Causal Inference and Designed Experiments

In 506 & 506, we have used predictive language for discussing linear models. In particular, we have used words like, "the expected difference between a unit with factor y and another unit with factor x" that focus on differences between units rather than differences within a unit.

Now we introduce causal language focusing on differences within a unit. In other words, we are interested in what would happen to a unit as a result of a different treatment or predictor values.

Causal inference is generally focused on a comparison of potential outcomes. These potential outcomes could be factual or counterfactual (what might have happened).

The textbook has a running example of taking fish oil supplements, we will consider a clinical trials setting with COVID-19 vaccination looking at antibody measurements for individuals with and without a vaccination.

Let y_i^0 be the antibody measurement (IgG) for individual i having received a control and y_i^1 be the antibody measurement (IgG) for individual i having received a vaccine. These two values, y_i^0 and y_i^1 represent the *potential outcomes.*

In the experiment, individual i either receives the control or the vaccine, hence, only one potential outcome is observed, which is denoted as the *factual outcome.* The other potential outcomes is a *counterfactual outcome* and cannot be observed (without a time machine).

The causal effect of the treatment for unit i is defined as $\tau_i = y_i^1 - y_i^0$.

Unfortunately, we cannot observe both y_i^1 and y_i^0, which is referred to as the fundamental problem of causal inference. Thus, understanding causal effects requires additional assumptions.

If treatments are randomly assigned, we can estimate an "average causal effect" across the respondents, but can't say anything about unit i. In other words, the average cause effect requires assuming that the effects are constant across units.

Another approach would be to attempt to have a replacement for one of the counterfactuals. For instance, could a pre-score be used in place of y_i^0? It would be imperfect as it wouldn't account for external impacts that occurred during the study (such as exposure to SARS-CoV-2).

An experiment could also be conceived that randomizes the ordering of the assignment such that each unit receives both treatments over the course of the study, this is called a crossover design.

The same idea applies for multiple treatment levels, continuous treatments...

Again, we cannot estimate $\tau_i = y_i^1 - y_i^0$, but we can estimate the sample average treatment effect (SATE)
$\tau_{SATE} = \frac{1}{n}\sum_i^n y_i^1 - \frac{1}{n}\sum_i^n y_i^0 \approx \frac{1}{n}\sum_i^n (y_i^1 - y_i^0)$

The treatment effects can be stratified by a particular group, for which is known as a *conditional average treatment effect* (CATE)

With statistics, the interest is rarely just the sample itself, but rather a broader population. Hence the target is often

$\tau_{PATE} = \frac{1}{N}\sum_i^N (y_i^1 - y_i^0)$

again this quantity requires knowing both potential outcomes for a given unit.

If the treatment and control groups are not similar, then $\tau_{SATE} = \frac{1}{n}\sum_i^n y_i^1 - \frac{1}{n}\sum_i^n y_i^0 \neq \frac{1}{n}\sum_i^n (y_i^1 - y_i^0)$. Ideally, the similarity between groups should be based on the potential outcomes rather than other measured variables.

Randomized experiments

Randomization can ensure treatment and control groups are balanced, on average (in expectation).

In a completely randomized experiment, the probability of being assigned any given treatment is the same for all units.

Note, that a completely randomized experiment does not guarantee a balanced sample for any particular realization.

```
set.seed(10)
tibble(factor = sample(rep(c('blue','gold'), each =4),4)) %>%
  bind_cols(tibble(treatment =rep('treat', 4))) %>%
  kable()
```

factor	treatment
blue	treat
blue	treat
blue	treat
blue	treat

Design Notation An *unbiased estimate* is correct, on average. In other words, the mean of the sampling distribution is equal to the estimand. Where the sampling distribution is based on repeated samples.

The sampling distribution of an *efficient* estimate has small variance.

Similarly, using the randomization distribution (based on repeated allocation of treatments) of the estimate

$$d^k = \frac{1}{n_1}\sum_{i,z_i^k=1} y_i^k - \frac{1}{n_2}\sum_{i,z_i^k=0} y_i^k,$$

where k denotes the k^th sample and is analogous to the sampling distribution.

If there are other observable factors that would be expected to result in different outcomes, this can be used in the experimental design. Recall the completely randomized design that resulted in all of th "blue" units being assigned the treatment.

```
set.seed(10)
tibble(factor = sample(rep(c('blue','gold'), each =4),4)) %>%
  bind_cols(tibble(treatment =rep('treat', 4))) %>%
  kable()
```

factor	treatment
blue	treat
blue	treat
blue	treat
blue	treat

A randomized block design allocates treatments and controls within each group of similar units.

```
tibble(block = rep(c('blue','gold'), each =4),
       level = c(sample(rep(c('treat','control'), each = 2)),sample(rep(c('treat','control'), each = 2))
```

block	level
blue	control
blue	treat
blue	treat
blue	control
gold	control
gold	control
gold	treat
gold	treat

Blocking structure can reduce variance in the treatment effect (*avoids "bad" random samples*)

Blocks are defined by pre-treatment variables. Blocks can be defined by anything that might be expected to predictive of the outcome.

Other experimental design structures include

- Matched pairs example: look for a "twin" or pair for each unit
- Cluster designs: Consider a treatment at the "teacher" level (such as training) when the sampling units are students.

Ideally any information about differences in units would be accounted for in the design phase, but it can also be included

Ignorability With ignorability the assignment of a treatment is independent of the potential outcomes. With a complete randomized design, this can be written as

$$z \perp y^0, y^1$$

The property of ignorability doesn't mean anything about a given randomization, but rather, there is no imbalance, on average.

Randomized blocks use conditional ignorability

$$z \perp y^0, y^1 | w$$

In other words, conditional on a block or within a block, the probability of a treatment is independent of the potential outcomes.

Efficiency Efficiency is a measure of the variability in the estimator.

Using blocking variables can result in a more efficient estimator if the units within the blocks are similar, but the blocks are different.

Regression methods can also be used to account for pre-treatment variables and result in a more efficient estimator.