MS&E 327: Topics in Causal Inference

Samuel Wong
Department of Statistics
Stanford University

Abstract

The course starts with the common terminology and assumptions for causal inference using the potential outcomes framework. It then delves into the world of randomized experiments using Neymanian and Fisherian frameworks. Covariate adjustments and interference are covered. The course then moves to observational studies, where matching and sensitivity analysis are discussed. After that, there is a section on causal identification and the use of DAGs as another framework to perform causal analysis. Lastly, some extra topics such as instrumental variables, regression discontinuity design, difference in differences, and synthetic controls are discussed.

Contents

1 Introduction **3**
1.1 SUTVA (Stable Unit Treatment Value Assumption) . . . 3
1.2 "The Science" . . . 3
1.3 Causal Estimands and Estimators . . . 4
1.4 Superpopulation Estimands . . . 4
1.5 No Causation without Manipulation . . . 4

2 Randomized Experiments **5**
2.1 The Randomization-Based Framework . . . 5
2.1.1 Assignment Mechanisms . . . 6
2.2 Neymanian Inference . . . 6
2.2.1 Neymanian Inference in Practice . . . 7
2.2.2 Horvitz-Thompson Estimator . . . 7
2.3 Fisher Randomization Test . . . 8
2.3.1 Power and Choice of Test Statistic . . . 8
2.3.2 Confidence Intervals and Point Estimates . . . 8
2.4 Covariate Adjustment . . . 9
2.4.1 Discrete Covariates . . . 9
2.4.2 Continuous Covariates . . . 10
2.5 Interference . . . 11
2.5.1 SDP (School District of Philadelphia) Experiment . . . 11
2.5.2 The Interference Machinery . . . 12
2.5.3 Neymanian Inference . . . 13
2.5.4 Fisherian Approach . . . 13
2.5.5 The Design Problem . . . 14

3 Observational Studies **15**
3.0.1 Strong Ignorability . . . 15
3.1 Exact Pairwise Matching . . . 15
3.2 Analysis . . . 16
3.3 Sensitivity Analysis . . . 16
3.3.1 Latent Confounder Representation . . . 17
3.4 Alternatives to Exact Pairwise Matching . . . 17
3.4.1 Optimal Pair Matching . . . 18
3.4.2 Non-pairwise Matching . . . 18
3.4.3 Constraints . . . 18
3.5 Alternatives to Matching and FRT . . . 18
3.5.1 Subclassification on the Propensity Score . . . 19
3.6 Inverse Propensity Weighted Estimators . . . 19

4 Causal Identification and Graphical Causal Models **20**
4.1 Identification . . . 20
4.1.1 Simple Identification Condition . . . 21
4.1.2 Identification Region . . . 21
4.1.3 Identification with Covariates . . . 21
4.2 Directed Acyclic Graphs (DAGs) . . . 21
4.2.1 Structural Causal Model . . . 22
4.2.2 Applications of DAGs . . . 23
4.2.3 Graphical Rules for Conditional Independence . . . 23
4.2.4 Causal Identification . . . 25
4.2.5 Causal Search . . . 26

5 Common Observational Studies Scenarios **28**
5.1 Instrumental Variables . . . 28
5.1.1 Instruments . . . 28
5.1.2 Non-parametric Instrumental Variables . . . 29
5.1.3 Average Causal Effect of Treatment on Compliers . . . 29
5.1.4 Additional Topics in Instrumental Variables . . . 30
5.2 Regression Discontinuity Design (RDD) . . . 31
5.2.1 Sharp Regression Discontinuity . . . 31
5.2.2 Estimation and Inference . . . 32
5.3 Difference in Differences (DiD) . . . 32
5.3.1 Derivation of DiD . . . 33
5.4 Synthetic Controls . . . 33

1 Introduction

What is a causal effect?

- Example: What is the effect of taking aspirin on Guillaume's headache? There are two scenarios:
  1. He takes the aspirin and we measure the state of his headache one hour later, denote it $Y_G(A)$
  2. He does not take the aspirin and we measure the state of his headache one hour later, denote it $Y_{NG}(A)$
- The causal effect is $\tau_G = Y_G(A) - Y_{NG}(A)$
- The fundamental problem of causal inference is that we can never observe both $Y_G(A)$ and $Y_{NG}(A)$. Either he took the aspirin or he did not!
- Someone may say that he should take the aspirin the first time he has a headache and then not take the aspirin the next time he has a headache and then we can measure the difference. However, this is not true. If we do this, we are calculating $\tau_G = Y_{G,t_1}(A) - Y_{G,t_2}(NA)$ instead, and we have NO guarantees that G at t_1 is the same as G at t_2.

We can start our discussion of causal inference using the Rubin Causal model, which is also sometimes called the Neyman-Rubin CM or the Potential Outcomes framework. Under this model, we have N units, so $i = 1, ..., N$. For each of these units, we perform a binary intervention where we assign them either to the treament group or the control group. Let $Z_i \in \{0, 1\}$ be the indicator for the treatment assignment of unit i (usually 1 means treatment and 0 means control). Then we denote $\vec{Z} = (Z_1, ..., Z_N)$, so it is the vector of assignments. We then denote $Y_i(\vec{Z})$, the potential outcome of unit i under $\vec{Z}$.

So why should Y_i be a function of the whole vector $\vec{Z}$? Shouldn't the outcome only depend on which group the ith unit is assigned to, regardless of the others? In other words, shouldn't $Y_i(\vec{Z}) = Y_i(Z_i)$? The answer most of the time is yes, it does only depend on itself, and this assumption is called SUTVA.

1.1 SUTVA (Stable Unit Treatment Value Assumption)

- No interference between units
  - For example, "you taking the aspirin has no effect on my headache"
  - $Y_i(\vec{Z}) = Y_i(Z_i)$ for $i = 1, ..., N$
- No hidden version of the treatment
  - The treament group should all take the same units, for example everyone takes 1 pill for the headache. No one takes half a pill or a quarter or a pill
  - In other words, $Y_i(Z_i)$ is a well defined function

The consequence of SUTVA is that each unit only has 2 potential outcomes, $Y_i(0)$ and $Y_i(1)$ (assuming control and only one treatment option).

1.2 "The Science"

What is called "The Science" in causal inference? It is the table/schedule of potential outcomes. For example, if we have N units, and we denote $\vec{Y}(0) = (Y_1(0), Y_2(0), ..., Y_N(0))$ and $\vec{Y}(1) = (Y_1(1), Y_2(1), ..., Y_N(1))$, then the science is $\underline{Y} = (\vec{Y}(0), \vec{Y}(1))$. "The Science" can be more clearly seen below in this table.

i	$Y_i(0)$	$Y_i(1)$
1	$Y_1(0)$	$Y_1(1)$
2	$Y_2(0)$	$Y_2(1)$
$\vdots$	$\vdots$	$\vdots$
N	$Y_N(0)$	$Y_N(1)$

Note that we will never be able to fully complete "The Science." This is just what we wish we had. As mentioned before, we can only see one outcome per unit i. Therefore, what we actually observe is something like the table below.

For each unit i, we either observe $Y_i(0)$ or $Y_i(1)$. Note that this can be rewritten as $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.

i	$Y_i(0)$	$Y_i(1)$
1	?	$Y_1(1)$
2	$Y_2(0)$	?
$\vdots$	$\vdots$	$\vdots$
N	?	$Y_N(1)$

1.3 Causal Estimands and Estimators

- Estimand: a quantity of interest (what is to be estimated; ie the ground truth we want to find). It is denoted $\tau = \tau(\underline{Y})$
- Estimator: our data-driven guess for the estimand. It is denoted $\hat{\tau} = \hat{\tau}(\hat{Y}^{obs}, \hat{Z}^{obs})$

Now that we have defined an estimand, what is a casual estimand? It is an estimand that is based on the contrast between potential outcomes for given units. Some example of causal estimands are:

- Individual Causal Effect (for unit i): $\tau_i = Y_i(1) - Y_i(0)$
- Average Treatment Effect: $\tau^{ATE} = \bar{Y}(1) - \bar{Y}(0) = \frac{1}{N}\sum_{i=1}^{N}\sum_i(Y_i(1) - Y_i(0))$
- Conditional Average Treatment Effect: for example, say X = "male", then $\tau_X = \frac{1}{N_X}\sum_{i=1}^{N}\mathbb{1}(X_i - X)\tau_i$
- Lift: $L = \frac{\bar{Y}(1)-\bar{Y}(0)}{\bar{Y}(0)}$

Some other points:

1. τ generally does not depend on Z^{obs} (the true values do not depend on the assignment of the units
2. $\tau = \tau(\underline{Y})$ depends on $\tau(\underline{Y}, x)$ (the other covariates)

1.4 Superpopulation Estimands

So far, we have assume a fixed and finite population with no mode, and that τ is specific to the population of size N. However, sometimes, it is reasonable to assume that the units we see are an iid draw from a superpopulation. Therefore, we have that $(Y_i(1), Y_i(0)) \overset{iid}{\sim} P$, with a well-defined first moment, so $E_P[Y_i(0)] = \mu_0$ and $E_P[Y_i(1)] = \mu_1$. We then have that $\theta_i = E_P[\tau_P] = \mu_1 - \mu_0$, and thus that $\theta = E_P[\tau]$, where τ_i and τ are the finite population quantities. For more clarity, τ_i is the individual effect and τ is the ATE.

In a randomized experiment, the distinction with or without the superpopulation does not matter most of the time. We can just assume we start with the finite population assumption.

1.5 No Causation without Manipulation

In order for us to determine causation, we need to have manipulation. However, we want to think carefully about how to design the experiment. For example, say we want to find the effect of BMI on cardiovascular risk. What does it mean to manipulate BMI?

We could think of ways to manipulate BMI (the treatment) by sawing off people's legs, force feeding people, or dehydrating people, but for obvious reasons those are infeasible. It is better instead to have a clearly defined and manipulable intervention. For example, we could look at the impact of following diet x for 3 days or the impact of following fitness program y.

One more example which could help illustrate the point. Say we want to find the effect of race on prosecutorial charging decisions. We need to clearly define what it means to manipulate race. Are we changing skin color, culture, etc.? Is it manipulable? Audit studies have been performed to manipulate people's "perception" of race, such as in studying discrimination in hiring, submitting the same resume with different ethnic names attached.

2 Randomized Experiments

For this section, we will assume the classic SUTVA assumptions.

2.1 The Randomization-Based Framework

- Why randomize?
  - Avoid confounding
  - Objectivity
- Randomization can remove the need for most assumptions
  - Don't need to assume iid
  - No modeling assumption
  - Other assumptions not required
- "In the absence of difficulties, assumptions play a minor role in randomized experiments, and no role at all in randomization tests of the hypothesis of no effect."

Recall that we can define $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. In classical statistics, we would make assumptions such as assuming that $Y_i|_{Z_i=1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and that $Y_i|_{Z_i=0} \overset{iid}{\sim} N(\mu_0, \sigma_0^2)$. To calculate the effect, we could just calculate an MLE such as $\hat{\tau}^{MLE} = \hat{\mu}_1^{MLE} - \hat{\mu}_0^{MLE}$. The problem with this approach is that we make a lot of assumptions. For example we assume iid-ness and also normality in this case, which may or may not be accurate to our problem.

On the other hand, in randomization-based inference, we take the view that "The Science" is fixed. The ONLY randomization in this framework is how we assign the units to the groups. In mathematical notation, we have that $\underline{Y} = (\vec{Y}(0), \vec{Y}(1))$ is fixed and that $\vec{Z}$ is random. Therefore, what we observe, $\vec{Y} = \vec{Y}(\vec{Z})$ is random because it is a function of $\vec{Z}$. For example, if our true "The Science" table was below:

i	$Y(0)$	$Y(1)$
1	2	4
2	1	3
3	0	4
4	1	5

a possible outcome of what we would actually observe could be

Z	$Y_i(Z)$
0	2
0	1
1	4
1	5

or possibly

Z'	$Y_i(Z')$
0	2
1	3
0	0
1	5

Since all the randomness comes from $\vec{Z}$, the distribution of $\vec{Z}$ will play a crucial role. We will assign based on a distribution/design P. Therefore, $\vec{Z} \sim P(\vec{Z})$.

2.1.1 Assignment Mechanisms

There are multiple ways to assign treatment and control to the different units. Two popular ways are shown below:

- Bernoulli with probability p
  - $P(Z_i = 1) = p$
  - $P(\vec{Z}) = \prod_{i=1}^{N} p^{Z_i}(1-p)^{1-Z_i}$
  - Note that the number in each group is a random quantity
  - This method would be like flipping coin for each unit and assign the heads to one group and the tails to another
- Completely Randomized Design (CRD (N_1, N))
  - $P(Z_i = 1) = \frac{N_1}{N}$
  - $P(\vec{Z}) = \binom{N}{N_1}^{-1}$ if $N_1(Z) = N_1$ and 0 otherwise
  - This method would be like predetermining having 50 people out of 100 in treatment group and 50 in control and then picking 50 names out of a hat without replacement (and those names would all go to treatment group)

If generally, we have that $P(\vec{Z}|\vec{X}, \vec{Y}(0), \vec{Y}(1))$, in other words that which group a unit is assigned to depends on possibly their outcome or other covariates, then we can define a randomized experiment as one that satisfies the following properties:

- Probabilistic: we have that $0 < P(Z_i|\vec{X}, \vec{Y}(0), \vec{Y}(1)) < 1$. That is, that there is a non-zero probability of getting some assignment into a group
- Known assignment mechanism: the probability is known
- Individualistic: An assignment for a unit does not depend on other units. $P(Z_i|\vec{X}, \vec{Y}(0), \vec{Y}(1)) = P(Z_i|\vec{X}_i, Y_i(0), Y_i(1))$
- Unconfoundedness: The assignment does not depend on the outcome. $P(\vec{Z}|\vec{X}, \vec{Y}(0), \vec{Y}(1)) = P(\vec{Z}|\vec{X})$

We can see that both the Bernoulli with probability p and the CRD(N_1, N)) fulfill all of these properties and can be used to create randomized experiments.

2.2 Neymanian Inference

Our estimand in Neymanian inference is the ATE (Average Treatment Effect). Recall that we have $\tau^{ATE} = \bar{Y}(1) - \bar{Y}(0)$. If we observe $\vec{Z}^{obs}$ and $\vec{Y}^{obs}$, then we can calculate the estimator as $\hat{\tau} = \bar{Y}^{obs}(1) - \bar{Y}^{obs}(0)$, the difference in means.

Even though we only observe one $\vec{Z}$, it comes from some distribution, call it η. Therefore, $Z \sim \eta$ induces a distribution on $\hat{\tau} = P_\eta(\hat{\tau})$. Similar to classical statistics, we are interested in the bias and the variance of our estimator $\hat{\tau}$. The difference is that in classical stats, the randomness is caused by the modeling assumptions (such as Normal distribution and iid-ness), whereas here, the randomness comes from the randomness of assignment only! The definition of bias is:

- $\text{Bias}_\eta(\hat{\tau}, \tau : \underline{Y}) = E_\eta[\hat{\tau}(\vec{Z}, \underline{Y})] - \tau(\vec{Y})$
- $\hat{\tau}$ is called unbiased for τ under η if for all $\underline{Y}$, $\text{Bias}_\eta(\hat{\tau}, \tau, \underline{Y}) = 0$

Our $\hat{\tau}$ in this type of inference is the difference in means estimator. We claim that if η is CRD(N_1, N), that $\hat{\tau}$ is unbiased. In other words, $\hat{\tau}^{DIM}$ is unbiased for τ^{DIM} under a CRD. This is an amazing result since we've made no assumption beyond SUTVA, and that this is true for any $\underline{Y}$! The proof of this is below.

1. Recall
   - $\tau = \tau^{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$
   - $\hat{\tau} = \hat{\tau}^{DIM} = \frac{1}{N_1(Z)}\sum_{i=1}^{N} Z_i Y_i - \frac{1}{N_0(Z)}\sum_{i=1}^{N}(1-Z_i)Y_i$
   - Under CRD(N_1, N), $N_1(Z) = N_1$ and $N_0(Z) = N - N_1 = N_0$, which are both constants
   - Therefore, $\hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N} Z_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1-Z_i)Y_i$
2. Linearity of Expectation
   - $E_\eta[\hat{\tau}] = \frac{1}{N_1}\sum_{i=1}^{N} E_\eta[Z_i Y_i] - \frac{1}{N_0}\sum_{i=1}^{N} E_\eta[(1-Z_i)Y_i]$
3. Simplification

- $Z_iY_i = Z_i[Z_iY_i(1) + (1-Z_i)Y_i(0)] = Z_i^2Y_i(1) + Z_i(1-Z_i)Y_i(0) = Z_iY_i(1)$, since $Z_i(1-Z_i) = 0$
- Similarly, $(1-Z_i)Y_i = (1-Z_i)Y_i(0)$
- Plugging this in, we have $E_\eta[\hat{\tau}] = \frac{1}{N_1}\sum_{i=1}^N E_\eta[Z_i]Y_i(1) - \frac{1}{N_0}\sum_{i=1}^N E_\eta[1-Z_i]Y_i(0)$

4. Probability Bridge

- Z_i is an indicator random variable so $E[Z_i] = P[Z_i = 1] = \frac{N_1}{N}$
- Similarly, $(1-Z_i) = \frac{N_0}{N}$
- Plugging this in, we have $E_\eta[\hat{\tau}] = \frac{1}{N_1}\sum_{i=1}^N \frac{N_1}{N}Y_i(1) - \frac{1}{N_0}\sum_{i=1}^N \frac{N_0}{N}Y_i(0) = \frac{1}{N}\sum_{i=1}^N Y_i(1) - \frac{1}{N}\sum_{i=1}^N Y_i(0) = \tau^{ATE}$

How about variance? The definition of variance is $Var_\eta(\hat{\tau}) = E_\eta[(\hat{\tau} - E_\eta[\hat{\tau}])^2]$. We make the following claim: if $\tau = \tau^{ATE}$, $\hat{\tau} = \hat{\tau}^{DIM}$, and $\eta = \text{CRD}(N_1, N)$, then

$Var(\hat{\tau}) = \frac{V_1}{N_1} + \frac{V_0}{N_0} - \frac{V_{10}}{N}$ where
$V_1 = \frac{1}{N-1}\sum_{i=1}^N (Y_i(1) - \bar{Y}(1))^2$
$V_0 = \frac{1}{N-1}\sum_{i=1}^N (Y_i(0) - \bar{Y}(0))^2$
$V_{10} = \frac{1}{N-1}\sum_{i=1}^N (\tau_i - \tau)^2$

2.2.1 Neymanian Inference in Practice

In practice, although Neymanian Inference only assumes SUTVA, in order to perform inference such as create a confidence interval, we need to assume normality. For example, we would assume we have a large enough sample for normality, and thus we claim that $\frac{\hat{\tau}-\tau}{\sqrt{Var(\hat{\tau})}} \sim N(0,1)$. We can then create our confidence interval as $CI_{1-\alpha} = [\hat{\tau} - q_{1-\frac{\alpha}{2}}\sqrt{Var(\hat{\tau})}, \hat{\tau} + q_{1-\frac{\alpha}{2}}\sqrt{Var(\hat{\tau})}]$. Then, by definition of CI, we have that $P_\eta(\tau \in CI_{1-\alpha}) = 1 - \alpha$.

However, $Var(\hat{\tau})$ depends on unknown quantities τ_i and τ, and the values in the science table (as seen from the variance formula above), so we need to find an estimator $\hat{Var}(\hat{\tau})$ instead. The general strategy is to be conservative, and choose a $\hat{Var}(\hat{\tau})$ such that $E[\hat{Var}(\hat{\tau})] \geq Var(\hat{\tau})$. That way, we have the same or larger confidence interval (more conservative).

For example, recall that $Var(\hat{\tau}) = \frac{V_1}{N_1} + \frac{V_0}{N_0} - \frac{V_{10}}{N}$. We can therefore write $\hat{Var}(\hat{\tau}) = \frac{\hat{V_1}}{N_1} + \frac{\hat{V_0}}{N_0} - \frac{\hat{V_{10}}}{N}$ where
$\hat{V_1} = \frac{1}{N_1-1}\sum_{i=1}^N Z_i(Y_i^{obs} - \bar{Y}(1)^{obs})^2$
$\hat{V_0} = \frac{1}{N_0-1}\sum_{i=1}^N (1-Z_i)(Y_i^{obs} - \bar{Y}(0)^{obs})^2$
$\hat{V_{10}} = 0$

Therefore, the Neyman estimate of variance is $\hat{Var}(\hat{\tau}) = \frac{\hat{V_1}}{N_1} + \frac{\hat{V_0}}{N_0}$. Then under $\text{CRD}(N_1, N)$, $E_\eta[\hat{Var}(\hat{\tau})] \geq Var(\hat{\tau})$.

The flip-side of the CI is hypothesis testing. Our null hypothesis is $H_0 : \bar{Y}(1) - \bar{Y}(0)$, which also can be stated as $\tau^{ATE} = 0$. We then calculate our test statistic $T^{obs} = \frac{\hat{\tau}}{\sqrt{\hat{Var}(\hat{\tau})}}$. Using a Z-table, we can calculate our p-value as $1 - \Phi(T^{obs})$ (1-sided), or $2 * (1 - \Phi(T^{obs}))$ (2-sided). By definition, our p-value is $P_\eta(pval \leq \alpha | H_0) \leq \alpha$.

2.2.2 Horvitz-Thompson Estimator

So far, we have been using $\hat{\tau}^{DIM}$ as our estimator for the estimand τ^{ATE}. However, all the proof done above has been for a $\text{CRD}(N_1, N)$ design. What happens if we are provided an exotic randomized design? We can instead use the Horvitz-Thompson estimator (the hammer that works for all situations)! Note however, that it may not always be the "best" estimator, though.

The definition of the Horvitz-Thompson estimator is as follows. Let η be any design that is a randomized experiment, so $\pi_i = P_\eta(Z_i = 1)$, where $0 < \pi_i < 1$.

Then $\hat{\tau}^{HT} = \frac{1}{N}\sum_{i=1}^N \frac{Z_i}{\pi_i}Y_i - \frac{1}{N}\sum_{i=1}^N \frac{1-Z_i}{1-\pi_i}Y_i$.

Note that if $\eta = \text{CRD}(N_1, N)$, then $\pi_i = \frac{N_1}{N}$ and then $\hat{\tau}^{HT} = \hat{\tau}^{DIM}$. However, the beauty of the HT estimator is that is extends beyond CRD. Let η be any design that is a randomized experiment, then the HT estimator $\hat{\tau}^{HT}$ is unbiased for τ^{ATE}. Note that just because the HT estimator is unbiased though, does not always mean it is the best estimator since it could have very high variance! Proof below:

- Recall $\hat{\tau}^{HT} = \frac{1}{N}\sum_{i=1}^N \frac{Z_i}{\pi_i}Y_i - \frac{1}{N}\sum_{i=1}^N \frac{1-Z_i}{1-\pi_i}Y_i$

- Denote $\hat{\tau_1}^{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i}{\pi_i}Y_i$ and $\hat{\tau_0}^{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{1-Z_i}{1-\pi_i}Y_i$
- By linearity of expectation, $E[\hat{\tau}^{HT}] = E[\hat{\tau_1}^{HT}] - E[\hat{\tau_0}^{HT}]$
- $E_\eta[\hat{\tau_1}^{HT}] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}E[Z_iY_i] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}E[Z_iY_i(1)] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}Y_i(1)E[Z_i] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}Y_i(1)P[Z_i = 1] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}Y_i(1)\pi_i = \bar{Y}(1)$
- Similarly, $E_\eta[\hat{\tau_0}^{HT}] = \bar{Y}(0)$
- Therefore, $E[\hat{\tau}^{HT}] = E[\hat{\tau_1}^{HT}] - E[\hat{\tau_0}^{HT}] = \bar{Y}(1) - \bar{Y}(0) = \tau^{ATE}$

2.3 Fisher Randomization Test

Neymanian inference makes very few assumptions, but it does require an asymptotic assumption of normality in order to perform inference such as Confidence Intervals and Hypothesis Testing. It also may be difficult to perform Neymanian inference for arbitrary designs and estimators (proof of unbiasedness). Therefore, we can turn to the Fisher randomization test, which makes no assumptions whatsoever, except for SUTVA.

Recall that the Neyman null hypothesis is $H_0^{Neyman} : \bar{Y}(1) = \bar{Y}(0)$. Therefore, we are saying that if the null holds, it is possible that some units have positive effect and some have negative, but on average they cancel out. With the Fisher null hypothesis, we make a much stronger statement. The Fisher null hypothesis is $H_0^{Fisher} : Y_i(1) = Y_i(0)$ for all i. We are saying that if the null holds, there is no effect for every single unit. Therefore, Fisher null implies the Neyman null and is a stronger hypothesis. There is some criticism that the Fisher null is too strong (and therefore by rejecting the null we don't learn much).

Similar to before, for FRT, we assume $\underline{Y}$ is fixed but unknown, $Z^{obs} \sim \eta$ and $Y^{obs} = Y(Z^{obs})$. We compute our test statistic $T(Z^{obs}, Y^{obs})$. We want to know how likely it is that we have observed a value of T^{obs} as extreme as we did. We know there is an underlying null distribution of T, and we calculate our pval as $P_\eta(T(Z, Y^{obs}) \geq T^{obs}|H_0) = \sum_Z \mathbb{1}\{T(Z, Y^{obs}) \geq T^{obs}\}P_\eta(Z)$. The FRT is a sharp test. That means that if H_0 is true, we can recover the $\underline{Y}^{(0)}$.

The algorithm for the FRT is as follows. We have FRT$(Z^{obs}, Y^{obs}, T(\cdot), \eta)$.

1. Having observed Z^{obs}, Y^{obs}, we compute $T^{obs} = T(Z^{obs}, Y^{obs})$
2. For $k = 1, ..., K$, we draw $Z^{(k)} \sim \eta$. We then compute $T^{(k)} = T(Z^{(k)}, Y(Z^{(k)}))$.
3. Compute a Monte Carlo approximation of the pval. We calculate $\hat{pval} = \frac{1}{K}\sum \mathbb{1}\{T^{obs} \leq T^{(k)}\}$. By the LLN, as $K \to \infty$, $\hat{pval} \to pval$

Note again that for FRT, we make no assumptions whatsoever (outside of SUTVA). Our Gaussian assumption does not come from an assumption about the underlying data, but rather the repeated draws that we perform. Note also that the validity of this test does not depend on our choice of $T(\cdot)$.

2.3.1 Power and Choice of Test Statistic

A good test must both control the Type I Error (valid) as well as have high power to detect certain violations of the null hypothesis. Recall that if we test at level α, then the Type I Error is defined as $P_\eta(pval \leq \alpha|H_0)$. The Power is defined as $P_\eta(pval \leq \alpha|H_1)$.

What affects the power of a test? The design η and the choice of test statistic $T(\cdot)$ do. For example, the Wilcoxian rank sum statistic is more robust to outliers than a T statistic.

2.3.2 Confidence Intervals and Point Estimates

In Neymanian Inference, the CI and point estimate was quite clear. However, in the Fisher Randomization test, we have to do a bit more work.

We had examined the null hypothesis where $H_0 : Y_i(1) = Y_i(0)$ for all units i. However, we can also choose other sharp nulls, where we define $H_0^{(\theta)} : Y_i(1) = Y_i(0) + \theta$, where the θ is a fixed constant. In this scenario, we can fill in the science table $\underline{Y}^{(\theta)}$ (now adding and subtracting θ as appropriate) and run the same Fisher algorithm as before. However, the only difference is that before we had that $Y(Z^{(k)}) = Y^{obs}$, which we used to calculate our T. Now, instead, we add or subtract θ as appropriate.

The reason we examine other sharp nulls is the following. We define our Confidence Interval as all the θ where the test does not reject $H_0^{(\theta)}$ In other words, $CI_{1-\alpha}(Z^{obs}) = \{\theta : pval_{H_0^\theta}(Z^{obs}) \geq \alpha\}$.

For a point estimate, we can use the Hodges-Lehman point estimate. The intuition behind this is that for a symmetric null distribution, we can keep additively shifting it over until we find the θ' that maximizes the likelihood. In other words, where our statistic is the mode of the $H_0^{(\theta')}$ distribution. With $0 < \theta < \theta'$, we find $\hat{\tau}^{HL} = \theta'$. More formally, we define $m(\theta) = E[T(Z, Y(Z))|H_0^{(\theta)}]$. The definition of our HL estimate is $\hat{\tau}^{HL} = zero_\theta(T^{obs} - m(\theta))$.

We claim that if $\hat{\tau}^{HL}$ exists, then $\hat{\tau}^{HL} \xrightarrow{P} \theta$. Note that in the special case where T is defined to be the difference in means, that $\hat{\tau}^{HL} = T^{obs} = \hat{\tau}^{DIM}$.

2.4 Covariate Adjustment

When we ran a randomized experiment, it is possible that all (or many) of one type of unit is randomly assigned to treatment or control, while all of another type of unit is randomly assigned to the other group. Then, the result we observe may not be the true result that we are trying to learn about. For example, let's say we are trying to find the effect of a fertilizer on the weight of fruits, and we perform a CRD(N_1, N) on a set of fruits that contain both apples and melons. In an unlucky scenario, all the apples could be randomly assigned to the control and the melons assigned to the treatment group. Then, when we look at the average treatment effect, we think that the fertilizer has a positive effect on the weight of the fruit, when in reality it may just be due to the fruit itself. Note that this issue is a variance issue, not a bias issue. The problem is that the randomization can put all of one type of unit into a group, even thought the expected value is a balanced distribution.

There are two main ways to address this issue. The first is from the design approach. We can change the design of the experiment to force balance within groups. For example, we can require that each of the treatment or control has half apples and half melons. The second is from the analysis approach. We could change our estimator. Instead of $\hat{\tau}^{DIM}$, we could calculate a $\hat{\tau}_M^{DIM}$ and a $\hat{\tau}_A^{DIM}$ (one for melons and one for apples). Then our estimator is the weighted average of these two values. $\hat{\tau} = w_A\hat{\tau}_A + w_M\hat{\tau}_M$. Note that the design and analysis can be interconnected. Also the design is more important that the analysis. If the experiment is designed well, the analysis will be easy (and simple). If the design is awful, no amount of clever analysis can save us.

2.4.1 Discrete Covariates

Recall that CRD(N_1, N) can be thought of drawing N_1 units without replacement out of a hat and assigned them to treatment. With a covariate adjustment, we perform one CRD per covariate adjustment, rather than as a whole.

More concretely, assume our population has males and females. In a CRD(N_1, N), we would randomly select N_1 units from all the people (regardless of gender) and assign them to the treatment group. Now, instead, for the males, we use CRD$(N_1(M), N(M))$ and for the females, we use CRD$(N_1(F), N(F))$. This is called a Stratified Completely Randomized Design (SCRD). Note that now we have $\vec{N_1}$ and $\vec{N}$, instead of scalars. In this case, $\vec{N_1} = (N_1(M), N_1(F))$ and $\vec{N} = (N(M), N(F))$.

Now let us denote $S_i \in \{1, ..., K\}$, where each of the 1 through K represent a different stratum. Then with SCRD, we say that the two important properties are:

- $P(Z_i = 1|S_i = k) = \frac{N_1(k)}{N(k)}$
- If $S_i \neq S_j$, then Z_i is independent of Z_j (note that this is not the case for CRD).

With Neymanian inference with SCRD, we do the following. Recall $\tau^{ATE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$. Let $\tau_k = \frac{1}{N(k)}\sum_{i=1}^{N}\mathbb{1}\{S_i = k\}(Y_i(1) - Y_i(0))$ (the average treatment effect for units in stratum k). Then we have that $\tau^{ATE} = \sum_{k=1}^{K} w_k\tau_k$ where $w_k = \frac{N(k)}{N}$ and $\sum_{k=1}^{K} w_k = 1$. This suggests that the estimator we can use is $\hat{\tau}^{SRCD} = \sum_{k=1}^{K} w_k\hat{\tau}_k$.

We claim that this estimator in unbiased. In other words, when η is SCRD, then $E_\eta[\hat{\tau}_k] = \tau_k$. For clarity, we emphasize that $\hat{\tau}_k = \frac{1}{N_1(k)}\sum_{i=1}^{N}\mathbb{1}\{S_i = k\}Z_iY_i - \frac{1}{N_0(k)}\sum_{i=1}^{N}\mathbb{1}\{S_i = k\}(1 - Z_i)Y_i$. Further more, we claim that when η is SCRD, $E_\eta[\hat{\tau}^{SCRD}] = \tau^{ATE}$. We prove that as follows:

$E_\eta[\hat{\tau}^{SCRD}] = E_\eta[\sum_{k=1}^{K}\frac{N(k)}{N}\hat{\tau}_k]$
$= \sum_{k=1}^{K}\frac{N(k)}{N}E_\eta[\hat{\tau}_k]$
$= \sum_{k=1}^{K}\frac{N(k)}{N}\tau_k$

$= \tau^{ATE}$

How about the variance? We claim that by leveraging the fact that we have "independent experiments across strata," that we reduce the variance of the estimator. We know that in η is SCRD, then $\hat{\tau}_k$ is independent of $\hat{\tau}_{k'}$. Thus, $Var(\hat{\tau}) = \sum_{k=1}^{K}[\frac{N(k)}{N}]^2[\frac{v_1(k)}{N_1(k)} + \frac{v_0(k)}{N_0(k)} - \frac{v_{10}(k)}{N(k)}]$, where each of the items in the sum is $Var(\hat{\tau}_k)$. Similar to before, we need to estimate this variance since it contains unknown parameters. Thus, we use a conservative variance estimate, and $\hat{Var}(\hat{\tau}) = \sum_{k=1}^{K}[\frac{N(k)}{N}]^2[\frac{\hat{v_1}(k)}{N_1(k)} + \frac{\hat{v_0}(k)}{N_0(k)}]$.

The intuition as to why the variance decreases when we stratify can be thought of as follows. If we did not stratify, we would be looking at v_1 (or v_0), and when we stratify we have $v_1(k)$ (or $v_0(k)$). Supposed $Y_i(1)$ is constant within each stratum but not across strata. This would be like our fruit case, where all apples weigh the same and all melons weigh the same, but apples do not weigh the same as melons. Then if we did not stratify, we would have $v_1 > 0$. When we stratify, each of the $v_1 \approx 0$ and when summed up, gives up a total of ≈ 0.

Similar to regular CRD, the CLT (asymptotics) follows in SCRD so we can do inference. An important point is that we have not made any assumption about the relationship between X_i (the covariate) and $Y_i(1)$ and $Y_i(0)$. Thus, it is always in our best interest to stratify when possible.

With a Fisher Randomization Test, we have no change in the approach. The only difference is that before we had $\text{FRT}(Z^{obs}, \eta^{CRD}, T(\cdot), H_0)$, and now we simply replace η^{CRD} with η^{SCRD}. Note that we should also consider switching to a test statistic that incorporates the covariate, but this is not required. For example, by switching from $T = \hat{\tau}^{DIM}$ to $T = \hat{\tau}^{SCRD}$.

In summary, SCRD allows for very simple analysis that generally reduces variance. It doesn't require any additional assumptions.

2.4.2 Continuous Covariates

With a continuous covariate (such as age) as opposed to a discrete covariate (such as gender), we cannot just separate the units into groups using the covariate, since most likely there will be at most one observation with that value of the continuous covariate. We will look at two different strategies for continuous covaraites, rerandomization and agnostic regresion.

In rerandomization, we apply CRD until we obtain a Z that we deem to be balanced. If we randomly draw a Z that we deem to be unbalanced, we reject it and redraw the assignment. More formally, we measure the covariate balance as $\hat{\tau}_X = \frac{1}{N_1}\sum_{i=1}^{N} Z_i X_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - Z_i)X_i$. We calculate $\rho(Z^{obs}; X) = |\hat{\tau}_X|$ and if it is too large, we redraw. The formal procedure is as follows:

- Draw $Z^{obs} \sim \eta$
- Compute the balance $\rho(Z^{obs}; X)$.
  - If $\rho(Z^{obs}; X) < gamma$, then keep Z^{obs}
  - If $\rho(Z^{obs}; X) \geq \gamma$, then reject Z^{obs} and go back to the previous step and redraw Z^{obs}.

Note that with rerandomization, we go from η to η_γ. Therefore, we have that $P_{\eta_\gamma}(Z) = P_\eta(Z|\rho(Z; X) < \gamma) = \mathbb{1}\{Z \in Z_\gamma\}\frac{P_\eta(Z)}{P_\eta(Z_\gamma)}$. This is a type of rejection sampling.

In terms of analysis, we can use the $\text{FRT}(Z^{obs}, \eta, T(\cdot), H_0)$, or the Neymanian Inference.

The second method is to use Agnostic Regression. In standard regression, we would have N units, and let us denote our covariate/s as X_i. Then we would run a regression with our linear model being $Y_i = \mu + Z_i\theta + \beta X_i + \epsilon_i$. We would assume $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$, and θ would be our estimand.

A few issues with this approach is firstly that we assume parametric/semi-parametric model and independence. Additionally, we've defined causal estimands as contrasts between potential outcomes and it is a bit unclear how that connects to θ, since if the model changes, θ also changes.

Agnostic regression takes a different view, where we consider $\hat{\theta}^{OLS}$ from that same linear model, but then we forget about the model completely. We have that $\hat{\theta}^{OLS} = \hat{\theta}^{OLS}(Z^{obs}, Y^{obs})$ is a random variable. Our new estimator $\hat{\tau}^{OLS} = \hat{\theta}^{OLS}$.

Under mild conditions (only assuming asymptotic normality), it can be shown that $\hat{\tau}^{OLS} \sim N(\tau^{ATE}, Avar(\hat{\tau}^{OLS}))$, where $Avar$ means asymptotic variance. The fact that $\hat{\tau}^{OLS} \to \tau^{ATE}$ is amazing (although not unbiased for a finite sample, only for

infinite)! However, $Avar(\hat{\tau}^{OLS})$ may be greater than $Avar(\hat{\tau}^{DIM})$, which is not what we want.

Instead, we use the linear model $Y_i = \mu = Z_i\theta + \beta X_i + \gamma Z_i(X_i - \bar{X}) + \epsilon_i$. Call this $\hat{\tau}^{REG} = \hat{\theta}^{OLSnew}$. This is the linear model we had before except with interaction terms between all the covariates and the assignment and zero-mean centered.

It can be shown that if the asymptotic normality condition holds, if $\frac{1}{N} max\{Y_i(Z) - \bar{Y}(Z)\}^2 \to 0$, and if $\frac{1}{N} max\{X_i - \bar{X}\}^2 \to 0$, then $\frac{\hat{\tau}^{REG} - \tau^{ATE}}{\sqrt{Avar(\hat{\tau}^{REG})}} \sim N(0,1)$ and $Avar(\hat{\tau}^{REG}) \leq Avar(\hat{\tau}^{DIM})$! This is an amazing result since it says nothing about how X and Y are related. In fact, if X is more related to Y, then we get a lower $Avar$. Therefore, the takeaway is that if you include interaction terms and center the covariates, then the regression estimator is never worse than DIM asymptotically.

We can estimate the variance of the estimator by the Huber-White sandwich (asymptotically conservative).

Summary:

- Without assuming the linear model, we can justify linear regression in randomized experiments
- If you do regression adjustments:
  - Center the covariates
  - Include interactions
  - Use Huber-White standard errors

2.5 Interference

So far, everything that has been done has assumed SUTVA. Recall SUTVA has two assumptions: no interference between units, and no hidden version of the treatment. We will now relax the no interference assumption (while keeping the no hidden version assumption).

Some real life examples where SUTVA may not hold include getting out votes, vaccines, policing, and marketplaces. In each of these, putting one member in a treatment can affect others around them, that may be placed in control.

Why is interference a problem? Well if SUTVA holds, we can randomly assign treatment and control and then compare them. If we have interference, then it is possible that some that are assigned to treatment may influence those around them that may be assigned to control. Then, when we compare, we would be comparing treatment to a mix of control and treatment-affected control. In the most extreme case where treatment has a strong effect of those around them, we might be comparing treatment to a mix of control and treatment. This would lead to an underestimation of the effects of the treatment.

The intuition for how to solve this would be to identify 3 groups, treatment, control, and control who was affected by neighbors receiving treatment. To analyze our primary effect, we would compare control to treatment. We would then analyze a spillover effect by comparing control to controls that were affected by neighbors receiving treatment.

Mathematically, we can see this as follows. Under SUTVA, we have that $Y_i(\vec{Z}) = Y_i(Z_i)$, so each unit has 2 potential outcomes. If SUTVA does not hold, then $Y_i(\vec{Z}) = Y_i(\vec{Z})$ and we have 2^N potential outcomes.

Without SUTVA, our estimand now is $\tau^{OE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(\vec{1}) - Y_i(\vec{0}))$. Note that this is different than ATE since we are now comparing the vector of all 1's to the vector of all 0's. If SUTVA did hold, then then would be equal.

Let's look at an example. Assume $N = 3$. Assume I observe $Z^{obs} = (1,0,0)$ and I also observe Y^{obs}. Under SUTVA, this would give me some information about $\vec{Y}(\vec{1})$ and $\vec{Y}(\vec{0})$. Out of the 6 numbers in my Science table, I observe 3 of them. Under arbitrary inference, I have no information at all about $\vec{Y}(\vec{1})$ and $\vec{Y}(\vec{0})$! This is because the vector of $(1,0,0)$ tells me nothing about the vector of $(1,1,1)$ or the vector of $(0,0,0)$, which I need for my estimator. Thus, I have learned 0 of the 6 missing values in my Science table.

2.5.1 SDP (School District of Philadelphia) Experiment

In order to better understand interference let's look at the SDP (School District of Philadelphia) Experiment. This experiment was looking for the effect of sending "late notices" to parents of absent children from school. In this experiment, the selection was done in 2 stages. In the first stage, CRD was done on the household level, where certain households with children who missed class were selected. Each of these households could have a different number of children. In the second stage, out of these selected households, exactly one child was selected from that household as "treatment" where the letter was sent to

their parents. Essentially there is interference because if one parent received a notice for one of their absent children, it likely would affect the behavior of the other absent children in that same household.

Notation:

- $Z_{ij} \in \{0,1\}$, where i = household and j = a unit within a household
- n_i = total number of units in household i
- $\vec{Z}_i = (Z_{i1}, ..., Z_{in_i})$ is the vector of assignments for a household
- $\vec{Z} = (\vec{Z}_1, ..., \vec{Z}_M)$
- $Y_{ij}(\vec{Z})$ is what we measure (no SUTVA so cannot simplify)

In order to reduce the number of potential outcomes (even without SUTVA), we make two assumptions:

- Assumption 1: Partial Interference
  - Interference can happen within households but not outside of the household
  - $Y_{ij}(\vec{Z}) = Y_{ij}(\vec{Z}_i)$ for all i, j
- Assumption 2: Stratified Interference
  - $Y_{ij}(Z_{ij} = 0, \vec{Z}_{i,-j}) = Y_{ij}(Z'_{ij} = 0, \vec{Z'}_{i,-j})$
  - This is for all $\vec{Z}_{i,-j}, \vec{Z'}_{i,-j}$ such that $\sum_{k\neq j} Z'_{ik} = \sum_{k\neq j} Z_{ik} = 1$
  - If I am unit j and I am untreated and one of my siblings is treated, it doesn't matter to me which one of my siblings is treated, so long as 1 is treated
  - More clearly, if there are 3 siblings, $Y_{i1}(0,1,0) = Y_{i1}(0,0,1)$

Putting these two assumptions together, we have that $Y_{ij}(\vec{Z}) = Y_{ij}(W_i, Z_{ij})$, that this only depends on if a household is treated or not, and whether j is treated or not. In order words, we have 3 potential outcomes: $Y_{ij}(1,1)$ (treated), $Y_{ij}(1,0)$ (spillover), and $Y_{ij}(0,0)$ (control).

Assuming households have the same size (this can easily be generalized to households of different sizes), our causal estimands are:

- Primary effect: $\tau^P = \frac{1}{N}\sum_{i,j}\{Y_{ij}(1,1) - Y_{ij}(0,0)\}$
- Spillover effect: $\tau^S = \frac{1}{N}\sum_{i,j}\{Y_{ij}(1,0) - Y_{ij}(0,0)\}$

To estimate this estimand, we can use the IPW (inverse probability weighted estimator). Let's define $\pi_{ij}(a,b) = P(W_i = a, Z_{ij} = b)$. Then our estimator can be the generalization of the Horvitz-Thompson estimator:

$\hat{\tau}^P = \frac{1}{N}\sum_{i,j}\frac{\mathbb{1}(W_i=1,Z_{ij}=1)}{\pi_{ij}(1,1)}Y_{ij} - \frac{1}{N}\sum_{i,j}\frac{\mathbb{1}(W_i=0,Z_{ij}=0)}{\pi_{ij}(0,0)}Y_{ij}$

In our setting, suppose the first stage of selection (with households) is CRD(M_1, M). Then by Bayes rule, $P(W_i = 1, Z_{ij} = 1) = P(Z_{ij} = 1|W_i = 1)P(W_i = 1)$. Thus, we have that $\pi_{ij}(1,1) = \frac{1}{n_i} * \frac{M_1}{M}$ and $\pi_{ij}(0,0) = \frac{M_0}{M}$.

2.5.2 The Interference Machinery

1. Exposure Mappings
   - Interference structure is constrained via exposure mappings
   - For every i, we define $h_i : \{0,1\}^N \to H$
     - For $\vec{Z}$, we have that $h_i(\vec{Z}) \in H$
     - For example, in the SDP Experiment, we have $h_{ij}(\vec{Z}) = (\sum_{k=1}^{n_i} Z_{ik}, Z_{ij})$, where $H = \{(1,1),(1,0),(0,0)\}$
   - We assume a well specified exposure mapping. Then for every $\vec{Z}, \vec{Z'}$, if $h_i(\vec{Z}) = h_i(\vec{Z'})$, then $Y_i(\vec{Z}) = Y_i(\vec{Z'})$
   - Thus, $Y_i(\vec{Z}) = Y_i(h_i(\vec{Z}))$, and since $H_i = h_i(\vec{Z})$, then $Y_i(\vec{Z}) = Y_i(H_i)$
   - Insight: Think of exposure H_i as effect treatment. Under this assumption, the potential outcomes are $\{Y_i(k)\}_{k\in H}$, and there are H potential outcomes for each unit

- Examples:
  (a) SUTVA: $h_i(\vec{Z}) = Z_i$, and $H = \{0, 1\}$
  (b) Network (1-local exposure) with $G(V, E)$: Let $e_{ij} = 1$ if i, j connected, else 0, then $N_i = \{j \in V : e_{ij} = 1\}$ (neighborhood of i). Then $h_i(\vec{Z}) = (W_i, Z_i)$ where $W_i = \frac{1}{|N_i|} \sum_{j \in N_i} Z_j$
  (c) Spatial Spillovers: Let V be a set of units in space (e.g. coordinates) and let d_{ij} be the distance between unit i and unit j. Then we have the same result as the Network mapping but where $e_{ij} = \mathbb{1}\{d_{ij} \leq d\}$, where d is some predetermined distance.

2. Causal Estimands

- Under SUTVA, the causal estimand is the contrast between potential outcomes under different treatments. Under interference, the casual estimands are the contrasts between potential outcomes under different exposures.
- $\tau_{kk'} = \frac{1}{N} \sum_{i=1}^{N} \{Y_i(H_i = k) - Y_i(H_i = k')\}$, where $k, k' \in H$
- For example, in SDP, we could denote $k = (1, 1)$ and $k' = (0, 0)$ and $\tau_{kk'} = \tau^P$. In the Network, we could denote $k = (\frac{1}{2}, 0), k' = (0, 0)$. We could also have denoted $k = (\frac{1}{3}, 0)$. Note that there isn't only ONE single "spillover" effect estimand
- Caveat:
  (a) Consider the Network setting, $h_i(\vec{Z}) = (\sum_{N_i} Z_j, Z_i)$, where W_i is the number of treated neighbors, so $W_i \in \{0, 1, ..., |N_i|\}$
  (b) Suppose we take $k = (5, 0)$ and $k' = (0, 0)$. Then take caution that $Y_i(k)$ is only defined for units with $|N_i| \geq 5$
  (c) To be more rigorous: $I_{kk'} = \{i : P(H_i = k) > 0 \text{ and } P(H_i = k') > 0)\}$
  (d) $\tau_{kk'} = |I_{kk'}|^{-1} \sum_{i \in I_{kk'}} (Y_i(k) - Y_i(k'))$

3. Insight: multi-arm trials on the exposure scale

- Consider a design $Z \sim \eta_Z$, this induces a distribution $H \sim \eta_H$.
- Under properly specified exposure, $Y_i(1), ..., Y_i(k)$ gives us the estimand $\tau_{kk'}$
- Conceptually, with multi-arm trial on the exposure scale, we can forget about Z and work with H only
- However, we randomize Z, NOT H (we still have some control over H since we determine the exposure mapping)

2.5.3 Neymanian Inference

We can use the Horvitz-Thompson estimator on the exposure contrasts between two exposures k and k':

$\hat{\tau}_{kk'} = |I_{kk'}|^{-1} \sum_{i \in I_{kk'}} \frac{\mathbb{1}\{H_i = k\}}{P(H_i = k)} Y_i - |I_{kk'}|^{-1} \sum_{i \in I_{kk'}} \frac{\mathbb{1}\{H_i = k'\}}{P(H_i = k')} Y_i$

As we have stated before, the HT estimator is unbiased, so $E[\hat{\tau}_{kk'}] = \tau_{kk'}$. We can then have $Var(\hat{\tau}_{kk'})$ and $\hat{Var}(\hat{\tau}_{kk'})$ available to us to perform confidence intervals. However, CLTs are a bit more complicated in this scenario.

Although the HT estimator is unbiased, it can have a large variance, and therefore the Hajek estimator is commonly used. In this estimator,

- $w_{ik} = \frac{1}{P(H_i = k)}$
- $w_k^+ = \sum_{i \in I_{kk'}} \mathbb{1}\{H_i = k\} w_k$
- $\hat{\tau}^H_{kk'} = \sum_{i \in I_{kk'}} \mathbb{1}\{H_i = k\} \frac{w_{ik}}{w_k^+} Y_i - \sum_{i \in I_{kk'}} \mathbb{1}\{H_i = k'\} \frac{w_{ik'}}{w_{k'}^+} Y_i$

Note that the Hajek estimator is no longer unbiased (has a small bias) but has dramatically lower variance than the HT estimator.

2.5.4 Fisherian Approach

If we consider the null of having no effect whatsoever, we have $H_0 : Y_i(k_1) = Y_i(k_2) = ... = Y_i(k_{|H|})$ for all units i. This is a sharp null! Thus, we can use the basic FRT.

For testing, we can test difference between two exposures by having $H_0 : Y_i(k) = Y_i(k')$ for all units i. However, note that this is no longer a sharp null (since we are not explicitly stating values for the other exposure mapping in the Science table). Thus the analysis can be complicated and require a tricky conditional FRT.

2.5.5 The Design Problem

We have that the design influences the estimator and that the estimator influences the design. For example, our interference structure and choice of estimand influence each other. Thus, we have some heuristics.

Assume households of size 2 for all units and assume partial interference. Therefore, $Y_i(Z) = Y_i(Z_i, Z_{[i]})$, where $[i]$ denotes the other unit in the household of unit i. Then our estimand for the primary effect is $\tau = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1,1) - Y_i(0,0))$. Then, our estimator can be the HT estimator, so $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}(H_i=(1,1))}{P(H_i=(1,1))}Y_i - \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}(H_i=(0,0))}{P(H_i=(0,0))}Y_i$.

Then which design should we use? We could use Bernoulli($\frac{1}{2}$), but then the problem is we have to throw away the data for all households with one treatment unit and one control unit (spillovers). Thus, a better way to do things is to use a Clustered Randomized trial, where half of the households are assigned to treatment. Then each household will either be all treatment or all control, and we can use all the data points in computing the estimate.

3 Observational Studies

In a randomized design, we are allowed to select which units go into treatment and which units go into control. In an observational study, we do not have this ability. Therefore, in order to perform inference, we need to make assumptions.

Take a Hormone Replacement Therapy (HRT) example where the goal is find the efficacy of the HRT on the rate of heart disease. If we just compare the results as we see them, we don't consider the underlying factors that determine whether the unit is in control or treatment. For example, say those who took the HRT drug had lower rate of Coronary Heart Disease (CHD). We may be inclined to state that HRT decreases the chance of CHD. However, say that people who were more likely to take the HRT drug also were more likely to live in an urban area (the same people who heard about the drug are also more likely to get better medical care in general). Then it is possible that the HRT drug does not help CHD at all, but rather its the access to medical care in general that helps.

Note why that this factor is specifically a problem. It is a problem because the urban/rural divide affects BOTH the CHD rate AND the propensity to seek HRT. If it only affected the CHD rate (outcome variable), then we would have equal urban/rural ratio in both treatment and control and our comparison would be fine. If it only affected the propensity of seek HRT, we may have many urbans in one group vs many rurals in another, but it would not affect the results of CHD, so again our comparison would be fine.

Now suppose we actually observe the urban/rural status. Would we be able to fix the problem then? What we could do is to actually match the units, so that each treatment and urban is matched 1-1 with a control and urban, and each treatment and rural is matched 1-1 with the control and rural. The remaining data that is not matched is not used. Then now are are able to compare apples to apples.

3.0.1 Strong Ignorability

The matching in order to create an apples-to-apples comparison is sufficient if we observe all the covariates that are related to BOTH the potential outcomes AND the probability of receiving treatment. More formally, our assumptions are:

Random Experiment	Strong Ignorability
Probabilistic: $0 < P(Z_i = 1 \mid X, Y(0), Y(1)) < 1$	Probabilistic
Known	X
Individualistic: $P(Z_i \mid X, Y(0), Y(1))$	Individualistic
Unconfounded: $P(Z_i = 1 \mid X_i)$	Unconfounded

The individualistic and unconfounded assumptions together are known as "selection on observables." It is important to distinguish that in a randomized experiment, we could enforce all 4 of these assumptions. In an observational study, we assume 3 of these assumptions, but it is not under our control. Note that ignorability is the single most widely used and abused assumption in all of causal inference.

3.1 Exact Pairwise Matching

In an exact pairwise match, such as the example earlier with HRT and matching urban with rural units, we match 1-1 and the unmatched units are discarded. More formally, if we let $I = \{1, ..., N\}$, then $M \subseteq I \times I$ such that:

- For all $(i, j) \in M$, $Z_i^{obs} + Z_j^{obs} = 1$ (exactly one treated unit per pair
- For all $(i, j), (i', j')$, we have that $i \neq i', j'$ and $j \neq i', j'$ (pairs strictly distinct)
- For all $(i, j) \in M$, $X_i = X_j$ (match is exact)

Below is the notation used for the exact pairwise matching. Assume we have 12 units and we are able to exactly match 4 pairs of them (8 total). Then give our sample, our notation is:

- $N = 12$
- $M = \{(5, 7), (3, 9), (4, 10), (11, 12)\}$ (matched pairs)
- $I_M = \{5, 7, 3, 9, 4, 10, 11, 12\}$ (all the units that have been matched)
- pairs $= k \in \{1, 2, 3, 4\}$, we often write $k \in M$
- Relabel units as (Z_{k1}, Z_{k2})

- Outcomes are (Y_{k1}, Y_{k2})

The intuition for matching is to formalize an apples-to-apples comparison. Under ignorability, if we take $k \in M$, we can denote $P(Z_{k1} = 1|X, Y(1), Y(0)) = \pi(X_{k1}) = \pi_{k1}$. Similarly, we have π_{k2}. However, we don't know the function $\pi(\cdot)$. But since $k \in M$, we have $X_{k1} = X_{k2}$, so therefore $\pi(X_{k1}) = \pi(X_{k2})$, or in other words $\pi_{k1} = \pi_{k2}$.

Therefore, if we let M be an exact pairwise match, then under ignorability, $\pi_{k1} = \pi_{k2}$ for each pair. Under the same setup, therefore, $P(Z_{k1} = 1|Z_{k1} + Z_{k2} = 1) = \frac{1}{2}$ for all $k \in M$. So although we don't know $\pi(\cdot)$, once we are matched, the π doesn't matter. Within each pair, it's as if the treatment has been randomized! We have turned our problem into the equivalent of a pairwise randomized design!

3.2 Analysis

By doing exact pairwise matching, we have turned our problem into a paired randomized design and we can treat it as if it is a randomized design. Let $p_k = P_\eta(Z_{k1} = 1|Z_{k1} + Z_{k2} = 1; M)$, where η is the unknown assignment mechanism. If η is ignorable, then $p_k = \frac{1}{2}$ for all $k \in M$.

If we focus on I_M, and let $\vec{Z}_M$ be the assignment vector for those units, then we have that $P_\eta(\vec{Z}_M|Z_{k1} + Z_{k2} = 1, \forall k)$. Then under ignorability, $\vec{Z}_M \sim \eta_M = PRD(\frac{\vec{1}}{2})$. Therefore to analyze, we do the following steps:

1. Obtain an exact pair match M
2. Pretend that $Z_M \sim PRD(\frac{\vec{1}}{2})$
3. Analyze as you would a PRD (paired randomized design)

Under Neymanian analysis, our estimand for the matched population (only) is $\tau = \frac{1}{|I_M|}\sum_{i\in I_M}[Y_i(1) - Y_i(0)]$. Therefore, our unbiased estimator for this is $\hat{\tau}^{DIM} = \frac{2}{N}\sum_{i\in I_M} Z_i Y_i - \frac{2}{N}\sum_{i\in I_M}(1 - Z_i)Y_i$.

Under Fisherian analysis, we simply perform FRT(η_M), an FRT as a PRD($\frac{\vec{1}}{2}$). This allows for sensitivity analysis.

In the words of Cochran, "The planner of an observational study should always ask himself the question: How would the study be conducted if it were possible to do it by controlled experimentation." In the words of Rubin, "Make the study as close as possible to a randomized experiment."

3.3 Sensitivity Analysis

We should always perform sensitivity analysis along with our results since we never know if we have captured all the influential covariates in our matching. Therefore, we may have deviation from our assumption of ignorability. We first define the $odds_{k1} = \frac{\pi_{k1}}{1-\pi_{k1}}$ and $odds_{k2} = \frac{\pi_{k2}}{1-\pi_{k2}}$. Then we define our odds ratio, $\nu_k = \frac{odds_{k1}}{odds_{k2}}$. In general, we have that $p_k = P(Z_{k1} = 1|Z_{k1} + Z_{k2} = 1) = \frac{\nu_k}{\nu_k+1}$. Under ignorability, we know that $\pi_{k1} = \pi_{k2}$ and thus $\nu_k = 1$, and $p_k = \frac{1}{2}$.

Sensitivity analysis asks about deviations from ignorability. In the words of Rosenbaum, "violations of ignorability are not like falling off a cliff, more like going down a gradual slope." Therefore, with sensitivity analysis, we are asking what happens if $\pi_{k1} \neq \pi_{k2}$ slightly. It is convenient to parameterize deviations in an odds ratio scale:

$\frac{1}{\Gamma} \leq \nu_k \leq \Gamma$; for all k, for $\Gamma \geq 1$

Note that if $\Gamma = 1$, then $\nu_k = 1$, which is ignorability.

Let's say that we want to perform a sensitivity analysis on the p-values. We've seen from before that $p_k = \frac{\nu_k}{\nu_k+1}$. By algebra, we have that $\frac{1}{\Gamma} \leq \nu_k \leq \Gamma$ is equivalent to $\frac{1}{1+\Gamma} \leq \frac{\nu_k}{\nu_k+1} \leq \frac{\Gamma}{1+\Gamma}$.
Recall that by performing exact pairwise matching, we essentially have k different PRDs. Therefore, let $\vec{\nu} = (\nu_1, ..., \nu_k)$, one for each of the PRDs. Then our p-values are $\vec{p}(\vec{\nu}) = (p_1(\nu_1), ..., p_k(\nu_k))$. Therefore our sensitivity analysis is $\vec{\nu} \in [\frac{1}{\Gamma}, \Gamma]^k$ which implies that $\vec{p}(\vec{\nu}) \in [\frac{1}{1+\Gamma}, \frac{\Gamma}{1+\Gamma}]^k$.

We fix any value in that interval, $\vec{\nu}_0 \in [\frac{1}{\Gamma}, \Gamma]^k$. This implies a design $\eta^*_{\vec{\nu}_0} = PRD(\vec{p}(\vec{\nu}))$. Then we can run a FRT($\eta^*_{\vec{\nu}_0}$) and obtain a pvalue, call it $pval_{\vec{\nu}_0}$. We can then compute the $pval_{\vec{\nu}}$ for any other $\vec{\nu} \in [\frac{1}{\Gamma}, \Gamma]^k$. We then compute the maximum of these pvalues, denoted as

$M(\Gamma) = max_{\vec{\nu} \in [\frac{1}{\Gamma}, \Gamma]^k}(pval_{\vec{\nu}})$

One possibility from running this analysis is that the pvalue under ignorability, call it $pval^I$ is less than our threshold, say $\alpha = 0.05$, but $M(\Gamma) > 0.05$. In this case, we would say that our conclusion to reject the null hypothesis based on ignorability is sensitive to violations of ignorability of up to Γ.

We can see that $M(\Gamma) \to 1$ as $\Gamma \to \infty$, so all observational studies are sensitive to sufficiently large deviations. Therefore, we want to find the largest Γ at which we go from rejecting the test to not rejecting. We denote this as $\Gamma_\alpha = inf(\Gamma \geq 1 : M(\Gamma) > \alpha)$. How we interpret this is as follows. If $\Gamma_\alpha = 2$, then there would need to be a hidden covariate making one of the units twice as likely to be treated as the other in the same pair, in order to change my conclusion. As the summary, the sensitivity analysis steps are:

1. Perform the exact match
2. Focus on I_M and pretend that $\vec{Z}_M \sim PRD(\frac{\vec{1}}{2})$
3. Compute $pval^I$ from FRT(η_M). This is the pvalue assuming ignorability, so when $\vec{\nu} = \vec{1}$
4. Fix Γ, and for every $\nu \in [\frac{1}{\Gamma}, \Gamma]^k$:
   - Consider $\eta_{\vec{\nu}} \sim PRD(\vec{p}(\vec{\nu}))$
   - Compute $pval_{\vec{\nu}}$ from the FRT(η_ν)
5. $M(\Gamma) = max_{\vec{\nu}}(pval_{\vec{\nu}})$
6. Calculate $\Gamma_\alpha = inf(\Gamma \geq 1 : M(\Gamma) > \alpha)$

3.3.1 Latent Confounder Representation

We can view sensitivity analysis in a different representation, which can give us some more insight and intuition. We can represent the sensitivity analysis as a logistic regression. Look at the two representations below:

- $\frac{1}{\Gamma} \leq \frac{\frac{\pi_{k1}}{1-\pi_{k1}}}{\frac{\pi_{k2}}{1-\pi_{k2}}} \leq \Gamma$ (our sensitivity analysis formula from before)
- $log(\frac{\pi_{k1}}{1-\pi_{k1}}) = K(X_{k1}) + \gamma U_{k1}$ (logistic regression, where K is some function of the covariates, and U_{k1} is an unmeasured confounder between 0 and 1

It can be shown that with $\Gamma = e^\gamma \geq 1$, that both of these statements are equivalent.

3.4 Alternatives to Exact Pairwise Matching

Although in an ideal case, exact pairwise matching can remove the confounders, in practice, there are many covariates and likely the covariates are a mix between continuous and discrete (not just discrete). Therefore exact pairwise matching is not feasible. Instead, we do our best and perform approximate covariate matching. We define the optimal matching M^{opt} as

$M^{opt} = argmin_M \sum_{i \in I_1} \sum_{j \in M(i)} d(\vec{X}_i, \vec{X}_j)$

What this formula is saying is that we want to find the optimal matching in the space of all matches. These matches are summing up the distance metric d for all the matches (unit i is in the treatment group I_1 and $j \in M(i)$, where $j \in M(i)$ is the set of the control units matched to unit i). Once we find the M^{opt}, then we perform the usual analysis and sensitivity analysis. Note that the sensitivity analysis accounts for both how the unmeasured covariate affects the outcome, as well as how the unmeasured covariate affects the assignment probability, which in turn can affect the outcome.

Let us define the propensity score as $\pi(X_i) = P(Z_i = 1|X_i)$, or in other words the probability of being in the treatment group given the covariates X_i. Let us also define the balancing score $b(X_i)$ as follows. If Z_i is independent of X_i once conditioned on $b(X_i)$, then $b(X_i)$ is a balancing score. This is saying that once we've conditioned on $b(X_i)$, that X_i contains no more information about the probability of assignment to treatment. We can see that the propensity score is one example of a balancing score. Therefore, conditioned on $\pi(X_i)$, we have that Z_i is independent of X_i. Thus, we only need to match on $\pi(X_i)$ to remove the bias. The steps for matching on estimated propensity score are:

- Estimate $\hat{\pi}$ with logistic regression. Have $logit(\pi_i) = \beta^T X_i + \epsilon_i$
- Define $d(X_i, X_j) = (\hat{\pi}(X_i) - \hat{\pi}(X_j))^2$

Essentially, this method is solving the problem by reducing the dimensionality of the large dimensional covariate space to the lower dimensional space of the propensity score, and then matching on this.

More recently, the more common way to solve this problem is to directly target the problem of high dimensional covariate space. In the single dimensional case, we have $d(X_i, X_j) = (X_i - X_j)^2$, let's denote it d_{ij}. In the multivariate case, we want to make sure to normalize each variable appropriately so one variable in large units does not take over the calculation. For example, with two covariates, we have $d_{ij} = [\frac{X_i^{(1)} - X_j^{(1)}}{SD(X^{(1)})}]^2 + [\frac{X_i^{(2)} - X_j^{(2)}}{SD(X^{(2)})}]$. This way, each covariate is in the same units of measurement. However, we can still have many correlated covariates, which can dominate this calculation. We therefore turn to the Mahalanobis distance, which takes into account the correlation. The Mahalanobis distance is defined as

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T \hat{\Sigma}^{-1} (X_i - X_j)}$$

A couple weaknesses that still remain with the Mahalanobis distance include the fact that outliers can skew the calculation as well as binary covariates. There are more ways to make the Mahalanobis distance more robust, including using ranks instead of the raw values.

3.4.1 Optimal Pair Matching

In order to find M^{opt}, assuming we have a defined distance metric d, we can create a matrix $D = (d_{ij})_{i \in I_1, j \in I_0}$. Essentially, our matrix D will have N_1 rows and N_0 columns and each entry of the matrix will be the distance between the ith member of the treatment group and the jth unit of the control group. For example, the entry d_{21} is the distance between the second treated unit and the first control unit. Once this is defined, we can use optimization techniques (software) in order to find the minimum of the sum of these distances, and therefore find M^{opt}. Notice that in optimal pair matching, the matching is done 1:1 (one treatment with one control).

3.4.2 Non-pairwise Matching

Instead of matching 1:1, we can match 1:n, and also make n fixed or variable. The first common way of non-pairwise matching is matching each treated unit to multiple (non-overlapping) control units. In the fixed ratio case, we match 1:n, so each treated unit is matched with n control units. In the variable ratio case, each treated unit is matched to one OR more control.

The pros of using the fixed ratio are that it is easier to interpret and slightly more efficient. The cons are that it is inflexible and can be biased (one treated unit may only have one good control match, but if we fix n to be 3, then we will be forced to have 2 additional bad matches). The pros of using the variable ratio are that it is better balanced (less bias) and more flexible. The cons are that it is harder to interpret. A rule of thumb is to use a fixed ratio with 1:2. This can get us a good trade-off between bias, efficiency, and interpretability.

Note that we can also match one control to many treated units. In the most general case, called optimal full match, it allows both variable ratio matching of 1:n and n:1.

3.4.3 Constraints

We can put constraints on our matrix D in order to obtain the desired match. For example, if we definitely do not want two units to be matched, we can make their distance equal ∞ in our matrix D, and the optimization algorithm will never match those two units.

If instead we just want to impose a larger penalty on certain matches, we can define $\tilde{d_{ij}} = d_{ij} + \gamma * |X_i - X_j|$, where γ is a parameter determined by us.

We can also use a propensity score caliper, which ensures that matched units have similar propensity scores. In this case we solve M^{opt} with the constraint that $|\hat{\pi}(X_i) - \hat{\pi}(X_j)| \leq c * SD(\hat{\pi}(X))$. This ensures that all matches are within a pre-determined amount of standard deviations from each other. We can use a soft caliper to impose a penalty, rather than a strict cutoff. In this case, we would define $\tilde{d_{ij}} = d_{ij} + \gamma * \frac{|\hat{\pi}(X_i) - \hat{\pi}(X_j)|}{SD(\hat{\pi}(X))}$,

3.5 Alternatives to Matching and FRT

We still assume unconfoundedness: $Z_i \perp (Y_i(1), Y_i(0))|_{X_i}$. We still assume probabilistic: $0 < \pi(X_i) < 1$.

3.5.1 Subclassification on the Propensity Score

Suppose we have $\pi(X_i) \in \{\pi_1, \pi_2, ..., \pi_K\}$. We have K different propensity scores and we can stratify based on these scores. We will have K strata, each with units with that propensity score. More formally, we define $I_k = \{i : S_i = k\}$, where S_i is the strata. Then $N(k) = |I_k|; N_1(k) = \sum_{i=1}^n \mathbb{1}\{S_i = k\}Z_i; N_0(k) = \sum_{i=1}^n \mathbb{1}\{S_i = k\}(1 - Z_i)$

We claim the proposition that for all $i \in I_k$ and $\pi(X_i) = \pi_k$, that $\vec{Z}_{I_k}|_{\sum_{i\in I_k} Z_i = N_1^{obs}(k)} \sim CRD(N_1^{obs}(k), N(k))$. In other words, the assignment for each strata is a CRD.

Secondly, we claim the proposition that $\vec{Z}|_{\sum_{i\in I_k} Z_i = N_1^{obs}(k)} \sim SCRD(\vec{N_1}^{obs}, \vec{N})$. In other words, the assignment across the whole experiment is SCRD.

Due to these two propositions, we can view this as an SCRD and use the same approaches as in a randomized experiment. For example, we can use the estimator $\hat{\tau}^{SCRD} = \sum_{k=1}^K \frac{N(k)}{N}\hat{\tau}_k$, where $\hat{\tau}_k$ is the DIM estimator within stratum k. From before, we know that $E_\eta[\hat{\tau}^{SCRD}] = \tau$, so the estimator is unbiased.

Recall from randomized experiments, we can use agnostic regression in the CRD case to perform covariate adjustments. Recall our formula $Y_i = \mu + Z_i\theta + \beta X_i = \gamma Z_i(X_i - \bar{X}) + \epsilon_i$. Then our $\tau^{\hat{REG}} = \hat{\theta}^{OLS}$, and our $Avar(\hat{\tau}^{REG}) \leq Avar(\hat{\tau}^{DIM})$. With our scenario here with K strata, we can perform this agnostic regression within each strata, and our final estimator $\hat{\tau}^{REG,SCRD} = \sum_{k=1}^K \frac{N(k)}{N}\hat{\tau}_k^{REG}$, where $\hat{\tau}_k^{REG}$ is the estimator from the regression for that stratum k. Under mildish assumptions (large strata, asymptotic convergence, etc.), $E[\hat{\tau}^{REG,SCRD}] = \tau$ and $Var[\hat{\tau}^{REG,SCRD}] \leq Var[\hat{\tau}^{SCRD}]$.

All the analysis that we have done so far has assumed that there are k different propensity scores. However, in practice, $\pi(X_i)$ is closer to continuous. In that case, assume all we have many different $\pi(X_i)$ with the lowest propensity score being π_0 and the highest being π_K. Then we can discretize the scores into bins. For example the first stratum will contain all the units with propensity scores from π_0 to some value π_1. The last stratum will contain all the units with propensity scores from some value π_{K-1} to π_K. However, in this case, even under ignorability, we no longer have a CRD within each stratum, and no longer have SCRD for the experiment as whole. Therefore $\hat{\tau}_k$ will be biased for τ_k even in large samples. $\hat{\tau}_k^{REG}$ will also be biased, but will generally be less biased that $\hat{\tau}$. As a consequence, $|bias(\hat{\tau}^{REG,SCRD})| \leq |bias(\hat{\tau}^{SCRD}, \tau)|$.

3.6 Inverse Propensity Weighted Estimators

We now switch to the superpopulation perspective. In this perspective, rather than think about The Science table as fixed, we instead think of $(Y_i(0), Y_i(1), Z_i, X_i) \overset{iid}{\sim} P$. We think of the values as an iid random draw from a superpopulation P. Under this framework, $\mu_1 = E[Y_i(1)]$, $\mu_0 = E[Y_i(0)]$ and $\tau_{SP} = \mu_1 - \mu_0$.

Under this framework, we have a Horvitz-Thompson-style estimator called the Inverse Propensity Weighted Estimator (IPW).

$\hat{\tau}^{IPW} = \frac{1}{N}\sum_{i=1}^N \frac{Z_i}{\pi(X_i)}Y_i - \frac{1}{N}\sum_{i=1}^N \frac{1-Z_i}{1-\pi(X_i)}Y_i$

The difference between the IPW estimator and the Horvitz-Thompson estimator is that we have replaced π_i with $\pi(X_i)$. Under ignorability, the IPW estimator is unbiased, since $E_P[\hat{\tau}^{IPW}] = \tau_{SP}$. Note that $Y_i(1)$ is a random variable now under the superpopulation framework. The proof is here:

$E[\frac{Z_iY_i}{\pi(X_i)}] = E[\frac{Z_iY_i(1)}{\pi(X_i)}]$
$= E[E[\frac{Z_iY_i(1)}{\pi(X_i)}|X_i]]$
$= E[\frac{1}{\pi(X_i)}E[Z_iY_i(1)|X_i]]$
$= E[\frac{1}{\pi(X_i)}E[Z_i|X_i]E[Y_i(1)|X_i]]$, since $Z_i \perp Y_i(1)|_{X_i}$ due to unconfoundedness
$= E[\frac{1}{\pi(X_i)}\pi(X_i)E[Y_i(1)|X_i]]$
$= E[E[Y_i(1)|X_i]]$
$= E[Y_i(1)]$
$= \mu_1$

Similarly we can show that $E[\frac{(1-Z_i)Y_i}{1-\pi(X_i)}] = \mu_0$. Then we have that $E[\hat{\tau}^{IPW}] = \frac{1}{N}\sum_{i=1}^N \mu_1 - \frac{1}{N}\sum_{i=1}\mu_0 = \mu_1 - \mu_0 = \tau_{SP}$. In practice, we estimate $\pi(X_i)$ with $\hat{\pi}(X_i)$, which is roughly fine if $\hat{\pi}(X_i) \xrightarrow{P} \pi(X_i)$.

An alternative to IPW is to use the Hajek estimator. Again, the Hajek estimator is not unbiased, but it is consistent and has a smaller variance than $\hat{\tau}^{IPW}$. Another alternative is to use a doubly-robust estimator (e.g. AIPW).

4 Causal Identification and Graphical Causal Models

The intuition in identification can be shown in a game. In this example, person A either writes some text in black ink, or writes the text in red ink, and then converts the red ink to black ink. Person B, only seeing the end result, tries to guess which methodology Person A used. It is clear that Person B cannot guess well no longer how many observation he sees, since what Person B observes is the same thing (text written in black). The issue here is that there are two different models, methodologies A and B, yet the observed data D is the same despite which methodology was used. Therefore, we can never discriminate between A and B on the basis of D.

A more formal example is to consider $W, S \sim P(W, S)$. In other words, W and S are drawn iid from a distribution P. Suppose $S \in \{0, 1\}$ and $S_i \sim Bern(\pi), 0 < \pi < 1$. We observe (W_i, S_i). We have that:

- $W_i(1)|_{S_i=1} \sim N(m_1, 1)$
- $W_i(1)|_{S_i=0} \sim N(\gamma_1, 1)$
- $W_i(0)|_{S_i=1} \sim N(\gamma_0, 1)$
- $W_i(0)|_{S_i=0} \sim N(m_0, 1)$
- $W_i = W_i(1)S_i + W_i(0)(1 - S_i)$

Say that $\pi, m_1, m_0, \gamma_1, \gamma_0$ are unknown parameters. Say we observed 1000 observations of (W_i, S_i) and we want to estimate these unknown parameters. To estimate π, we could use $\hat{\pi} = \frac{\sum_{i=1}^{1000} S_i}{1000}$. To estimate m_1, we could use $\hat{m_1} = \frac{\sum_{i=1}^{1000} S_i W_i}{\sum_{i=1}^{1000} S_i}$. However, how would we estimate γ_1? We cannot do that since we have no data that helps us in doing so. It doesn't matter if we have 1000 observations or 1M observations, we will cannot estimate it. It also doesn't matter how complex our model is for the estimator; it will not help. The problem here is that $W_i(1)|_{S_i=1} \sim N(m_1, 1); W_i(0)|_{S_i=0} \sim N(m_0, 1); S_i \sim Bern(\pi)$ completely characterize $P(W, S)$. γ_0 and γ_1 don't actually appear in the specification of $P(W, S)$. Therefore, we say that γ_0, γ_1 are not identifiable from the observed data.

Identification should not be confused with estimation or prediction. Identification answers the question: "If I had enough data, could I estimate this as precisely as I wish?"

4.1 Identification

The general setup for identification is as follows:

- P = latent distribution
- $Q = q(P)$ = some observed distribution
- $\tau = \tau(P)$ = some estimand

Identifiability asks: "Is there a function f such that $\tau(P) = f(q(P))$? Often, we want to find under what assumptions does there exist such a function f.

Looking back at our previous example, we have $(W(1), W(0), S) \sim P$. We observe $(W, S) \sim Q = q(P)$. Our estimands $\tau_{m_1} = \tau_{m_1}(P) = E[W(1)|S = 1]$ and $\tau_{\gamma_1} = \tau_{\gamma_1}(P) = E[W(1)|S = 0]$. We can rewrite the estimand $\tau_{m_1} = f_{m_1}(Q) = E[W|S = 1]$. Since both W, S are observed, we say that τ_{m_1} is identifiable. On the other hand, for τ_{γ_1}, we can have two different latent distributions P and P', which would lead to two different estimands τ_{γ_1} and $\tau_{\gamma_1'}$, but $q(P) = q(P') = Q$. Therefore τ_{γ_1} is not identifiable.

Identification is about setting the limits of what can be known using the data that can be observed. We have non-parametric identification when P is non-parametric. We have causal identification when $\tau(P)$ is a causal estimand.

Back to the causal example we had before. Assume we are in a superpopulation setting, so $(Y_i(1), Y_i(0), Z_i) \overset{iid}{\sim} P$. Under what conditions can we identify the ATE: $\tau(P) = E[Y(1) - Y(0)] = \mu_1 - \mu_0$ from the observed values $(Y_i, Z_i) \overset{iid}{\sim} Q$?

If we look at $\mu = E[Y(1)]$, we have that

$\mu_1 = E[Y(1)]$
$= E[E[Y(1)|Z]]$
$= E[Y(1)|Z = 1]P[Z = 1] + E[Y(1)|Z = 0]P[Z = 0]$

$= E[Y|Z=1]P[Z=1] + E[Y(1)|Z=0]P[Z=0]$
$= m_1 * \pi + \gamma_1 * (1-\pi)$

We know that $P[Z=1]$ and $P[Z=0]$ are identifiable since they are functions of Z_i, which is observed. We also know that $E[Y|Z=1]$ is identifiable (assuming $P[Z=1] > 0$). However, $E[Y(1)|Z=0]$ is not identifiable. Similarly, if we were trying to find μ_0, then we would find that $E[Y(0)|Z=1]$ is also not identifiable. Thus, we conclude that the ATE is not identifiable without assumptions.

4.1.1 Simple Identification Condition

We need to make assumptions in order to estimate the ATE. We can make some simple identification conditions. Note that these are not "simple" in the sense that the assumption is benign, but rather "simple" as in easy to state.

The assumption we make is that we supposed $E[Y(1)|Z=0] = E[Y(1)|Z=1] = E[Y|Z=1]$. Assuming $P(Z=1) > 0$, then we have that $\gamma_1 = m_1$. Thus, we have that $\mu_1 = m_1 * \pi + \gamma_1 * (1-\pi) = m_1 * \pi + m_1 * (1-\pi) = m_1 = f_{m_1}(Q)$. Similarly, if we assume that $E[Y(0)|Z=1] = E[Y(0)|Z=0] = E[Y|Z=0]$ assuming $P(Z=0) > 0$, then $\mu_0 = m_0$.

Therefore, our assumption of unconfoundedness, that $(Y(0), Y(1)) \perp Z$, along with our assumption of probabilistic assignment $0 < P(Z=1) < 1$, is sufficient to identify the ATE. In other words, we just proved that ignorability is sufficient to identity the ATE.

In order for this assumption to likely be true, randomization helps make this assumption credible.

4.1.2 Identification Region

Rather than looking at identification as binary, we may want to see the region as to which our estimand is identifiable. For example, recall our formula $\mu_1 = m_1 * \pi + \gamma_1 * (1-\pi)$. Rather than say that γ_1 is not identifiable and therefore μ_1 is not identifiable, we can examine what values that γ_1 could possibly take, and that information can give us a region as to where μ_1 could lie. More formally, the identification region is defined as $H\{\mu_1\} = \{m_1 * \pi + \gamma_1 * (1-\pi); \gamma_1 \in \Gamma\}$.

If we had no constraints at al, then $H\{\mu_1\} = \mathbb{R}$. However, suppose we know that $Y(0), Y(1) \in \{0,1\}$. Then we even observing the data, we know that $\mu_1 \in [0,1]$. Thus, $\gamma_1 \in [0,1]$. Therefore, we have that $H\{\mu_1\} = \{m_1 * \pi + \gamma_1 * (1-\pi); \gamma_1 \in [0,1]\} = [m_1\pi, m_1\pi + (1-\pi)]$.

For example, if we knew that $m_1 = 0.5$ and that $\pi = 0.9$, then the identification region $H\{\mu_1\} = [0.45, 0.55]$, which may be good enough. Note that this is not a Confidence Interval.

4.1.3 Identification with Covariates

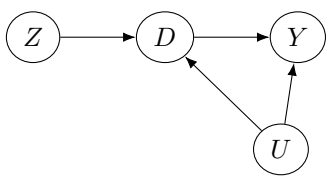
We obtain similar results as before when we have covariates, only that the results are now conditioned on the covariates. Now we have that $(Y_i(0), Y_i(1), Z_i, X_i) \sim P$. Now, we have that:

$\mu_1 = E[E[Y(1)|X]]$
$= \sum_{x \in X} E[Y(1)|X]P[X=x]$
$= E[Y(1)|X, Z=1]P[Z=1|X] + E[Y(1)|X, Z=0]P[Z=0|X]$

The sufficient condition for identifiability is that for any X such that $P(X) > 0$, $E[Y(1)|X, Z=0] = E[Y(1)|X, Z=1]$ as well as $E[Y(0)|X, Z=1] = E[Y(0)|X, Z=0]$; that $Y(0), Y(1) \perp Z|_X$; and that $0 < P(Z|X) < 1$. In other words, if we have ignorability conditional on the covariates X, that is a sufficient condition for identifiability.

4.2 Directed Acyclic Graphs (DAGs)

Directed Acyclic Graphs are used in causal inference as a visual tools than can encode causal relationships. For example, let's look at the DAG below:



The nodes in the DAG represent random variables. The arrows represent causal relationships. For example, we would say that in the DAG above, that Z causes D and D causes Y. If we were examining the relationship between D and Y, then U would be a confounder (and if U is unobserved, it is an unobserved confounder). From the DAG above, we have that:

- The parents of Y are $\{D, U\}$
- The children of U are $\{D, Y\}$
- The ancestors of Y are $\{D, U, Z\}$
- The ancestors of U is $\{\emptyset\}$
- The descendants of Z are $\{D, Y\}$
- The root nodes are $\{Z, U\}$

In a DAG, we assume the random variables represented by the root nodes are independent. In this case, $P(Z, U) = P(Z)P(U)$. The arrows that represent the causal relationships can be thought of functions that take input from the parent nodes and output the child nodes. For example, we would have that $Y = f_Y(D, U)$ and that $D = f_D(U, Z)$. Note that these marginal functions are not assumed to be known, but when put together, they induce a joint distribution. We do observe a joint distribution from the data. For example, $(P(Z), P(U), f_Y, f_D)$ are unknown, but they induce a joint distribution $P(Z, U, D, Y)$, which is observed.

4.2.1 Structural Causal Model

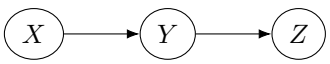
The definition of a structural causal model is as follows:

1. $N = R \cup V$, where N is the set of all random variables, R is the set of root random variables, and V is the set of non-root random variables
2. $P(\{X_r\}_{r \in R}) = \prod_{r \in R} P(X_r)$, where X_r are root random variables. What this is saying is that we have independence for root random variables.
3. $\{f_v\}_{v \in V} : X_v = f_V(Pa[X_v])$, where $Pa[X_v]$ are the parents of X_v. What this is saying is that there is a function that maps the parents of v to the value of v

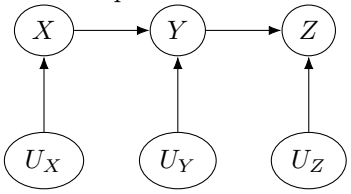
A Structural Causal Model (SCM) fully specifies a joint distribution. In other words $P(N)$ is fully specified. Consider an SCM with $X, Y \in N$. Then X is a direct cause of Y if $Y = f_Y(X, ...)$. Thus, if X appears in the argument of the structural equation defining Y, then it is a direct cause of Y. There is a clear relationship between SCMs and DAGs.

- A SCM implies a DAG!
  - N are the nodes in a DAG. R are the root nodes and V are the non-root nodes
  - For all $v \in V, u \in N$, we have that $u \to v$ iff u is a direct cause of v in the SCM. This specifies a unique DAG
- DAGs contain less info than the SCM. Multiple SCMs can have the same DAG structure. Imagine different SCMs with the same causality structure but different values of the function f. They would all have the same DAG
- DAGs inherit the notion of causality from the SCM. Let G be a DAG. Then X is a direct cause of Y if $X \in Pa(Y)$. Also, X is a cause (may not be direct) of Y if $X \in An(Y)$.

By convention, when we write



it is often implied that the DAG actually looks like



where U_X, U_Y, U_Z are independent. We add these in order to make things stochastic. For example, now we would have that $Y = f_Y(X, U_Y)$. This (generally) does not affect the key relations between the main variables of interest (X, Y, Z).

4.2.2 Applications of DAGs

Two types of causal questions people care about in the DAG-world are causal identification (the effect of interventions) and causal discovery.

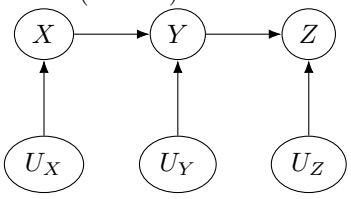
In causal identification, say we have N variables and we know the true DAG G and the joint distribution $P(X_1, ..., X_N)$ from the observed process. The question we want to answer is if the causal effect of X_i on X_j is identifiable from the data. We cannot get a causality from $P(\cdot)$ alone since it is an observational study. The DAG will tell us whether X_i is a cause of X_j, but does not tell us the effect.

In causal discovery, say we have N variables and we know the joint distribution $P(X_1, ..., X_N)$ from the observed process. We want to know how much of the true DAG we can recover. We also want to know if we posit G', can we test whether or not it is the true DAG?
The benefit of DAGs is that they allow us to address these questions using mostly graphical criteria!

4.2.3 Graphical Rules for Conditional Independence

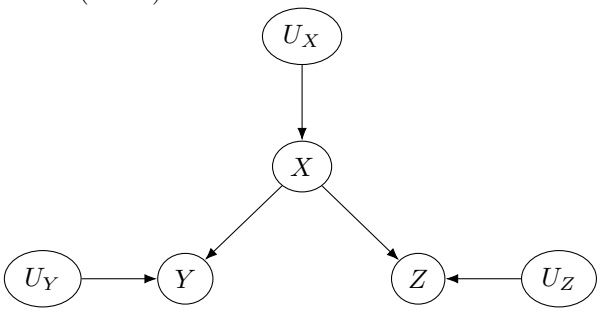
DAGs contain information about the dependency structure of the joint distribution $P(X_1, ..., X_N)$. For example, if we have $X \to Y$, then we know $X \not\perp\!\!\!\perp Y$. (The $\not\perp\!\!\!\perp$ symbol means not independent).

RULE 1: (Chains)



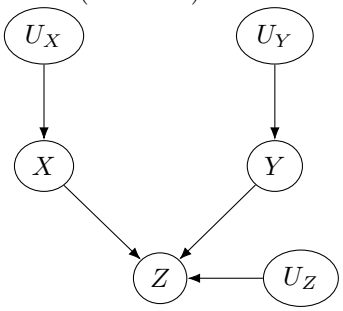
We have that U_X, U_Y, U_Z are all independent since they are root nodes. The rule is that $X \perp\!\!\!\perp Z|_Y$. This is because if we fix a value for Y, say $Y = a$, we know that $Y = f_Y(X, U_Y)$, so a change in X will be offset by a change in U_Y since $Y = a$ (fixed). We have that $Z = f_Z(Y, U_Z) = f(a, U_Z)$, so the value of Z only depends on U_Z, since a is fixed. Thus, X and Z are conditionally independent on Y.

RULE 2: (Forks)



The rule is that $Y \not\perp\!\!\!\perp Z$ since both are affected by X, but $Y \perp\!\!\!\perp Z|_X$. Why is there conditional independence? If we fix $X = a$, then $Z = f_Z(a, U_Z)$ and $Y = f_Y(a, U_Y)$. Since a is fixed, Z is only a function of U_Z and Y is only a function of U_Y. Since we know that root nodes $U_Z \perp\!\!\!\perp U_Y$, thus Z and Y are conditionally independent on X.

RULE 3: (Colliders)



The rule is that $X \perp\!\!\!\perp Y$, but $X \not\perp\!\!\!\perp Y|_Z$. This is because we have $Z = f_Z(X, Y, U_Z)$. Therefore, if we fix $Z = a$, then a change of X implies a change of Y and U_Z. An example of this is if we know that to get into college you need to either be a gifted musician or have a high GPA, or both, and we learn that a person was admitted and know that they are a poor musician,

then this tells us a lot about their GPA.

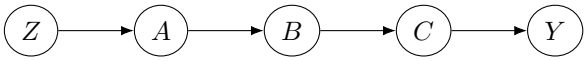
Given these three rules, we want to find out that if we are provided an arbitrary DAG G, and we take $X, Y \in N$ and $Z \subseteq N \backslash \{X, Y\}$, is there a graphical rule to decide whether or not $X \perp\!\!\!\perp Y|_Z$?

To answer this, we first need to define a blocked path. A path p is blocked by Z if either of the two conditions hold:

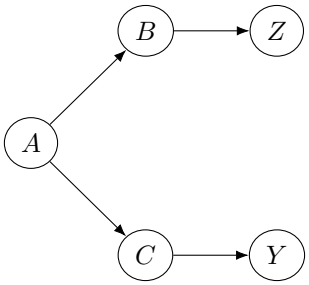
1. p contains $A \to B \to C$ or $A \leftarrow B \to C$ such that $B \in Z$
2. p contains a collider $A \to B \to C$ such that $B \notin Z$ AND $Desc(B) \cap Z = \{\emptyset\}$

We then define d-separation as follows. If Z blocks every path from X to Y, then X and Y are d-separated. If X, Y are d-separated by Z, this is equivalent to $X \perp\!\!\!\perp Y|_Z$.

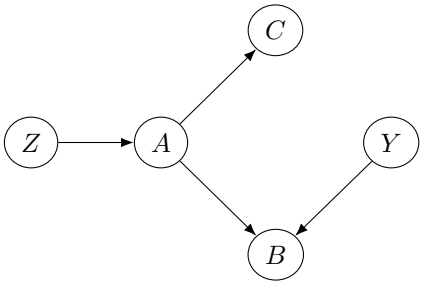
Below are some examples.



In this example, $Z \not\perp\!\!\!\perp Y$, since the chain induces dependence. However, $Z \perp\!\!\!\perp Y|_B$ using rule 1. Similarly, $A \perp\!\!\!\perp C|_B$.

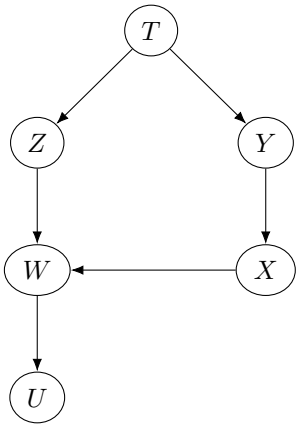


In this example, $Z \not\perp\!\!\!\perp Y$, since forks induce dependence. However, $B \perp\!\!\!\perp C|_A$, and thus $Z \perp\!\!\!\perp Y|_A$. Also, we have that $Z \perp\!\!\!\perp Y|_B$.



In this example, $Z \perp\!\!\!\perp Y$, by rule 3 (colliders). However, by that rule, $Z \not\perp\!\!\!\perp Y|_B$.

Now imagine we have that following DAG:

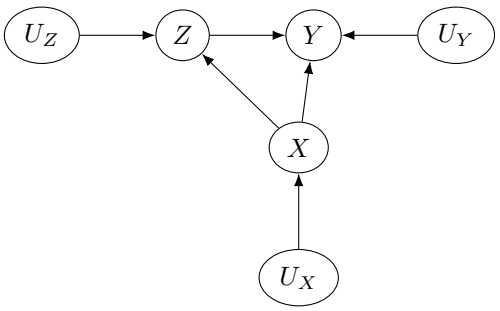


Let's say that we want to d-separate Z and Y. We see that there are two paths between Z and Y. The first path is $Z \leftarrow T \to Y$. The second path is $Z \to W \leftarrow X \leftarrow Y$. To block the first path, we can just condition on T. To block the

second path, we can either condition on: nothing, $\{W, X\}$, $\{U, X\}$, $\{U, W, X\}$, or $\{X\}$. Thus, all together, we can d-separate Z and Y is we condition on $\{T\}$, or $\{T, X\}$, or $\{T, W, X\}$, etc. Note that conditioning on $\{T, W\}$ is NOT a valid d-separated set since conditioning on $\{W\}$ alone does not block the second path.

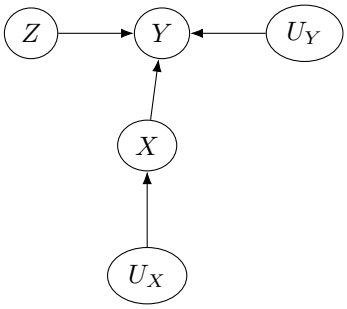
4.2.4 Causal Identification

Let's say we want to find the effect of Hormone Replacement Therapy (HRT) on Coronary Heart Disease (CHD) from an observational study. Assume we have one confounder, "fitness," where more fit people tend to also try HRT, and also have lower rates of CHD. Let HRT be denoted as $Z \in \{0, 1\}$. Let CHD be denoted as $Y \in \{0, 1\}$. Let X denote fitness. From the data in our observational study, we observe the joint distribution $P(X, Z, Y)$ and we also have the DAG:



The question we want to answer is can we get the causal effect of Z on Y from $P(X, Z, Y)$ and the DAG? First, we need to state the estimand. The estimand is NOT $P(Y|Z = 1) - P(Y|Z = 0)$ since this is not a causal effect because X is a confounder. Stated explicitly, $P(Y|Z = 1)$ is the probability of CHD for women who happen to take HRT. What we want is the probability of CHD if we GIVE HRT to women. We denote that latter as $P(Y|do(Z = 1))$. Note that we can view the difference between $P(Y|Z = 1)$ and $P(Y|do(Z = 1))$ as the difference between correlation and causation. Thus, the estimand that we want is $P(Y = 1|do(Z = 1)) - P(Y = 1|do(Z = 0))$. Note that $P(Y|do(Z = 1)) = P(Y(1))$ in the potential outcomes framework. (Also note that for ease of notation we are using P for both discrete and continuous, we can just replace summations with integrals etc.)

How do we get this do-operator? We use graph manipulation. Whereas previously we have the DAG G look as above, our new graph G_M cuts the edges from Z. Thus, our DAG G_M looks like:



Thus, we have changed our structural model from using G and P to using G_M and P_M.

SCM: (G, P)	SCM: (G_M, P_M)
$P(U_X, U_Y, U_Z) = P(U_X)P(U_Y)P(U_Z)$	$P(U_X, U_Y, U_Z) = P(U_X)P(U_Y)P(U_Z)$
$Z = f_Z(U_Z, X)$	$Z = z$
$X = f_X(U_X)$	$X = f_X(U_X)$
$Y = f_Y(U_Y, Z, X)$	$Y = f_Y(U_Y, z, X)$

Therefore, we can define $P(Y|do(Z = 1)) = P_M(Y|Z = 1)$. However, we don't observe P_M, only P, so how can be ge this quantity. We can see from the SCM above that we have a connection between P and P_M. We know that $P(X) = P_M(X)$ and we know that $P(Y|Z, X) = P_M(Y|Z, X)$. Thus, we have that

$P(Y|do(Z = 1)) = P_M(Y|Z = 1)$
$= \int P_M(Y|Z = 1, X)P_M(X|Z = 1)dx$ (by Law of Total Probability)
$= \int P(Y|Z = 1, X)P_M(X|Z = 1)dx$
$= \int P(Y|Z = 1, X)P_M(X)dx$ since Y is a collider so $X \perp\!\!\!\perp Z$

$= \int P(Y|Z=1,X)P(X)dx$

Now we can obtain this from the observed data, since this only depends on $P(\cdot)$, which we observe. Thus, we have that $P(Y|do(Z=1))$ is identifiable, and we now have a blueprint for computing this effect. We now have derived the "covariate adjustment formula":

$P(Y|do(Z=1)) - P(Y|do(Z=0)) = \int \{P(Y|Z=1,X) - P(Y|Z=0,X)\}P(X)dx$

One way to look at this formula is that we take the difference between the treated and non-treated units with a certain set of covariates, and then take the weighted average of this effect across all these different groups of covariates, essentially, an SCRD.

We can generalize this by using "The Backdoor Criterion." Let (Z,Y) be two nodes in a DAG, and let $X \subseteq N \backslash \{Z,Y\}$. We have that X satisfies the backdoor criterion relative to (Z,Y) if:

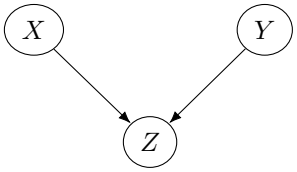
- For all $x \in X$, $x \notin Desc(Z)$
- X blocks every path between Z and Y that contains an arrow into Z

If X satisfies the backdoor criterion with respect to (Z,Y), then $P(Y|do(Z=z)) = \int P(Y|(Z=z),X)P(X)dx$. For example, in our previous example, X is not a descendant of Z and X also blocks all paths (there is only one path) between Z and Y that contains an arrow into Z and therefore X satisfies the backdoor criterion.

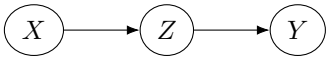
4.2.5 Causal Search

For causal search, we may be interested in model testing. We assume that we know $P(X_1,...,X_N)$ and that we can test conditional independence using the rules stated before. Our goal is to check whether a candidate DAG G_1 is compatible with the data and test that we observed.

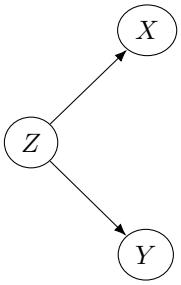
For example, say we ran an observation study and from the data we obtained $P(X,Y,Z)$ and found that $X \perp\!\!\!\perp Y|_Z$ from this joint distribution. Say we propose the following candidate DAG G_1.



We know that G_1 cannot be the correct DAG because if G_1 were true, then by rule 3 (colliders), then $X \not\perp\!\!\!\perp Y|_Z$. However, say we propose G_2 as the following:



and we also propose G_3 as the following:

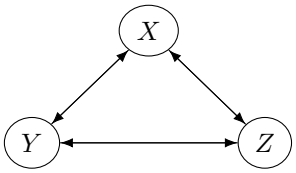


Then, we find that both G_2 and G_3 have the dependence footprint that $X \not\perp\!\!\!\perp Z$ and also $X \perp\!\!\!\perp Y|_Z$. Thus we find that G_2 and G_3 are indistinguishable from $P(\cdot)$ alone. We say that G_2 and G_3 are Markov equivalent. Since both are possible DAGs that satisfy our requirements, we cannot rule either one out as the true DAG.

Say we want to find all the DAGs that are compatible with the observed joint distribution $P(\cdot)$. There are multiple families of algorithms that accomplish this. One family of algorithms works by starting with a fully connected undirected graph, and then applying successive marginal/conditional independence tests and collections of rules to deduce the presence/absence of edges and the direction of the edges.

More specifically, we would eliminate edges among variables that are unconditionally independent. After that for all edges connecting two nodes A and B and a connected node C, we would eliminate the edge between A and B if $A \perp\!\!\!\perp B|_C$.

For example, say we started with a fully connected undirected graph below:



Say from our tests, we found that $X \not\perp\!\!\!\perp Z$, $X \not\perp\!\!\!\perp Y$, and $Z \not\perp\!\!\!\perp Y$. This does not help us rule out any candidate DAGs. However, say we also found that $X \not\perp\!\!\!\perp Y|_Z$, $X \not\perp\!\!\!\perp Z|_Y$ and $Z \perp\!\!\!\perp Y|_X$. Then from this information, we could rule out $Y \rightarrow X \leftarrow Z$. However, we cannot rule out $Y \leftarrow X \leftarrow Z$, $Y \rightarrow X \rightarrow Z$, or $Y \leftarrow X \rightarrow Z$. All 3 of these DAGs are Markov equivalent and satisfy all of our independence tests.

5 Common Observational Studies Scenarios

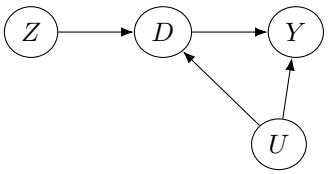
5.1 Instrumental Variables

In real life, often even though we may perform a randomized experiment, the units do not always comply with what they are assigned to do. Say we are performing a clinical trial on the effect of a pill on blood pressure. We perform a randomized experiment and have the usual framework where $Z_i \in \{0, 1\}$, SUTVA assumption so we have $Y_i(1)$ and $Y_i(0)$, and our estimand and estimator are τ^{ATE} and $\hat{\tau}^{DIM}$ respectively. However, say one of the units assigned to the treatment group doesn't like the taste of the pill and says he won't take it. Say another unit moves to NYC halfway through the trial, and the treatment can only be administered in Palo Alto, so she also stops the treatment. These are two examples of non-compliance (where units that are assigned to treatment do not do the treatment, or vice versa where units that are assigned to control do the treatment).

Therefore, we see a difference in between the assignment effect and the treatment effect. We denote $D_i \in \{0, 1\}$ as the indicator for whether unit i takes the treatment and denote $Z_i \in \{0, 1\}$ as the indicator for whether unit i is assigned to the treatment. With non-compliance, τ^{ATE} is the effect of being assigned to receive treatment, not the effect of taking the treatment.

So for non-compliance studies, what should be the estimand? If we denote $N_1^* = \sum_{i=1}^N D_i$ and $N_0^* = \sum_{i=1}^N (1 - D_i)$, then could we use the estimator $\hat{\tau}^{BAD} = \frac{1}{N_1^*} \sum_{i=1}^N D_i Y_i - \frac{1}{N_0^*} \sum_{i=1}^N (1 - D_i) Y_i$? As it turns out, this is not a good estimand to use. The reason is because although Z_i was randomized, D_i was not randomized. There may be underlying differences in who actually decided to take the treatment vs. who did not. The people who take treatment when assigned to treatment might be very different from people who don't take the treatment when assigned to treatment.

Therefore, we can estimate the assignment effect because it was randomized, but for the treatment effect, we need to go back to the techniques from observational studies. Our DAG is this situation looks as follows:



where Z is the assignment variable, D is the treatment variable, Y is the outcome variable, and U are unobserved confounder variables. We have that $P(Y|do(Z = z))$ is identifiable because by the backdoor criterion, we have no arrows going into Z, so if we take the empty set, we can see that it satisfies the backdoor criterion. On the other hand, $P(Y|do(Z = z))$ is non-identifiable because U is unobserved and therefore unblockable. Since it does not satisfy the backdoor criterion, this quantity is non-identifiable.

We could view the effect of D on Y as an observational study and ignore Z completely. We would have to assume ignorability, and then we could proceed with previously discussed methods such as matching, etc. However, ignorability is a strong assumption, and we also know that Z was randomized, so we are in a better situation than if this was purely an observational study. Can we use the randomization of Z to help us avoid using the assumption of ignorability?

5.1.1 Instruments

The answer to the question is that we can use the randomization of Z to help us avoid using the assumption of ignorability if Z satisfies:

- Z is random
- Z is an actual push (Z affects D). For example, assigning someone to treatment actually makes them more likely to take the treatment
- Z must affect Y only through D (this one is tough to prove)

An example is if in an encouragement study, we may provide units in the treatment groups vouchers to encourage the treatment behavior of attending a course. The assignment is random, and it makes logical sense that giving people vouchers would incentivize them to actually take the class. However, it is harder to prove that the assignment only affects the outcome through the voucher.

5.1.2 Non-parametric Instrumental Variables

In the framework, we have that $Z_i \in \{0,1\}$ is the assignment/encouragement/instrument, whereas $D_i \in \{0,1\}$ is the treatment of interest. Let $D_i(\vec{Z})$ and the outcome be $Y_i(\vec{Z}, \vec{D})$.

Assuming SUTVA, we have that $D_i(\vec{Z}) = D_i(Z_i)$ and $Y_i(\vec{Z}, \vec{D}) = Y_i(Z_i, D_i)$. Therefore, under SUTVA

- Potential Treatments: $\{D_i(0), D_i(1)\}$
- Potential Outcomes: $\{Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1)\}$
- Observed Outcomes: $Y_i(Z_i, D_i(Z_i))$

We can look at the Intent to Treat Effect (ITT). The estimand for the outcome is $\tau_Y^{ITT} = \frac{1}{N}\sum_{i=1}^N (Y_i(1, D_i(1)) - Y_i(0, D_i(0)))$ and the estimand for the treatment is $\tau_D^{ITT} = \frac{1}{N}\sum_{i=1}^N (D_i(1) - D_i(0))$. If we let $Z_i \sim \text{CRD}(N_1, N)$, then we have these unbiased estimators: $\hat{\tau}_Y^{ITT} = \frac{1}{N_1}\sum_{i=1}^N Z_i Y_i - \frac{1}{N_0}\sum_{i=1}^N (1-Z_i) Y_i$ and $\hat{\tau}_D^{ITT} = \frac{1}{N_1}\sum_{i=1}^N Z_i D_i - \frac{1}{N_0}\sum_{i=1}^N (1-Z_i) D_i$. We care about ITT because it is more robust inference even if not of primary interest and it also tells us the relevance of assigning on treating.

We can think of the difference between Z_i and D_i in the following way. If we denote $G_i = (D_i(0), D_i(1)) \in \{(0,0), (0,1), (1,0), (1,1)\}$, we can think of G_i as a partially observed latent covariate. We can define the different groups of G_i in the table below:

	$D_i(0)$		
$D_i(1)$		0	1
	0	never takers	defiers
	1	compliers	always takers

When we randomize Z_i, we may think that we have balancing G_i across the two different groups. However, we can see that for units i where $D_i = 1$, this includes the compliers (C), defiers(D), and always takers (AT). For units i where $D_i = 0$, this includes the compliers (C), defiers (D), and the never takers (NT). If AT and NT are very different units, we will then be comparing different types of units in the two groups and creating bias. This is why the $\hat{\tau}^{BAD} = \frac{1}{N_1^*}\sum_{i=1}^N D_i Y_i - \frac{1}{N_0^*}\sum_{i=1}^N (1-D_i) Y_i$ is not a good estimator.

We can also look at stratum-specific ITT. For example, for the compliers (C), we have $\tau_Y^{ITT,C} = \frac{1}{N_C}\sum_{i=1}^N (\mathbb{1}\{G_i = C\}(Y_i(1, D_i(1)) - Y_i(0, D_i(0))))$ and we have $\tau_D^{ITT,C} = \frac{1}{N_C}\sum_{i=1}^N (\mathbb{1}\{G_i = C\}(D_i(1) - D_i(0)))$. We can obtain similar formulas for the ITT estimands for the AT, NT, and D groups.

Therefore, we can combine these stratum-specific ITT to calculate the overall ITT as follows:

$\tau_Y^{ITT} = \frac{1}{N}\sum_{i=1}^N \tau_i$
$= \frac{1}{N}[\sum_{i=1}^N \mathbb{1}\{G_i = C\}\tau_i + \sum_{i=1}^N \mathbb{1}\{G_i = D\}\tau_i + \sum_{i=1}^N \mathbb{1}\{G_i = AT\}\tau_i + \sum_{i=1}^N \mathbb{1}\{G_i = NT\}\tau_i]$
$= p_C \tau_Y^{ITT,C} + p_D \tau_Y^{ITT,D} + p_{AT}\tau_Y^{ITT,AT} + p_{NT}\tau_Y^{ITT,NT}$, where p denotes the proportion

We can do the same for τ_D^{ITT}. Notice that for $G_i \in \{AT, NT\}$, that $D_i(0) = D_i(1)$ for these units. This is because always takers and never takers always react the same no matter which group they are assigned to. Therefore, we have that $\tau_D^{ITT,NT} = \tau_D^{ITT,AT} = 0$. Thus, we have that $\tau_D^{ITT} = p_C \tau_D^{ITT,C} + p_D \tau_D^{ITT,D}$.

5.1.3 Average Causal Effect of Treatment on Compliers

Given that we don't want to use $\hat{\tau}^{BAD}$ as our estimator, we want to instead look at the effect of the treatment on compliers. Below, we will explain why, but in order to do so, we must make some assumptions.

We make an assumption in working with instrumental variables, which is the Exclusion-Restriction assumption. This assumption is that $Y_i(Z_i, D_i) = Y_i(D_i)$. This means that the only effect that the assignment has on the outcome is through the treatment. The assignment directly affects the treatment, and the treatment directly affects the outcome, so the assignment does not directly affect the outcome.

The next assumption we make is that $\tau_D^{ITT} > 0$. In other words, the assignment is a push. Receiving the encouragement has to make a unit more likely to receive treatment.

The final assumption we make is called monotonicity. It says that $D_i(1) \geq D_i(0)$. What this says is that encouragement cannot discourage a unit from taking the treatment. It can only help or do nothing. The significance of this is that is assumes no defiers.

We are able to prove that if all of these assumptions hold, that $\tau^C = \frac{\tau_Y^{ITT}}{\tau_D^{ITT}}$. This is amazing because we can estimate both the numerator and the denominator of the right hand side, which will give us an estimate of the effect of the treatment on the compliers! The proof of this is below.

Proof:

- Step 0: Using the assumption that $\tau_D^{ITT} > 0$, we know that $\frac{\tau_Y^{ITT}}{\tau_D^{ITT}}$ is well-defined (denominator non-zero)
- Step 1: Using the assumption of monotonicity, there are no defiers, so $N_D = 0$
- Step 2: Let g be a group (either C, NT, AT, D). Then we know that $\tau_Y^{ITT,g} = \frac{1}{N_g}\sum_{i=1}^{N} \mathbb{1}\{G_i = g\}(Y_i(1, D_i(1)), Y_i(1, D_i(0)))$. Using the Exclusion-Restriction assumption, this equals $\frac{1}{N_g}\sum_{i=1}^{N} \mathbb{1}\{G_i = g\}(Y_i(D_i(1)), Y_i(D_i(0)))$. Note that for $g \in \{AT, NT\}$, that $D_i(1) = D_i(0)$, and therefore $Y_i(D_i(1)) = Y_i(D_i(0))$, which implies that $\tau_Y^{ITT,NT} = \tau_Y^{ITT,AT} = 0$.
- Step 3: Therefore, for compliers, $\tau_Y^{ITT,C} = \frac{1}{N_C}\sum_{i=1}^{N} \mathbb{1}\{G_i = C\}(Y_i(D_i(1)) - Y_i(D_i(0)))$. However, for compliers, $D_i(0) = 0$ and $D_i(1) = 1$, so this quantity equals $\frac{1}{N_C}\sum_{i=1}^{N} \mathbb{1}\{G_i = C\}(Y_i(1) - Y_i(0))$. Therefore, we have that $\tau_Y^{ITT,C} = \tau^C$.
- Step 4: The overall ITT for Y is $\tau_Y^{ITT} = p_C\tau_Y^{ITT,C} + p_D\tau_Y^{ITT,D} + p_{AT}\tau_Y^{ITT,AT} + p_{NT}\tau_Y^{ITT,NT}$. From Step 2, we know that $\tau_Y^{ITT,NT} = \tau_Y^{ITT,AT} = 0$, so those terms go to 0, and from Step 1, we know that $N_D = 0$, so $p_D = 0$ and that term goes away. Thus, we are left with $\tau_Y^{ITT} = p_C\tau^{ITT,C}$ and from Step 3, we have that this is equal to $p_C\tau^C$.
- Step 5: The overall ITT for D is $\tau_D^{ITT} = p_C\tau_D^{ITT,C} + p_D\tau_D^{ITT,D} + p_{AT}\tau_D^{ITT,AT} + p_{NT}\tau_D^{ITT,NT}$. Again, the last 3 times go to 0, and the result is that $\tau_D^{ITT} = p_C\tau_D^{ITT,C}$. However, we have that $\tau_D^{ITT,C} = \frac{1}{N_C}\sum_{i=1}^{N} \mathbb{1}\{G_i = C\}(D_i(1) - D_i(0))$. We know that for compliers, $D_i(1) = 1$ and $D_i(0) = 0$. Therefore, this term equals $\frac{1}{N_C}\sum_{i=1}^{N} \mathbb{1}\{G_i = C\} = \frac{1}{N_C}N_C = 1$. By plugging this in, this implies then that $\tau_D^{ITT} = p_C$.
- Step 6: Plugging this result into the result in Step 4, we get that $\tau^C = \frac{\tau_Y^{ITT}}{p_C} = \frac{\tau_Y^{ITT}}{\tau_D^{ITT}}$.

Using this formula, we have an estimand. We can estimate τ_Y^{ITT} using a DIM estimator $\hat{\tau}_Y^{ITT}$ and we can estimate τ_D^{ITT} using a DIM estimator $\hat{\tau}_D^{ITT}$. Thus, we have our final IV estimator, $\hat{\tau}^{IV} = \frac{\hat{\tau}_Y^{ITT}}{\hat{\tau}_D^{ITT}}$. This is an amazing result! Although we can't identify who the compliers are, we can estimate the effect of treatment on the population.

The benefit for using instrumental variables is that we don't need to make the assumption of ignorability, which is great. However, we do need to make other assumptions (often considered weaker than ignorability though). The other downside is that we've moved the goalposts. We originally wanted to estimate the treatment effect for the population, but ended up with an estimate of the treatment effect for compliers.

In performing inference for IVs, we can get the point estimate as stated above, and to get the variance estimate, we can do so using the Delta method.

5.1.4 Additional Topics in Instrumental Variables

One term used in IV literature is the concept of Weak Instruments. The Exclusion-Restriction assumption is one of the most difficult assumptions to prove holds for a particular study. If we suppose that the Exclusion-Restriction assumption does not hold, then we end up with bias. Take, for example, below, where we assume that it does not hold for NT. Let's say that for all units i, where $G_i = NT$, that $Y_i(0,0) - Y_i(1,0) \neq 0$, which implies that $\tau_Y^{ITT,NT} \neq 0$.

Then, we have that

$$\frac{\tau_Y^{ITT}}{\tau_D^{ITT}} = \frac{p_C\tau_C + p_{NT}\tau^{ITT,NT}}{p_C}$$
$$= \tau_C + \frac{p_{NT}}{p_C}\tau^{ITT,NT}$$

Therefore, we have a bias of $\frac{\tau_Y^{ITT}}{\tau_D^{ITT}} - \tau_C = \frac{p_{NT}}{p_C}\tau^{ITT,NT}$. Now, assuming that $\tau^{ITT,NT}$ is fixed, the bias will be very large then p_C is very small. Basically this would happen when Z has a small effect on D.

By definition, an instrument Z is said to be weak when Z has a small effect on D. Weak instruments are not a problem when the Exclusion-Restriction holds. However, weak instruments make IV analysis more sensitive to Exclusion-Restriction violations, and the bias can blow up very quickly. Note also that the formula for our estimate is $\hat{\tau}^{IV} = \frac{\hat{\tau}_Y^{ITT}}{\hat{\tau}_D^{ITT}}$, tha the variance may be very large if p_C is small.
Another topic in IV analysis is whether or not we have one-sided or two-sided non-compliance. In two-sided non-compliance (the previous analysis), we assume that some people who are assigned to treatment are not taking treatment, and some who are not assigned to treatment are taking it. In one-sided non-compliance, we say that people assigned to treatment may or may not take it, but people not assigned to treatment CANNOT take it. For example, in a drug trial, people who are not assigned to treatment would not be able to get a hold of the drug. In one-sided non-compliance, we only have Compliers (C) and Never Takers (NT). We do not have Defiers (D) or Always Takers (AT).

Another topic in IV analysis is the use of 2 stage least squares. In 2 stage least squares, we start with the model $D_i = \alpha + \beta Z_i + \gamma^T X_i + \epsilon_i$, where Z_i is the instrument and X_i are the covariates. From this first regression, we are able to get the predictions $\hat{D}_i$. We then have our second regression $Y_i = \mu + \eta \hat{D}_i + \nu^T X_i + \delta_i$. We then run this second regression and our estimate $\hat{\eta}$ is the estimate of τ_C under some conditions.

5.2 Regression Discontinuity Design (RDD)

We explain the intuition behind Regression Discontinuity Design (RDD) using an example. We want to find the effect of receiving a National Merit Fellowship on the probability of graduating from college. We denote $Z_i \in \{0, 1\}$ where 0 is not receiving the fellowship and 1 is receiving the fellowship. We denote $Y_i \in \{0, 1\}$ where 0 is not graduating and 1 is graduating. This study looks like a usual observational study since the NMF is not assigned at random, it is earned. What is different in this case is that we have a forcing variable X_i, which is the PSAT score. How the NSF works is that a PSAT score above a certain threshold, c, results in an NSF. Therefore, $P(Z_i|X_i) = 1$ if $X_i \geq c$ and 0 if $X_i < c$. Thus, X_i is a forcing variable because it's value is the sole determiner of the assignment Z_i.

Now if we want to perform the usual observational methods, we need to assume ignorability and also perform matching. However, ignorability is violated here because of the violation of probabilistic assignment. We do not have $0 < \pi(X_i) < 1$, it is either 0 or 1. Also, we cannot have a good match. We cannot match when $Z_i = 0$ and $Z_i = 1$, because they will never have the same X_i, since $Z_i = 0$ only when $X_i < c$ and $Z_i = 1$ only when $X_i \geq c$.

Now say we plot on a graph the x-axis being X_i, the score on the PSAT, and on the y-axis $P(Y = 1, X)$. For values on the x-axis where $X_i < c$, we may see a straight line trend, but right around c, we see a sudden increase. Then after that jump, we see another straight line trend when $X_i > c$. This is regression discontinuity and we can see that the treated units have higher outcomes from this jump. Intuitively, we are assuming there is nothing special about $X = c$ except that it triggers treatment, and we can see that there is some effect on treatment on graduation. How do we perform detailed analysis to quantify this effect?

5.2.1 Sharp Regression Discontinuity

Let us take the superpopulation framework, so we have that $(X_i, Z_i, Y_i(0), Y_i(1)) \sim P$. The probability of assignment, $\pi(X)$ is 1 if $X \geq c$ and 0 otherwise. Thus all the values below c are not treated, and we have $E[Y(0)|X]$ and all the values above c are treated, and we have $E[Y(1)|X]$. To perform analysis, we assume that we have continuity at the point c. In other words, both $\mu_0(x) = E[Y(0)|X = x]$ and $\mu_1(x) = E[Y(1)|X = x]$ are continuous at c. We can also write this as $lim_{x \to c^-} \mu_0(x) = E[Y(0)|X = x]$ and $lim_{x \to c+} \mu_1(x) = E[Y(1)|X = x]$, and we assume they both exist.

Our estimand is $\tau^{SRD} = E[Y(1) - Y(0)|X = c]$. Then under the continuity assumption, $\tau^{SRD} = lim_{x \to c^+} E[Y|X = x] - lim_{x \to c^-} E[Y|X = x]$. Both of these quantities on the right hand side are observed. We prove this equality below.

$\mu_0(x) = E[Y(0)|X = c] = \lim_{x \to c^-} E[Y(0)|X = x]$ by continuity
$= \lim_{x \to c^-} E[Y(0)|X = x, Z = 0]$ by SRD
$= \lim_{x \to c^-} E[Y|X = x, Z = 0]$ by consistency
$= \lim_{x \to c^-} E[Y|X = x]$ by SRD

We can do similar calculation to obtain $\mu_1(x)$. Thus, we have that $\tau^{SRD} = \mu_1(x) - \mu_0(x) = lim_{x \to c^+} E[Y|X = x] - lim_{x \to c^-} E[Y|X = x]$.

Some caveats to this approach:

- We have moved the goalpost. We want to observe the effect of treatment, but now we are estimating the effect of treatment at point c. Our estimand is $\tau^{SRD} = E[Y(1) - Y(0)|X = c]$.

- In certain cases, we can see behaviors of sorting/bunching
  - For example, an experiment in Colombia, a poverty index score determined eligibility to free social programs
  - When the algorithm to determine the score was not revealed, then there was no problem
  - When the algorithm was revealed, those barely on the other side of the threshold c crossed barely over to the other side to get the free program.
  - Thus there is some form of self selection and you have bunching right below the threshold c from the "cheaters."

5.2.2 Estimation and Inference

The challenge with estimation and inference is that the regression is at the boundary point. Therefore, we likely do not have much data near the boundary, and we also will have bias.

First, we always plot $Y \sim X$ on a graph first. If there is nothing that jumps out at us there, then we have no analysis. If we do see something we can attack the problem using two different methods.

1. Method 1: Non-parametric Kernel approaches
   - Define our estimate to be $\hat{\tau}^{unif} = \frac{\sum Y_i \mathbb{1}\{c \leq X_i \leq c+h\}}{\sum Y_i \mathbb{1}\{c-h \leq X_i \leq c+h\}} - \frac{\sum Y_i \mathbb{1}\{c-h \leq X_i \leq c\}}{\sum Y_i \mathbb{1}\{c-h \leq X_i \leq c+h\}}$
   - Intuitively, we are taking a small amount of distance (h) below and above the threshold c, taking the average of these below threshold points and above threshold points and then taking the difference of these points
   - This estimate is biased, it would likely have an overestimating effect.
2. Method 2: Local Linear Regression
   - Solve for $\hat{\alpha}_0, \hat{\beta}_0 = argmin \sum_{i:c-h \leq X_i \leq c} (Y_i - \alpha - \beta(X_i - c))^2$
   - Solve for $\hat{\alpha}_1, \hat{\beta}_1 = argmin \sum_{i:c \leq X_i \leq c+h} (Y_i - \alpha - \beta(X_i - c))^2$
   - Calculate $\hat{\tau}^{LLR} = \hat{\mu}_1(X = c) - \hat{\mu}_0(X = c) = \hat{\alpha}_1 - \hat{\alpha}_0$
   - Inference can be done with OLS with robust standard errors with some conditions on bandwidth

5.3 Difference in Differences (DiD)

We obtain intuition for Difference in Differences (DiD) from the following examples. Say we are trying to find the effect of minimum wage law on employment in the state of New Jersey. We want to see if increasing minimum wage will lead to a decrease in the number of employees hired. Each unit i is a fast food employer and Y_i is the number of employees at a fast food restaurant i.

To figure this out, we may have an idea to use time (before/after the increase in minimum wage). Say the day of the minimum wage increase was T_0, and define t as the time before T_0 and $t+1$ as the time after T_0. Then, in our superpopulation framework, we have $(Z_{i,t}, Z_{i,t+1}, Y_{i,t}, Y_{i,t+1}) \sim P$. Our estimand is $E[Y_{i,t+1}(1) - Y_{i,t+1}(0)] = \mu_{t+1}(1) - \mu_{t+1}(0)$. In order to estimate this estimand, our estimator will be $\hat{\tau} = \bar{Y}_{t+1} - \bar{Y}_t$, which implies that $E[\hat{\tau}] = \mu_{t+1}(1) - \mu_{t+1}(0)$ which is all we observe. However, almost certainly, $\mu_t(0) \neq \mu_{t+1}(0)$. There is no guarantee that the only change in number of employees from time t to $t+1$ is the minimum wage law. There are almost certainly other factors at work that could also affect the number of employees in between those times. Thus, to use this estimand, we have to use a strong assumption of no time-confounding, which may not be realistic.

Another idea may to be use a different state, say Pennsylvania (PA), that does not have this change in minimum wage as a control. Now we have another variable $G_i \in \{0, 1\}$, which is 0 for Pennsylvania and 1 for New Jersey. Then, using the super population framework, we have that $(Y_i, Z_i, G_i) \sim P$ at time $t + 1$. Our estimand in this case is $\tau = E[Y_i(1) - Y_i(0)|G_i = 1] = \mu^{NJ}(1) - \mu^{NJ}(0)$. Again, because we don't observe $\mu^{NJ}(0)$, our estimator will be defined as $\hat{\tau} = \bar{Y}^{NJ} - \bar{Y}^{PA}$, which implies that $E[\hat{\tau}] = \mu^{NJ}(1) - \mu^{PA}(0)$. However, we run into problems again because we have no guarantee that $\mu^{PA} = \mu^{NJ}$. Thus, we need to make a strong assumption again.

We may be tempted to do observational study analysis on this, but we cannot make the ignorability assumption because we violate the probabilistic assignment requirement, since $P(Z_i = 1|G_i)$ equals 1 if $G_i = 1$ and 0 otherwise, and never anything in between.

Therefore, both ideas previously discussed make unrealistic assumptions and different types of data. The first approach uses 1 state, but 2 time periods, while the second approach uses 2 states, but 1 time period. Within each approach, observing more of the same kind of data would not help (reduce variance but does not help with bias), since this is an identification problem. Thus, we combine the two approaches and use 2 states with 2 time periods. We can get identification under arguably weaker conditions than approach 1 or 2 alone. This is the Difference in Differences.

5.3.1 Derivation of DiD

In DiD, we have 2 states as well as 2 time periods. Notation is shown below.

- Outcomes are $Y_{i,t}(0), Y_{i,t}(1), Y_{i,t+1}(0), Y_{i,t+1}(1)$
- Which state we are in is $G_i \in \{0,1\}$, where 0 is for PA and 1 is for NJ
- $P(Z_{i,t'} = 1|G_i = g) = \mathbb{1}\{t' = t+1, G_i = 1\}$. The only time that the assignment is 1 is in the state of NJ at time $t+1$. In NJ in time t, as well as in PA for either time t or $t+1$, we are assigned to $Z_i = 0$, the control.

The estimand is $\tau = E[Y_{i,t+1}(1) - Y_{i,t+1}(0)|G_i = 1] = \mu^{NJ}_{t+1}(1) - \mu^{NJ}_{t+1}(0)$. Our assumption using DiD is a weaker assumption, but we assume that in the absence of treatment, that NJ and PA would have evolved, but in the same way. In other words, the conjectural trends affecting PA and NJ are the same except for the effect of the treatment. This assumption is called "parallel trends." We are assuming that $E[Y_{i,t+1}(0) - Y_{i,t}(0)|G_i = 1] = E[Y_{i,t+1}(0) - Y_{i,t}(0)|G_i = 0]$. We can also denote this assumption as $\Delta_t\mu^{NJ}(0) = \Delta_t\mu^{PA}(0)$, where $\Delta_t\mu^{NJ}(0) = \mu^{NJ}_{t+1}(0) - \mu^{NJ}_t(0)$.

However, what do we observe? We observe $m^G_{t'} = E[Y_{i,t'}|G_i = g]$, where $t' \in \{t, t+1\}$ and $G_i \in \{0,1\}$. The proposition is as follows. Our DiD estimand is $\tau = \Delta_t m^{NJ} - \Delta_t m^{PA}$. The proof is below.

- Step 1: Show that $\mu^{NJ}_{t+1}(1) = m^{NJ}_{t+1}$; $\mu^{NJ}_t(0) = m^{NJ}_t$; $\mu^{PA}_{t+1}(0) = m^{PA}_{t+1}$; $\mu^{PA}_t(0) = m^{PA}_t$. The first of the four is shown below, the calculation is similar for all four
  - $\mu^{NJ}_{t+1}(1) = E[Y_{i,t+1}(1)|G_i = 1]$
    $= E[Y_{i,t+1}(1)|Z_i = 1, G_i = 1]$ from the setup $= E[Y_{i,t+1}|Z_i = 1, G_i = 1]$ from consistency assumption
    $= E[Y_{i,t+1}|G_i = 1]$ from the setup
    $= m^{NJ}_{t+1}$
- Step 2: By assumption, $\Delta_t\mu^{NJ} = \Delta_t\mu^{PA}$. This implies that:
  - $\mu^{NJ}_{t+1}(0) = \mu^{NJ}_t(0) + \Delta_t\mu^{PA}$
    $= m^{NJ}_t + \Delta_t m^{PA}$
- Step 3:
  - $\tau = \mu^{NJ}_{t+1}(1) - \mu^{NJ}_{t+1}(0)$
    $= m^{NJ}_{t+1} - (m^{NJ}_t + \Delta_t m^{PA})$
    $= (m^{NJ}_{t+1} - m^{NJ}_t) - \Delta_t m^{PA}$
    $= \Delta_t m^{NJ} - \Delta_t m^{PA}$

Now that we have this relationship, in practice we have our estimator $\hat{\tau}^{DiD} = \Delta_t\bar{Y}^{NJ}_t - \Delta_t\bar{Y}^{PA}_t$.

One of the advantages of using DiD is that we don't need to have Panel Data to perform the analysis. We don't need to track the same fast food restaurants over time, but instead can use different samples of fast food restaurants in time t and in time $t+1$. However, one difficulty with DiD is that the parallel trend assumption is NOT scale invariant. Another difficulty is if there are multiple time points.

5.4 Synthetic Controls

The synthetic controls technique can be explained using an example. Let's say that there was a law that increased cigarette taxes by 25 cents a pack, and we want to find the impact of this law on smoking. However, this law is only passed in one state, say California. Therefore, let us denote $i = 0, 1, ..., 49$ as the 50 different states, where $i = 0$ is California. Let us denote the time $t = 1, ..., T_0, T_0 + 1, ..., T$, where the law is passed at time T_0. Then let us denote Z_{it} as 1 if $i = 0$ and $t \geq T_0 + 1$ and 0 otherwise. In other words, where the law is enacted (California after time T_0) we assign 1 and otherwise 0. Then our assignment table looks like below.

		t	
		$1, ..., T_0$	$T_0 + 1, ..., T$
	0	$Z = 0$	$Z = 1$
i	1	$Z = 0$	$Z = 0$
	$\vdots$	$\vdots$	$\vdots$
	49	$Z = 0$	$Z = 0$

We want to know the effect of the law (which is passed only in California). Therefore, our estimand is $\tau_t = Y_{0,t}(1) - Y_{0,t}(0)$, where $t = T_0 + 1, ..., T$. However, we don't observe $Y_{0,t}(0)$, and so we need to estimate it. In theory, we could estimate it using DiD, where we pick a "similar" state which hasn't passed the law and assume parallel trends to do our analysis. However, say in this case, none of the other 49 states are similar to California. Thus we need to use the Synthetic Controls approach.

In synthetic controls, we don't have a "similar" state to California, so we create a synthetic state using the data from the other 49 states. The idea is that a weighted combination of the other 49 states can be used (when combined with the appropriate weights), to create a fake "state" that is similar to California. Mathematically, we want $\sum_{i=1}^{N} w_i Y_{i,t}(0) \approx Y_{0,t}(0)$, where the weights are positive and the sum of the weights is 1 (although these constraints are sometimes relaxed). We observe $Y_{i,t}(0)$ for $i \geq 1$ (all states except California) all time t.

Thus, our estimator is $\hat{\tau}_t = Y_{0,t}(1) - \sum_{i=1}^{N} w_i Y_{i,t}$ for $t \geq T_0 + 1$.

But how do we get the weights w_i? We do this from the time period prior to the law $t = 1, ...T_0$. We try to fit the weights such that they are as close to California as possible. We solve

$\vec{w}^* = argmin_{\vec{w}} \sum_{i=1}^{T_0} (Y_{0,t}(0) - \sum_{i=1}^{N} w_i Y_{0,t}(0))^2$

Once we solve for the optimal weights, then we can just use these weights for our estimator for the time periods PAST T_0 in order to obtain our estimate.

Synthetic controls is slightly complicated and likely requires optimization software. The difficulty though, is that it requires very very strong assumptions and the fitting of the weights has to be very precise. In addition, how to perform inference such as confidence intervals and p-values is not yet settled. One idea proposed is to do a FRT-like analysis. We can examine the values before and after for all of the "non-Californian" states and see what percentage of the time the change in smoking in California is greater than the change in these "placebo" states.