Review Session 6

API-202 Sophie Hill 03/04/2022

Agenda

- Feedback

- Review:
  - Potential outcomes notation
  - Causal estimands (ATE, ATT, ATC)

- PSet 2
  - Loading in the data
  - New concepts
  - New R Skills

Feedback

Thanks to everyone who has filled out feedback surveys for the class. Here are a couple of things that were relevant for me:

- "Lack of connection between review sessions and lecture"
  - Valid! Now we have covered R basics, we will integrate concepts from lecture in the review sessions. Today is a good example!

- "Too much time spent helping individuals with R problems during review session"
  - Always a tricky balance! If you have an R issue during session, I may ask you to see me after or come to office hours to debug.

Potential outcomes notation

What is $Y_i(0)$?

- Probability that individual i has an outcome of 0
- Potential outcome for individual i under control
- Probability that individual i is assigned to control
- Potential outcome for individual i under treatment

What is $Y_i(0)$?

Probability that individual i has an outcome of 0

Potential outcome for individual i under control

Probability that individual i is assigned to control

Potential outcome for individual i under treatment

What is $Y_i(0)$?

Probability that individual i has an outcome of 0

Potential outcome for individual i under control

Probability that individual i is assigned to control

Potential outcome for individual i under treatment

What is $Y_i(1) - Y_i(0)$?

- Average treatment effect
- Average treatment effect on the treated
- Average treatment effect on the control
- Individual treatment effect

What is $Y_i(1) - Y_i(0)$?

- Average treatment effect
- Average treatment effect on the treated
- Average treatment effect on the control
- Individual treatment effect

What is $Y_i(1) - Y_i(0)$?

- Average treatment effect
- Average treatment effect on the treated
- Average treatment effect on the control
- Individual treatment effect

The average potential outcome under treatment among the units assigned to control is...

- $E[Y_i(0)|T_i = 1]$
- $E[Y_i(1)|T_i = 0]$
- $E[Y_i|T_i = 1]$
- $E[Y_i|T_i = 0]$

The average potential outcome under treatment among the units assigned to control is...

$$E[Y_i(0)|T_i = 1]$$
$$E[Y_i(1)|T_i = 0]$$
$$E[Y_i|T_i = 1]$$
$$E[Y_i|T_i = 0]$$

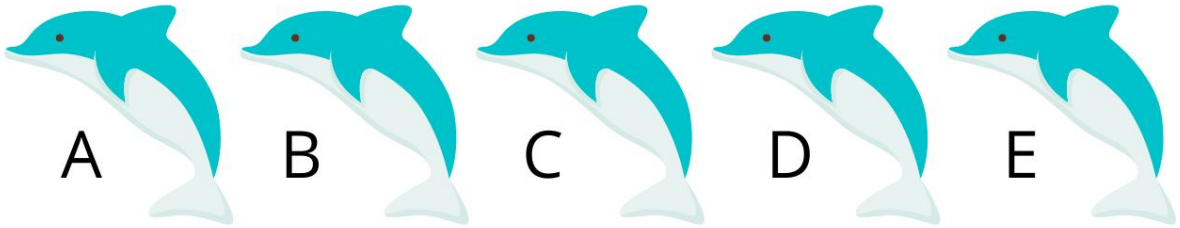
The average potential outcome under treatment among the units assigned to control is...

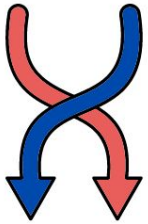
$$E[Y_i(0)|T_i = 1]$$
$$E[Y_i(1)|T_i = 0]$$
$$E[Y_i|T_i = 1]$$
$$E[Y_i|T_i = 0]$$

Causal estimands (ATE, ATT, ATC)

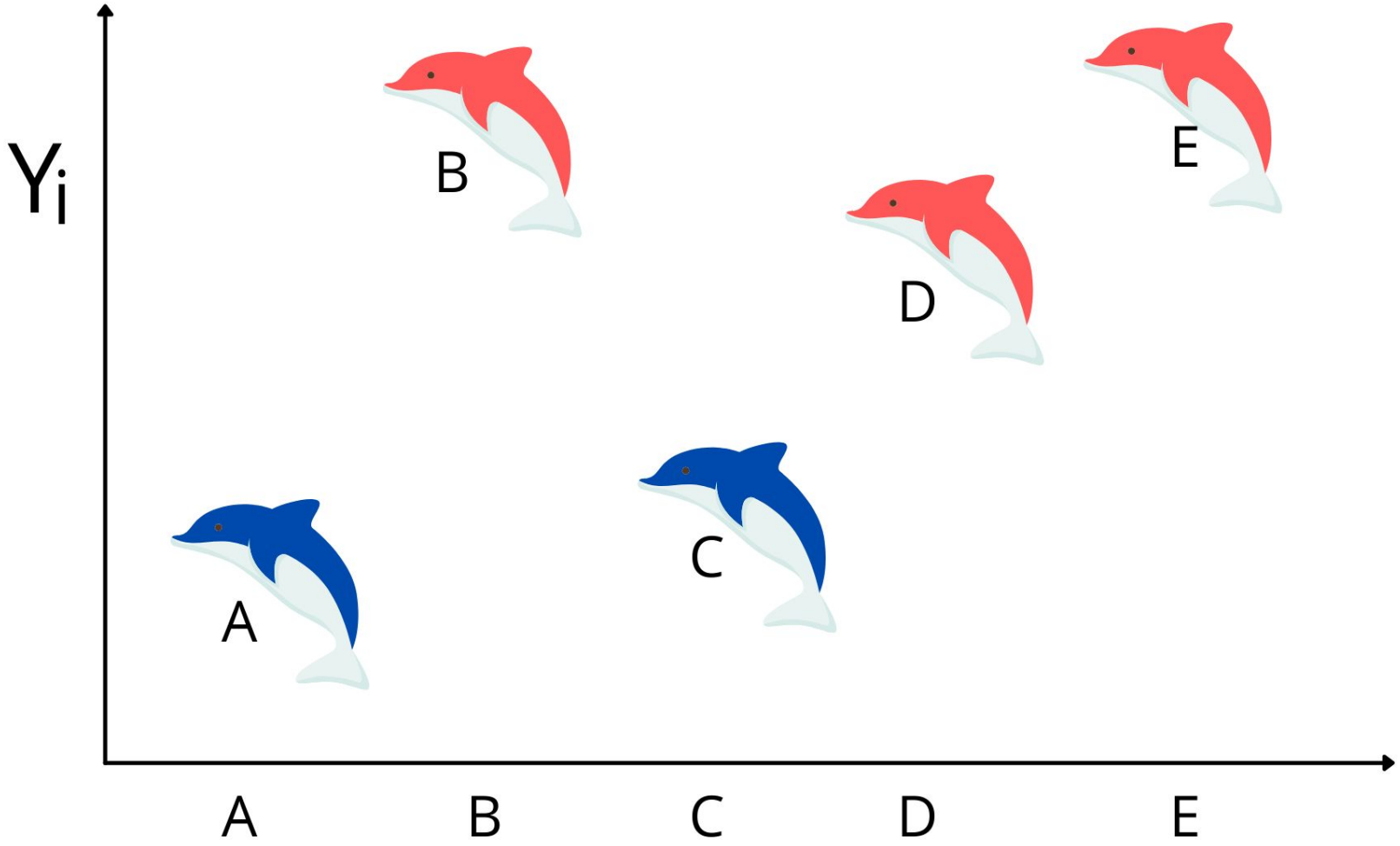
Causal estimands* (ATE, ATT, ATC)

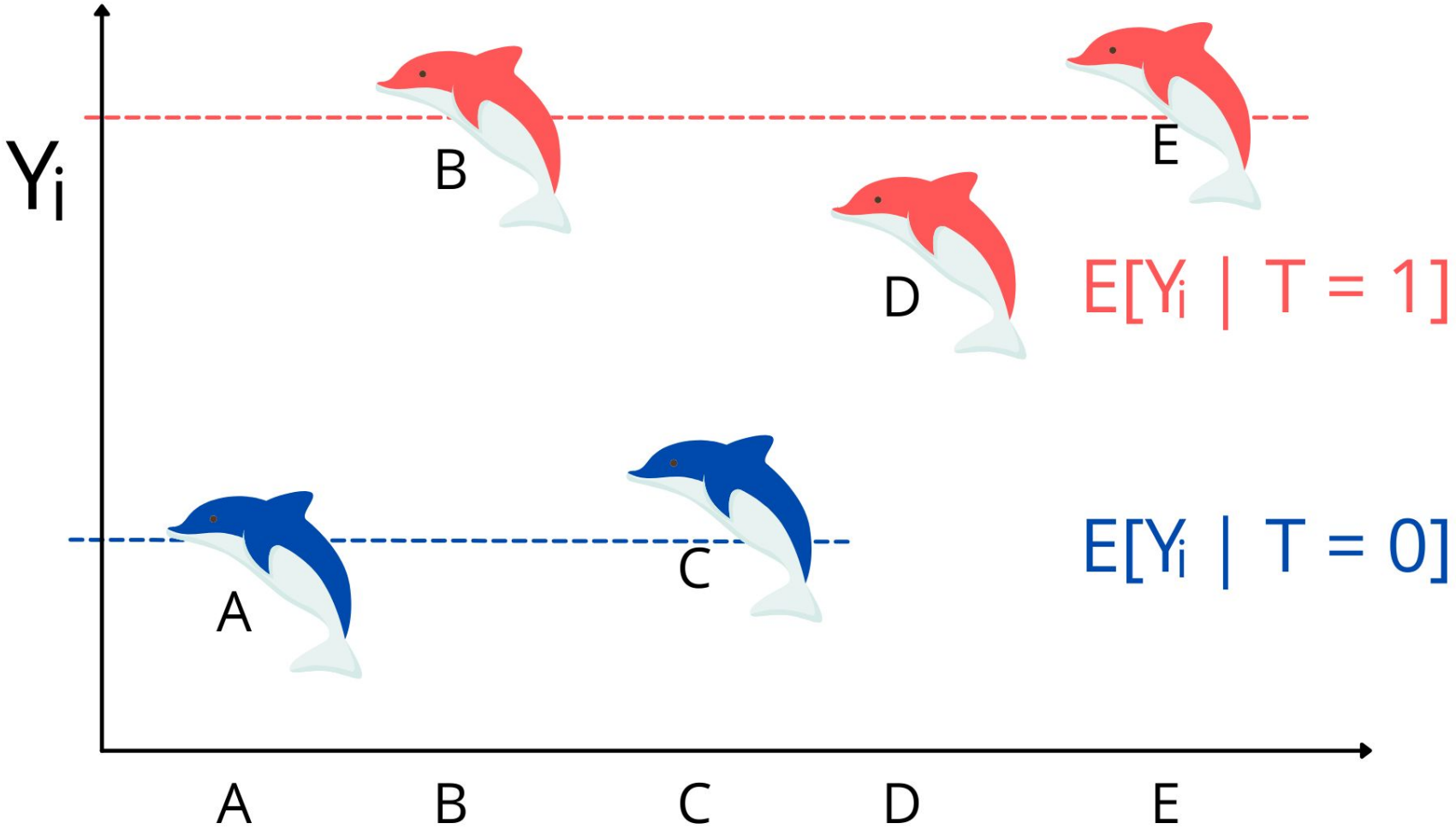
*with dolphins

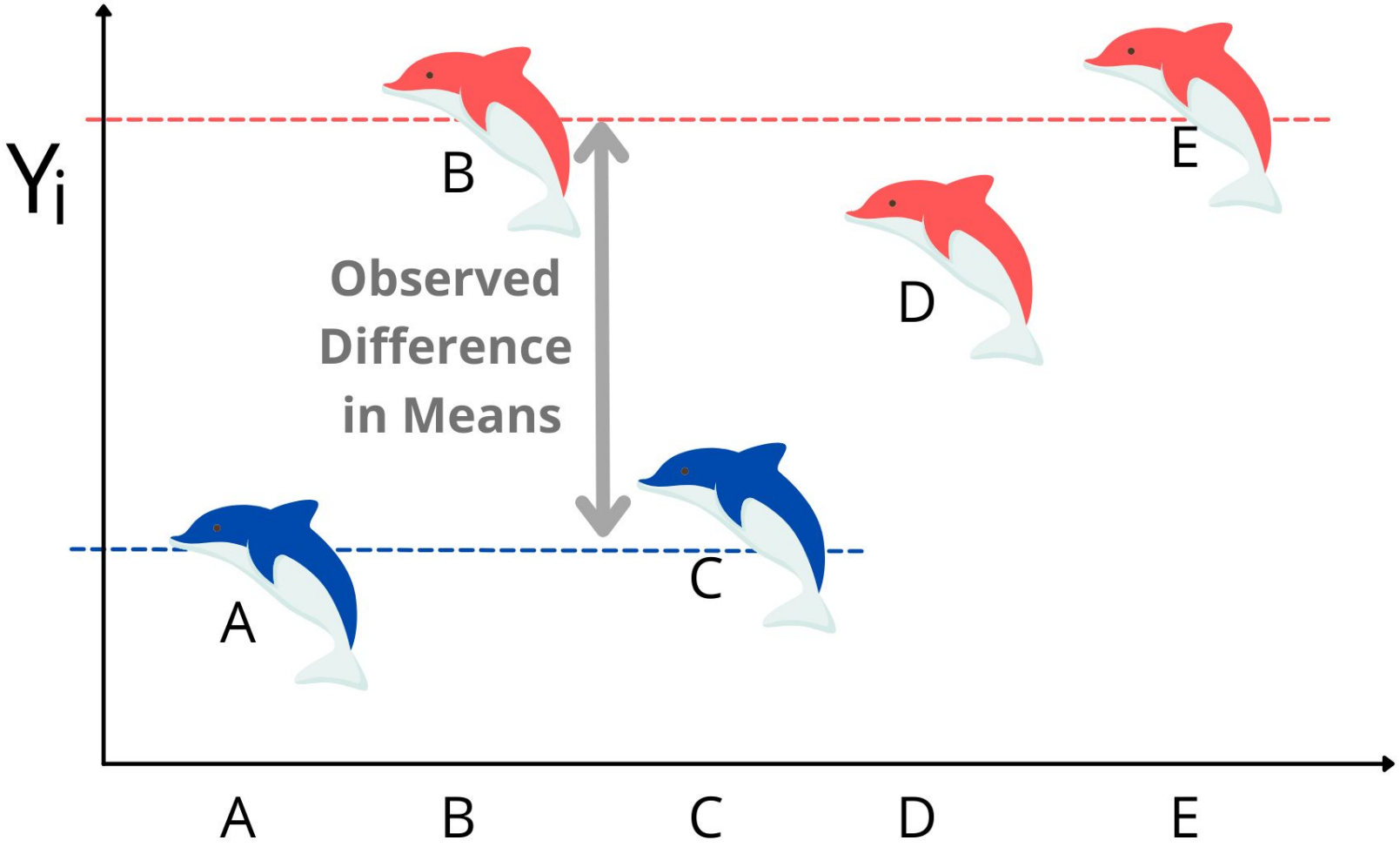




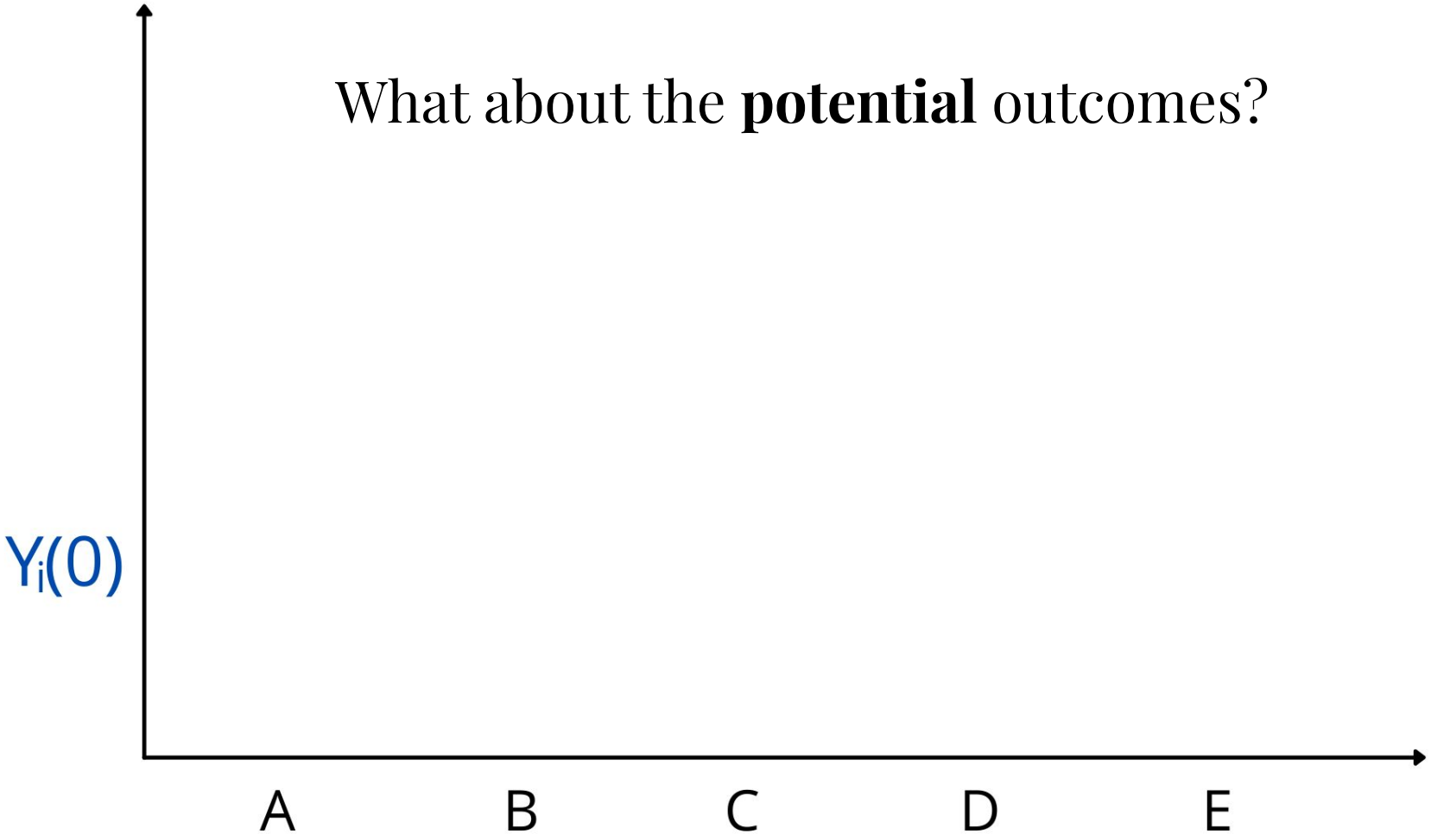
Treatment **Control**


Y_i
A
B
C
D
E
A
B
C
D
E

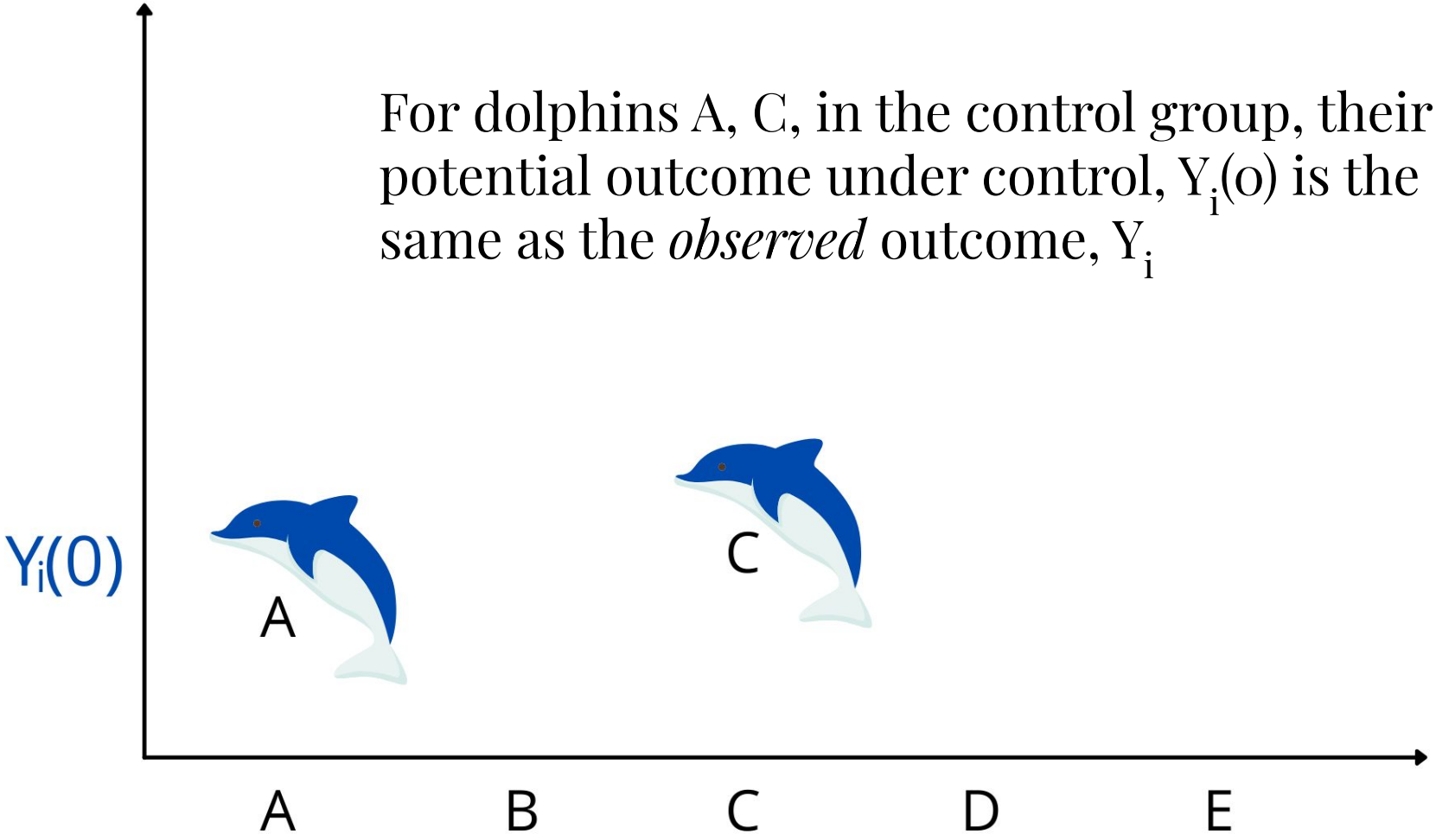

Y_i
B
E
D
$E[Y_i \mid T = 1]$
A
C
$E[Y_i \mid T = 0]$
A
B
C
D
E


Y_i
B
E
D
Observed
Difference
in Means
C
A
A
B
C
D
E

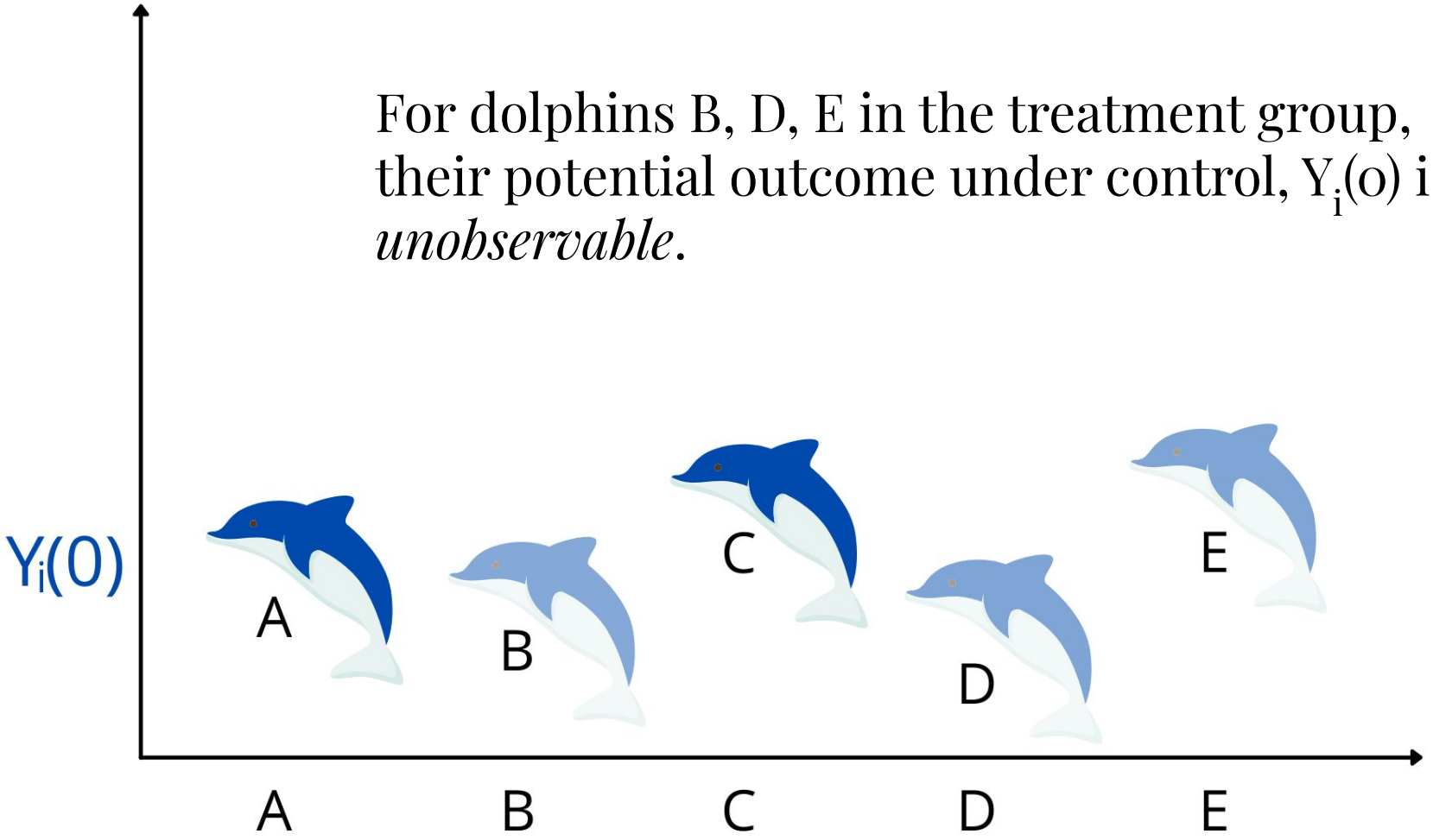
What about the **potential** outcomes?

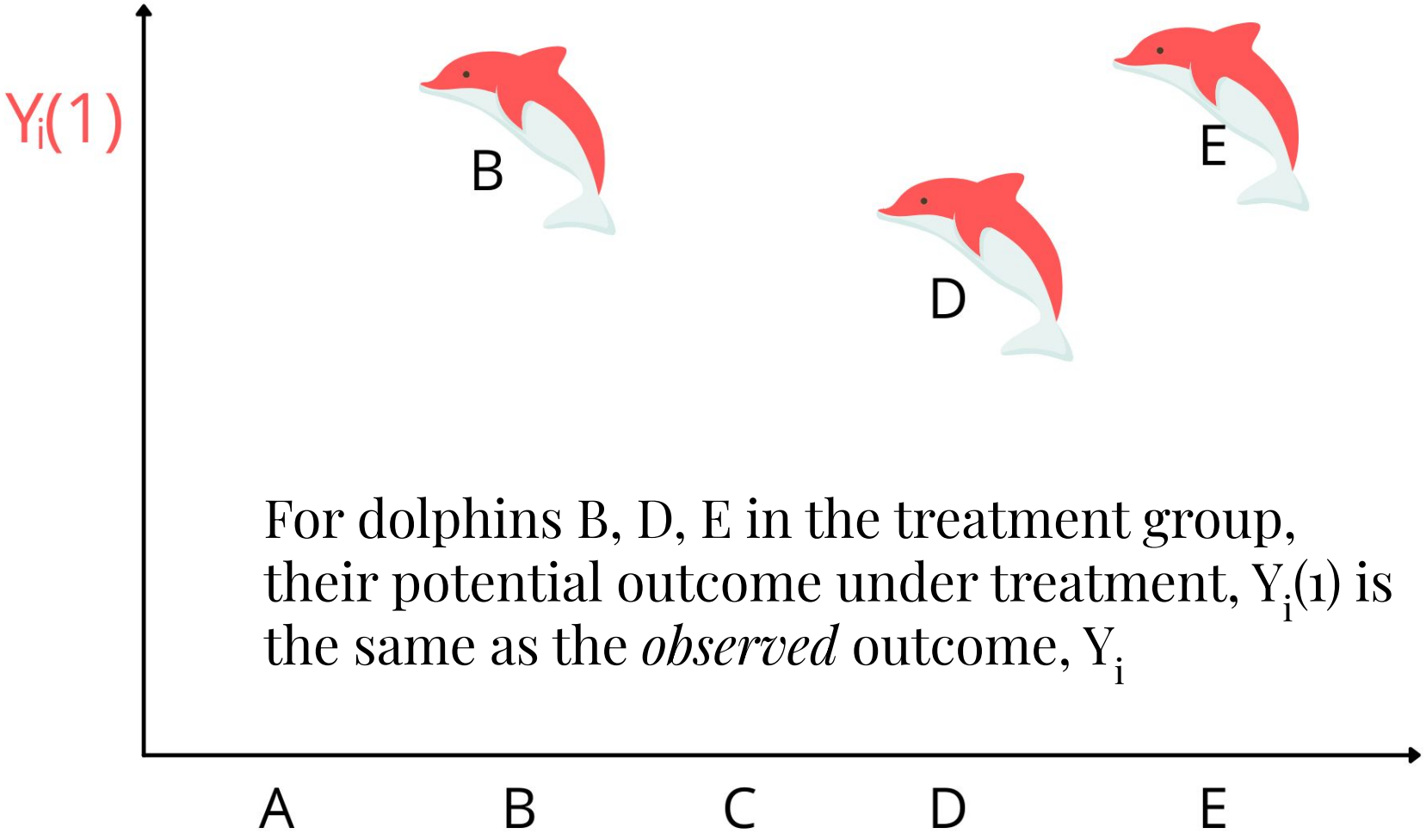


For dolphins A, C, in the control group, their potential outcome under control, $Y_i(o)$ is the same as the *observed* outcome, Y_i

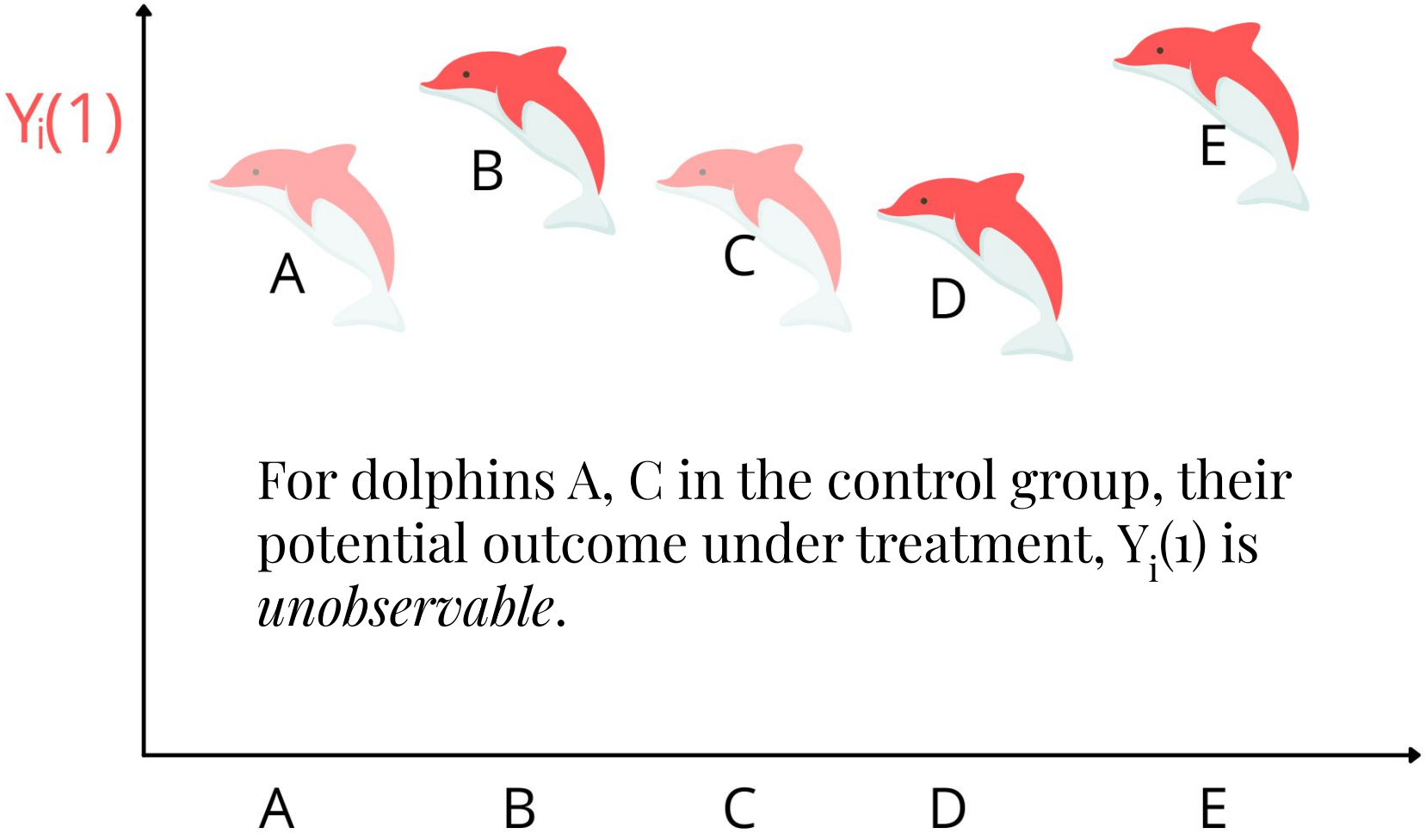


For dolphins B, D, E in the treatment group, their potential outcome under control, $Y_i(0)$ is *unobservable*.





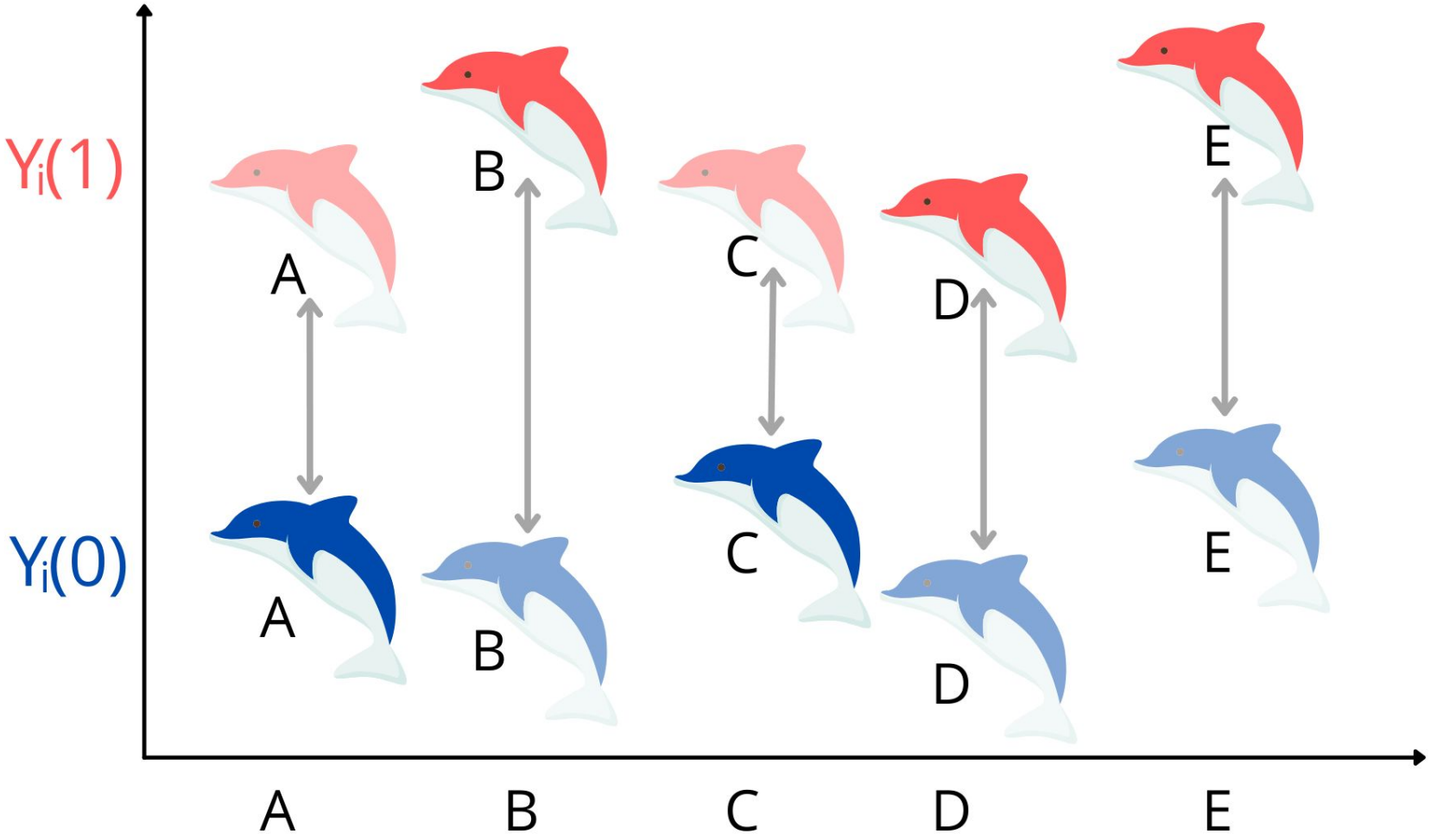
For dolphins B, D, E in the treatment group, their potential outcome under treatment, $Y_i(1)$ is the same as the *observed* outcome, Y_i



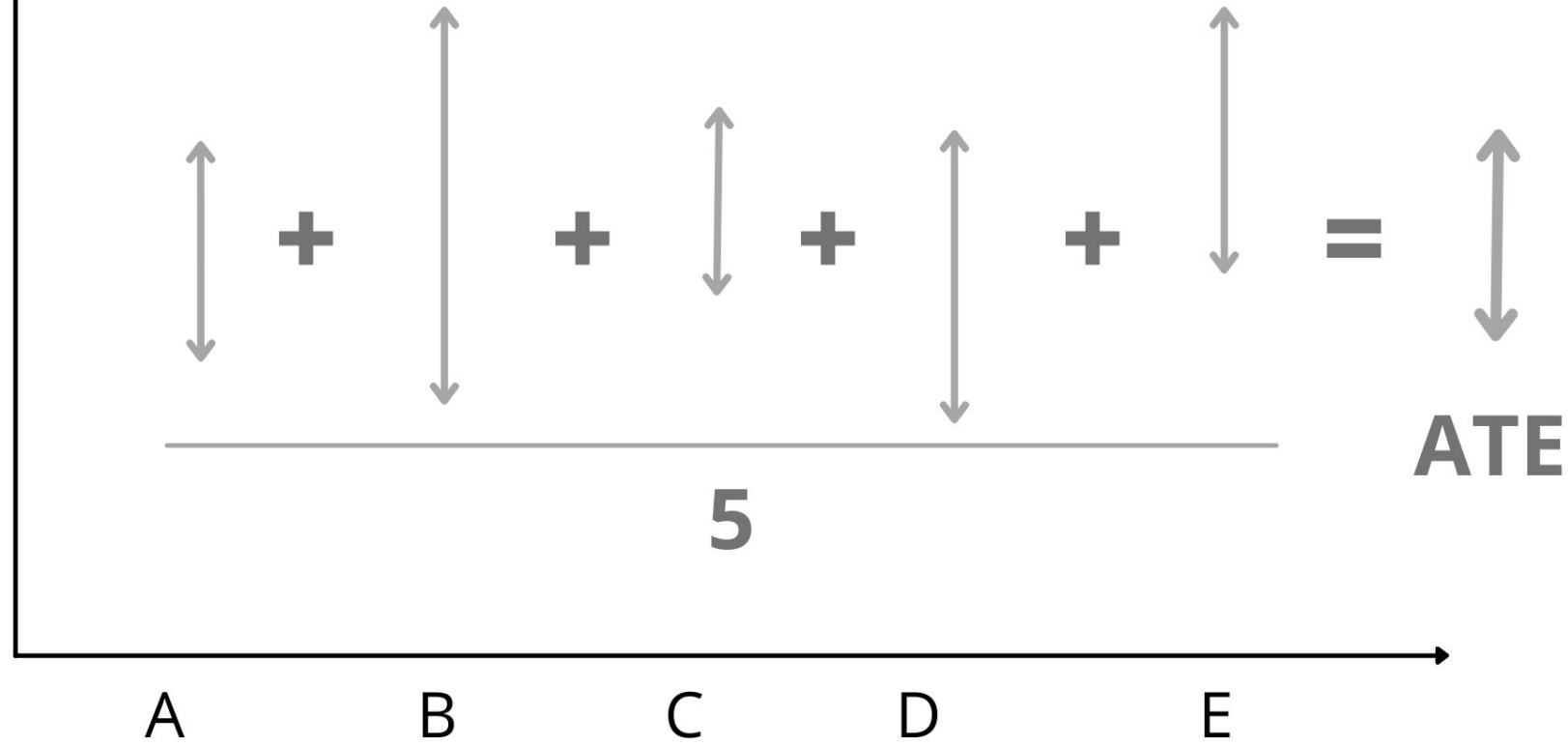
For dolphins A, C in the control group, their potential outcome under treatment, $Y_i(1)$ is *unobservable*.

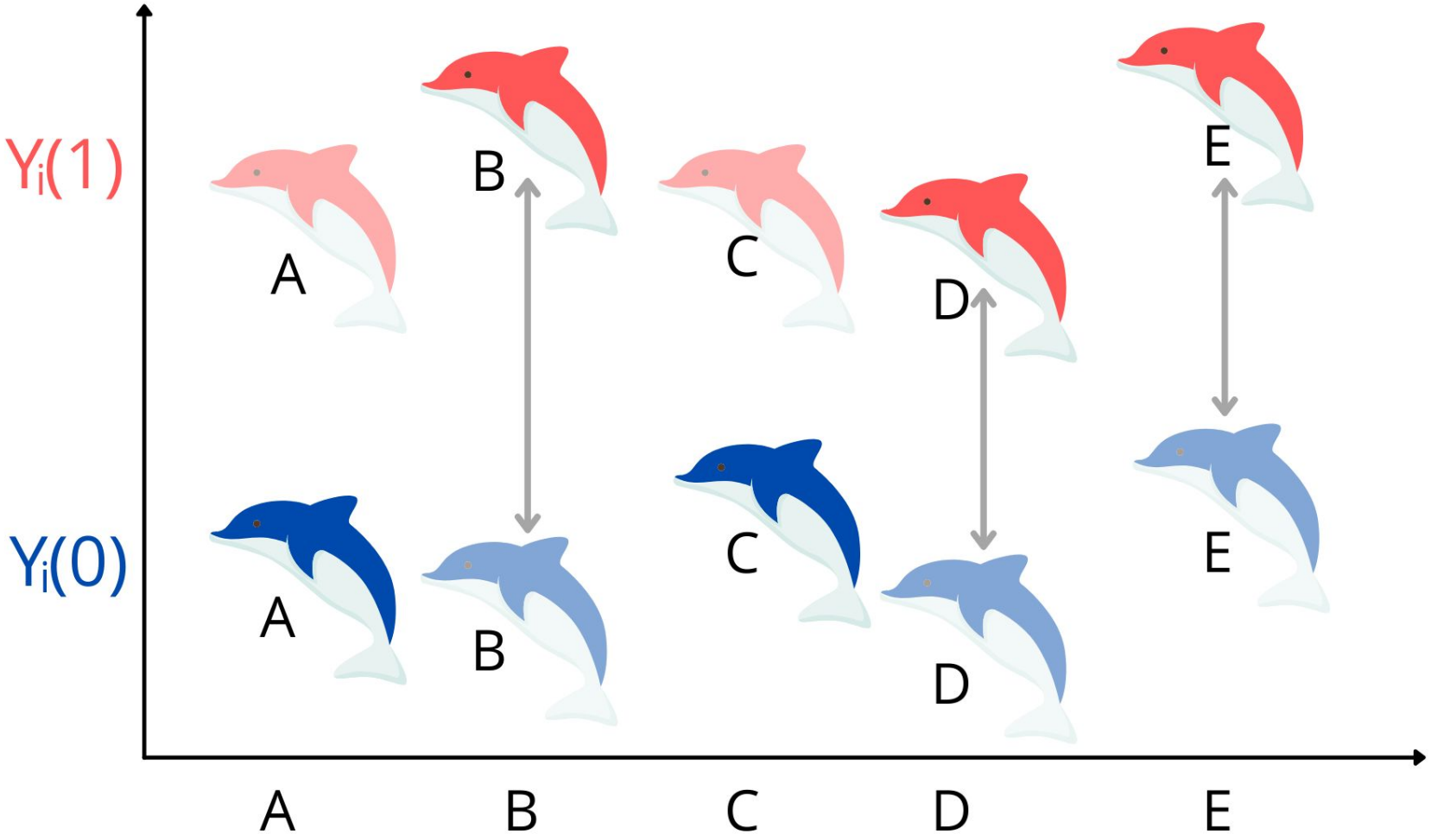
Let’s imagine for a second that we *could* observe all the potential outcomes…

… in that case, we could calculate **individual-level** treatment effects!

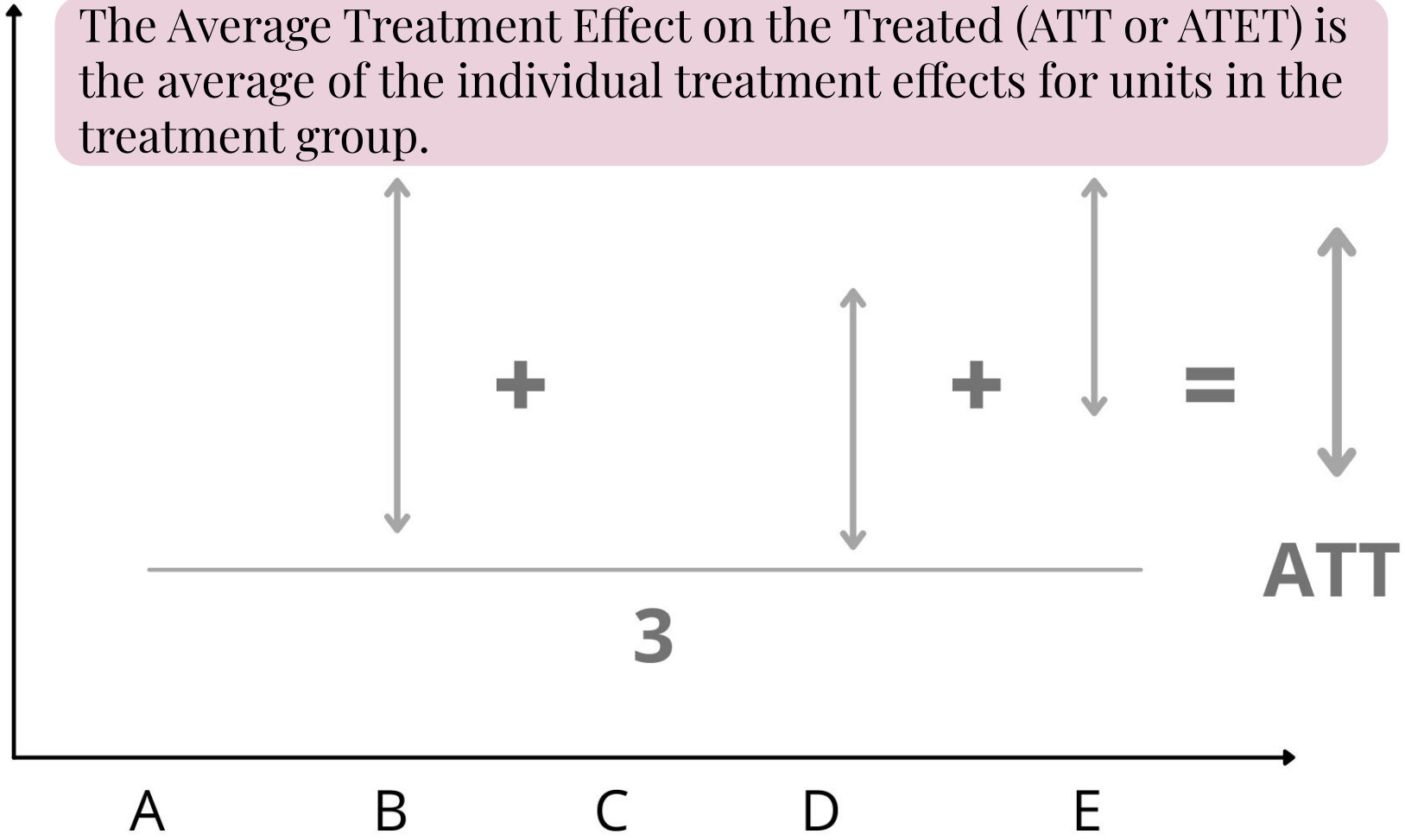

$Y_i(1)$
$Y_i(0)$
A
B
C
D
E
A
B
C
D
E
A
B
C
D
E

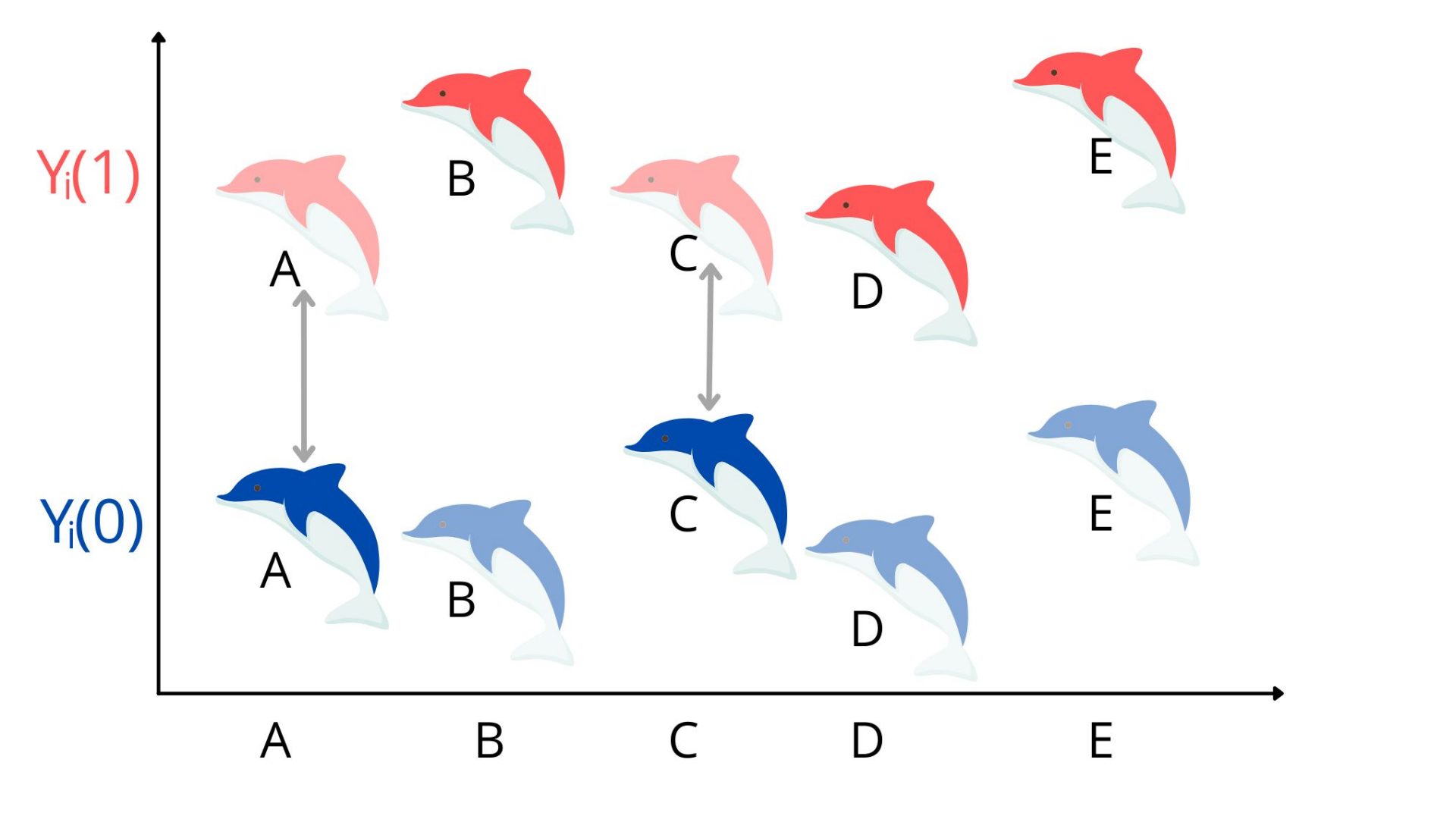
The Average Treatment Effect (ATE) is the average of all the individual treatment effects.



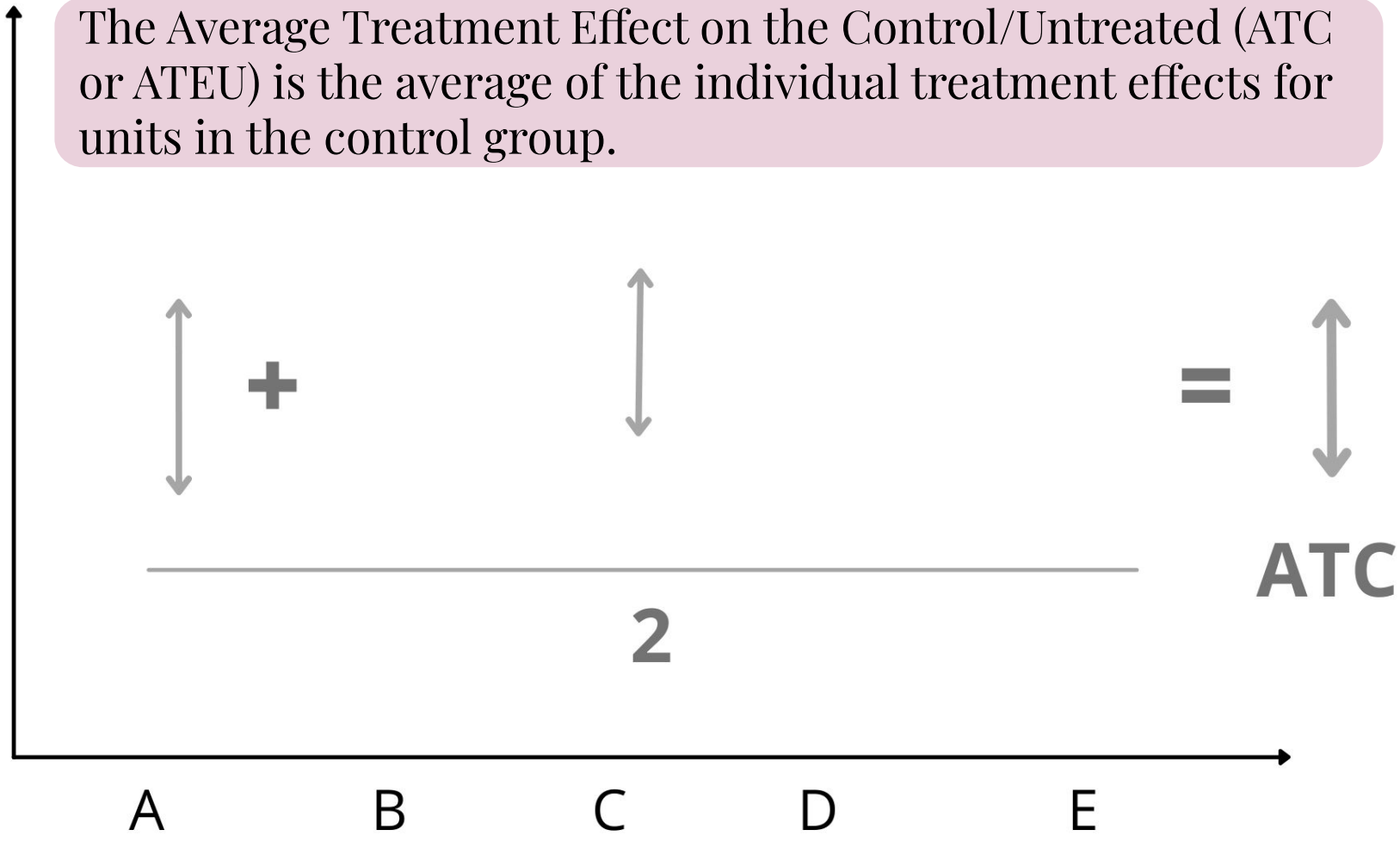

$Y_i(1)$
$Y_i(0)$
A
B
C
D
E
A
B
C
D
E
A
B
C
D
E

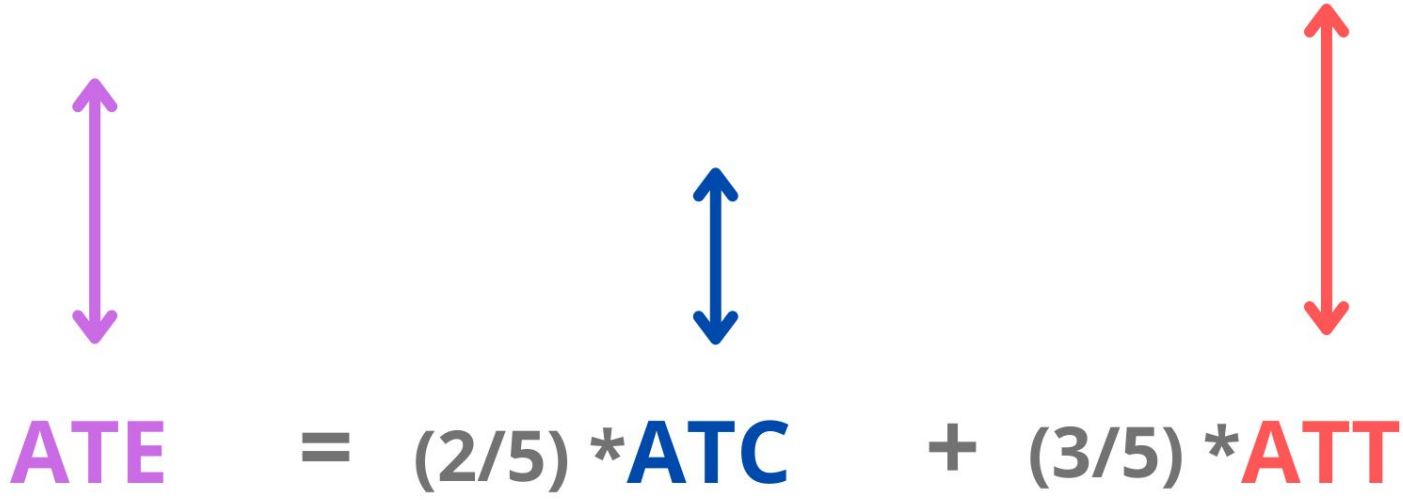
The Average Treatment Effect on the Treated (ATT or ATET) is the average of the individual treatment effects for units in the treatment group.




$Y_i(1)$
$Y_i(0)$
A
B
C
D
E
A
B
C
D
E
A
B
C
D
E

The Average Treatment Effect on the Control/Untreated (ATC or ATEU) is the average of the individual treatment effects for units in the control group.





The **Average Treatment Effect (ATE)** can be written as the weighted average of the **ATC** and the **ATT**. The weights (2/5 and 3/5) correspond to the proportion in control and the proportion in treatment.

Causal estimands: Examples

Individual-level Treatment Effects:

- What is the effect on earnings of enrolling in a job training program on *you* specifically?

Average Treatment Effect:

- What is the average effect on earnings of enrolling in a job training program across the population?

Average Treatment Effect on the Treated:

- What is the average effect on earnings of enrolling in a job training program among the unemployed?

Causal estimands: Examples

We would *love* to be able to estimate individual-level effects!!

But this requires some HUGE assumptions...

Individual-level Treatment Effects:

- What is the effect on earnings of enrolling in a job training program on *you* specifically?

Average Treatment Effect:

- What is the average effect on earnings of enrolling in a job training program across the population?

Average Treatment Effect on the Treated:

- What is the average effect on earnings of enrolling in a job training program among the unemployed?

Causal estimands: Examples

The ATE is easier to estimate, and is often useful when the intervention could, in principle be given to anyone (e.g. GOTV flyer).

Individual-level Treatment Effects:

- What is the effect on earnings of enrolling in a job training program on *you* specifically?

Average Treatment Effect:

- What is the average effect on earnings of enrolling in a job training program across the population?

Average Treatment Effect on the Treated:

- What is the average effect on earnings of enrolling in a job training program among the unemployed?

Causal estimands: Examples

The ATT is often the focus in medical research (you don't care about the effect of chemotherapy on people who don't have cancer!), or targeted policy programs.

Individual-level Treatment Effects:

- What is the effect on earnings of enrolling in a job training program on *you* specifically?

Average Treatment Effect:

- What is the average effect on earnings of enrolling in a job training program across the population?

Average Treatment Effect on the Treated:

- What is the average effect on earnings of enrolling in a job training program among the unemployed?

Can we estimate any of these quantities using the **observed** difference-in-means (ODM)?

Yes!

If there is no selection bias:

ODM = ATT

If there is no selection bias and no heterogeneous effects bias:

ODM = ATT = ATE

Definitions:

$$\text{ODM} = E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0]$$

Definitions:

$$\text{ODM} = E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0]$$

$$\text{ATE} = E[Y_i(1) - Y_i(0)]$$

Definitions:

$$\text{ODM} = E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0]$$

$$\text{ATE} = E[Y_i(1) - Y_i(0)]$$

$$\text{ATT} = E[Y_i(1) - Y_i(0) \mid T_i = 1]$$

We can decompose the observed difference in means (ODM) like this:

$$\text{ODM} = E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0]$$

We can decompose the observed difference in means (ODM) like this:

$$\begin{aligned}\text{ODM} &= E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0] \\ &= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]\end{aligned}$$

We can decompose the observed difference in means (ODM) like this:

$$
\begin{aligned}
\text{ODM} &= E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0] \\
&= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] \\
&= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] + \underbrace{(E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 1])}_{\text{Adds to zero}}
\end{aligned}
$$

We can decompose the observed difference in means (ODM) like this:

$$
\begin{aligned}
\text{ODM} &= E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0] \\
&= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] \\
&= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] + \underbrace{(E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 1])}_{\text{Adds to zero}} \\
&= (E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 1]) + (E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0])
\end{aligned}
$$

We can decompose the observed difference in means (ODM) like this:

$$
\begin{aligned}
\text{ODM} &= E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0] \\
&= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] \\
&= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] + \underbrace{(E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 1])}_{\text{Adds to zero}} \\
&= (E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 1]) + (E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]) \\
&= \underbrace{E[Y_i(1) - Y_i(0) \mid T_i = 1]}_{\text{ATT}} + \underbrace{(E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0])}_{\text{Selection bias}}
\end{aligned}
$$

How does the ATE relate to the ATT?

$$\text{ATE} = q(\text{ATT}) + (1 - q)(\text{ATC})$$

where q = proportion in treatment group

How does the ATE relate to the ATT?

$$\begin{aligned} \text{ATE} &= q(\text{ATT}) + (1-q)(\text{ATC}) \\ &= \text{ATT} - (1-q)(\text{ATT}) + (1-q)(\text{ATC}) \end{aligned}$$

where q = proportion in treatment group

How does the ATE relate to the ATT?

$$
\begin{aligned}
\text{ATE} &= q(\text{ATT}) + (1-q)(\text{ATC}) \\
&= \text{ATT} - (1-q)(\text{ATT}) + (1-q)(\text{ATC}) \\
&= \text{ATT} - \underbrace{(1-q)(\text{ATT} - \text{ATC})}_{\text{Heterogeneous effects bias}}
\end{aligned}
$$

where q = proportion in treatment group

Putting it all together:

$$\begin{aligned} \text{ODM} &= \text{ATT} + (\text{Selection bias}) \\ \text{ATT} &= \text{ATE} + (\text{Heterogeneous effects bias}) \\ \implies \text{ODM} &= \text{ATE} + (\text{Heterogeneous effects bias}) + (\text{Selection bias}) \end{aligned}$$

$$\begin{aligned} \text{ODM} &= \text{ATT} + (\text{Selection bias}) \\ \text{ATT} &= \text{ATE} + (\text{Heterogeneous effects bias}) \\ \implies \text{ODM} &= \text{ATE} + (\text{Heterogeneous effects bias}) + (\text{Selection bias}) \end{aligned}$$

This is why randomization is so important!!

Randomization means:

- no heterogeneous effects bias
- no selection bias

Can we estimate any of these quantities using the **observed** difference-in-means (ODM)?

Yes!

If there is no selection bias:

ODM = ATT

If there is no selection bias and no heterogeneous effects bias:

ODM = ATT = ATE

PSet 2: Out now!

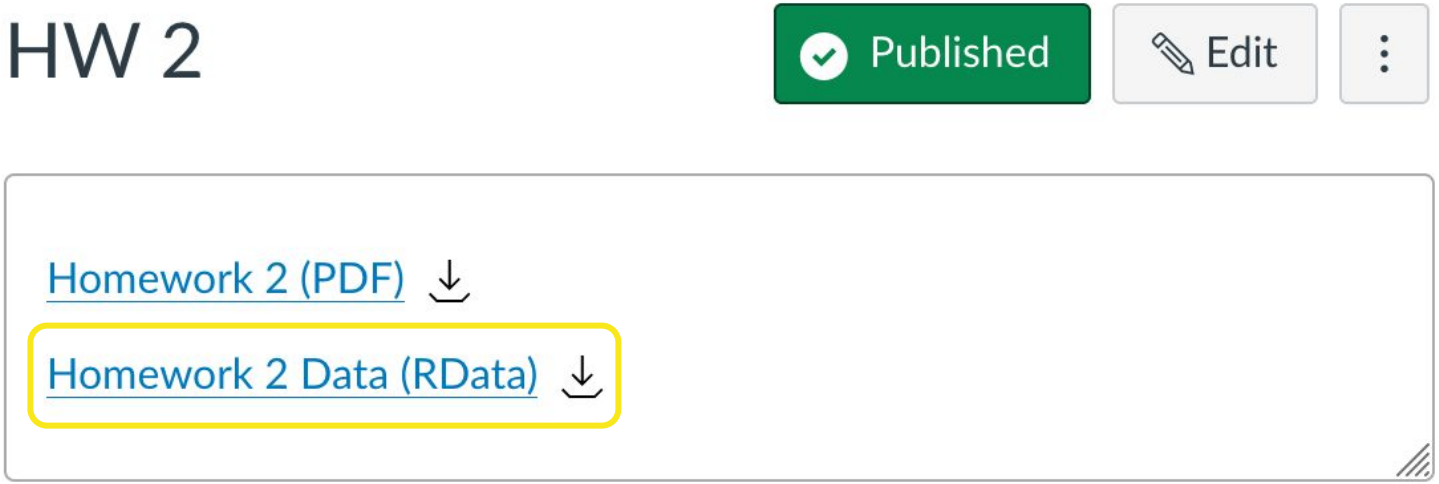
Due: Thursday 10th March 8:30am

In the problem set, you will use real data from the Cooperative Congressional Election Surveys (CCES) to investigate voting patterns in 2016 and 2020.

You will also analyze the results of an RCT..

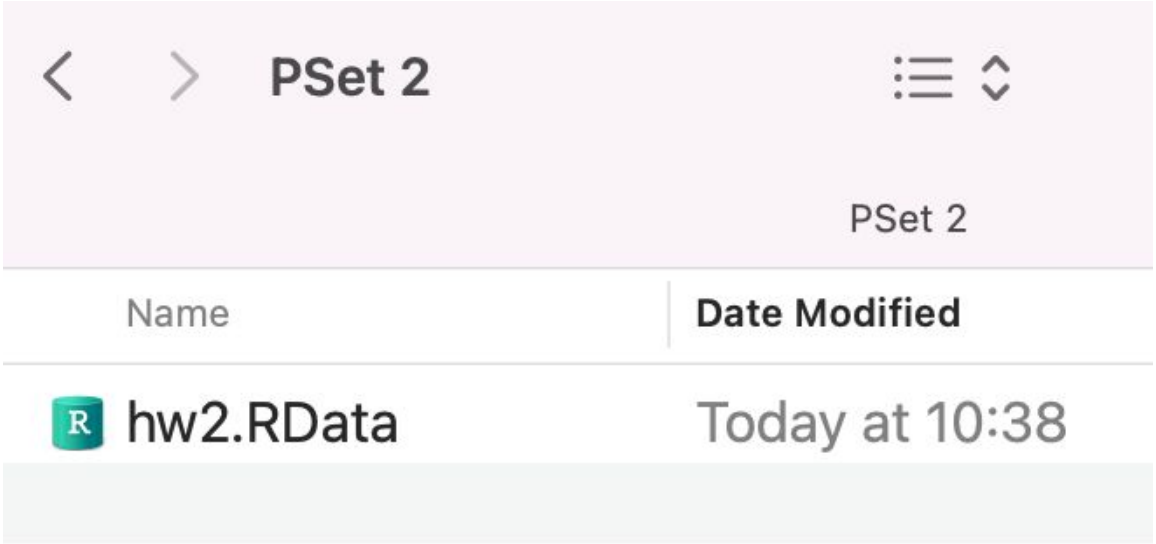
PSet 2: Set up

1. Download the RData file from Canvas



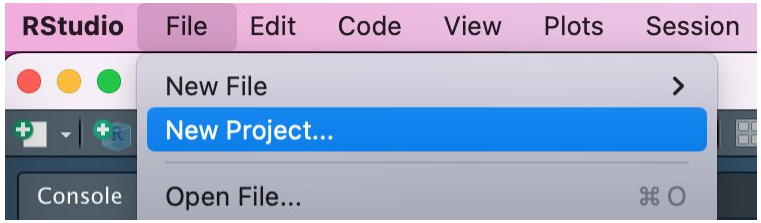
PSet 2: Set up

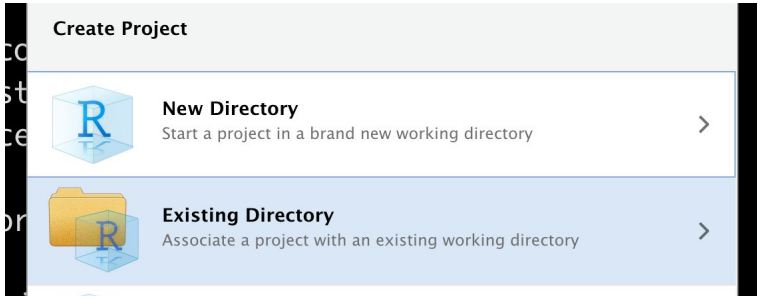
2. Save the RData file in a new folder for PSet 2

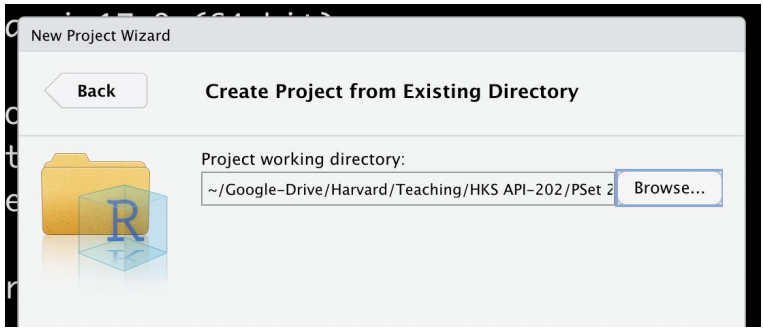


PSet 2: Set up

3. Open RStudio, File > New Project > Existing Directory > Pick the folder you created in step (2)



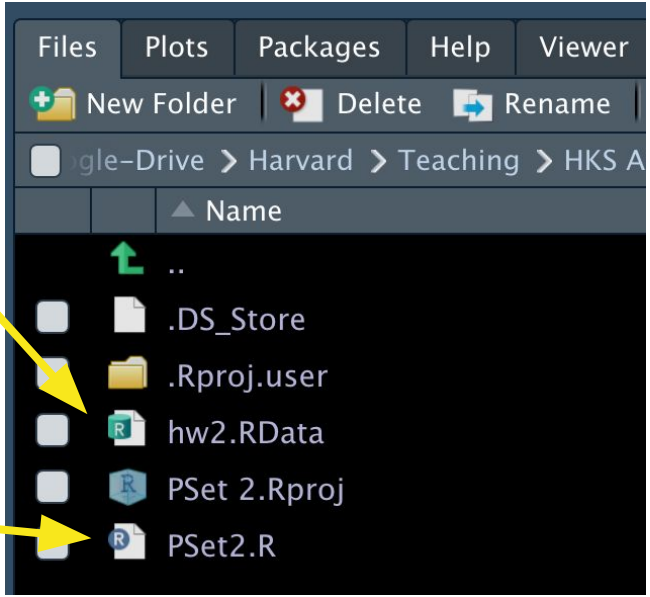




PSet 2: Set up

4. RStudio will open up this new project. You should be able to see the RData file in the Files pane.

Go File > New File > R Script to open up a new script. Save it as “PSet2.R” (or whatever). You can now see it in the Files pane too!



PSet 2: Set up

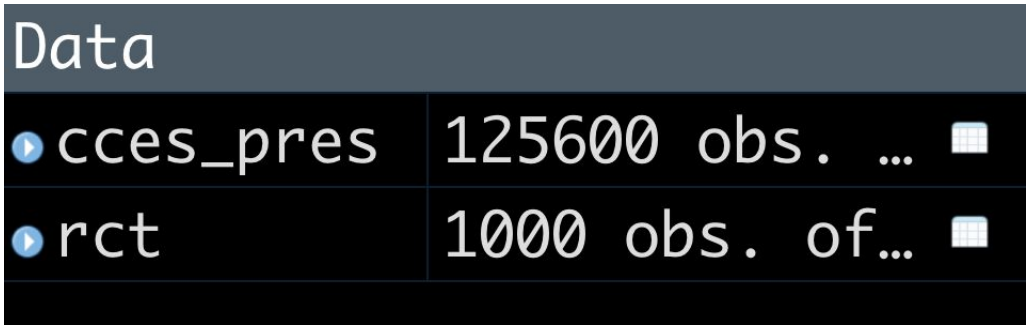
5. In the script file, paste in the code provided in PSet 2:

```
library(tidyverse) # load tidyverse package
options(scipen=4) # set numbers to print as decimals
theme_set(theme_bw()) # set plot theme

load("hw2.RData") # load RData object
```

PSet 2: Set up

After loading "hw2.RData", you will see **two** datasets appear in the Environment pane: `cces_pres` and `rct`.



RData is R's own data format and you can save multiple objects (datasets, values, model output) in a single RData file. Cool!

PSet 2: What you need to know

This stuff should be familiar-ish

Concepts:

- Survey weights
- Bivariate regression

R Skills:

- Creating new variables
- Creating group-level averages using group_by() and summarize()

PSet 2: What you need to know

Concepts:

- Survey weights
- Bivariate regression

R Skills:

- Creating new variables
- Creating group-level averages using group_by() and summarize()

- Multiple regression
- Linear probability model

- Using survey weights to get weighted frequencies
- Merging two datasets using left_join()

This stuff is new-ish

New concept: Multiple regression

Bivariate regression looks like this:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_i + + \epsilon_i$$

We estimate it in R like this:

```
lm(y ~ x, data = dat)
```

We interpret the coefficient $\boldsymbol{\beta}_1$ like this:

"a 1 unit increase in X_i is associated with a $\boldsymbol{\beta}_1$ increase in Y_i"

New concept: Multiple regression

Multiple regression, with more than one predictor, looks like this:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{ii} + \boldsymbol{\beta}_2 X_{2i} + \boldsymbol{\beta}_3 X_{3i} + \epsilon_i$$

We estimate it in R like this:

```
lm(y ~ x1 + x2 + x3, data = dat)
```

We interpret the coefficient $\boldsymbol{\beta}_1$ like this:

"holding X_2 and X_3 constant, a 1 unit increase in X_i is associated with a $\boldsymbol{\beta}_1$ increase in Y_i"

New concept: Multiple regression

Section B / Class 11 / Slide #27

Coefficients in multiple regression

The coefficient size for each predictor variable depends on the other variables in the model

- The coefficient for each predictor variable is the predicted change in the mean of Y associated with a one unit difference in X, **given the other predictors variables in the model** ("holding other predictors constant")
- If predictor variables are associated with each other and the outcome, then coefficients will change depending on what is included in the model

Section C / Class 6 / Slide #24

Multivariate regression analysis

- Extending to multivariate setting,

The slope coefficient $\hat{\beta}_1$ tells us the average change in Y associated with a one-unit increase in X_1, holding X_2 constant.

- $\hat{\beta}_1$ measures the relationship between "the part of X_1 unexplained by X_2" and "the part of Y unexplained by X_2." ($\hat{\beta}_2$ can be interpreted similarly, switching "1" and "2")

New concept: Linear Probability Model

The Linear Probability Model is the same as the standard linear model you've seen, except the outcome Y_i is **binary (0/1)** rather than continuous:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

New concept: Linear Probability Model

The Linear Probability Model (LPM) is the same as the standard linear model you've seen, except the outcome Y_i is **binary (0/1)** rather than continuous: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$

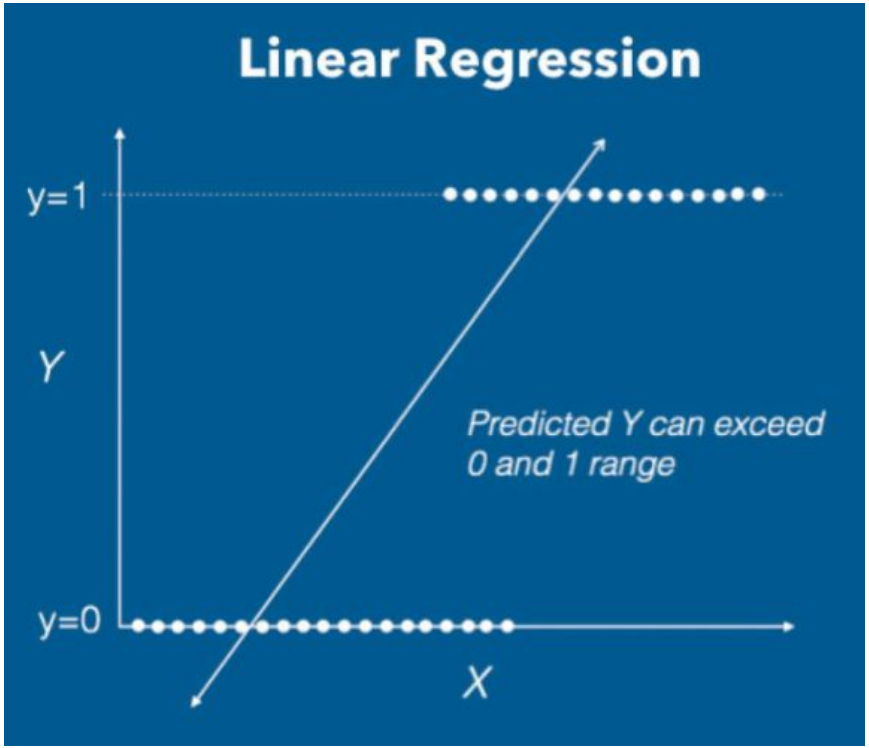
The coefficients from a LPM can be interpreted in terms of **percentage points**. For example, if $\beta_1 = 0.15$, then we could say:

"holding X_2 and X_3 constant, a 1-unit increase in X_1 is associated with a 15 percentage point increase in the probability of Y = 1"

New concept: Linear Probability Model

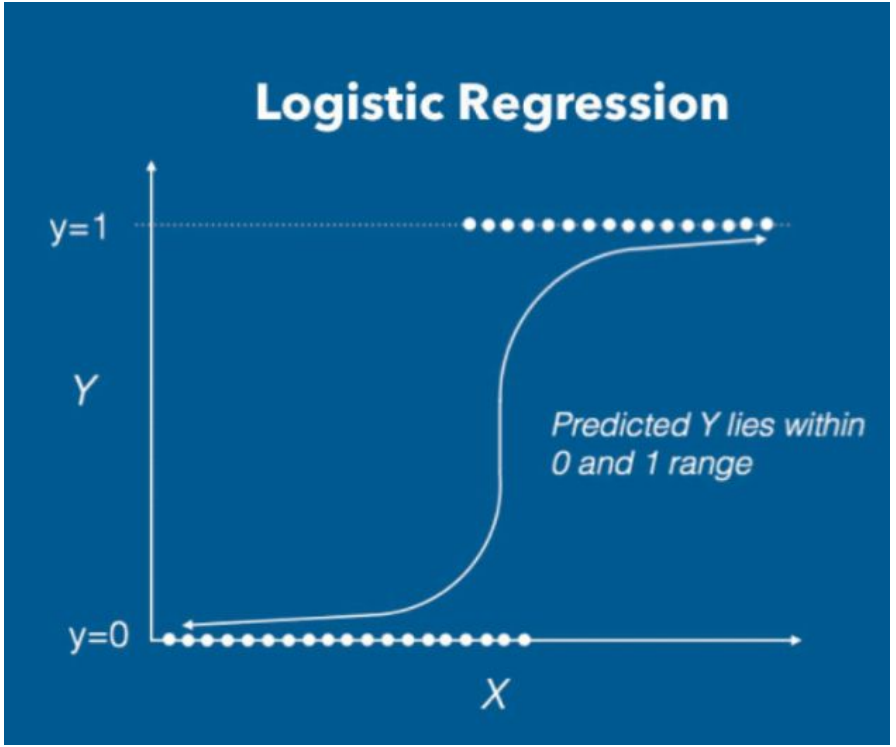
The predicted values from the Linear Probability Model can be interpreted as the predicted probability that Y = 1.

Note: it is possible to get a predicted probability from a LPM of **less than 0** or **more than 1**, which obviously doesn't make sense.



New concept: Linear Probability Model

Later in the course we will learn about logistic regression, which fixes this problem by stopping the predicted probabilities from going outside of [0, 1].



New R Skill: weighted frequencies

This code chunk is provided in PSet 2:

```
cces_pres %>%
    filter(year == 2020) %>%
    count(pid3, wt = weight) %>%
    mutate(n = round(n, 1),
      pct=scales::percent(prop.table(n)))
```

New R Skill: weighted frequencies

```
cces_pres
```

```
# A tibble: 125,600 × 24
    year   case_id weight st      age
   <dbl>     <dbl>  <dbl> <fct> <dbl>
 1  2016 222168628  1.34  NH       47
 2  2016 273691199  1.18  LA       22
 3  2016 284214415  0.217 MO       52
 4  2016 287557695  0.532 AL       28
 5  2016 290387662  1.26  CO       34
 6  2016 290932100  0.531 AL       53
 7  2016 292860642  1.45  TX       54
 8  2016 295367942  1.69  PA       25
 9  2016 295717127  2.40  GA       53
10  2016 295859014  0.888 PA       59
# … with 125,590 more rows, and 13 mor
```

New R Skill: weighted frequencies

```
cces_pres %>%
filter(year == 2020)
```

```
# A tibble: 61,000 × 24
    year    case_id weight st      age
   <dbl>      <dbl>  <dbl> <fct> <dbl>
 1  2020 1232318697  0.783 CT       54
 2  2020 1231394915  1.34  FL       65
 3  2020 1232451503  0.404 IA       74
 4  2020 1232494791  0.958 MA       58
 5  2020 1232494829  0.195 IL       53
 6  2020 1232473675  1.06  OH       59
 7  2020 1232514745  0.756 VA       70
 8  2020 1232532211  1.16  SD       73
 9  2020 1232856347  0.845 OH       50
10  2020 1232498693  0.920 NY       57
# … with 60,990 more rows, and 13 more
```

New R Skill: weighted frequencies

```
cces_pres %>%
filter(year == 2020) %>%
count(pid3, wt = weight)
```

```
# A tibble: 5 × 2
  pid3              n
  <fct>         <dbl>
1 Independent 16198.
2 Democrat    20387.
3 Republican  17380.
4 Not Sure     4906.
5 Other        2129.
```

New R Skill: weighted frequencies

```
cces_pres %>%
filter(year == 2020) %>%
count(pid3, wt = weight) %>%
mutate(n = round(n, 1),
  pct=scales::percent(prop.table(n)))
```

```
# A tibble: 5 × 3
  pid3              n pct
  <fct>         <dbl> <chr>
1 Independent 16198. 26.6%
2 Democrat    20387. 33.4%
3 Republican  17380. 28.5%
4 Not Sure     4906. 8.0%
5 Other        2129  3.5%
```

New R Skill: merging two datasets

R has several functions* for merging two datasets. These are equivalent to the VLOOKUP() function in Excel.

- inner_join(x, y): keep rows with matches in both x and y
- left_join(x, y): keep rows with matches in x or both x and y
- right_join(x, y): keep rows with matches in y or both x and y
- full_join(x, y): keep all rows

*You need to load `tidyverse` in order to use these functions. They are in the `dplyr` package, which is part of the `tidyverse` suite.

New R Skill: merging two datasets

This code is given to you in PSet 2:

```
cces_pres <- cces_pres %>% left_join(rct)
```

Since the pipe operator pipes the object into the first slot of a function, this is equivalent to:

```
cces_pres <- left_join(cces_pres, rct)
```

New R Skill: merging two datasets

R will merge the data based on columns with the same name.

There is one shared column between `cces_pres` and `rct`: it is `case_id`, which uniquely identifies each respondent.

```
> unique(length(cces_pres$case_id))
[1] 125600
> unique(length(rct$case_id))
[1] 1000
```

There are 125,600 unique IDs in `cces_pres` and 1,000 unique IDs in `rct`.

New R Skill: merging two datasets

So by using `left_join(cces_pres, rct)`, we are saying:

- Start with the `cces_pres` dataset
- Add a new column called `treat`, which is the only column in `rct` besides the ID column `case_id`
- Fill in the values of `treat` from `rct` by matching on `case_id`
- For the CCES respondents who **weren't** in the experiment, put `NA` in the column `treat`

New R Skill: merging two datasets

`cces_pres` has 125,600 rows and 24 columns

`rct` has 1,000 rows and 2 columns

If we use `left_join()`, we end up with 125,600 rows and 25 columns.

If we use `inner_join()`, for example, we end up with 1,000 rows and 25 columns.

```
> cces_pres %>% dim()
[1] 125600     24
> rct %>% dim()
[1] 1000    2
> cces_pres %>%
+   left_join(rct) %>%
+   dim()
Joining, by = "case_id"
[1] 125600     25
> cces_pres %>%
+   inner_join(rct) %>%
+   dim()
Joining, by = "case_id"
[1] 1000   25
```

Accountability check

Key concepts / skills:

1. Interpreting regression coefficients
2. Understanding potential outcomes notation
3. Being able to define different causal estimands (ATE, ATT)

Accountability check

If you are **not** feeling comfortable with any of these key concepts/skills, then…

Accountability check

If you are **not** feeling comfortable with any of these core concepts/skills, then…

Accountability check

We are going to be building on this foundation over the next few weeks.

Take the extra 30 mins *now* to get comfortable with notation and definitions. It will pay off!!