Proof of Slides 93-96

Properties for $\mathbb{E}[X]$

1. $\mathbb{E}[A - B] = \mathbb{E}[A] - \mathbb{E}[B]$
2. $\mathbb{E}[X] = \Pr(B) \cdot \mathbb{E}[X \mid B] + \Pr(\text{not } B)\, \mathbb{E}[X \mid \text{not } B]$

Proof of Property (2):

- First, we know $\Pr(A) = \Pr(B) \Pr(A \mid B) + \Pr(\text{not } B) \Pr(A \mid \text{not } B)$, because:

$$\begin{aligned}
\Pr(A) &= \Pr(A \cap B) + \Pr(A \cap \text{not } B) \\
&= \Pr(B) \cdot \frac{\Pr(A \cap B)}{\Pr(B)} + \Pr(\text{not } B) \cdot \frac{\Pr(A \cap \text{not } B)}{\Pr(\text{not } B)} \\
&= \Pr(B) \cdot \Pr(A \mid B) + \Pr(\text{not } B) \cdot \Pr(A \mid \text{not} B)
\end{aligned}$$

- Applying this to the definition of $\mathbb{E}[X]$:

$$\begin{aligned}
\mathbb{E}[X] &= \sum x \Pr(X = x) \\
&= \sum x \{\Pr(B) \Pr(X = x \mid B) + \Pr(\text{not } B) \Pr(X = x \mid \text{not } B)\} \\
&= \sum x \Pr(B) \Pr(X = x \mid B) + \sum x \Pr(\text{not } B) \Pr(X = x \mid \text{not } B) \\
&= \Pr(B) \left\{\sum x \Pr(X = x \mid B)\right\} + \Pr(\text{not } B) \left\{\sum x \Pr(X = x \mid \text{not } B)\right\} \\
&= \Pr(B)\, \mathbb{E}[X \mid B] + \Pr(\text{not } B)\, \mathbb{E}[X \mid \text{not } B]
\end{aligned}$$

- *Q.E.D.*

> Note: Some textbooks will write "not B" as $\neg B$, B' or B^c

Now, we can expand the expression for ATE using Properties 1 and 2:

$$\begin{aligned}
\text{ATE} &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] && \text{prop. 1} \\
&= \Pr(D_i = 1)\, \mathbb{E}[Y_i(1) \mid D_i = 1] + \Pr(D_i = 0)\, \mathbb{E}[Y_i(1) \mid D_i = 0] && \text{prop. 2} \\
&\quad - \{\Pr(D_i = 1)\, \mathbb{E}[Y_i(0) \mid D_i = 1] + \Pr(D_i = 0)\, \mathbb{E}[Y_i(0) \mid D_i = 0]\}
\end{aligned}$$

For our convenience, let's define $\Pr(D_i = 1) = p$, such that $\Pr(D_i = 0) = (1 - p)$. We arrive at the expression on slide 94:

$$\begin{aligned}
\text{ATE} = & p\, \mathbb{E}[Y_i(1) \mid D_i = 1] + (1 - p)\, \mathbb{E}[Y_i(1) \mid D_i = 0] \\
& - p\, \mathbb{E}[Y(0) \mid D_i = 1] - (1 - p)\, \mathbb{E}[Y_i(0) \mid D_i = 0]
\end{aligned}$$

Recall that out of the four terms in the previous equation, only the first and fourth terms (in red) are observable (and are indeed the constituent terms of SDO). Our goal here is to try to express the SDO as a combination of ATE and the other two unobservable terms in order to quantify the unobserved biases.

For brevity, I will reexpress the conditional probabilities as

$$\begin{cases} \mathbb{E}[Y_i(1) \mid D_i = 1] & = e_{11} \\ \mathbb{E}[Y_i(1) \mid D_i = 0] & = e_{10} \\ \mathbb{E}[Y_i(0) \mid D_i = 1] & = e_{01} \\ \mathbb{E}[Y_i(0) \mid D_i = 0] & = e_{00} \end{cases}$$

Such that

$$\begin{cases} \text{SDO} = e_{11} - e_{00} \\ \text{ATE} = pe_{11} + (1-p)e_{10} - pe_{01} - (1-p)e_{00} \\ \text{ATT} = e_{11} - e_{01} \\ \text{ATU} = e_{10} - e_{00} \end{cases}$$

Now, we can try rearranging the terms, with a bit of brute force:

$$\begin{aligned} \text{ATE} &= pe_{11} + (1-p)e_{10} - pe_{01} - (1-p)e_{00} \\ 0 &= \text{ATE} - pe_{11} - (1-p)e_{01} + pe_{10} + e_{00} - pe_{00} \\ e_{11} - e_{00} &= \text{ATE} + (1-p)e_{11} - (1-p)e_{10} + pe_{01} - pe_{00} \qquad (+e_{11} - e_{00} \text{ on both sides}) \end{aligned}$$

Our next goal is to simplify the RHS, by factorizing most terms with $(1-p)$ and expressing them as ATT or ATU

$$\begin{aligned} e_{11} - e_{00} &= \text{ATE} + (1-p)e_{11} - (1-p)e_{10} + pe_{01} - pe_{00} \\ &= \text{ATE} + (1-p)e_{11} - (1-p)e_{10} + (p-1)e_{01} - (p-1)e_{00} + e_{01} - e_{00} \\ &= \text{ATE} + (1-p)e_{11} - (1-p)e_{10} - (1-p)e_{01} + (1-p)e_{00} + e_{01} - e_{00} \\ &= \text{ATE} + \underbrace{(1-p)e_{11} - (1-p)e_{01}}_{(1-p)\text{ATT}} \underbrace{-(1-p)e_{10} + (1-p)e_{00}}_{-(1-p)\text{ATU}} + e_{01} - e_{00} \\ &= \text{ATE} + \{e_{01} - e_{00}\} + (1-p)\text{ATT} - (1-p)\text{ATU} \\ &= \text{ATE} + \{e_{01} - e_{00}\} + (1-p)(\text{ATT} - \text{ATU}) \end{aligned}$$

Q.E.D.

> This proof can also be completed by factorizing the terms with p, but you will arrive at a slightly different expression for the biases:
>
> $$e_{11} - e_{01} = \text{ATE} + (e_{11} - e_{10}) - p(\text{ATT} - \text{ATU})$$
>
> The proof is left as an exercise for the reader.

Thus, we now know that,

$$\text{SDO} = \text{ATE} + e_{01} - e_{00} + (1-p)(\text{ATT} - \text{ATU})$$

For the SDO to be an unbiased estimator of the ATE (i.e. $\text{SDO} = \text{ATE}$), we will need that:

$$\begin{cases} e_{01} - e_{00} = 0 \\ \text{ATT} - \text{ATU} = 0 \end{cases}$$

This will only happen when:

1. The average of baseline potential outcomes of folks ($Y_i(0)$) in the treated group is the same as those in the control group. We call this no "selection bias."
    - This will be violated when, say, I run a study on whether revision increases grades, but I put all the rich kids in the treatment group, and all the poor kids in the control groups. Since we know that the rich kids will most probably have higher baseline grades than the poor kids even if they don't study, there will be a difference in the average $Y_i(0)$ s across the two groups - hence there is a selection bias.
2. The average (individual) treatment effect of the folks in the treatment group, is same as that of those in the control group. We call this no "heterogeneous treatment effect bias."
    - Using the same example, rich kids with higher baseline grades will receive less marginal benefit from studying than poor kids (i.e. it is much harder to go from 80 to 90 than 60 to 70). This means the average treatment effect of the two groups will not be the same (or $\text{ATT} \neq \text{ATU}$) - hence a heterogenous treatment effect bias.
    - As you might imagine, both biases will be resolved automatically if people are randomly assigned into the treatment and control groups. So these biases are most relevant for poorly-designed experiments or observational studies, where we have no control over the treatment assignment (e.g. an evaluation of historical educational policies in a rich vs poor country).