

# Ec140 - Selection Bias and Potential Outcomes

Fernando Hoces la Guardia

06/30/2022

# Housekeeping

- PS1 due Tomorrow at 5pm on Gradescope.
    - Last question ("Describe how an RCT...") is now optional.
    - Aim to submit at 4pm, to avoid any potential problem related with uploading.
- Late problem sets will not be graded.**

# Today's Lecture

- Selection Bias
- Potential Outcomes Framework

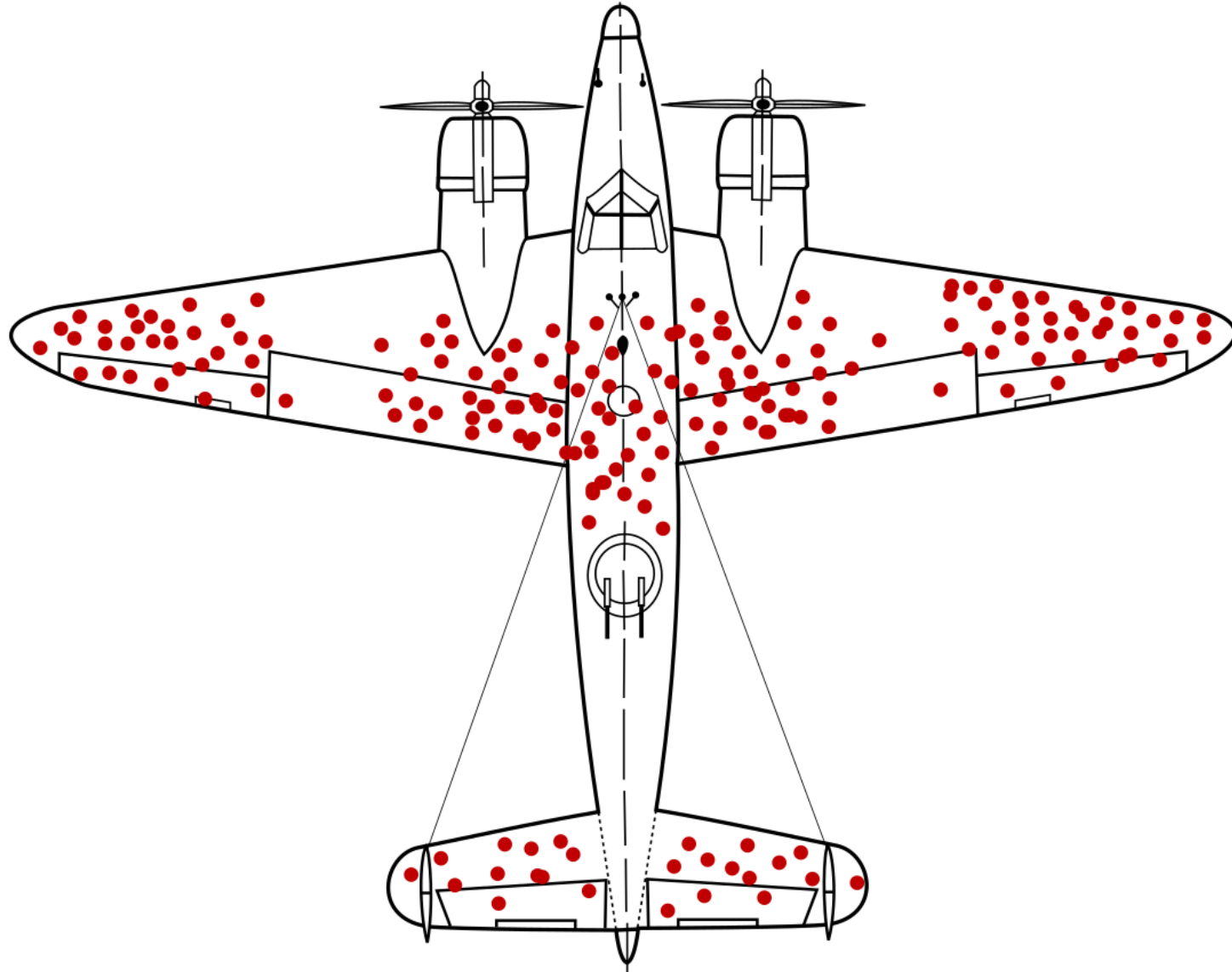
# Selection Bias

## Wikipedia Definition:

Selection bias is the bias introduced by the selection of individuals, groups, or data for analysis in such a way that proper randomization is not achieved, thereby failing to ensure that the sample obtained is representative of the population intended to be analyzed.

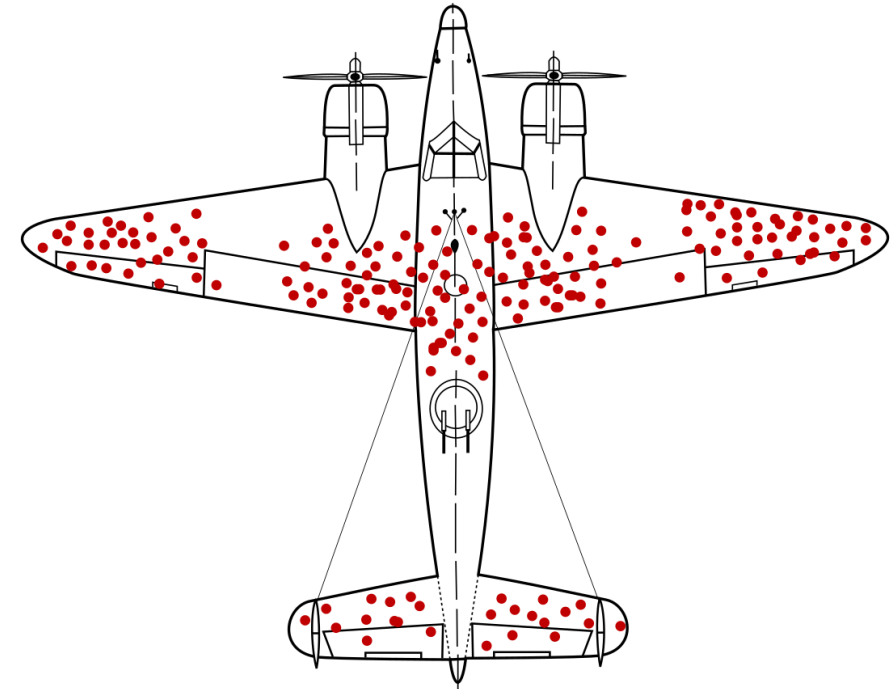
- Econometric textbooks, tend to define selection bias in term of a regression or (as MM) a randomized controlled trial.
- We will start from this more general definition to connect with the concept of **conditional expectation**.

# SB Example 1: Airplanes in World War II



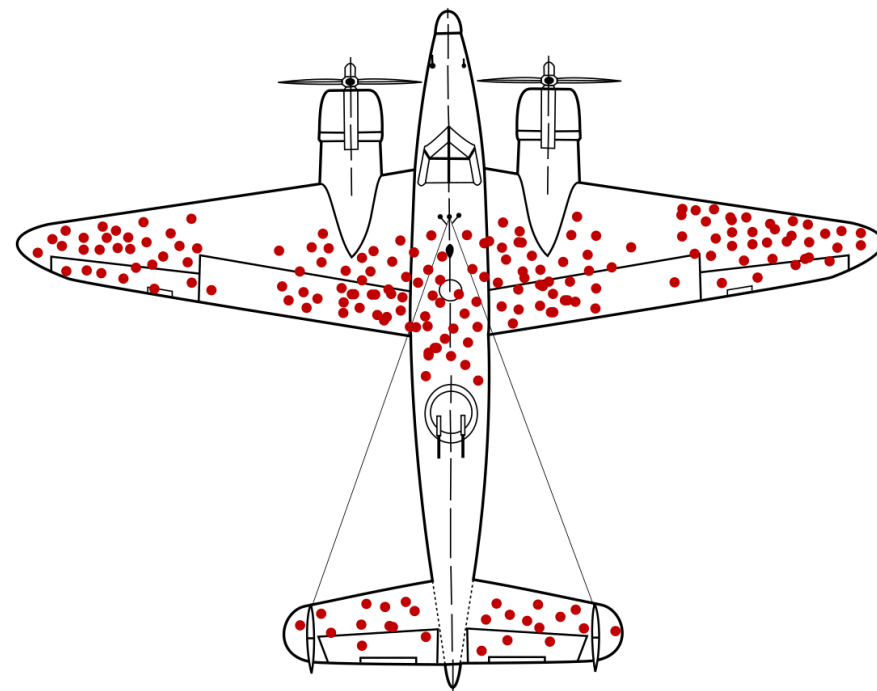
# SB Example 1: Airplanes in World War II. Using Expectation 1/2

- How would you use conditional expectations to characterize this problem?
- Let's start by simplifying the problem by assuming that each plane only had two sections. Now define two random variables: binary variables (bernulli) to indicate if the plane received damage in locations one, and two.  
( $DL1 : \{\text{No damaged in lct 1, Damaged in lct1}\} \rightarrow \{0, 1\}$   
, same for  $DL2$ ).
- We also need to define random variable for that we are conditioning on. In this case, let's use a binary variable for return  
( $R : \{\text{Plane didn't return, Plane returned}\} \rightarrow \{0, 1\}$ )



# SB Example 1: Airplanes in World War II. Using Expectation 2/2

- One way of characterizing the problem would be that the engineers thought they were observing  $\mathbb{E}(DL1)$  and  $\mathbb{E}(DL2)$  and concluding  $\mathbb{E}(DL1) > \mathbb{E}(DL2)$ .
- But in they were actually observing  $\mathbb{E}(DL1|R = 1)$  and  $\mathbb{E}(DL2|R = 1)$  and most likely  $\mathbb{E}(DL1|R = 0) < \mathbb{E}(DL2|R = 0)$
- If you don't like the math notation, you can provide the same answer, but in narrative form.
- This is called **survivorship bias**, and is a type of selection bias.



# SB Example 2: Health Insurance 1/2

- We can do something similar for our health insurance example.
- The "hidden" information could be many things. For example: maybe uninsured people are less have different standards of what constitutes good health, and for the same true health status, uninsured tend to report much higher scores than insured (thanks Andy!).

Randomized Trials 5

TABLE 1.1  
graphic characteristics of insured and uninsured couples in the NHIS

| Husbands           |                |                   | Wives          |                |                   |
|--------------------|----------------|-------------------|----------------|----------------|-------------------|
| I                  | No HI<br>(2)   | Difference<br>(3) | Some HI<br>(4) | No HI<br>(5)   | Difference<br>(6) |
| A. Health          |                |                   |                |                |                   |
|                    | 3.70<br>[1.01] | .31<br>(.03)      | 4.02<br>[.92]  | 3.62<br>[1.01] | .39<br>(.04)      |
| B. Characteristics |                |                   |                |                |                   |
|                    | .17            | -.01<br>(.01)     | .15            | .17            | -.02<br>(.01)     |
|                    | 41.26          | 2.71              | 42.24          | 39.62          | 2.62              |



# SB Example 2: Health Insurance 2/2

- Define a binary random variable that represents if an individual tends to over report good health or not ( $ORep : \{\text{no over report, over reports}\} \rightarrow \{0, 1\}$ )  
 . In this case the previous comparison translates into:
- $\mathbb{E}(H|HI = 1, ORep = 1)$  for column (4), and  $\mathbb{E}(H|HI = 0, ORep = 0)$  for column (5).
- This is a violation of *other things equal* assumption.

Randomized Trials 5

TABLE 1.1  
graphic characteristics of insured and uninsured couples in the NHIS

| Husbands           |                |                   | Wives          |                |                   |
|--------------------|----------------|-------------------|----------------|----------------|-------------------|
| I                  | No HI<br>(2)   | Difference<br>(3) | Some HI<br>(4) | No HI<br>(5)   | Difference<br>(6) |
| A. Health          |                |                   |                |                |                   |
|                    | 3.70<br>[1.01] | .31<br>(.03)      | 4.02<br>[.92]  | 3.62<br>[1.01] | .39<br>(.04)      |
| B. Characteristics |                |                   |                |                |                   |
|                    | .17            | -.01<br>(.01)     | .15            | .17            | -.02<br>(.01)     |
|                    | 41.26          | 2.71              | 42.24          | 38.62          | 2.62              |

# SB Example 3: Country Characterization by Foreign Visitors

- Characterization of Americans according to foreigners visiting Berkeley.
- Characterization of Chinese according to foreigner visiting a specific city.

## **Implications:**

- > Selection Bias is a key reason to promote diversity equity and inclusion (DEI)
- > Selection Bias is one of the main reasons it is so important that you ask questions in class. Especially questions like "I didn't understand that last concept, could you please explain it again?"

# More Examples

- Convention of Statisticians. [XQCD](#)
- [Heike Crabs](#)
- Appearance and Intelligence of Movie Stars (From [Causal Inference, The Mixtape](#))
- Think of at least two examples yourself!
- ([Hernan Cascicari on Surveys](#) [in Spanish, and strong language warning])



# Potential Outcomes Framework

# The Potential Outcomes Framework

**Key idea:** Each individual can be exposed to **multiple alternative treatment states**.

- smoking cigarettes, smoking cigars or not smoking,
- growing up in a poor vs a middle class neighborhood vs a rich neighborhood,
- being in a small or a big class.

For practicality, let this treatment variable  $D_i$  be a binary variable:

$$D_i = \begin{cases} 1 & \text{if individual } i \text{ is treated} \\ 0 & \text{if individual } i \text{ is not treated} \end{cases}$$

**Treatment group**

all the individuals such that  $D_i = 1$ .

**Control group**

all the individuals such that  $D_i = 0$ .

# The Potential Outcomes Framework

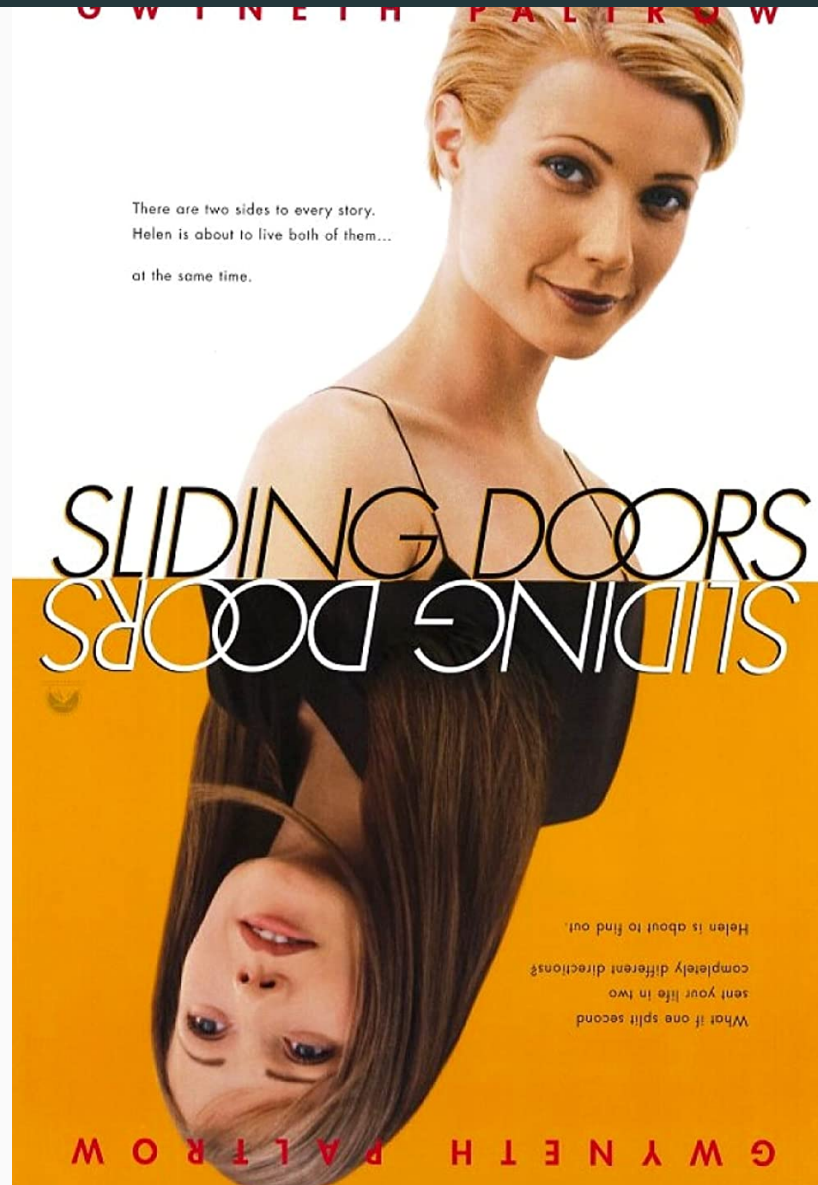
- In this framework, each individual has two **potential outcomes**, but only one **observed outcome**  $Y_i$ :
  - $Y_{1i}$ : potential outcome if individual  $i$  receives the treatment ( $D_i = 1$ ),
  - $Y_{0i}$ : potential outcome if individual  $i$  does not receive the treatment ( $D_i = 0$ ).
- In real life we only observe  $Y_i$  which can be written as:

$$Y_i = D_i \times Y_{1i} + (1 - D_i) \times Y_{0i}$$

- **Fundamental Problem of Causal Inference**: for any individual  $i$ , we only observe one of either potential outcomes (Holland, 1986).

# Example: Sliding Doors (1998)

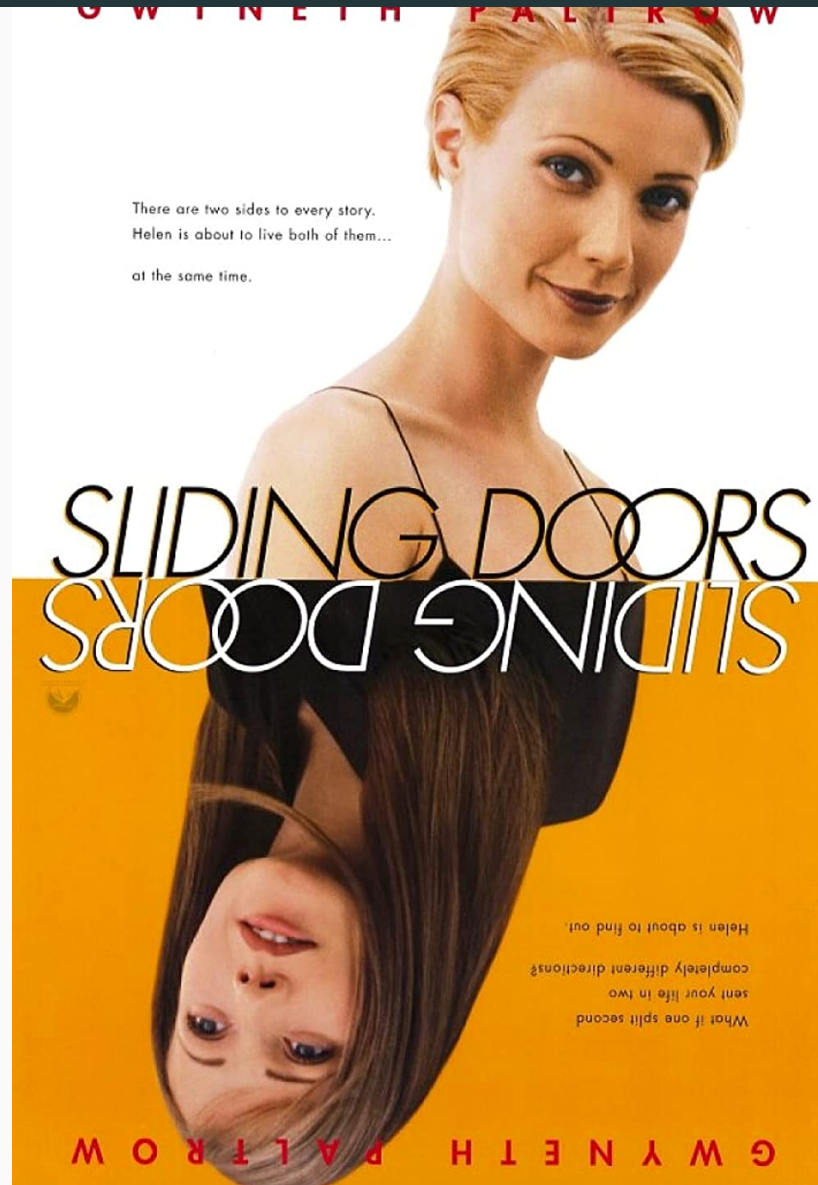
$Y_{1i}$



# Example: Sliding Doors (1998)

$Y_{1i}$

- Other examples:
  - Run Lola Run
  - Avenger's What If?
  - Midnight Library
  - Suggestions?





# The Potential Outcomes Framework

- The potential outcome that is not observed exists in principle, it is called the **counterfactual outcome**.

| Group                         | $Y_{1i}$            | $Y_{0i}$            |
|-------------------------------|---------------------|---------------------|
| Treatment group ( $D_i = 1$ ) | Observable as $Y_i$ | Counterfactual      |
| Control group ( $D_i = 0$ )   | Counterfactual      | Observable as $Y_i$ |

- From these we can define the **individual treatment effect**  $\kappa_i$ :

$$\kappa_i = Y_{1i} - Y_{0i}$$

- $\kappa_i$  measures the **causal effect of the treatment** ( $D_i$ ) on outcome  $Y$  for individual  $i$  (let's read this using the Sliding Doors example).
- Since the treatment effect *cannot* be observed at the individual level, we estimate averages across many individuals.

# Book Example 1/2

- In this ideal scenario, we can observe both worlds: the health and Maria and Khuzdar with and without health insurance.
- Here we can compute the individual treatment ( $\kappa_i$ ) effect for each. For example:
- $Y_{1,Khuzdar} - Y_{0,Khuzdar} = 1$
- In the real world, Khuzdar has HI, while Maria Doesn't, Hence the comparison between could be of interest:
- $Y_{Khuzdar} - Y_{Maria} = -1$
- Can we interpret this difference as causal?

TABLE 1.2  
Outcomes and treatments for Khuzdar and Maria

|   | Khuzdar<br>Khalat | Maria<br>Moreño |
|---|-------------------|-----------------|
| Potential outcome without insurance: $Y_{0i}$ | 3                 | 5               |
| Potential outcome with insurance: $Y_{1i}$    | 4                 | 5               |
| Treatment (insurance status chosen): $D_i$    | 1                 | 0               |
| Actual health outcome: $Y_i$                  | 4                 | 5               |
| Treatment effect: $Y_{1i} - Y_{0i}$           | 1                 | 0               |

From Mastering 'Metrics: The Path from Cause to Effect. © 2015 Princeton University Press. Used by permission.  
All rights reserved.

# Book Example 2/2

- *Other things equal* fails here, because their initial health was different. Let's look at the simple difference but adding and subtracting Khuzdar health without HI.
- The first parenthesis represents a (individual) causal effect. The second term represents the things that are not equal.

$$\begin{aligned} Y_{Khuzdar} - Y_{Maria} &= Y_{1,Khuzdar} - Y_{0,Maria} \\ &= (Y_{1,Khuzdar} - Y_{0,Khuzdar}) + (Y_{0,Khuzdar} - Y_{0,Maria}) \\ &= (1) + (-2) \end{aligned}$$

TABLE 1.2  
Outcomes and treatments for Khuzdar and Maria

|   | Khuzdar<br>Khalat | Maria<br>Moreño |
|---|-------------------|-----------------|
| Potential outcome without insurance: $Y_{0i}$ | 3                 | 5               |
| Potential outcome with insurance: $Y_{1i}$    | 4                 | 5               |
| Treatment (insurance status chosen): $D_i$    | 1                 | 0               |
| Actual health outcome: $Y_i$                  | 4                 | 5               |
| Treatment effect: $Y_{1i} - Y_{0i}$           | 1                 | 0               |

From Mastering Metrics: The Path from Cause to Effect. © 2015 Princeton University Press. Used by permission.  
All rights reserved.

- Now let's move into a (slightly) larger data set.

# Example: Book Example, But with $N = 10$

| $i$ | $Y^1$ | $Y^0$ | $D$ | $\kappa$ |
|-----|-------|-------|-----|----------|
| 1   | 5     | 2     | 1   | 3        |
| 2   | 1     | 4     | 0   | -3       |
| 3   | 3     | 1     | 1   | 2        |
| 4   | 2     | 1     | 1   | 1        |
| 5   | 4     | 4     | 0   | 0        |
| 6   | 5     | 4     | 1   | 1        |
| 7   | 1     | 2     | 1   | -1       |
| 8   | 2     | 3     | 0   | -1       |
| 9   | 4     | 1     | 0   | 3        |
| 10  | 3     | 1     | 0   | 2        |

- Potential outcomes for health of individual with ( $Y_1$ ) and without ( $Y_0$ ) health insurance.
- Analogous to the individual comparison, we could be interested in the average comparison:

$$Avg_n[Y_{1i} - Y_{0i}] = \frac{1}{n} \sum_{i=1}^n (Y_{1i}) - \frac{1}{n} \sum_{i=1}^n (Y_{0i})$$

- This mean is the the average causal effect
- Can we compute this mean?

# Example: Book Example, But with $N = 10$

| $i$ | $Y^1$ | $Y^0$ | $D$ | $\kappa$ |
|-----|-------|-------|-----|----------|
| 1   | 5     | NA    | 1   | 3        |
| 2   | NA    | 4     | 0   | -3       |
| 3   | 3     | NA    | 1   | 2        |
| 4   | 2     | NA    | 1   | 1        |
| 5   | NA    | 4     | 0   | 0        |
| 6   | 5     | NA    | 1   | 1        |
| 7   | 1     | NA    | 1   | -1       |
| 8   | NA    | 3     | 0   | -1       |
| 9   | NA    | 1     | 0   | 3        |
| 10  | NA    | 1     | 0   | 2        |

- Potential outcomes for health of individual with ( $Y_1$ ) and without ( $Y_0$ ) health insurance.
- Analogous to the individual comparison, we could be interested in the average comparison:

$$Avg_n[Y_{1i} - Y_{0i}] = \frac{1}{n} \sum_{i=1}^n (Y_{1i}) - \frac{1}{n} \sum_{i=1}^n (Y_{0i})$$

- This mean is the the average causal effect
- Can we compute this mean?
- **We can't.**

# The Problem of Causal Inference

- From the data, we can compute the *difference-in-group-means*:

$$\text{Difference in group means} = \underbrace{\frac{1}{N_T} \sum_{i=1}^{N_T} (Y_i | D_i = 1)}_{\text{Avg}_n[Y_i | D_i=1]} - \underbrace{\frac{1}{N_C} \sum_{i=1}^{N_C} (Y_i | D_i = 0)}_{\text{Avg}_n[Y_i | D_i=0]}$$

- Is  $\text{Avg}(Y_i | D_i = 1) - \text{Avg}(Y_i | D_i = 0)$  a *good* estimator for the average causal effect?

# Estimating Causal Effects

**Assumption:** Let  $\kappa_i = \kappa$  for all  $i$ .

- The treatment effect is equal (constant) across all individuals  $i$ .

**Note:** We defined

$$\kappa_i = \kappa = Y_{1,i} - Y_{0,i}$$

which implies

$$Y_{1,i} = Y_{0,i} + \kappa$$

# Simple Difference in Group Means

Is  $Avg(Y_i \mid D_i = 1) - Avg(Y_i \mid D_i = 0)$  a *good* estimator for the average causal?

$$\begin{aligned}\text{Difference in group means} &= Avg(Y_i \mid D_i = 1) - Avg(Y_i \mid D_i = 0) \\ &= Avg(Y_{1,i} \mid D_i = 1) - Avg(Y_{0,i} \mid D_i = 0) \\ &= Avg(\kappa + Y_{0,i} \mid D_i = 1) - Avg(Y_{0,i} \mid D_i = 0) \\ &= \kappa + \underbrace{Avg(Y_{0,i} \mid D_i = 1) - Avg(Y_{0,i} \mid D_i = 0)}_{\text{Selection bias}} \\ &= \text{Average causal effect} + \text{Selection bias}\end{aligned}$$

Our proposed difference-in-means estimator gives us the sum of:

1.  $\kappa$ , the average causal effect that we want.
2. **Selection bias** How much treatment and control groups differ, on average.



# Let's Bring Expectations Back!

We previously defined expectations as the population version of the mean. Hence, we can use expectation to represent this problem at the population level:

$$\mathbb{E}(\text{Difference in group means}) = \kappa + \underbrace{\mathbb{E}(Y_{i0}|D_i = 1) - \mathbb{E}(Y_{i0}|D_i = 0)}_{\text{Selection bias}}$$

- Looking at the expectations formulation of selection bias, it becomes clear that our problems would be solved if we could make  $Y_{i0}$  independent of  $D$ .

# Note: On Name of The Problem

- In addition to ***The Fundamental Problem of Causal Inference***, this problem is usually referred with the following terms:
- We are **missing data** on all the potential outcomes for which the treatment status did not happen in real world. Hence this is also referred as a missing data problem.
- This is also called a **identification** problem (as in: we cannot identify the average treatment effect).
- In this course, you will not be asked to memorize the different names, just be aware of the different terminology when consulting references.

# Acknowledgments

- Kyle Raze's Undergraduate Econometrics 1
- ScPoEconometrics
- XQCD
- MM
- Matt Hollian
- Causal Mixtape (Also Hanna Fry)
- Wikipedia (Survivorship Bias)
- MM [bookdown](#) and MM [blog post](#) on chapter 1

