Causality and the Potential Outcomes Framework

Alan Liang, Benjamin Lee

March 2018

1 Motivation and Goal

The goal of conducting an experiment is to discover a causal relationship between 2 variables, which we call causality. In an experiment, we attempt to isolate all confounding factors to show that only the independent variable (often through being administered a treatment or not) affects some result. For example, say we want to conduct an experiment on whether taking an aspirin will relieve a headache. We field 2 participants, Alice and Bob, and place Alice in the treatment group and Bob in the control group.

However, placing participants into either the treatment or control group leads to the **fundamental problem of causal inference**: we can only observe the result of the participant from being placed in one of the groups, but not the other! In our example, since Alice was placed in the treatment group, we can only observe the results of whether Alice had her headache relieved because she took an aspirin; we cannot compare it to whether Alice would have had her headache relieved if she did not take an aspirin. This problem is illustrated below:

	treatment	control
Alice	1	?
Bob	?	0

Perhaps if there were many parallel universes that we could play around with, we would be able to see all the values in this table. But, until we have harnessed that technology, there appears to be no way to get around this problem: Alice cannot both have taken an aspirin and not taken one at the same time.

2 Terminology and Variables

2.1 Outcome

We use the letter Y to denote the potential outcome. Specifically, Y_{i0} denotes the potential outcome for the subject i in the control group, and Y_{j1} denotes the potential outcome for some other subject j in the treatment group. Notably, due to the fundamental problem of causal inference, we cannot both know Y_{i0} and Y_{i1}!

2.2 Treatment

In treatment-control experiments, treatment is denoted by T. Specifically, T_i denotes whether the subject i received the treatment:

$$T_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise} \end{cases}$$

Generalizing our table above to accommodate these variables:

	$T=1$	$T=0$
Alice	Y_a1	?
Bob	?	Y_b0

2.3 Well Defined Questions

A causal question asks *what is the effect of the treatment on the outcome?* and *if the opposite treatment had been received, how would the outcome differ?*
In the potential outcomes framework, a well-defined causal question has a corresponding hypothetical experiment that can be set up and conducted. This means that there is some clear action (manipulation, treatment, intervention) that distinguishes the two potential outcomes.
Below are some traits for well-defined groups:

- Is the question **clear and specific**?
- Can you clearly **identify** the alternative treatment (the 'control' condition, what is the specific alternative to the treatment we are considering)?
- Can you create an experiment out of the causal question?
- "No Interference" or SUTVA: the treatment applied to one unit does not affect the outcomes of other units. Otherwise, there are ¿¿ 2 potential outcomes.

2.4 Conditional Probability

Conditional probability is a measure of the probability of event given that another event has occurred by assumption. It is represented in notation with a |, so that $A|B$ means A *given* B.
For example, to refer to the probability "what is the probability Alice has a headache given she took an aspirin?", we write: $P(A|B)$
Here, A refers to whether Alice has a headache, and B to whether Alice took an aspirin.

3 Running an Experiment

3.1 Average Treatment Effect

When we conduct an experiment, we would like to know if there is actually a difference between the treatment and the control group. This is known as the treatment effect, and is defined to be the difference between the 2 potential outcomes:

$$\text{Treatment effect} = Y_{i1} - Y_{i0}$$

However, these potential outcomes are random variables and do not possess any inferable meaning by themselves (we'll get into this another day). Instead, we seek the find the expected difference between these potential outcomes, which is also the average. This is known as the **Average Treatment Effect (ATE)**:

$$\text{ATE} = \text{E}[Y_{i1} - Y_{i0}]$$

More specifically, we are only looking to find the difference due to the treatment. This means that we are seeking the average treatment effect on those in the treatment group only. This is known as the **treatment effect on the treated (ToT)**. Notice that we have conditioned on $T_i = 1$, i.e. it is given that i is in the treatment group:

$$\text{Average ToT} = \text{E}[Y_{i1} - Y_{i0}|T_i = 1] = \text{E}[Y_{i1}|T_i = 1] - \text{E}[Y_{i0}|T_i = 1]$$

But in reality, since we do not have many parallel universes, we can only see 2 subgroups: the Y_{i0} for those in the control group, and the Y_{i1} for those in the treatment group. Expressed in terms of conditional probability:

$$\text{Observed treatment effect for those in treatment group} = \mathrm{E}[Y_{i1}|T_i = 1]$$
$$\text{Observed control effect for those in control group} = \mathrm{E}[Y_{i0}|T_i = 0]$$
$$\text{Observed difference} = \mathrm{E}[Y_{i1}|T_i = 1] - \mathrm{E}[Y_{i0}|T_i = 0]$$

This is actually going to be approximately equal to the average treatment effect on the treated. But why?

3.2 Selection Bias

We are looking for the average ToT: $\mathrm{E}[Y_{i1}|T_i = 1] - \mathrm{E}[Y_{i0}|T_i = 1]$
We know the observed difference: $\mathrm{E}[Y_{i1}|T_i = 1] - \mathrm{E}[Y_{i0}|T_i = 0]$
Notice that we can express one in terms of the other (and some more):

$$\text{Observed Difference} = \text{Average TOT} + (\mathrm{E}[Y_{i0}|T_i = 1] - \mathrm{E}[Y_{i0}|T_i = 0])$$

This extra term we've added, $\mathrm{E}[Y_{i0}|T_i = 1] - \mathrm{E}[Y_{i0}|T_i = 0]$, is the **selection bias**.
In words, the selection bias is the difference between the expected outcome if an individual were in the control given that he/she is in the treatment and the expected outcome if an individual were in the control given that he/she is in the control.
Intuitively, this makes sense: if there was no selection bias whatsoever, the expected outcome if nothing were to happen to you (i.e. you are in the control group) for both groups should be the same.
On the other hand, if there was selection bias, there would be a difference in expected outcomes between someone assigned to the control group and someone assigned to the treatment even if nothing were to happen to both of them.
Hence, our goal is now to make selection bias zero.

3.3 Randomization

Randomized assignment will eliminate selection bias. The idea is that randomizing assignment will make both groups similar, so that their potential outcomes would be the same:

$$\mathrm{E}[Y_{i0}|T_i = 1] = \mathrm{E}[Y_{i0}|T_i = 0] = \mathrm{E}[Y_{i0}]$$
$$\mathrm{E}[Y_{i1}|T_i = 1] = \mathrm{E}[Y_{i1}|T_i = 0] = \mathrm{E}[Y_{i1}]$$

This would mean that our selection bias is now:

$$\mathrm{E}[Y_{i0}|T_i = 1] - \mathrm{E}[Y_{i0}|T_i = 0] = 0$$

Hence, our average treatment effect on the treated would be equal to the observed difference:

$$\text{Observed Difference} = \text{Average TOT} + \text{Selection bias}$$
$$\text{Observed Difference} = \text{Average TOT}$$