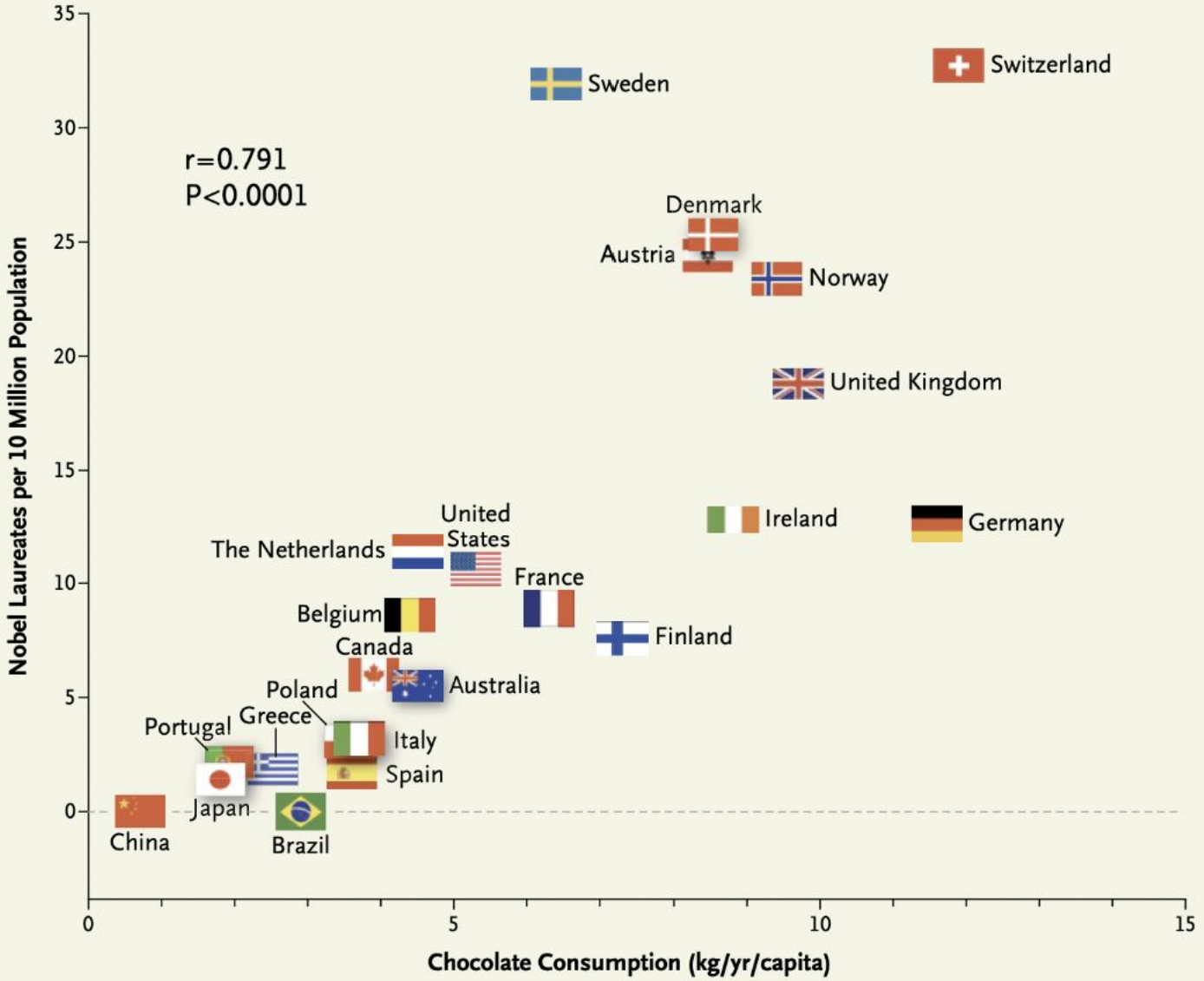
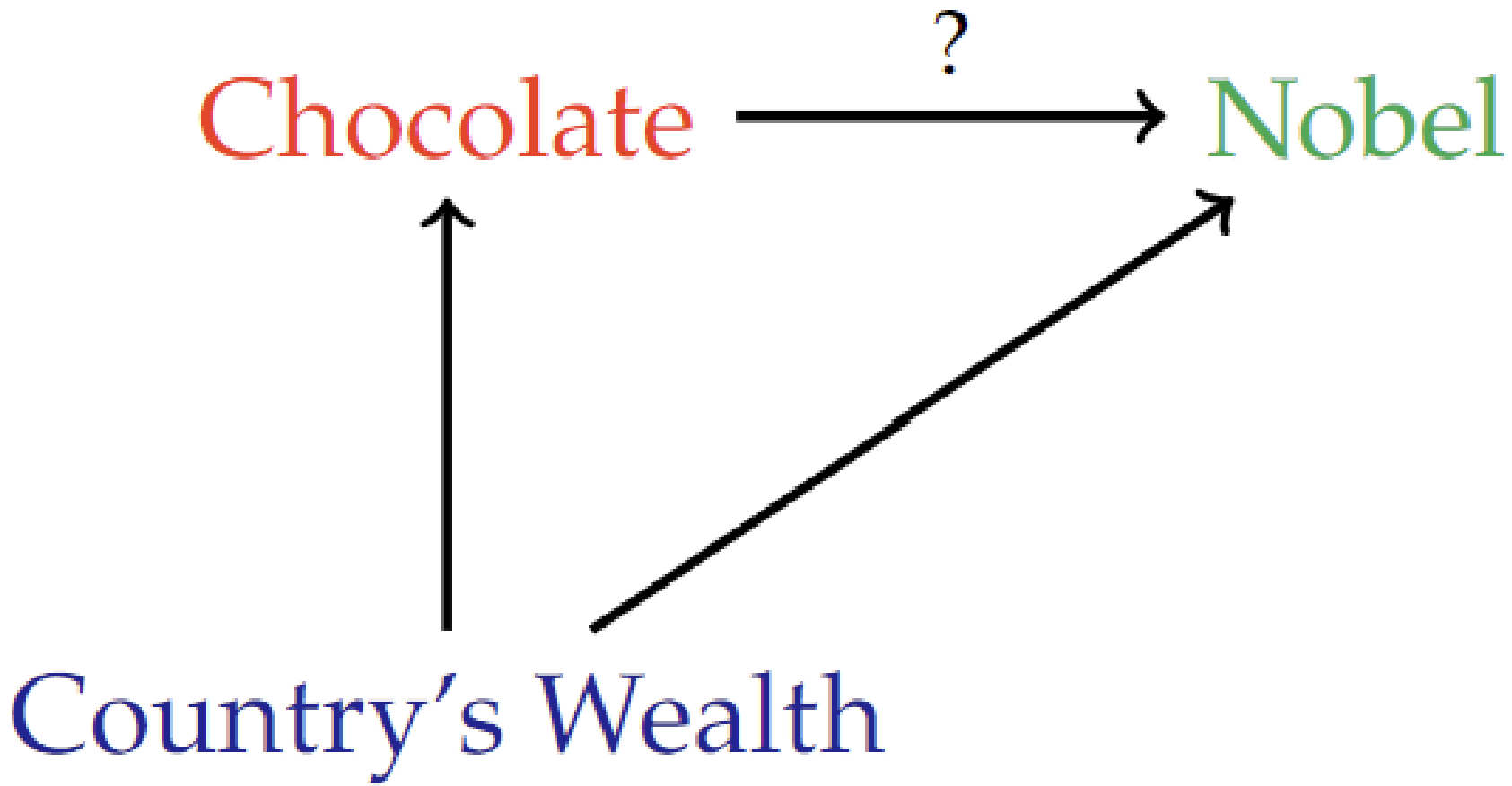
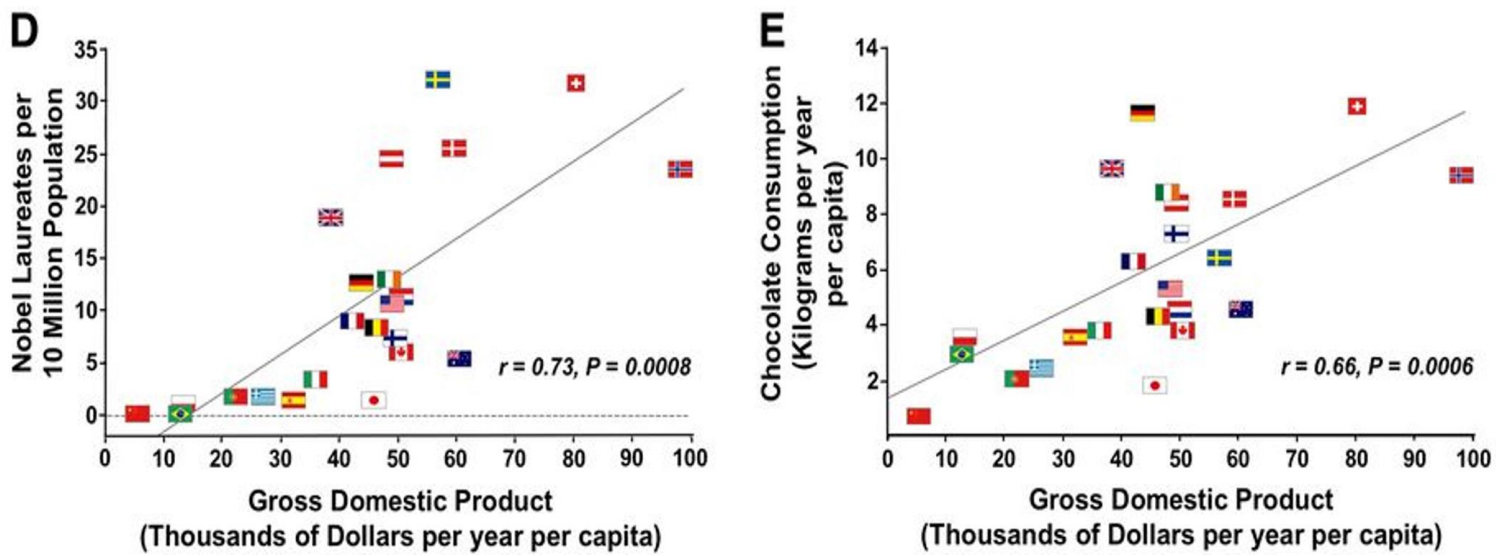
MS&E 228: Causality in Observational Data

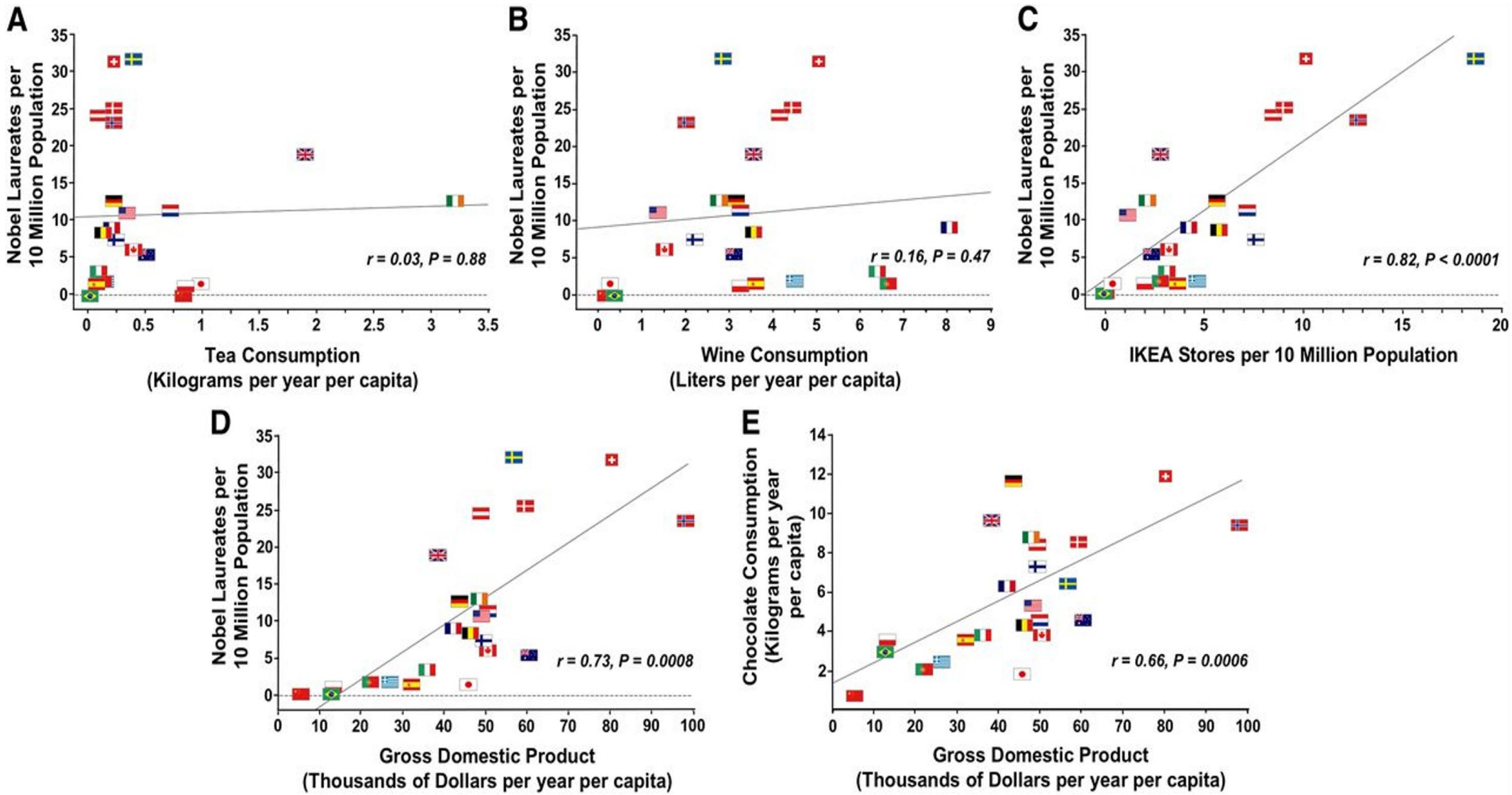
Vasilis Syrgkanis

MS&E, Stanford

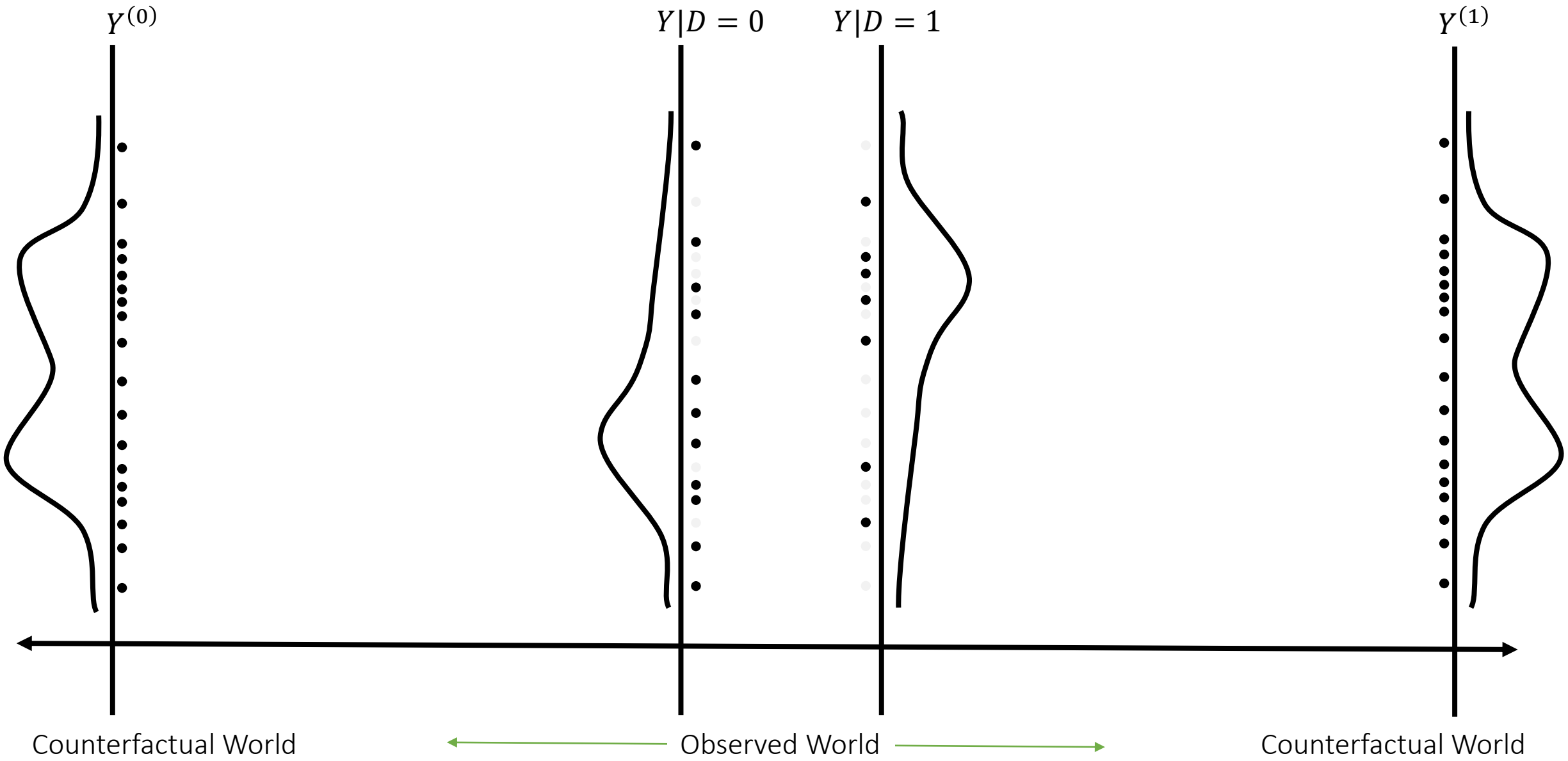

r=0.791
P<0.0001
Sweden
Switzerland
Denmark
Austria
Norway
United Kingdom
Ireland
Germany
United States
The Netherlands
France
Belgium
Finland
Canada
Australia
Poland
Greece
Portugal
Italy
Spain
Japan
China
Brazil
Nobel Laureates per 10 Million Population
Chocolate Consumption (kg/yr/capita)
0
5
10
15
20
25
30
35

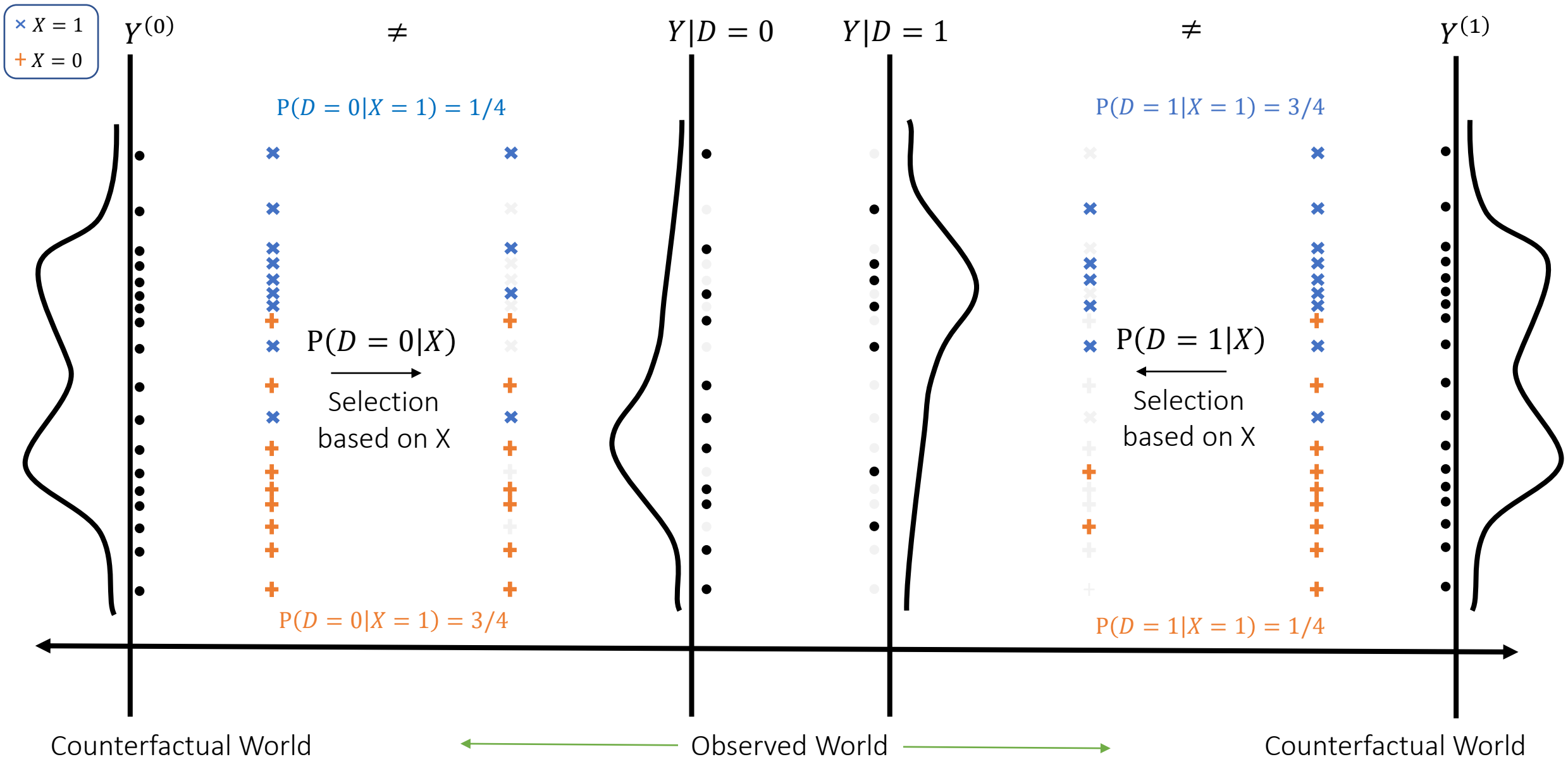

?
Chocolate
Nobel
Country's Wealth

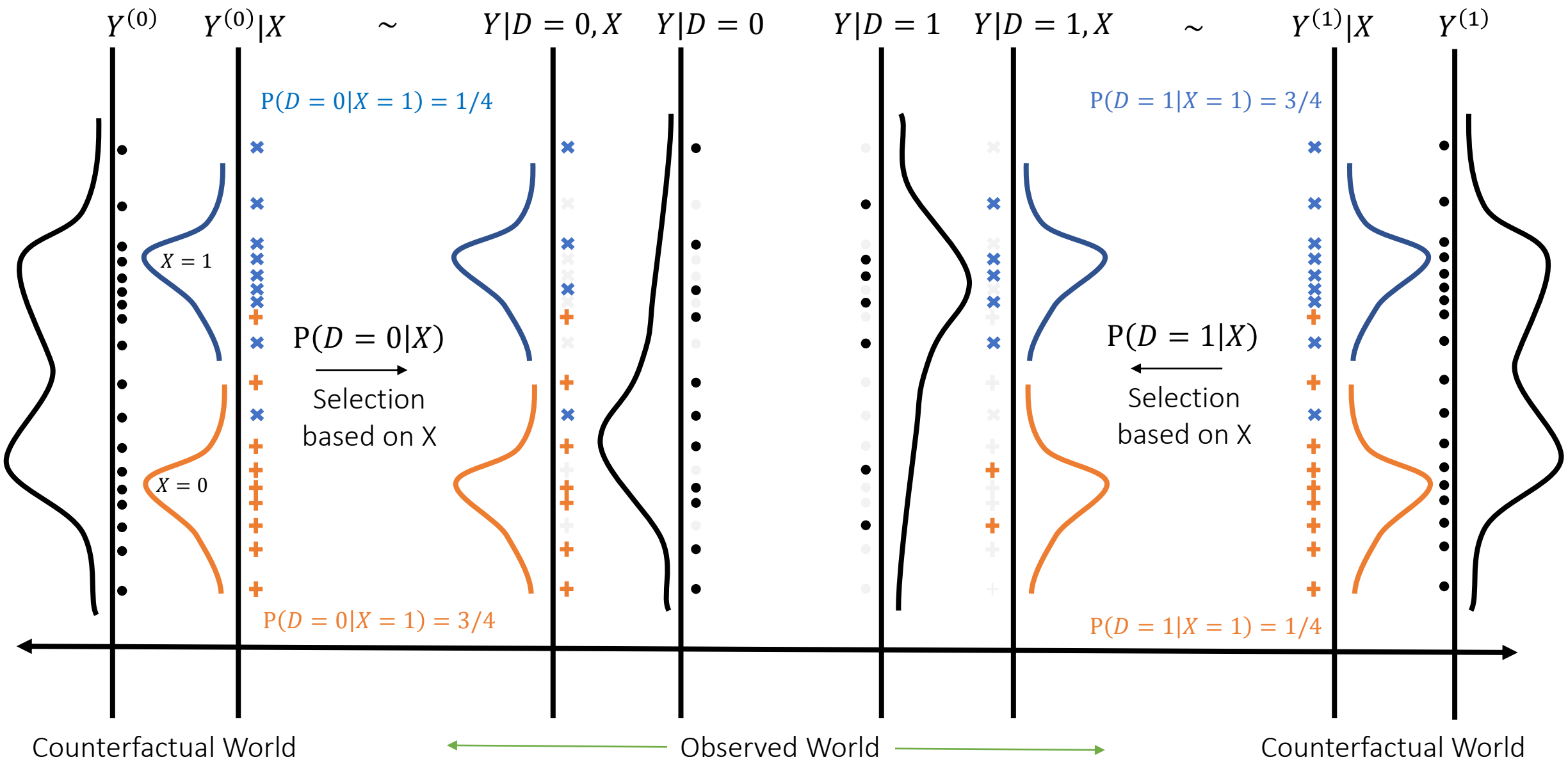

D
Nobel Laureates per
10 Million Population
35
30
25
20
15
10
5
0
0 10 20 30 40 50 60 70 80 90 100
r = 0.73, P = 0.0008
Gross Domestic Product
(Thousands of Dollars per year per capita)
E
Chocolate Consumption
(Kilograms per year
per capita)
14
12
10
8
6
4
2
0 10 20 30 40 50 60 70 80 90 100
r = 0.66, P = 0.0006
Gross Domestic Product
(Thousands of Dollars per year per capita)

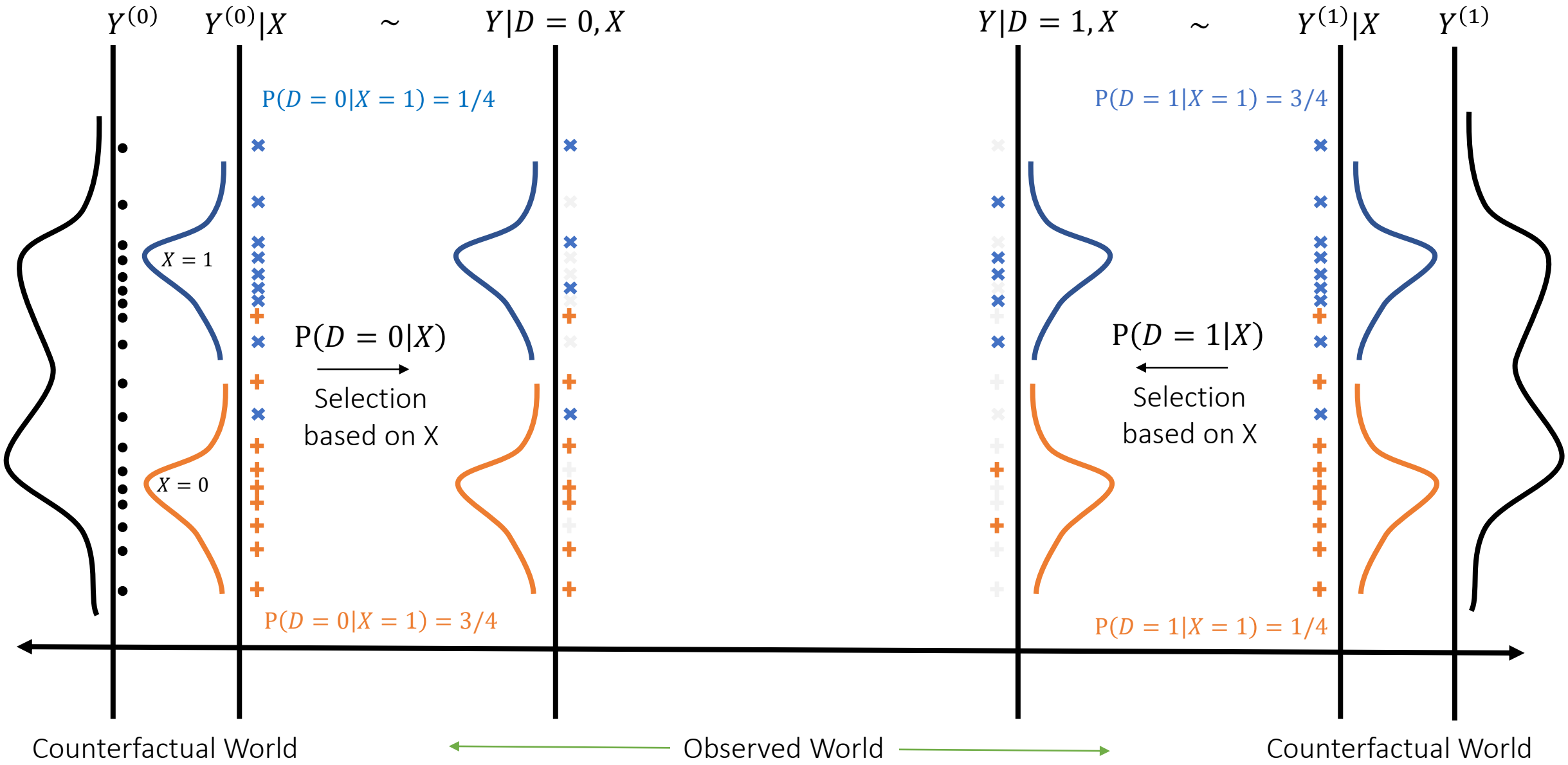

A
Nobel Laureates per 10 Million Population
Tea Consumption (Kilograms per year per capita)
r = 0.03, P = 0.88
B
Nobel Laureates per 10 Million Population
Wine Consumption (Liters per year per capita)
r = 0.16, P = 0.47
C
Nobel Laureates per 10 Million Population
IKEA Stores per 10 Million Population
r = 0.82, P < 0.0001
D
Nobel Laureates per 10 Million Population
Gross Domestic Product (Thousands of Dollars per year per capita)
r = 0.73, P = 0.0008
E
Chocolate Consumption (Kilograms per year per capita)
Gross Domestic Product (Thousands of Dollars per year per capita)
r = 0.66, P = 0.0006

Through the Lens of Potential Outcomes


$Y^{(0)}$
$Y|D=0$
$Y|D=1$
$Y^{(1)}$
Counterfactual World
Observed World
Counterfactual World


× $X = 1$
+ $X = 0$
$Y^{(0)}$
≠
$Y|D = 0$
$Y|D = 1$
≠
$Y^{(1)}$
$\mathrm{P}(D = 0|X = 1) = 1/4$
$\mathrm{P}(D = 0|X)$
Selection based on X
$\mathrm{P}(D = 0|X = 1) = 3/4$
$\mathrm{P}(D = 1|X = 1) = 3/4$
$\mathrm{P}(D = 1|X)$
Selection based on X
$\mathrm{P}(D = 1|X = 1) = 1/4$
Counterfactual World
Observed World
Counterfactual World

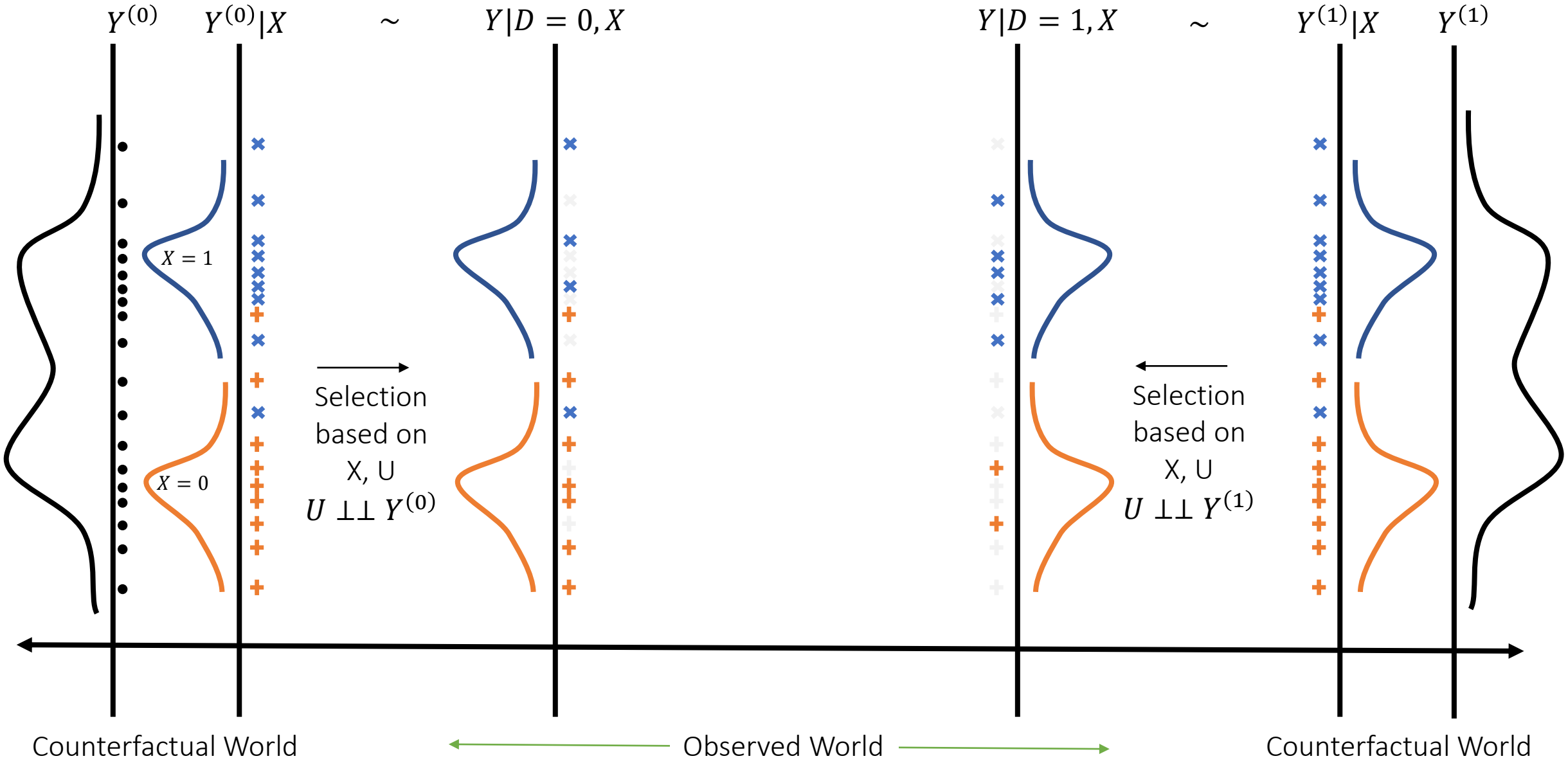

$Y^{(0)}$
$Y^{(0)}|X$
~
$Y|D=0,X$
$Y|D=0$
$Y|D=1$
$Y|D=1,X$
~
$Y^{(1)}|X$
$Y^{(1)}$
$\mathrm{P}(D=0|X=1)=1/4$
$\mathrm{P}(D=1|X=1)=3/4$
$X=1$
$X=0$
$\mathrm{P}(D=0|X)$
Selection based on X
$\mathrm{P}(D=1|X)$
Selection based on X
$\mathrm{P}(D=0|X=1)=3/4$
$\mathrm{P}(D=1|X=1)=1/4$
Counterfactual World
Observed World
Counterfactual World


$Y^{(0)}$
$Y^{(0)}|X$
$\sim$
$Y|D=0,X$
$Y|D=1,X$
$\sim$
$Y^{(1)}|X$
$Y^{(1)}$
$\mathrm{P}(D=0|X=1)=1/4$
$\mathrm{P}(D=1|X=1)=3/4$
$X=1$
$X=0$
$\mathrm{P}(D=0|X)$
Selection based on X
$\mathrm{P}(D=1|X)$
Selection based on X
$\mathrm{P}(D=0|X=1)=3/4$
$\mathrm{P}(D=1|X=1)=1/4$
Counterfactual World
Observed World
Counterfactual World

Conditional Ignorability

- For sub-populations with the same X, treatment is assigned as if RCT

$$Y^d \perp\!\!\!\perp D \mid X$$


$Y^{(0)}$
$Y^{(0)}|X$
$\sim$
$Y|D=0,X$
$Y|D=1,X$
$\sim$
$Y^{(1)}|X$
$Y^{(1)}$
$X=1$
$X=0$
Selection based on X, U
$U \perp\!\!\!\perp Y^{(0)}$
Selection based on X, U
$U \perp\!\!\!\perp Y^{(1)}$
Counterfactual World
Observed World
Counterfactual World

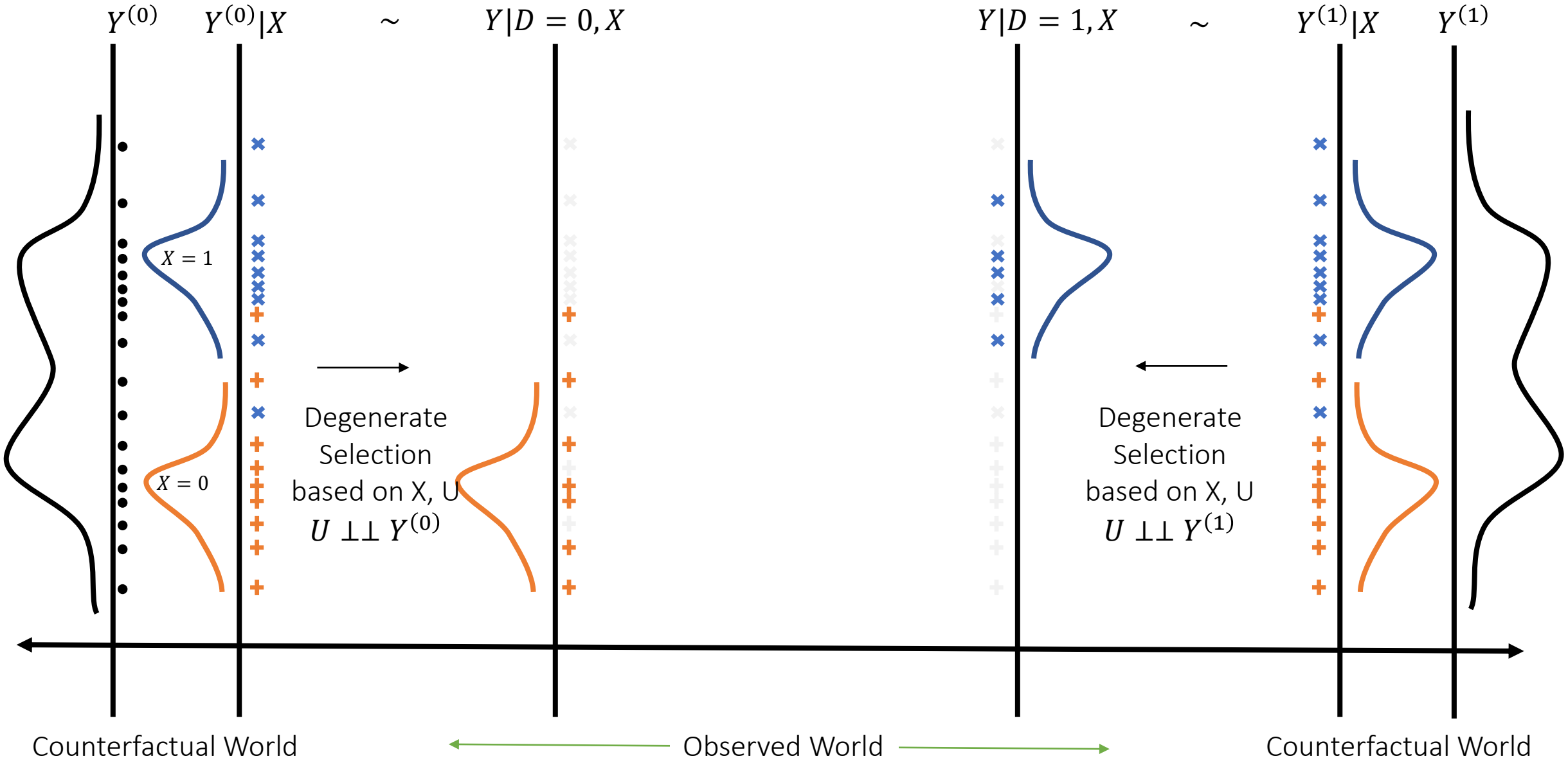
Conditional Ignorability

- For sub-populations with the same X, treatment is assigned as if RCT

$$Y^d \perp\!\!\!\perp D \mid X$$

- The probability of receiving treatment (propensity) is non-degenerate

$$0 < p(X) \coloneqq \Pr(D = 1|X) < 1$$


$Y^{(0)}$
$Y^{(0)}|X$
$\sim$
$Y|D=0,X$
$Y|D=1,X$
$\sim$
$Y^{(1)}|X$
$Y^{(1)}$
$X=1$
$X=0$
Degenerate Selection based on X, U
$U \perp\!\!\!\perp Y^{(0)}$
Degenerate Selection based on X, U
$U \perp\!\!\!\perp Y^{(1)}$
Counterfactual World
Observed World
Counterfactual World

Conditional Ignorability

- For sub-populations with the same X, treatment is assigned as if RCT

$$Y^{(d)} \perp\!\!\!\perp D \mid X$$

- The probability of receiving treatment (propensity) is non-degenerate

$$0 < p(X) \coloneqq \Pr(D = 1|X) < 1$$

- Conditional expectation of observed outcome given X recovers conditional expectation of potential outcome given X

$$E\left[Y^{(d)}\middle|X\right] = E\left[Y^{(d)}\middle|D = d, X\right] = E\left[Y^{(D)}\middle|D = d, X\right] = E[Y|D = d, X]$$

Identification of Conditional Average Treatment Effect

- Under conditional ignorability, Conditional Average Predictive Effect

$$\pi(X) \coloneqq E[Y|D=1,X] - E[Y|D=0,X], \qquad (\text{CAPE})$$

- Is equal to the Conditional Average Treatment Effect

$$\delta(X) \coloneqq E\left[Y^{(1)}\middle|X\right] - E\left[Y^{(0)}\middle|X\right], \qquad (\text{CATE})$$

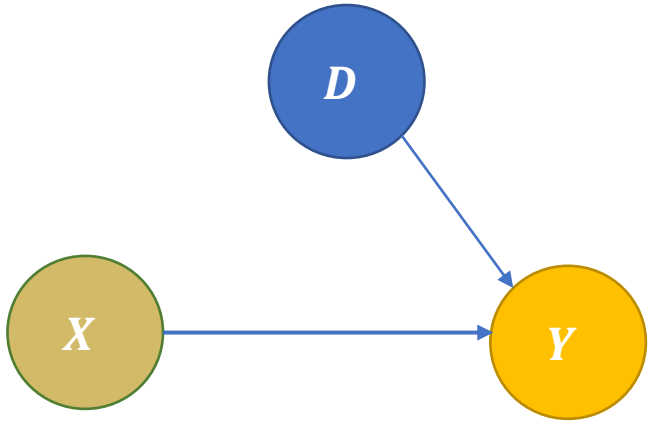
- Similarly for APE and ATE

$$\delta = E[\delta(X)] = E[\pi(X)] = \pi$$

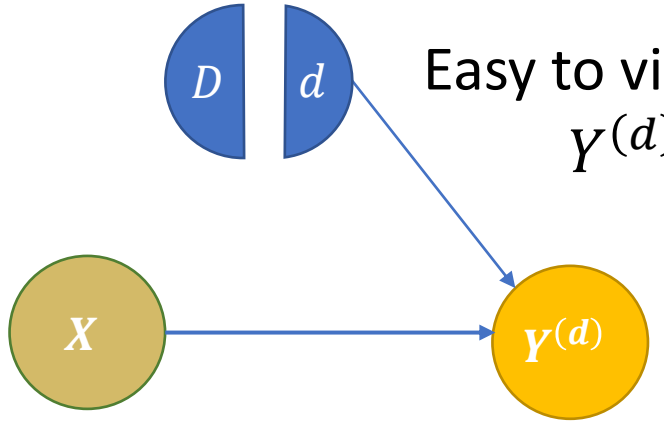
Causal Diagrams

RCTs and Causal Diagrams

- Causal diagrams can help visualize how our assumptions imply the identification of a causal effect
- First instances in work of Sewall and Philip Wright'28
- Pioneered and fully developed by Pearl and Robins [80s-90s]



Causal Diagram (Graph)

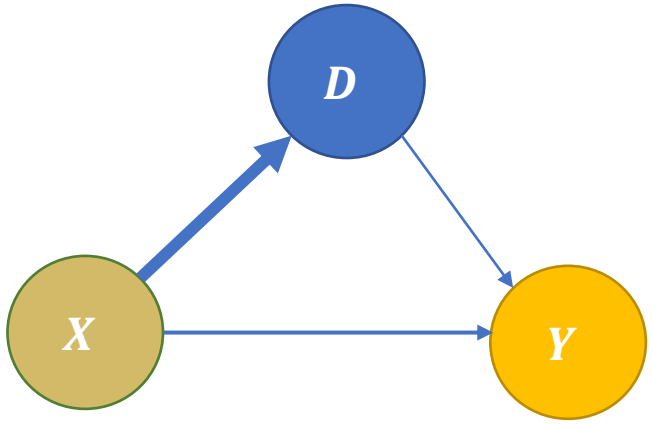


Easy to visualize that $Y^{(d)} \perp\!\!\!\perp D$

Single World Intervention Graph (SWIG)

Conditional Ignorability and Causal Diagrams

- Causal diagrams can help visualize how our assumptions imply the identification of a causal effect
- First instances in work of Sewall and Philip Wright'28
- Pioneered and fully developed by Pearl and Robins [80s-90s]

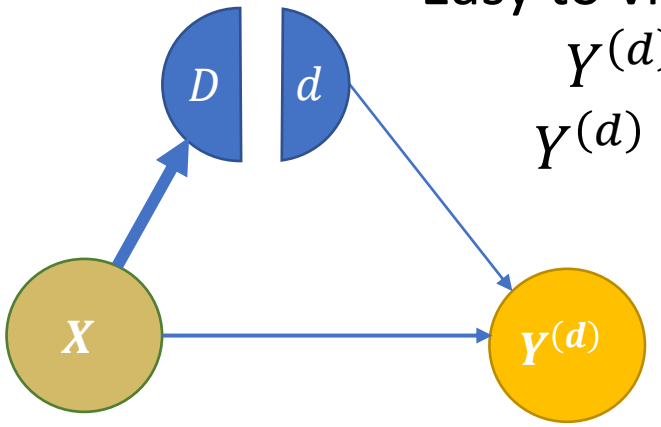


Causal Diagram (Graph)

Easy to visualize that

$Y^{(d)} \not\perp\!\!\!\perp D$

$Y^{(d)} \perp\!\!\!\perp D|X$



Single World Intervention Graph (SWIG)

Connection to Linear Regression

- If we further assume that CEF is linear in transformations W of raw X

$$E[Y|D,X] = \alpha D + \beta' W$$

- In other words, we assume the decomposition

$$Y = \alpha D + \beta' W + \epsilon, \qquad E[\epsilon|D,X] = 0$$

- Then note

$$\delta(X) = E\left[Y^{(1)} - Y^{(0)}\middle|X\right] = E[Y|D=1,X] - E[Y|D=0,X] = \alpha$$

- CATE is a constant and equal to the predictive effect of D
- Inference can be carried out via OLS or Double Lasso techniques

Connection to Linear Regression

- More reasonably we can relax and allow for effect heterogeneity
- Assume that CEF is linear in transformations W of raw X and interactions; with de-meaned W, i.e. $E[W] = 0$

$$E[Y|D, X] = \alpha_1 D + \alpha_2' WD + \beta' W + \beta_0$$

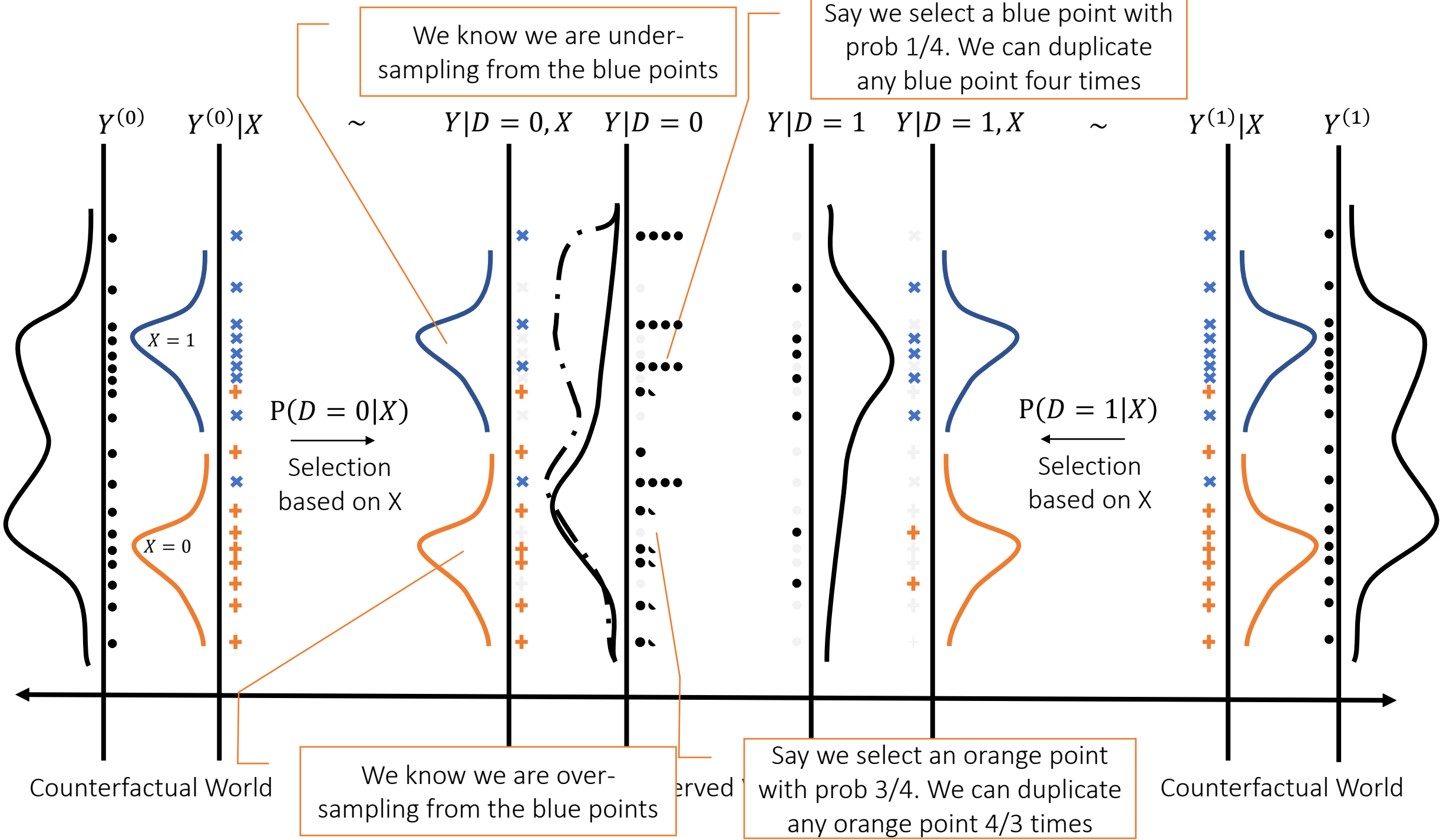
- Then note

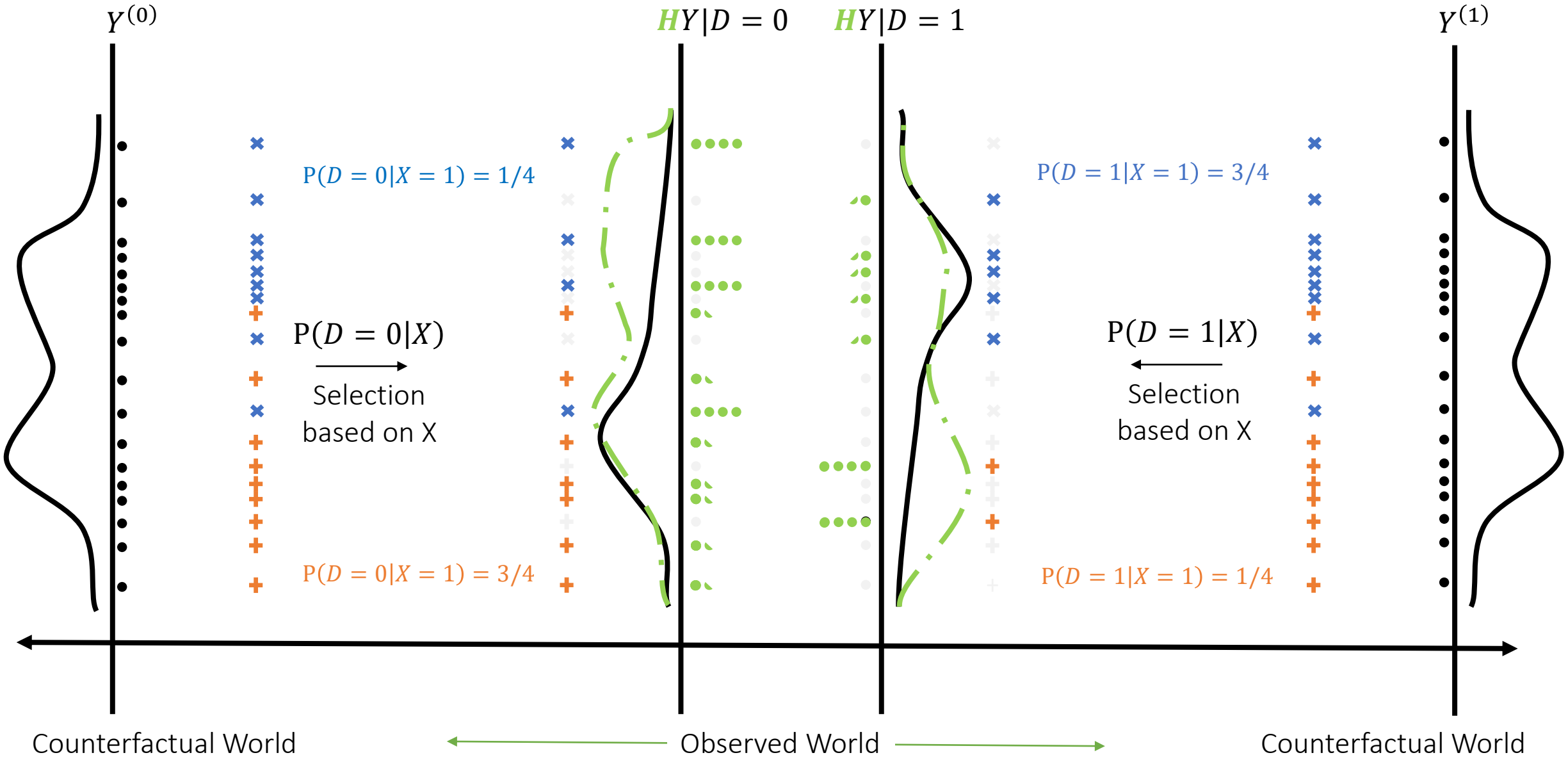
$$\delta(X) = E[Y|D = 1, X] - E[Y|D = 0, X] = \alpha_1 + \alpha_2' W$$

- And ATE also is recovered by: $\delta = \alpha_1$
- Inference on ATE and coefficients in CATE can be carried out via OLS or Double Lasso techniques

Identification via Propensity Scores

- The CAPE approach requires learning conditional expectation function $E[Y|D,X]$
- How outcome varies with treatment and observable characteristics
- In many settings we have more information about the selection process than the outcome process
- For instance, in stratified RCTs we know the propensity score $p(X)$
- In such cases, when we have a better grasp of the “selection process” can we avoid learning the “outcome process”; which could involve complex mechanisms in the real world


We know we are under-sampling from the blue points
Say we select a blue point with prob 1/4. We can duplicate any blue point four times
$Y^{(0)}$
$Y^{(0)}|X$
$\sim$
$Y|D=0,X$
$Y|D=0$
$Y|D=1$
$Y|D=1,X$
$\sim$
$Y^{(1)}|X$
$Y^{(1)}$
$X=1$
$X=0$
$\mathrm{P}(D=0|X)$
Selection based on X
$\mathrm{P}(D=1|X)$
Selection based on X
Counterfactual World
We know we are over-sampling from the blue points
Say we select an orange point with prob 3/4. We can duplicate any orange point 4/3 times
Counterfactual World


$Y^{(0)}$
$HY|D=0$
$HY|D=1$
$Y^{(1)}$
$P(D=0|X=1)=1/4$
$P(D=0|X)$
Selection based on X
$P(D=0|X=1)=3/4$
$P(D=1|X=1)=3/4$
$P(D=1|X)$
Selection based on X
$P(D=1|X=1)=1/4$
Counterfactual World
Observed World
Counterfactual World

Identification via Propensity Scores

- Re-weight observations based on the inverse of the propensity of their observed treatments and then take the difference in means
- Let W be inverse propensity weight we multiplied each observation by

$$\frac{E[1(D=d)\,W\,Y]}{E[1(D=d)W]} = \frac{E\left[\frac{1(D=d)}{\Pr(D=d|X)}Y\right]}{E\left[\frac{1(D=d)}{\Pr(D=d|X)}\right]} = E\left[\frac{1(D=d)}{\Pr(D=d|X)}Y\right]$$

Horvitz-Thompson Reweighting

- An inverse propensity re-weighted average observed outcome, identifies the average potential outcome

$$E\left[\frac{1(D=d)}{\Pr(D=d|X)}Y\right] = E[Y(d)]$$

- Same holds even conditioning on X

$$E\left[\frac{1(D=d)}{\Pr(D=d|X)}Y\middle| X\right] = E[Y(d) \mid X]$$

Horvitz-Thompson Reweighting

Tower Law. For any random variables Z, V, U
$E[Z|V] = E[E[Z|U,V]|V]$

- Simple proof: law of iterated expectations (tower law)

$$E\left[\frac{1(D=d)}{\Pr(D=d|X)}Y\middle|X\right] = E\left[\frac{1(D=d)}{\Pr(D=d|X)}E[Y|D,X]\middle|X\right]$$

$$= E\left[\frac{1(D=d)}{\Pr(D=d|X)}E[Y|D=d,X]\middle|X\right]$$

$$= E\left[\frac{1(D=d)}{\Pr(D=d|X)}\middle|X\right]E[Y|D=d,X]$$

$$= E[Y|D=d,X] = E\left[Y^{(d)}\middle|D=d,X\right] = E\left[Y^{(d)}\middle|X\right]$$

Horvitz-Thompson Reweighting

- Inverse propensity re-weighted average observed outcome, identifies the average treatment effect

$$\delta = E[H\,Y], \qquad H = \frac{1(D=1)}{\Pr(D=1|X)} - \frac{1(D=0)}{\Pr(D=0|X)}$$

- If we know the propensity $p(X)$ (stratified RCT), then we have an easy way to estimate the ATE (a simple average)
- However, not statistically efficient
- Ignores all extra information in X that can help explain Y
- IPW is similar qualitatively to two-means estimate; can have large variance because it does not remove the “explainable” variation in Y

Sensitive to Violations of Randomization

- Even if we know propensity, should perform co-variate balance checks
$$E[H|X] = 0$$
- Equivalently for any function f of X
$$E[H\ f(X)] = E[E[H|X]f(X)] = 0$$
- For any vector of transformations $\mathbf{W} = f(X)$, if we run a linear regression of $H \sim W$, then by BLP orthogonality
$$E[(H - \beta W)W] = 0 \Rightarrow \beta = E[WW']^{-1}E[HW] = 0$$
- Run linear regression predicting H from many transformations W of X and check if any coefficient is significant

Alternative: Conditioning on Propensity

- Rosenbaum and Rubin: instead of stratifying by X it suffices to stratify by $p(X)$

$$E[\,Y \mid D = d, p(X)\,] = E\left[Y^{(d)} \middle| p(X)\right]$$

- And therefore, average effect is identified as

$$\delta = E\Big[E[Y|D = 1, p(X)] - E[Y|D = 0, p(X)]\Big]$$

- If $p(X)$ is known and X is complex and high-dimensional, allows us to avoid the high-dimensional regression problem
- Suffices to run a (non-linear) regression on a single scalar co-variate to estimate $E[Y|D = d, p(X)]$ (e.g. run OLS on many engineered features of $p(X)$, or generic ML)

Alternative: Conditioning on Propensity

- Rosenbaum and Rubin: instead of stratifying by X it suffices to stratify by $p(X)$

$$E[\,Y \mid D = d, p(X)\,] = E\left[Y^{(d)} \middle| p(X)\right]$$

- Intuition: we can think of $D = 1\{U < p(X)\}$ with $U \perp\!\!\!\perp Y^{(d)}, X$; So D only correlates with $Y^{(d)}$ through $p(X)$
- Formally: by Horvitz-Thompson Theorem

$$E\left[Y^{(1)} \middle| p(X)\right] = E\left[Y \frac{1(D=1)}{p(X)} \middle| p(X)\right]$$

$$= E\left[E[Y|D=1, p(X)] \frac{1(D=1)}{p(X)} \middle| p(X)\right] = E[Y|D=1, p(X)]$$

Improving precision

- Extra co-variates W can easily be incorporated in the Rosenbaum-Rubin approach to increase precision

- Especially if we identify a W for which the co-variate balance check is violated, it is advisable to include it in the regression

- Run OLS for each treatment group, or equivalently interactive model

$$Y = \gamma_1' \phi\big(p(X)\big) D + \gamma_0' \phi\big(p(X)\big)(1 - D) + \beta' W + \epsilon$$

- Then take difference of average predictions of the model in treatment and control group

Clever Co-Variate Approach

- [Scharfstein-Rotnitzky-Robins] In fact it suffices to run a regression with the clever covariate H!
- Equivalently run an OLS

$$Y = \gamma\left(\frac{D}{p(X)} - \frac{(1-D)}{1-p(X)}\right) + \beta' W + \epsilon$$

- Even if the model is wrong, the BLP solution in the above decomposition will recover the correct ATE!

Clever Co-Variate Approach

- Let $H = \phi(D, X)$ and note that H guarantees for any f $(\mathbf{homework})$
$$E[f(D, X)H] = E[f(1, X) - f(0, X)]$$
- Then by the BLP orthogonality
$$E[\epsilon H] = 0 \Rightarrow E[YH] = E[\gamma\phi(D, X)\, H] = E\left[\gamma\left(\phi(1, X) - \phi(0, X)\right)\right]$$
- Thus, we have by the Horvitz-Thompson theorem:
$$\delta = E[YH] = E\left[\gamma\left(\phi(1, X) - \phi(0, X)\right)\right]$$
- Hence, if we use a BLP model as the CEF, we correctly recover the ATE
$$Y = \gamma\left(\frac{D}{p(X)} - \frac{(1 - D)}{1 - p(X)}\right) + \epsilon$$

Clever Target Outcome Approach

- We can also change our target outcome to be the re-weighted outcome
- Note that we showed that the CATE satisfies

$$\delta(X) = E[\, HY \mid X \,]$$

- So CATE can be thought as the solution to the prediction problem of predicting HY from X
- If we assume an interactive CEF model

$$E[Y|D, X] = \alpha_1 D + \alpha_2' WD + \beta' W + \beta_0$$

- Then note

$$E[HY|X] = \alpha_1 + \alpha_2' W$$

- OLS and Double Lasso can be used to perform inference on α_1 and α_2

Average Treatment Effect on the Treated ATT

- Many times we care about the effect for the people that actually received the treatment

$$ATT = E\left[Y^{(1)} - Y^{(0)} \middle| D = 1\right]$$

- Since we have observed data for one potential outcome, we can relax conditions

$$ATT = E[Y|D = 1] - E\left[Y^{(0)} \middle| D = 1\right]$$

- Conditional ignorability only for one potential outcome

$$Y^{(0)} \perp\!\!\!\perp D|X$$

- Weak overlap: $p(X) < 1$

Identification of ATT

- Under one-sided conditional ignorability and overlap

$$ATT = E[Y|D = 1] - E\Big[E[Y|D = 0, X]\Big]$$

$$\begin{aligned} E[Y(0)|D = 1] &= E[E[Y(0)|D = 1, X]|D = 1] \\ &= E\Big[E[Y(0)|D = 0, X]\Big] \\ &= E\Big[E[Y|D = 0, X]\Big] \end{aligned}$$

Identification of ATT

- Under one-sided conditional ignorability and overlap

$$ATT = E[Y|D = 1] - E\Big[E[Y|D = 0, X]\Big]$$

- We can also derive a Horvitz-Thompson style identification

$$ATT = E\big[Y\ \overline{H}\big], \qquad \overline{H} = H\,\frac{p(X)}{E[D]}$$