RUBIN & NEYMAN’S FRAMEWORK

- Average treatment effect
- IPW, the learners, AIPW
- Heterogeneous treatment effect

THE AVERAGE TREATMENT EFFECT

The Potential Outcomes Framework (Rubin-Neyman) – Recall of the main assumptions

1. **SUTVA:** No interference between units (your treatment doesn't affect my outcome), and only one version of the treatment – $Y = TY(1) + (1 - T)Y(0)$ → SUTVA establishes a link between potential and observed outcome, and impose independence between subjects

2. **Ignorability** (or **unconfoundedness**): Treatment assignment is "as good as random" given the adjustment set $\mathbf{Z}$, i.e., there is no unmeasured confounder – $(Y(0), Y(1)) \perp\!\!\!\perp T | \boldsymbol{Z}$ → Ignorability states that our causal estimand is identifiable and an estimate can be performed using available data

3. **Positivity:** Every unit has a non-zero probability of receiving any treatment – $\forall\, t, \boldsymbol{z}, i,\ 0 < \mathbb{P}(T_i = t | \boldsymbol{Z}_i = \boldsymbol{z}) < 1$
→ Positivity means that any subject in the population has a chance to be treated, irrespective of its characteristics ($\boldsymbol{z}$)

THE AVERAGE TREATMENT EFFECT

The Potential Outcomes Framework (Rubin-Neyman) – Recall of the main assumptions

1. **SUTVA:** No interference between units (your treatment doesn't affect my outcome), and only one version of the treatment – $Y = TY(1) + (1 - T)Y(0)$
2. **Ignorability** (or **unconfoundedness**): Treatment assignment is "as good as random" given the adjustment set $\mathbf{Z}$, i.e., there is no unmeasured confounder – $(Y(0), Y(1)) \perp\!\!\!\perp T|\boldsymbol{Z}$
3. **Positivity:** Every unit has a non-zero probability of receiving any treatment – $\forall\, t, \boldsymbol{z}, i,\ \ 0 < \mathbb{P}(T_i = t|\boldsymbol{Z}_i = \boldsymbol{z}) < 1$)

Given assumptions 1.-2.-3. The average treatment effect (ATE, $\tau \coloneqq \mathbb{E}[Y(1) - Y(0)]$) can be identified, simply by marginalizing out $\mathbf{Z}$:

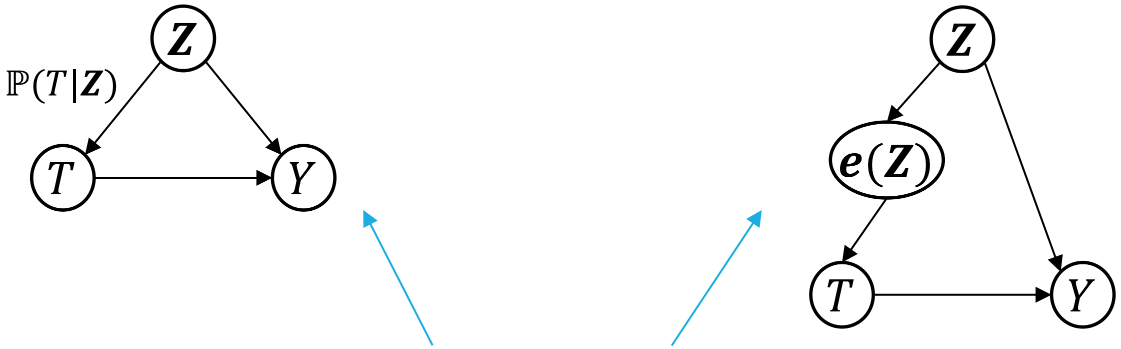
$$\begin{aligned}
\tau &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] && \text{Linearity of } \mathbb{E} \\
&= \mathbb{E}_{\mathbf{Z}}\mathbb{E}[Y(1)|\mathbf{Z}] - \mathbb{E}_{\mathbf{Z}}\mathbb{E}[Y(0)|\mathbf{Z}] && \text{Law of iterated } \mathbb{E} \\
&= \mathbb{E}_{\mathbf{Z}}\mathbb{E}[Y(1)|T = 1, \mathbf{Z}] - \mathbb{E}_{\mathbf{Z}}\mathbb{E}[Y(0)|T = 0, \mathbf{Z}] && \text{Ignorability and positivity} \\
&= \mathbb{E}_{\mathbf{Z}}\mathbb{E}[Y|T = 1, \mathbf{Z}] - \mathbb{E}_{\mathbf{Z}}\mathbb{E}[Y|T = 0, \mathbf{Z}] && \text{SUTVA}
\end{aligned}$$

INVERSE PROPENSITY WEIGHTING (IPW)

A first intuitive idea to estimate ATE, adjusting for confounding from Z

- We define the ***propensity score***: $e(\boldsymbol{Z}) \coloneqq \mathbb{P}(T = 1|\boldsymbol{Z})$ → The propensity score corresponds to the probability of being treated given our adjustment coordinates.

- Positivity and ignorability given $\boldsymbol{Z}$ implies also ignorability given $e(\boldsymbol{Z})$:


$\boldsymbol{Z}$
$\mathbb{P}(T|\boldsymbol{Z})$
T
Y
$\boldsymbol{Z}$
$\boldsymbol{e}(\boldsymbol{Z})$
T
Y

$\boldsymbol{Z}$ is a valid adjustment set $\Longrightarrow$ $e(\boldsymbol{Z})$ is also a valid adjustment set

INVERSE PROPENSITY WEIGHTING (IPW)

A first intuitive idea to estimate ATE, adjusting for confounding from Z

- We define the ***propensity score***: $e(\boldsymbol{Z}) \coloneqq \mathbb{P}(T = 1|\boldsymbol{Z})$ → The propensity score corresponds to the probability of being treated given our adjustment coordinates.

- Positivity and ignorability given $\boldsymbol{Z}$ implies also ignorability given $e(\boldsymbol{Z})$.

- We want to make T independent from $\boldsymbol{Z}$: this can be done by reweighting each subject using the inverse probability of receiving its value of treatment given its value of $\boldsymbol{Z}$. This is exactly the propensity score → Inverse propensity score matching!

Intuition

The edge we would like to remove describes $\mathbb{P}(T|\boldsymbol{Z})$, by weighting with $1/\mathbb{P}(T = 1|\boldsymbol{Z})$ we are effectively cancelling this out.

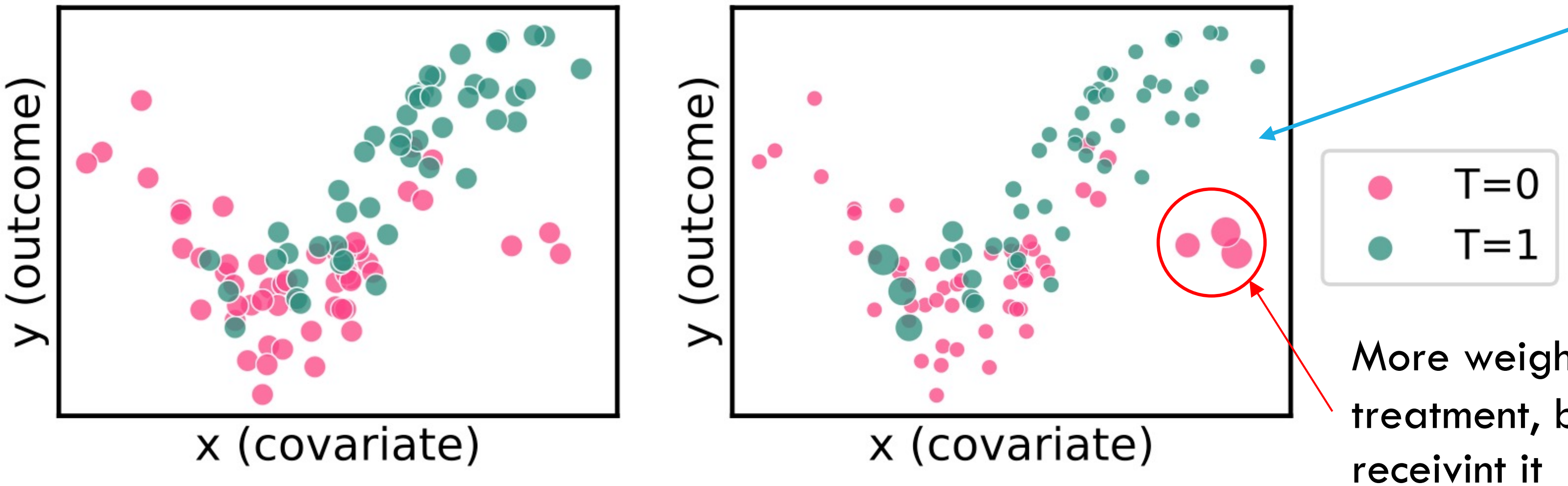
INVERSE PROPENSITY WEIGHTING (IPW)

A first intuitive idea to estimate ATE, adjusting for confounding from $\mathbb{Z}$

- We define the ***propensity score***: $e(\boldsymbol{Z}) := \mathbb{P}(T = 1|\boldsymbol{Z})$ → The propensity score corresponds to the probability of being treated given our adjustment coordinates.

- Positivity and ignorability given $\boldsymbol{Z}$ implies also ignorability given $e(\boldsymbol{Z})$.

- We want to make T independent from $\boldsymbol{Z}$: this can be done by reweighting each subject using the inverse probability of receiving its value of treatment given its value of $\boldsymbol{Z}$. This is exactly the propensity score → Inverse propensity score matching!



Reweighted population: as if treatment assignment was random

More weight to subjects who received treatment, bud had low probability of receivint it

INVERSE PROPENSITY WEIGHTING (IPW)

A first intuitive idea to estimate ATE, adjusting for confounding from Z

- We define the ***propensity score***: $e(\boldsymbol{Z}) \coloneqq \mathbb{P}(T = 1|\boldsymbol{Z})$
- Positivity and ignorability given $\boldsymbol{Z}$ implies also ignorability given $e(\boldsymbol{Z})$.
- We want to make T independent from $\boldsymbol{Z}$: this can be done by reweighting each subject using the inverse probability of receiving its value of treatment given its value of $\boldsymbol{Z}$. This is exactly the propensity score.
- This leads to the following statistical estimand for ATE:

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[\frac{\mathbb{I}(T=1)Y}{e(\mathbf{Z})}\right] - \mathbb{E}\left[\frac{\mathbb{I}(T=0)Y}{1-e(\mathbf{Z})}\right]$$

INVERSE PROPENSITY WEIGHTING (IPW)

A first intuitive idea to estimate ATE, adjusting for confounding from Z

- We define the ***propensity score***: $e(\boldsymbol{Z}) \coloneqq \mathbb{P}(T = 1|\boldsymbol{Z})$
- Positivity and ignorability given $\boldsymbol{Z}$ implies also ignorability given $e(\boldsymbol{Z})$.
- We want to make T independent from $\boldsymbol{Z}$: this can be done by reweighting each subject using the inverse probability of receiving its value of treatment given its value of $\boldsymbol{Z}$. This is exactly the propensity score.
- Replacing expectations by empirical means and the propensity score by a valid estimator $\hat{e}(\boldsymbol{Z})$ (e.g., using logistic regression), we can derive the IPW estimator for the ATE:

$$\hat{\tau}_{IPW} = \frac{1}{N}\sum_i \left(\frac{\mathbb{I}(t_i = 1)y_i}{\hat{e}(\mathbf{z}_i)} - \frac{\mathbb{I}(t_i = 0)y_i}{1 - \hat{e}(\mathbf{z}_i)} \right)$$

THE LEARNERS

- An alternative approach consist in fitting a statistical model $\hat{\mu}$ (a **learner** - from linear regression to ML model) directly to the conditional expectation $\mu(t, \boldsymbol{z}) \coloneqq \mathbb{E}[Y|T = t, \boldsymbol{Z} = \boldsymbol{z}]$
- $\mathbb{E}_{\mathbf{Z}}$ can be approximated through empirical mean, as before:

$$\hat{\tau}_{S-l} = \frac{1}{N}\sum_i (\hat{\mu}(1, \mathbf{z}_i) - \hat{\mu}(0, \mathbf{z}_i))$$

This is called the S-learner (S for single), since we train here a unique model of the conditional expectation!

THE LEARNERS

- To enforce the learner to correctly use T as a predictor, and not just $\boldsymbol{Z}$ (which can be high-dimensional), we can decide to train two different models, $\hat{\mu}_1$ and $\hat{\mu}_2$, one per intervention.

- Then we proceed as usual with the empirical mean:

$$\hat{\tau}_{T-l} = \frac{1}{N} \sum_{i} \left(\hat{\mu}_1(\mathbf{z}_i) - \hat{\mu}_0(\mathbf{z}_i) \right)$$

This is called the T-learner (T for two), since we train two distinct models of the conditional expectation!

✓ The T-learner has the advantage of mitigating the bias toward zero treatment effect issue the S-learner may have.

× The drawback is that the T-learner uses a significanlty reduced portion of data to train each learner

THE LEARNERS

- The learners proposed so far aim to estimate the ATE by 1. separately estimating (using either a single or two learners) the outcome for the treated and untreated, and 2. taking their difference.

- We can try to work from an ITE viewpoint, i.e. building an estimator for the individual treatment effect.

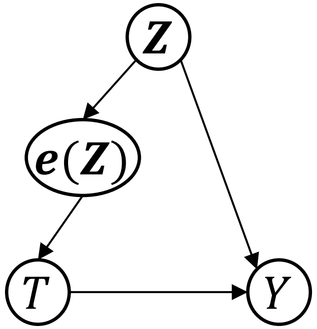
1. Consider a model to estimate the outcome of treated/untreated, given the adjustment variables $\boldsymbol{Z}$.
2. For an untreated subject j, the ITE is modelled as: $d_j^0 = \mu_1(\boldsymbol{z}_j) - y_j$
   For a treated subject i, the ITE is modelled as: $d_i^1 = y_i - \mu_0(\boldsymbol{z}_I)$
3. A final model can be trained to estimate directly $\hat{d}$.

- This is called the X-learner: has the advantage of requiring less data wrt to the T-learner (since it makes use of all data for fitting), while ensuring the variable T is still accounted for.

- X stands for cross: the model learned on treated is used to infer the ITE of untreated and vice versa.

AUGMENTED IPW (AIPW)

What about combining both IPW and learners for more robustness?



We can close the backdoor path between T and $\boldsymbol{Z}$ in 3 different ways:

- By controlling for $\boldsymbol{Z}$ → outcome model through the learners
- By controlling for $\mathrm{e}(\boldsymbol{Z})$ → treatment model through IPW
- By controlling for both

NOTE: Although the estimands obtained either by controlling for $\boldsymbol{Z}$ or by controlling for $\mathrm{e}(\boldsymbol{Z})$ are equivalent, their estimators (which quantify them) might have different errors. Importantly, one of the models might be correct and the other might be misspecified.

AUGMENTED IPW (AIPW)

What about combining both IPW and learners for more robustness?

We can close the backdoor path between T and $\boldsymbol{Z}$ in 3 different ways:

- By controlling for $\boldsymbol{Z}$ → outcome model through the learners
- By controlling for $\mathrm{e}(\boldsymbol{Z})$ → treatment model through IPW
- By controlling for both

NOTE: Although the estimands obtained either by controlling for $\boldsymbol{Z}$ or by controlling for $\mathrm{e}(\boldsymbol{Z})$ are equivalent, their estimators (which quantify them) might have different errors. Importantly, one of the models might be correct and the other might be misspecified.

This lead us to the AIPW estimator wich is a ***doubly robust*** estimator for the ATE, i.e. if the 3 assumptions hold and 1 among the outcome and the treatment models are consistent (not misspecified), the AIPW is consistent too.

$$\hat{\tau}_{AIPW} = \frac{1}{N}\sum_i \left(\frac{t_i(y_i - \hat{\mu}_1(\mathbf{z}_i))}{\hat{e}(\mathbf{z}_i)} + \hat{\mu}_1(\mathbf{z}_i) \right) - \frac{1}{N}\sum_i \left(\frac{(1-t_i)(y_i - \hat{\mu}_0(\mathbf{z}_i))}{1-\hat{e}(\mathbf{z}_i)} + \hat{\mu}_0(\mathbf{z}_i) \right)$$

AUGMENTED IPW (AIPW)

What about combining both IPW and learners for more robustness?

We can close the backdoor path between T and $\boldsymbol{Z}$ in 3 different ways:

- By controlling for $\boldsymbol{Z}$ → outcome model through the learners
- By controlling for $\mathrm{e}(\boldsymbol{Z})$ → treatment model through IPW
- By controlling for both

This lead us to the AIPW estimator wich is a ***doubly robust*** estimator for the ATE, i.e. if the 3 assumptions hold and 1 among the outcome and the treatment models are consistent (not misspecified), the AIPW is consistent too.

$$\hat{\tau}_{AIPW} = \underbrace{\boxed{\frac{1}{N}\sum_i \left(\frac{t_i(y_i - \hat{\mu}_1(\mathbf{z}_i))}{\hat{e}(\mathbf{z}_i)} + \hat{\mu}_1(\mathbf{z}_i) \right)}}_{\mathbb{E}_{\mathbf{Z}}\mathbb{E}[Y|T=1,\mathbf{Z}]} - \underbrace{\boxed{\frac{1}{N}\sum_i \left(\frac{(1-t_i)(y_i - \hat{\mu}_0(\mathbf{z}_i))}{1-\hat{e}(\mathbf{z}_i)} + \hat{\mu}_0(\mathbf{z}_i) \right)}}_{\mathbb{E}_{\mathbf{Z}}\mathbb{E}[Y|T=0,\mathbf{Z}]}$$

CONDITIONAL TREATMENT EFFECT - CATE

- The **Average Treatment Effect (ATE)** can be misleading, as it often masks heterogeneous responses across a population.

- Let us consider a set $\boldsymbol{X} \subseteq \boldsymbol{Z}$: we are interested in studying the heterogeneity of the treatment effect with respect to $\boldsymbol{X}$ → **Conditional Average Treatment Effect (CATE)** allows us to estimate the treatment effects for specific subgroups or individuals:

$$\tau(\boldsymbol{x}) \coloneqq \mathbb{E}[Y(1) - Y(0)|\boldsymbol{X} = \boldsymbol{x}]$$

This is the foundation for Personalized Medicine!

- The estimators seen so far can naively adapted to evaluate CATE simply by averaging only on subpopulations characterized by different level of $\boldsymbol{X}$, and/or explicitely introducing interaction terms between T and $\boldsymbol{X}$ (in the learners)

CAUSAL RANDOM FORESTS

- Identifying meaningful subpopulations with respect to $\boldsymbol{X}$ can be challenging, especially if $\boldsymbol{X}$ is multivariate and high-dimensional → Why don't do it in a data-driven manner?

- Some models have been explicitly designed to deal with heterogeneous effects. This is the case for the ***causal random forests***.

CAUSAL RANDOM FORESTS

Causal random forests use **causal trees**. In brief:

Causal trees aim to estimate CATE by **partitioning the covariate space into subgroups that are as homogeneous as possible in terms of treatment effects.** Each “leaf” of the tree represents a subgroup of observations with similar covariate profiles, and the estimated treatment effect for each leaf can be interpreted as an approximation of CATE for that covariate region.

$$\hat{\tau}(l) = \bar{Y}_{\mathbf{X}\in l, T=1} - \bar{Y}_{\mathbf{X}\in l, T=0}$$

Leaf l

CAUSAL RANDOM FORESTS

Standard CART (Classification and Regression Trees) are optimized to split the sample by minimizing outcome prediction errors, for instance by measuring MSE (Mean Squared Error). In our case, this would mean to minimize:

$$MSE = \mathbb{E}\left[\sum_{l\in\pi}(\hat{\tau}(l) - \tau(l))^2\right]$$

The partition

CAUSAL RANDOM FORESTS

Standard CART (Classification and Regression Trees) are optimized to split the sample by minimizing outcome prediction errors, for instance by measuring MSE (Mean Squared Error). In our case, this would men to minimize:

$$\cancel{MSE = \mathbb{E}\left[\sum_{l\in\pi}(\hat{\tau}(l) - \tau(l))^2\right]}$$

The fundamental problem of causal inference!

→ τ is never observed, only one outcome per unit is.

CAUSAL RANDOM FORESTS

Solution: causal trees'objective is to split based on **treatment effect heterogeneity,** i.e. the split should be done in such a way that the estimated treatment effects in the right and left leaves are as far apart as possible (while also ensuring each leaf has a mix of treated and control units for a valid comparison).

Problem: treatment effect being an estimated quantity, there is a great **risk of overfitting** if the same data are used both to optimize the split and to estimate the treatment effect within each leaf.

CAUSAL RANDOM FORESTS

Solution: Honest causal trees. Split the dataset into training and estimation sets. Use the training set for splitting and the estimation set to estimate treatment effects in each leaf.

→ The classical MSE is replaced by EMSE (Expected Mean Squared Error)
→ An estimator of EMSE can be derived: it accounts for both the variance of outcomes within leaves and an adjustment for the split proportions
→ The optimal splitting criterion (risk function) is to choose the split (only in the training data) that maximizes the reduction in estimated EMSE

CAUSAL RANDOM FORESTS

The forest: Causal random forests make use of B honest causal trees.

Causal forest perform something slightly more elaborate than just averaging over trees' outputs: it uses the forest to calculate ***similarity weights***. Briefly:

- Two units are similar if they often land in the same leaf.
- We estimate the CATE for a point x by a weighted average of outcomes of its neighbors in the forest.

→ This allows for the estimation of **individualized** treatment effects with asymptotic normality and confidence intervals.

IMPLEMENTATION IN PYTHON

Several implementation of the methods discussed in this class are available, both in R and in Pyhton.

Among the widely used Python packages:

- DoWhy, a complete framework for DAG-based causal inference.
- EconML, a library focused on modeling heterogeneous treatment effects using machine learning.