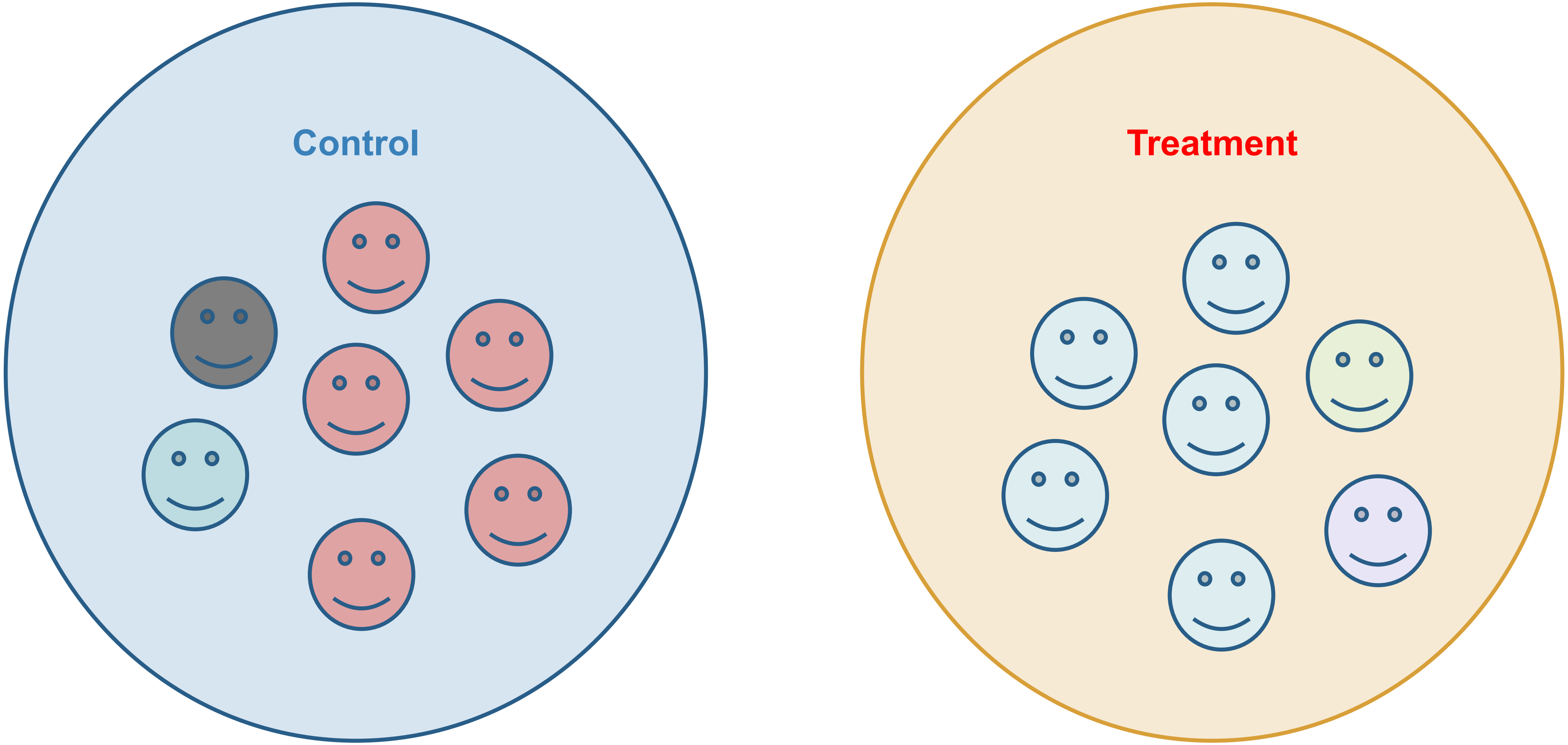
Can we only infer association?



Randomized Experiment



Observational Study



Inferring Causality using Regression

causal effects can be estimated using regression

- If the two conditions are met
  - The model is correct and
  - includes all “confounding covariates”.
- **Confounding covariates**: predictors that can affect
  - treatment assignment or
  - the outcome
- If the confounding covariates are all observed,
  - then accurate estimation comes down to proper modeling and
  - the extent to which the model is forced to extrapolate beyond the support of the data.
- If the confounding covariates are not observed
  - then they are “omitted” or “lurking” variables that complicate the quest to estimate causal effects.
  - for example, if we suspect that healthier patients received the treatment, but no accurate measure of previous health status is included in the model,

Bias Due to Confounding covariates

- Let's formalize the problem let T_i be binary indicator for some treatment in our usual regression setting.

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i$$

- If we don't control for x_i the model we fit will be

$$y_i = \beta_0^* + \beta_1^* T_i + \epsilon_i^*$$

- If we define another regression of treatment on x

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i$$

Bias Due to Confounding covariates

The model with confounding covariate

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i$$

The model without

$$y_i = \beta_0^* + \underline{\beta_1^*} T_i + \epsilon_i^*$$

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i$$

- Substituting and rearranging the terms we get

$$y_i = \beta_0 + \beta_2\gamma_0 + \underline{(\beta_1 + \beta_2\gamma_1)} T_i + \epsilon_i + \beta_2\nu_i$$

- Therefore, if we don't control for x_i the estimate we ge is

$$\beta_1^* = (\beta_1 + \beta_2\gamma_1)$$

Bias Due to Confounding covariates

$$y_i = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)T_i + \epsilon_i + \beta_2\nu_i$$

- If we study $\beta_1^* = (\beta_1 + \beta_2\gamma_1)$ we see that
  - If there is no association between the treatment and the x_i then $\gamma_1 = 0$ in

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i$$

  - if there is no association between the outcome and the x_i then $\beta_2 = 0$ in

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i$$

- In either case not controling for x_i will not induce bias since $\beta_2\gamma_1 = 0$ thus

$$\beta_1^* = \beta_1$$

- In summary, uncontrolled variable is a problem if it is associate with both the outcome and with the treatment.

Counter Factuals

It’s a little early but…

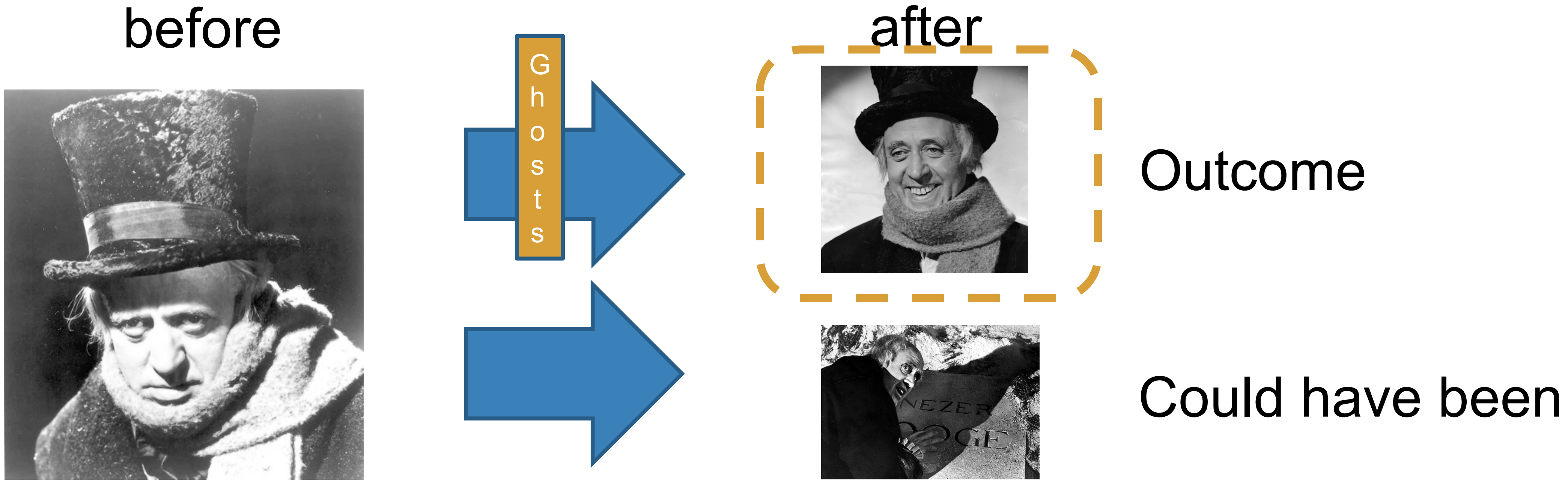
- let’s talk about The Christmas Carol

In the story of Christmas Carol…

- Ebenezer Scrooge was a man who cared only about his money.
- On the Christmas eve he was visited by the three spirits of Christmas.
- The experience changes him into a nice person.

Christmas Carol as experiment

- Experimental unit: Ebenezer Scrooge.
- Treatment: Visits by the three spirits of Christmas.
- Outcomes: His personality.
- Can we say the visits by the spirits caused him to become a nice person?



Rubin's causal model

- Causal inference is about comparison of potential outcomes.
- It's about comparing Scrooges if the ghosts did not visit vs Scrooge afte
- The two possible outcomes are called the *potential outcomes*.
- But since there is only one Scrooge, we can have only one outcome.
- The Scrooge that could have become is called the *counter factual*.

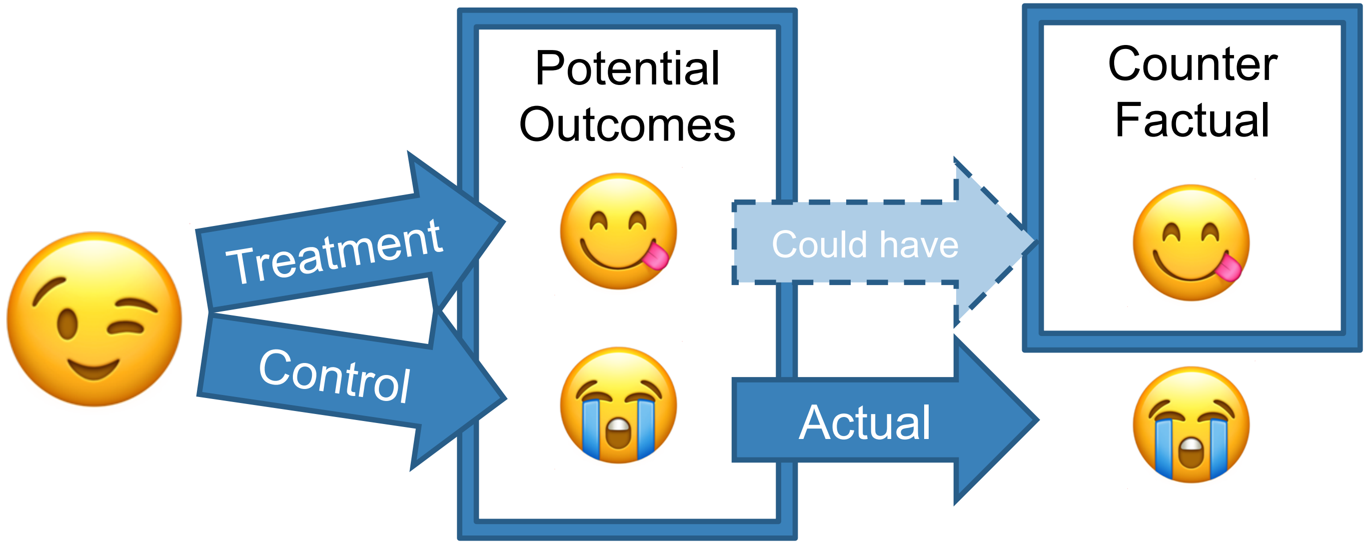
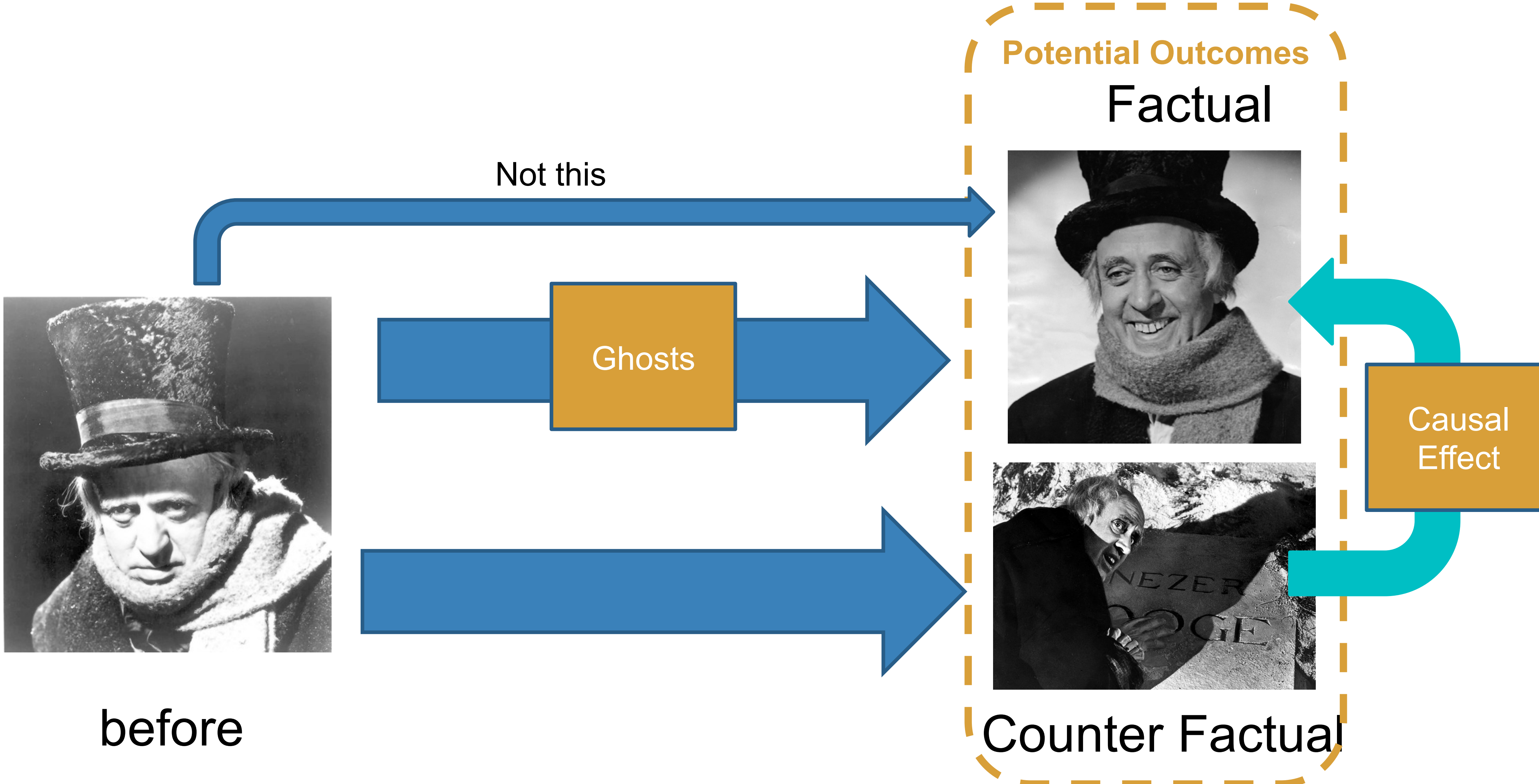


Figure 1: Scrooge under 3 doses of Christmas spirit

(b) Treated

(a) No treatment

Rubin's Causal Model



What's wrong with comparing Scrooge before and after?

(a) Before treatment (b) Treated

Figure 2: Scrooge before and after

Formalizing the potential outcome causality

- Treatment assignment indicator T_i.
- For an individual *i* let
  - Outcome under treatment: y_i^1 and
  - Outcome under no treatment: y_i^0
  - Individual treatment effect t_i:

$$t_i = \boxed{y_i^1 - y_i^0}$$

- Average treatment effect (ATE)

$$ATE = E\left(\boxed{y^1 - y^0}\right) = E\left(\boxed{y^1}\right) - E\left(\boxed{y^0}\right)$$

i	T_i	y_i^0 (Outcome under no treatment)	y_i^1 (Outcome under treatment)	$y_i^1 - y_i^0$
1	0	**69**	75	6
2	0	**111**	108	-3
3	1	92	**102**	10
4	1	112	**111**	-1
⋮		⋮	⋮	⋮
n	1	111	**114**	3

TRUE treatment effect

Limitation due to Reality

- In reality, however, in many settings we can only get one of the potential outcomes

i	T_i	y_i^0	y_i^1	$y_i^1 - y_i^0$
1	0	**69**	75	6
2	0	**111**	108	-3
3	1	92	**102**	10
4	1	112	**111**	-1
$\vdots$		$\vdots$	$\vdots$	$\vdots$
n	1	111	**114**	3

$\Rightarrow$

i	T_i	y_i^0	y_i^1	$y_i^1 - y_i^0$
1	0	69	?	?
2	0	111	?	?
3	1	?	102	?
4	1	?	111	?
$\vdots$		$\vdots$	$\vdots$	$\vdots$
n	1	?	114	?

- The quantity that we can calculate is

$$E(y^1 \mid T_i = 1) - E(y^0 \mid T_i = 0) \neq ATE = E(y^1 - y^0) = E(y^1) - E(y^0)$$

The fundamental problem

- This is what Rubin describes as the fundamental problem of causal inference.
- What Rubin did was to convert Causal inference into a missing data problem.

i	T_i	y_i^0	y_i^1	$y_i^1 - y_i^0$
1	0	**69**	75	6
2	0	**111**	108	-3
3	1	92	**102**	10
4	1	112	**111**	-1
$\vdots$		$\vdots$	$\vdots$	$\vdots$
n	1	111	**114**	3

$\Rightarrow$

i	T_i	y_i^0	y_i^1	$y_i^1 - y_i^0$
1	0	69	?	?
2	0	111	?	?
3	1	?	102	?
4	1	?	111	?
$\vdots$		$\vdots$	$\vdots$	$\vdots$
n	1	?	114	?

The fundamental problem

- Effect we observe are
  - *treatment on treated* and
  - *no treatment on not treated*.

	Outcome under treatment **y^1**	Outcome under no treatment **y^0**
Treatment (T_i=1)	$E(y_i^1 \mid T_i=1)$	?
No treatment (T_i=0)	?	$E(y_i^0 \mid T_i=1)$

- If we can impute the counter factuals we can calculate *ATE*.
  - y_i^1 for $T_i = 0$ (treatement on not treated) and
  - y_i^0 for $T_i = 1$ (no treatment on treated)

The fundamental problem (example)

- Recruited 16 moderately sad people
  - 😔😔😔😔😔😔😔😔 😔😔😔😔😔😔😔😔
- Gave half of them them happy drug and sugar pills to the other half.

	Outcome under drug **y^1**	Outcome under no drug **y^0**
Happy drug	🤣😂😃🤣🤣😁🤣😆	?
No drug	?	😩😫😤😩😫🙁😩😣

- If we could get at least the average of response for the gray boxes we can talk about the ATE.
- That is response if we did not give the drug takers no drug and if the no drug group the drug.

Clarification on couple of notations

- Different average treatment effects
  - *Average Treatment Effect (ATE)*

$$ATE = E(y^1) - E(y^0)$$

  - *Average Treatment Effect on Treated (*ATT)

$$\text{ATT} = E(y^1 | T_i = 1) - E(y^0 | T_i = 1)$$

  - *Average Treatment Effect on Not Treated* (ATNT)

$$\text{ATNT} = E(y^1 | T_i = 0) - E(y^0 | T_i = 0)$$

Special cases when we are able to impute through additional assumptions

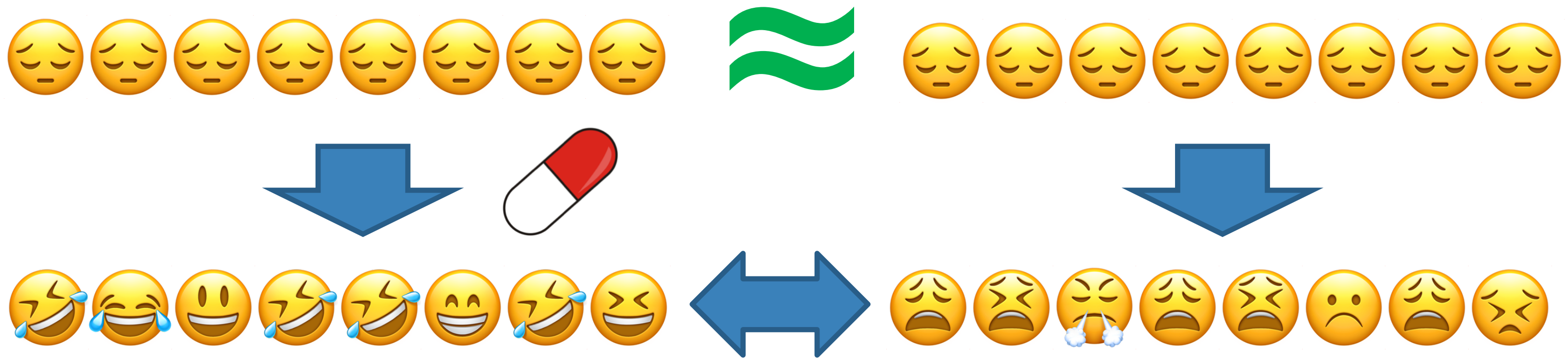
- We can substitute counterfactuals in certain cases.
  - Identical replicates such as twins or replicated material.
  - Treatment that does not alter the subject.
  - Under no treatment pretreatment condition persists.
- However, the only reasonable imputation is on response if the treated group was not treated ($y_i^0 | T_i = 1$).
- Therefore what you can get is ATT not ATE.
- Additional assumptions are necessary, especially dealing with human subjects.

Randomization

- Under randomization what you are missing is at random, thus

$$E(y^1 | T_i = 1) - E(y^0 | T_i = 0) = E(y^1) - E(y^0) = ATE$$

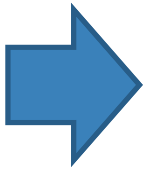
- What this means is: on average the y^1 and y^0 are comparable.



Can we fill that missing gap

- How about we put some values in and call it causal inference?

i	T_i	y_i^0	y_i^1	$y_i^1 - y_i^0$
1	0	69	?	?
2	0	111	?	?
3	1	?	102	?
4	1	?	111	?
$\vdots$		$\vdots$	$\vdots$	$\vdots$
n	1	?	114	?



i	T_i	y_i^0	y_i^1	$y_i^1 - y_i^0$
1	0	**69**	75	6
2	0	**111**	108	-3
3	1	92	**102**	10
4	1	112	**111**	-1
$\vdots$		$\vdots$	$\vdots$	$\vdots$
n	1	111	**114**	3

Let’s impute em missin values

- How about we put some values in and call it causal inference.
- There are ways to impute missing values by making strong assumption that missingness is random conditional on the observed covariates X_i .

Gender	y^0	y^1
F	1.2	?
F	1.1	?
M	2.2	?
M	2.1	?
F	?	3.1
F	?	3.0
M	?	2.0
M	?	2.2

Gender	y^0	y^1
F	1.2	3 ish
F	1.1	3 ish
M	2.2	2 ish
M	2.1	2 ish
F	1 ish	3.1
F	1 ish	3.0
M	2 ish	2.0
M	2 ish	2.2

Let's impute em missin values

- How about we put some values in and call it causal inference.
- There are ways to impute missing values by making strong assumption that missingness is random conditional on the observed covariates X_i .
- Even so it is non-trivial since we have T_i that divides the two groups.

Treatment	Gender	y^0	y^1
0	F	1	?
0	F	1	?
0	M	2	?
0	M	2	?
1	F	?	3
1	F	?	3
1	M	?	2
1	M	?	2

- We need a way to cross the *T* boundary.

Ignorability

- One big leap of faith is to say that

treatment assignment *T* is independent of the potential outcomes (y^0,y^1) given set of covariates X_i.

- Formally stated (conditional) ignorability assumption says
  - The distribution of the potential outcomes, (y^0,y^1), is the same across levels of the treatment variable, *T*, once we condition on confounding covariates *X*.
  - Any two classes at the same levels of the confounding covariates to have had the same probability of receiving the supplemental version of the treatment.

Ignorability as stratification

- Hypothetical example for illustration
  - Difference in gender distribution between the treated and control.
  - No treatment effect
  - Response differ by gender

i	T_i	gender	y_i^0	y_i^1	$y_i^1 - y_i^0$
1	0	M	10	10	0
2	0	M	10	10	0
3	0	M	10	10	0
4	0	F	100	100	0
5	0	F	100	100	0
6	1	M	10	10	0
7	1	F	100	100	0
8	1	F	100	100	0
9	1	F	100	100	0
10	1	M	10	10	0

$$E(y^1|T_i = 1) - E(y^0|T_i = 0) = 320/5 - 230/5 = 90/5$$

$$E(y^1|T_i = 1, M) - E(y^0|T_i = 0, M) = 30/3 - 20/2 = 0$$

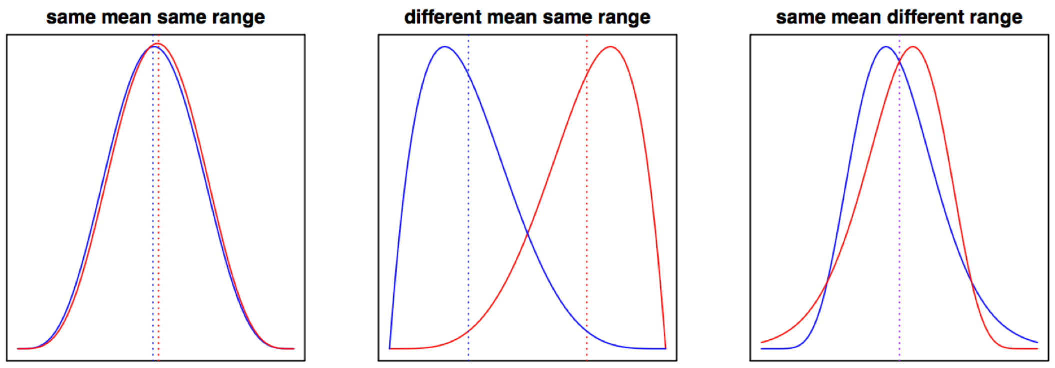
$$E(y^1|T_i = 1, F) - E(y^0|T_i = 0, F) = 200/2 - 300/3 = 0$$

Imbalance and overlap

- If the distribution of X to be same for each level of T then we can compare

$$E(y^1 | T_i = 1, X) - E(y^0 | T_i = 0, X).$$

- Imbalance:
  the distribution of pre-treatment variables differ for the treatment and control groups.
- Lack of overlap:
  when one of the group does not have value in certain region of pre-treatment variable.
- There are many ways to address the issue and correct for the imbalance.



Matching

- Matching is a general terminology for throwing away control samples so that they resemble the case samples in terms of pre-treatment covariates.
- You could say it is imputing $y_i^0 | T_i = 1$ using control patients $y_i^0 | T_i = 0$.
- By doing so we can calculate the treatment effect on the treated or the average treatment on treated (ATT)

$$ATT = E(y^1 | T_i = 1) - E(y^0 | T_i = 1)$$

i	T_i	y_i^0	y_i^1	$y_i^1 - y_i^0$
1	0	69	?	?
2	0	111	?	?
3	1	?	102	?
4	1	?	111	?
$\vdots$		$\vdots$	$\vdots$	$\vdots$
n	1	?	114	?

Propensity score

- Propensity score is probability of getting treated given pre-treatment covariates.

- Under the conditional ignorability assumption we have conditional ignorability wrt propensity score.

$$e(\mathbf{X}_i) = P(T_i = 1|\mathbf{X}_i)$$

- when a unit in the treatment group and a corresponding matched unit in the control group have the same propensity score, the two matched units will have, in probability, the same value of the covariate vector $\mathbf{X}_i$.

$$(y^0,y^1) \perp T_i \mid \mathbf{X}_i \Rightarrow (y^0,y^1) \perp T_i \mid e(\mathbf{X}_i)$$

- Very coarsely speaking we can use $\mathbf{X}_i$ to find twin among the controls that match the case.

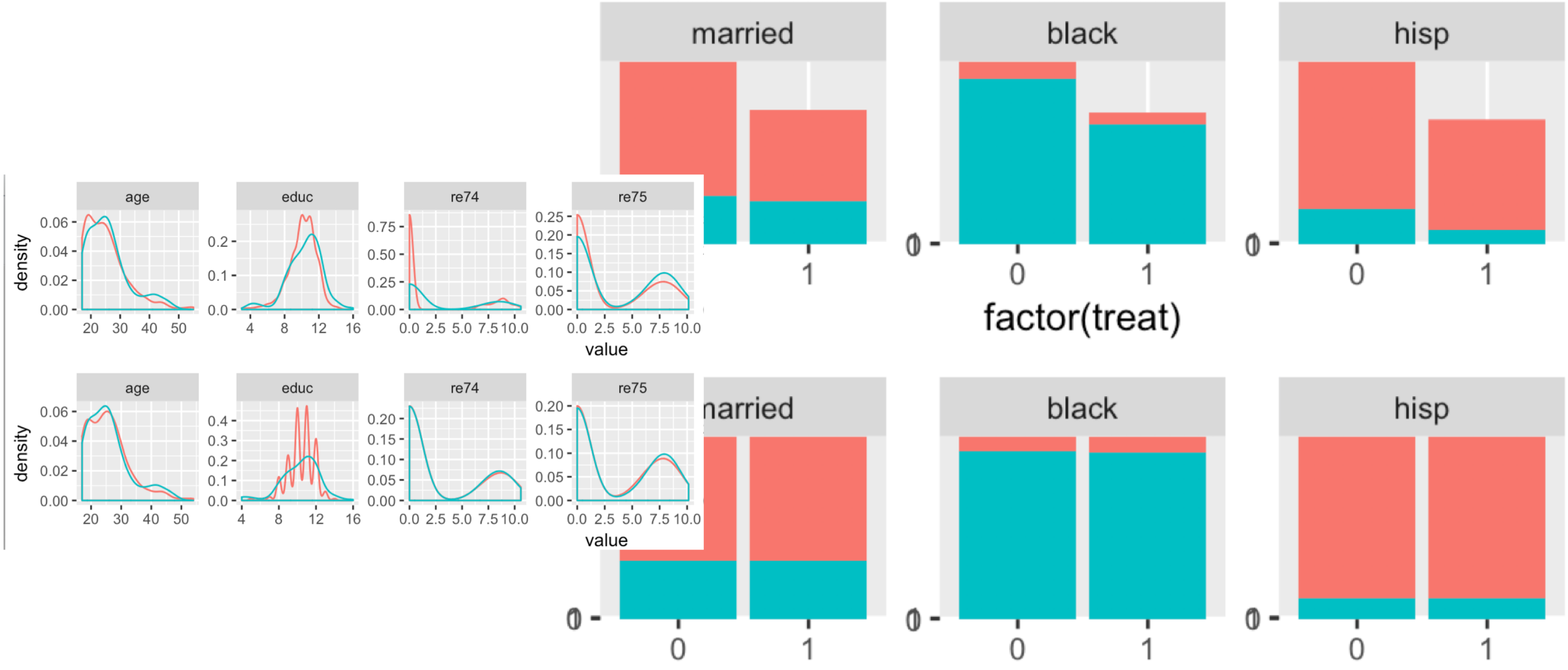
Propensity score matching

1. Using logistic regression model
   - $P(T_i = 1|\mathbf{X}_i) P(T_i = 1|\mathbf{X}_i) = logit^{-1}(\mathbf{X}_i\beta)$
2. Predict propensity score using the fitted model
   - $\hat{\boldsymbol{e}}(\mathbf{X}_i) = logit^{-1}(\boldsymbol{X}_i\widehat{\boldsymbol{\beta}})$
3. Find among controls n_1 samples that have similar propensity score.
4. Check the distribution of $\mathbf{X}$ are similar in both case and control.
5. Compare the means of each sample using regression.

Doing matching in R

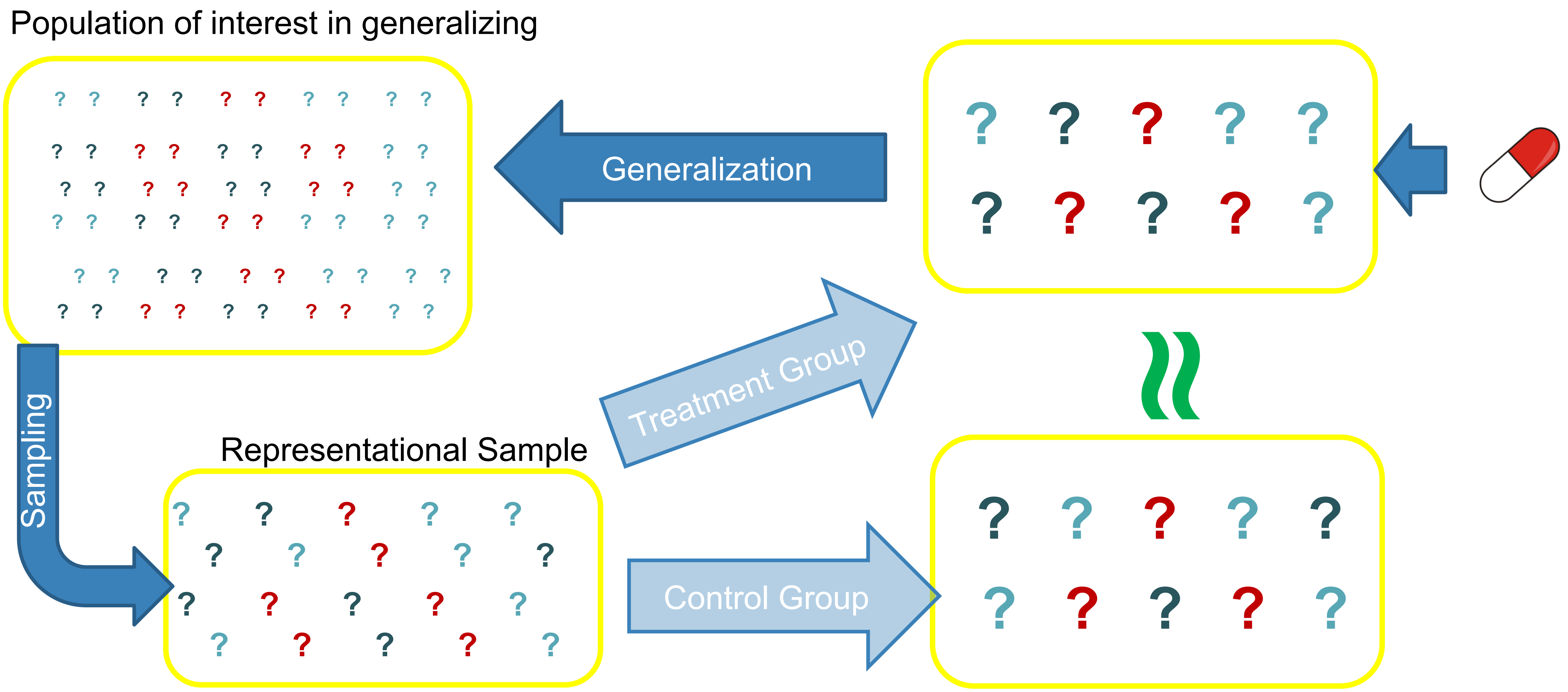
```
> data(lalonde)
> ps.fit <- glm(treat ~ age + educ + re74 + re75 + married +
black + hisp, data =lalonde, family = binomial())
> pscores<-predict (ps.fit, type="link")
> matches <- matching (z=lalonde$treat, score=pscores)
> matched <- lalonde[matches$matched,]
```

After matching

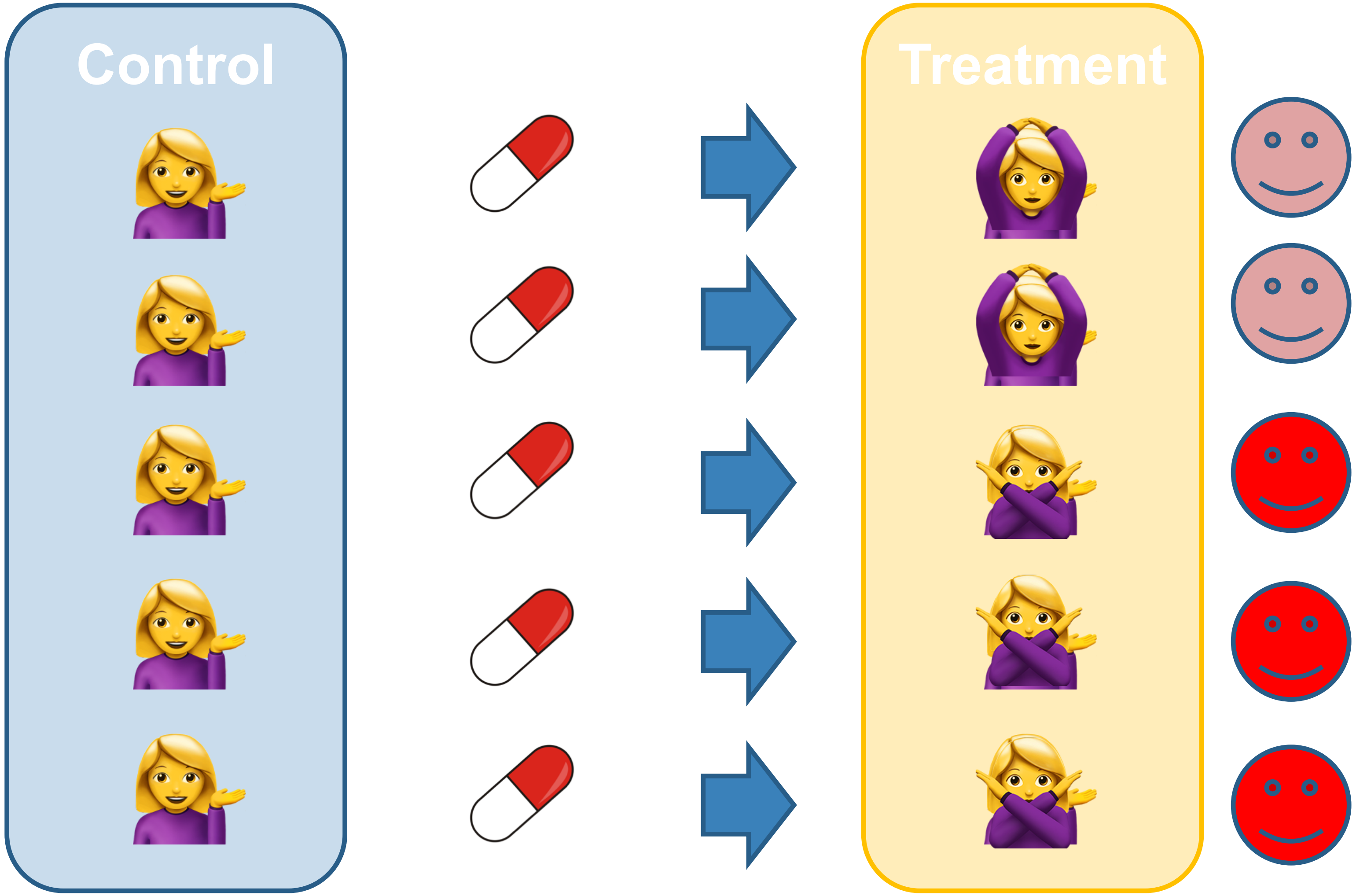


ITT and Instrumental Variable

The ideal experiment

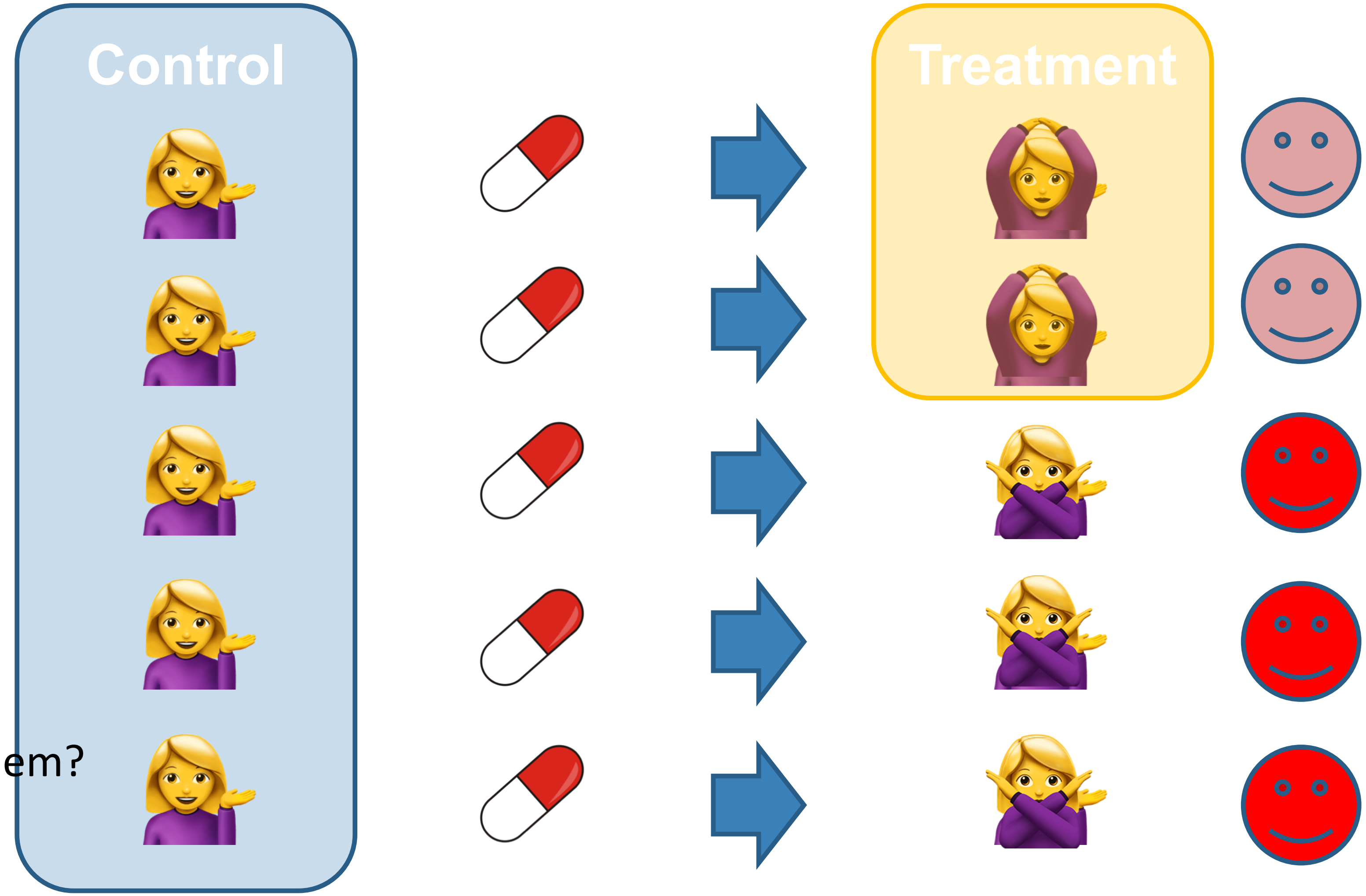


Partial Compliance



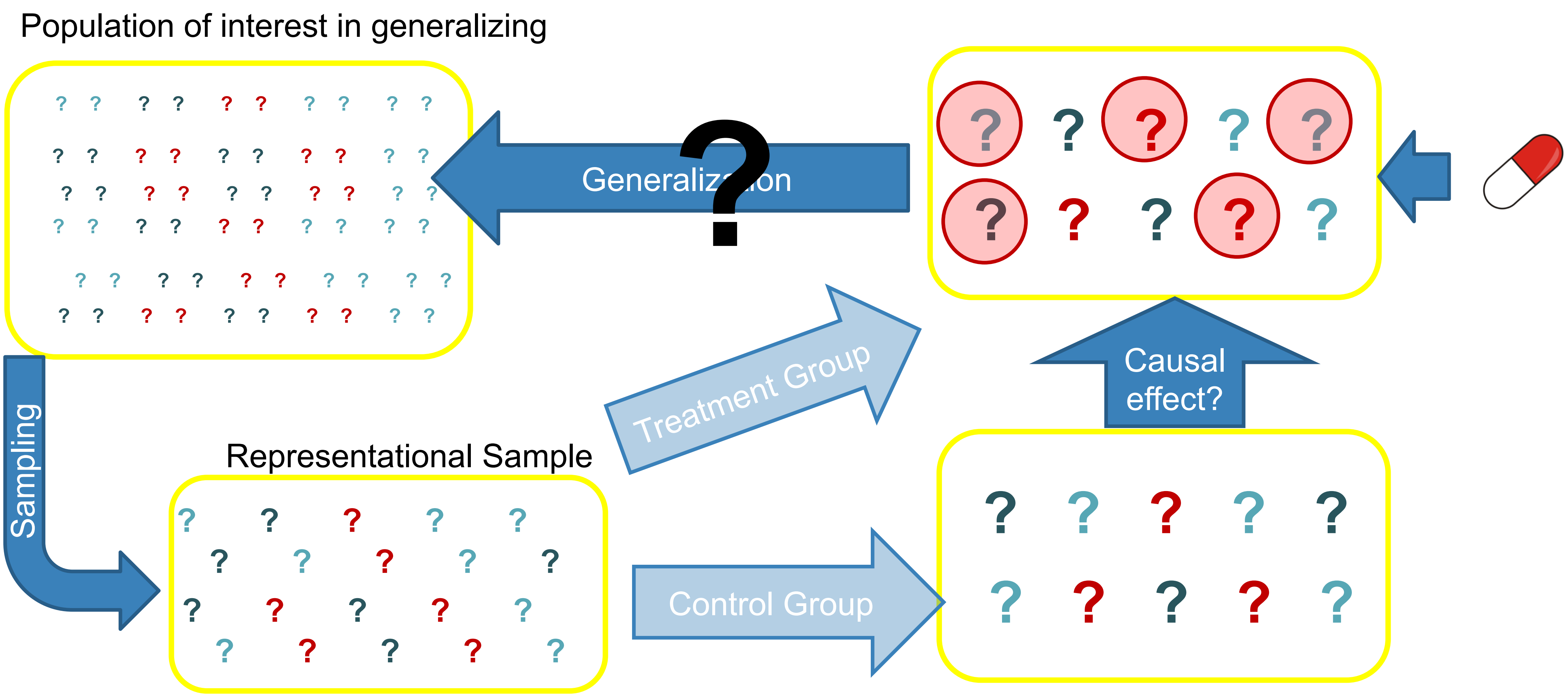
- Game over?

Subset analysis

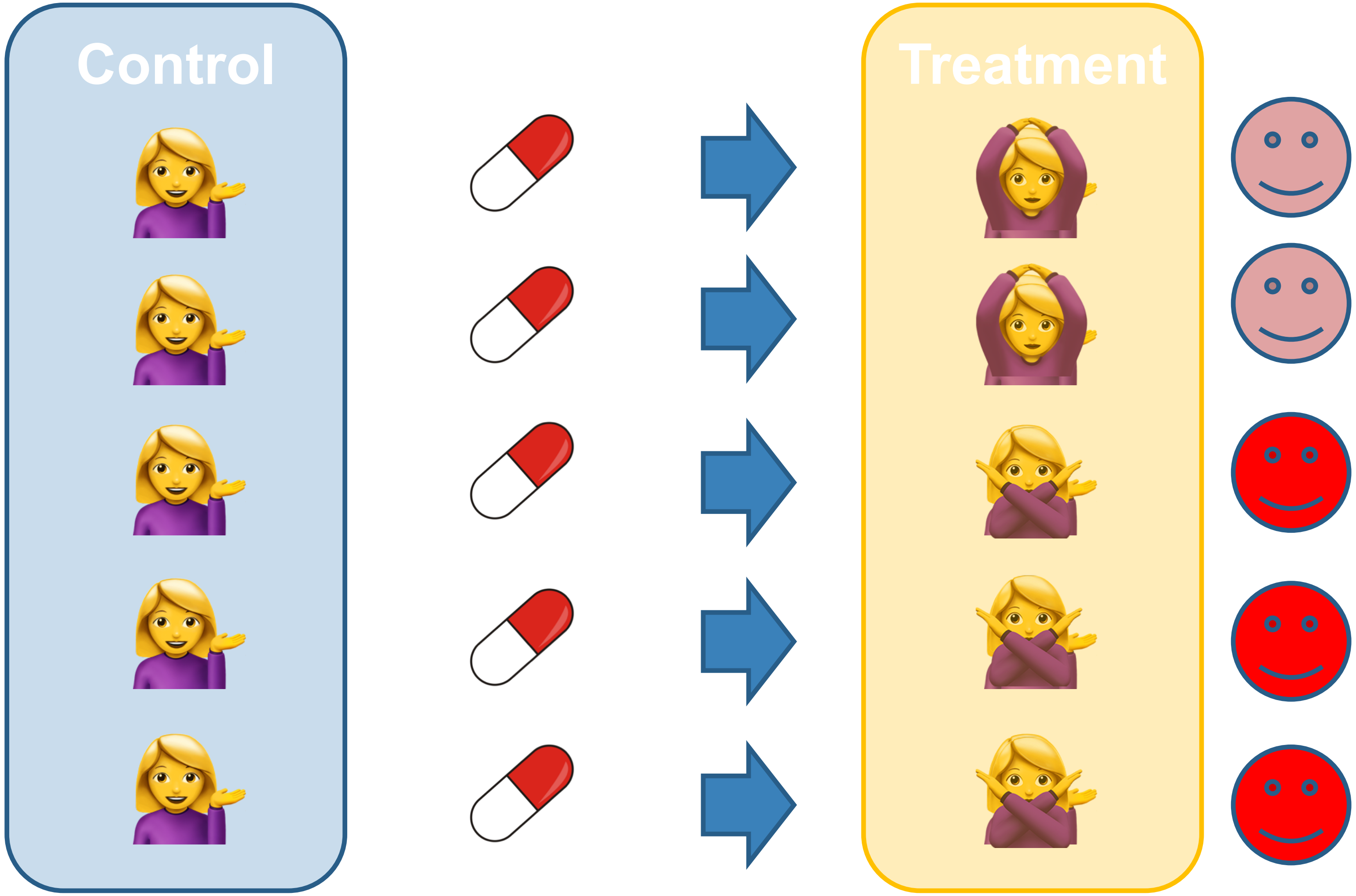


- What's the problem?

The ideal experiment



Intent to treat



intent-to-treat (ITT)

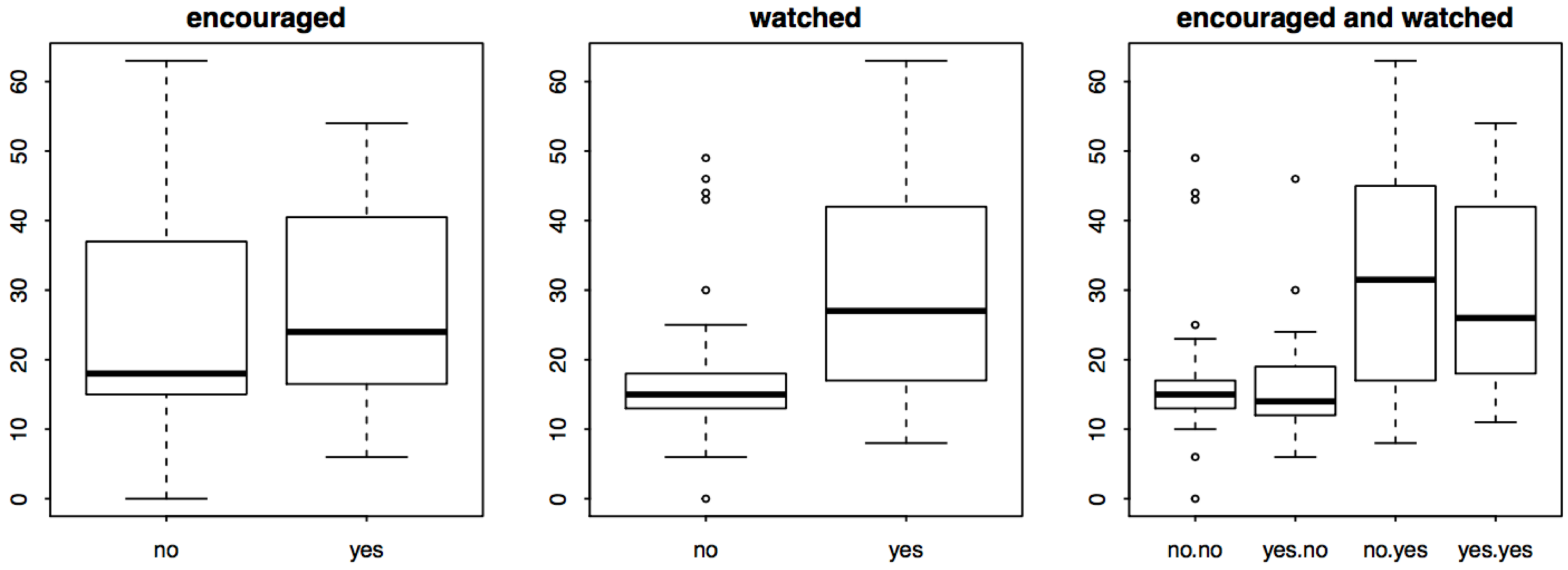
- Comparison of treated and control based on assignment only.
- Compliance is not considered leading to simpler analysis.
- May reflect the reality
  - A drug is 100% effective if taken but always 50% of patients will not take the drug, then maybe that should be the population effect.
  - In extreme case, if a drug has negative effect, removing dropouts will make the result seem effective.
- But if the sampling strategy does not reflect what will happen in population, interpretation of the result is unclear.

Effect of Sesame Street on letter recognition

- Does watching Sesame Street have an effect on letter recognition?
- Let’s think of an experiment you might conduct.
- Randomized experiment on preschool children.
  - Treatment: watching Sesame Street
  - Control: not watching
- What might be the problem with this design?

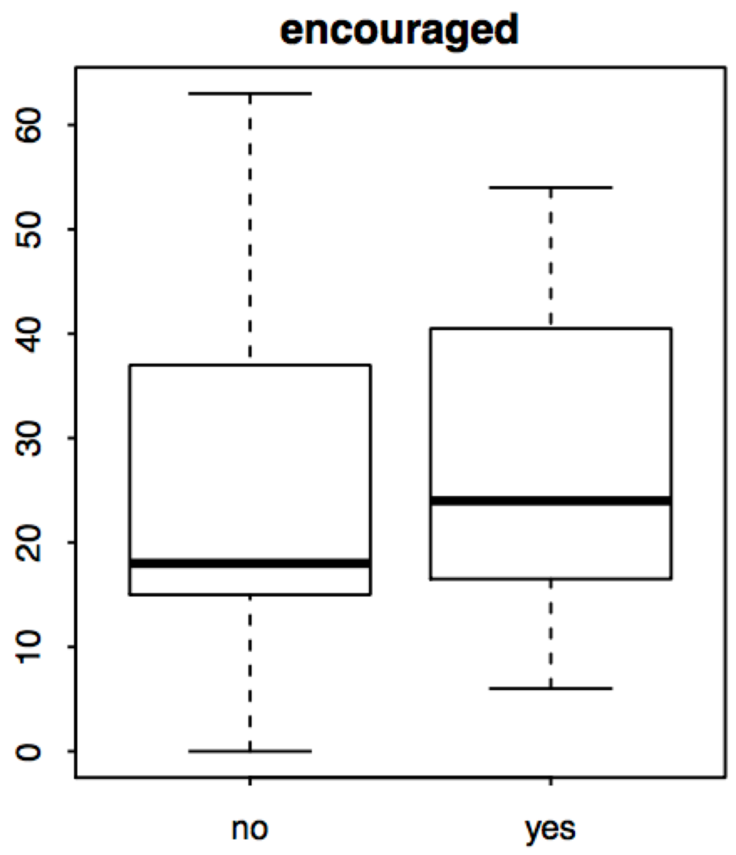
Randomized Encouragement Design

- Researchers cannot force kids to watch or not watch the show.
- Therefore, we cannot randomize watching.
- Instead, children were randomly *encouraged* to watch.



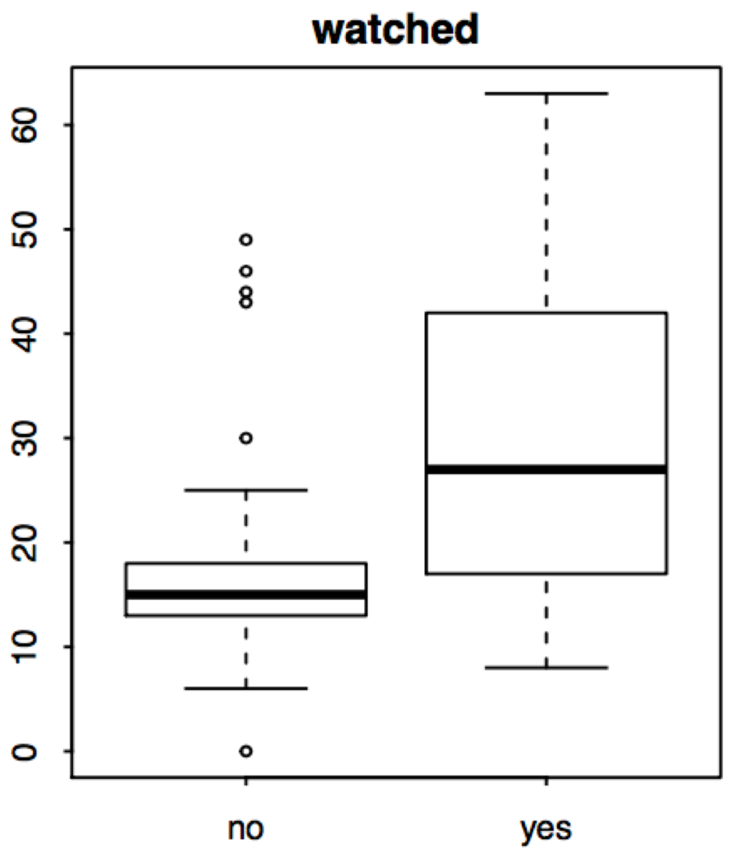
Encouraged vs not encouraged

- The simplest analysis is to compare the encouraged vs the none encouraged group.
- Inevitably this is going to have smaller effect size since we have both watchers and non-watchers in both groups.
- However, in terms of generalizability of the effect on encouragement, this is easier to argue.



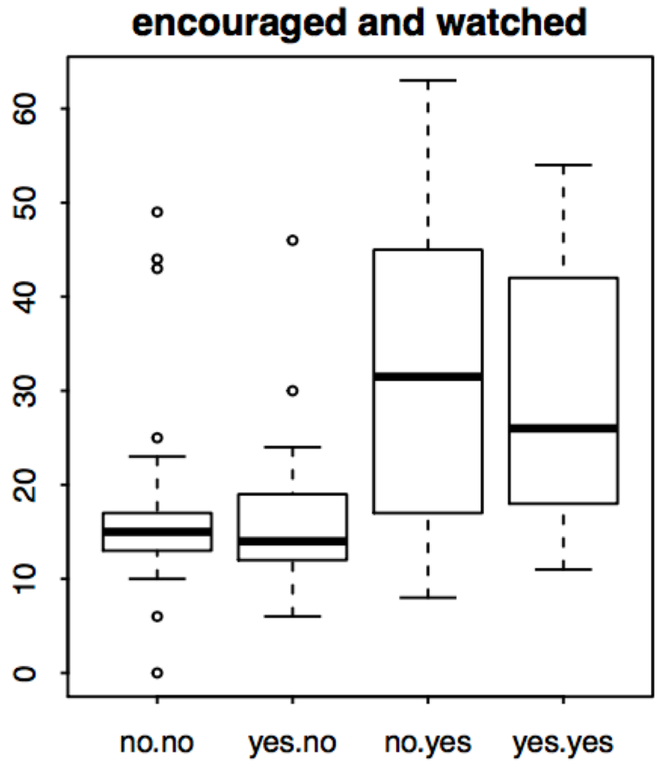
If we knew who watched

- Comparison will allow us to infer the association between watching and the outcome.
- However, it is not possible to draw causal conclusions from such analysis.



Focusing on subset of children

- never watcher
- Always watcher
- Induced watcher.



Instrumental variables estimate

- To formalize the Local average treatment effects to go beyond ITT, we introduce instrumental variable.
- Which require some assumptions
- Ignorability of the instrument with respect to the potential outcomes
- Nonzero association between instrument and treatment variable
- Monotonicity:
  - No children who would watch if they were not encouraged but who would NOT watch if they were encouraged.
- Exclusion restrictions:
  - Those children whose behavior would not have been changed by the encouragement, there is no effect of encouragement on outcome

Hypothetical example

Unit i	T_i^0	T_i^1	Potential viewing outcomes	Encouragement indicator z_i	Potential test outcomes y_i^0	y_i^1	Encouragement effect $y_i^1 - y_i^0$
1	**0**	1	(induced watcher)	0	**67**	76	9
2	**0**	1	(induced watcher)	0	**72**	80	8
3	**0**	1	(induced watcher)	0	**74**	81	7
4	**0**	1	(induced watcher)	0	**68**	78	10
5	**0**	0	(never-watcher)	0	**68**	68	0
6	**0**	0	(never-watcher)	0	**70**	70	0
7	**1**	1	(always-watcher)	0	**76**	76	0
8	**1**	1	(always-watcher)	0	**74**	74	0
9	**1**	1	(always-watcher)	0	**80**	80	0
10	**1**	1	(always-watcher)	0	**82**	82	0
11	0	**1**	(induced watcher)	1	67	**76**	9
12	0	**1**	(induced watcher)	1	72	**80**	8
13	0	**1**	(induced watcher)	1	74	**81**	7
14	0	**1**	(induced watcher)	1	68	**78**	10
15	0	**0**	(never-watcher)	1	68	**68**	0
16	0	**0**	(never-watcher)	1	70	**70**	0
17	1	**1**	(always-watcher)	1	76	**76**	0
18	1	**1**	(always-watcher)	1	74	**74**	0
19	1	**1**	(always-watcher)	1	80	**80**	0
20	1	**1**	(always-watcher)	1	82	**82**	0

- TRUE ITT

$$\begin{aligned} \text{ITT} &= \frac{9+8+7+10+9+8+7+10+0+\cdots+0}{20} \\ &= 8.5 \cdot \frac{8}{20} + 0 \cdot \frac{12}{20} \\ &= 8.5 \cdot 0.4. \end{aligned}$$

- TRUE effect of watching * proportion of induced watcher

Local Average treatment effects (LATE)

- Percent of children that were induced to watch

```
fit.1a <- lm (watched ~ encouraged)
```

- ITT estimate

```
fit.1b <- lm (y ~ encouraged)
```

- Estimated effect of watching

```
iv.est <- coef(fit.1a)[,"encouraged"]/coef(fit.1b)[,"encouraged"]
```