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.

Causal inference is generally focused on a comparison of potential outcomes.

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.

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.*

Unfortunately, we cannot observe both y_i^1 and y_i^0

If treatments are randomly assigned, we can estimate an "average causal effect" across the respondents, but can't say anything about unit i.

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?

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,

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)

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

If the treatment and control groups are not similar, then $\tau_{SATE} = \frac{1}{n}\sum_i^n y_i^1 - \frac{1}{n}\sum_i^n y_i^0$

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.

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

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

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

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

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

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

Randomized blocks use conditional ignorability

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.