MS&E228: Lecture 2
Causality via Experiments

Vasilis Syrgkanis

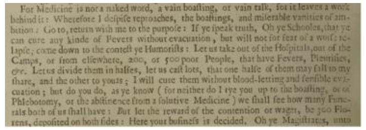
MS&E, Stanford

Randomized Control Trial (RCT)

- n units are drawn from a population
- Each unit is assigned at random with some probability p to one of two groups {control, treatment}
- Each unit in the control group receives treatment $\mathrm{D} = 0$
- Each unit in the treatment group receives treatment $\mathrm{D} = 1$
- At the end of the experiment an outcome y is measured for each unit

A Brief History of Experimentation

1662 — John Baptista Van Helmont proposes a randomized experiment for the effect of bloodletting



For Medicine is not a naked word, a vain boaſting, or vain talk, for it leaves a work behind it: Wherefore I deſpiſe reproaches, the boaſtings, and miſerable vanities of ambition: Go to, return with me to the purpoſe: If ye ſpeak truth, Oh ye Schooles, that ye can cure any kinde of Fevers without evacuation, but will not for fear of a worſe relapſe; come down to the conteſt ye Humoriſts: Let us take out of the Hoſpitals, out of the Camps, or from elſewhere, 200, or 500 poor People, that have Fevers, Pleuriſies, &c. Let us divide them in halfes, let us caſt lots, that one halfe of them may fall to my ſhare, and the other to yours; I will cure them without blood-letting and ſenſible evacuation; but do you do, as ye know (for neither do I tye you up to the boaſting, or of Phlebotomy, or the abſtinence from a ſolutive Medicine) we ſhall ſee how many Funerals both of us ſhall have: But let the reward of the contention or wager, be 300 Florens, depoſited on both ſides: Here your buſineſs is decided. Oh ye Magiſtrates, unto

1828 — Louis conducts a non-randomized controlled experiment showing strong evidence against bloodletting

1923 — Neyman publishes "On the Application of Probability Theory to Agricultural Experiments"

1925 — Fisher publishes Statistical Methods for Research Workers

1935 — Fisher publishes The Design of Experiments

1935 — Michigan Supreme Court authorizes RCTs as part of medical practice

1947 — First published medical RCT in Great Britain

1962 — FDA to prohibit the marketing of any new drug in the absence of "substantial evidence."

1969 — FDA: "…partially controlled studies are not acceptable evidence to support claims of effectiveness"

2000- — Abhijit Banerjee, Esther Duflo, and Michael Kremer, popularize the use of RCTs in economics (esp. for policy problems related to poverty).

2000- — 10s of thousands of RCTs run annually by companies like Airbnb, Amazon, Booking.com eBay, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yahoo! and Yandex

Credit: Ron Kohavi, [History of Controlled Experiments]

RCTs are the gold standard for measuring the “causal effect” of a “treatment” on an “outcome”

Goal #1: Mathematical Definition of Causality

- We want to formally (mathematically) define the causal effect of a binary treatment on a scalar outcome of interest

Examples

- Effect of seed A vs B on crop yield
- Effect of completion of a job training program on observed wage
- Effect of drug vs placebo on overall survival
- Effect of an ad impression on conversion (purchase)
- Effect of eating avocados on longevity

Causality via Potential Outcomes

- Nature generates two latent (unobserved) random outcomes $Y^{(1)}, Y^{(0)}$
- $Y^{(d)}$: potential outcome that would have been observed if unit received treatment $d \in \{0,1\}$

Example

- $Y^{(0)}$: wage if you don't participate in a training program
- $Y^{(1)}$: wage if you participate in a training program

Causality via Potential Outcomes

- $Y^{(0)}, Y^{(1)}$ are called “counterfactuals” as they can never be simultaneously observed
- Fundamental problem of causal inference

Example

- We don’t have two replicas of each unit
- We cannot observe your wage with the training program and without

Causality via Potential Outcomes

- The individual treatment effect is: $\mathrm{ITE} \coloneqq Y^{(1)} - Y^{(0)}$
- Un-attainable due to fundamental problem of causal inference
- Average Treatment Effect (ATE)

$$\delta \coloneqq E\left[Y^{(1)} - Y^{(0)}\right]$$

Example

- $ITE \coloneqq$ difference in wage of your potential outcome with the training program and without
- $ATE \coloneqq$ average difference in the population

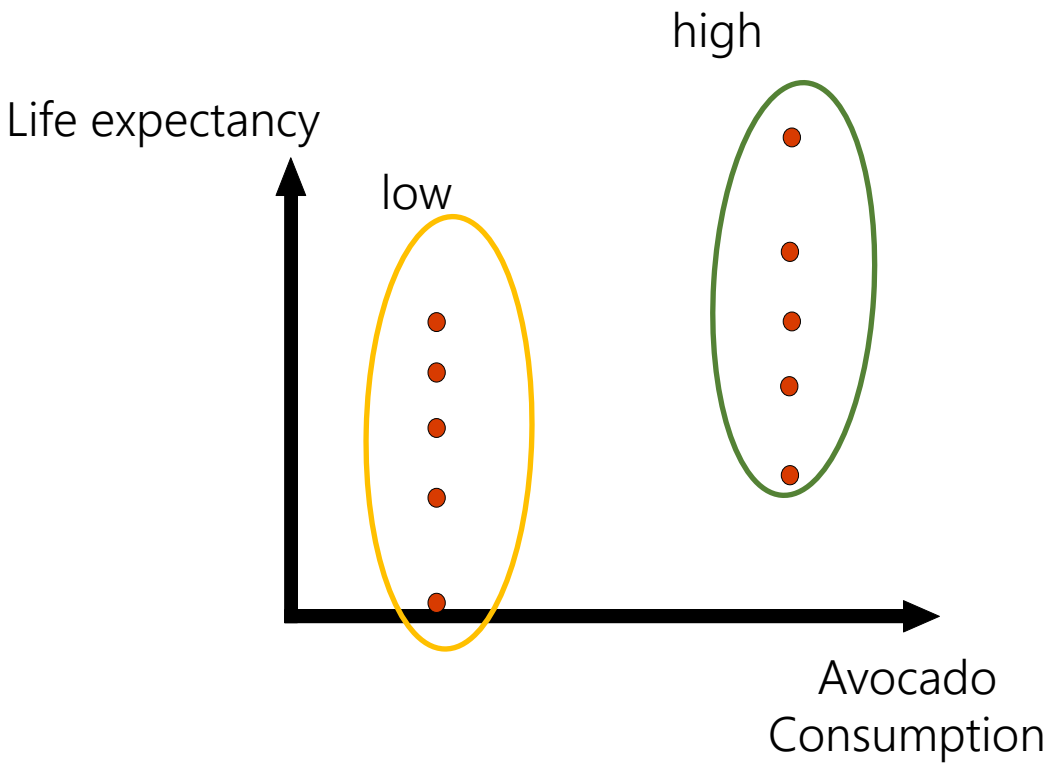
Treatment Assignment and Observed Outcome

- Each unit receives treatment $D \in \{0,1\}$, we observe
$$Y \equiv Y^{(D)}$$

- Given data (D, Y) what quantities can we "identify" $\equiv$ "measure if we had access to infinite data"

- Can we measure the ATE?

The Effect of Eating Avocados on Health



- Can identify average outcome within each treatment group

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

- When is this representative of average un-conditional counterfactual outcome

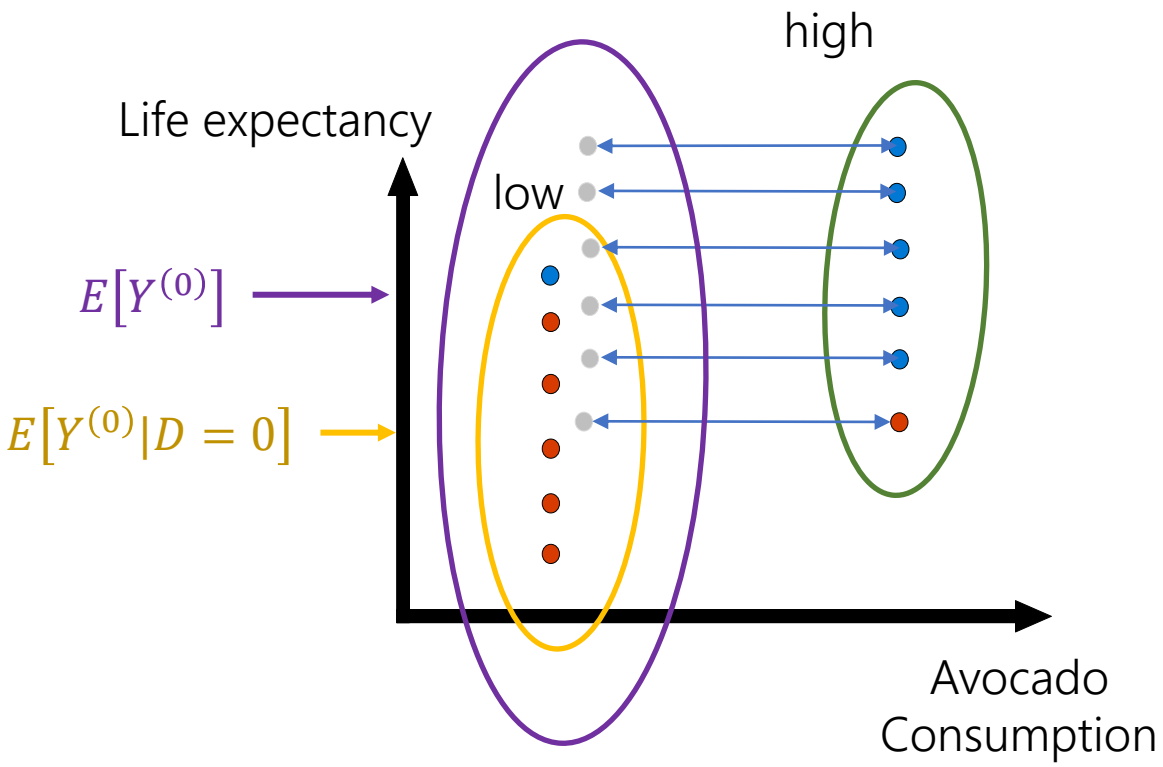
$$E\left[Y^{(d)}\right]$$

- In general

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

- Known as “selection bias”

The Effect of Eating Avocados on Health



- Can identify average outcome within each treatment group

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

- When is this representative of average un-conditional counterfactual outcome

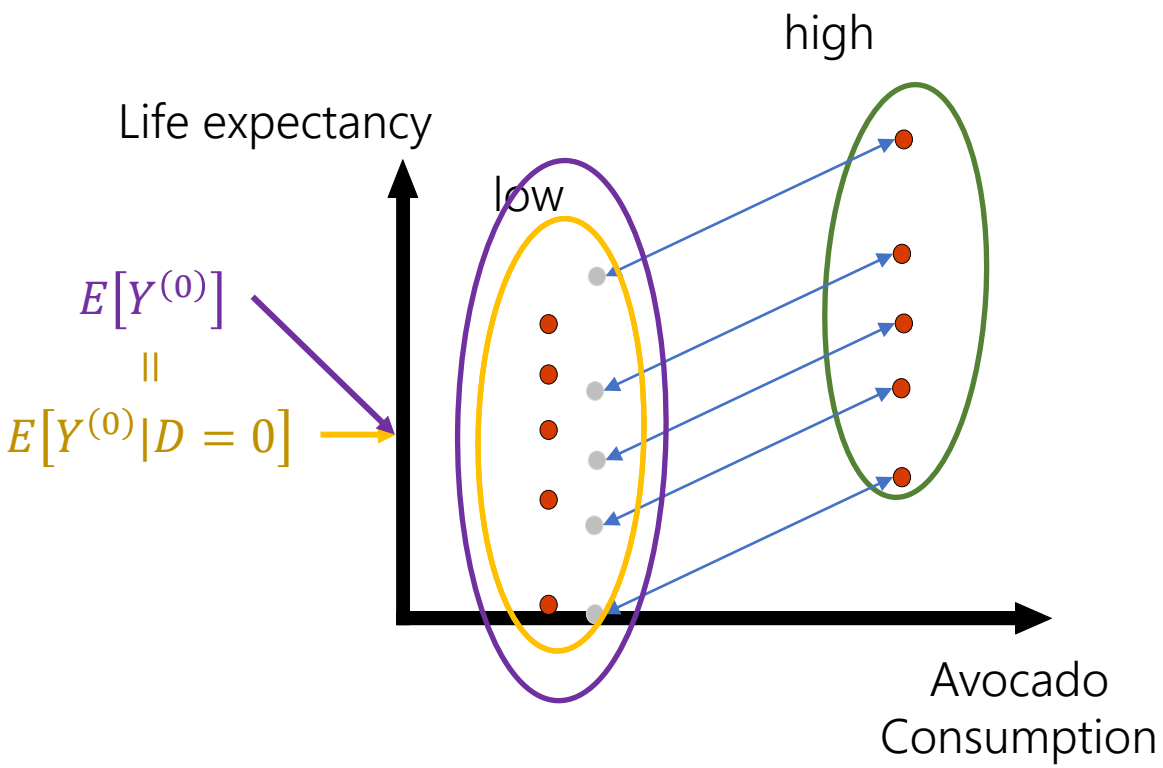
$$E\left[Y^{(d)}\right]$$

- In general

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

- Known as “selection bias”

The Effect of Eating Avocados on Health



- Can identify average outcome within each treatment group

$$E[Y|D=d] = E[Y^{(D)}|D=d] = E[Y^{(d)}|D=d]$$

- When is this representative of average un-conditional counterfactual outcome

$$E[Y^{(d)}]$$

- If treatment was fully randomly assigned (e.g. a randomized experiment)

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

- Then no selection bias:

$$E[Y^{(d)}|D=d] = E[Y^{(d)}]$$

Identification of ATE under Random Assignment

Suppose that treatment is randomly assigned (i.e. RCT)

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

and $0 < \Pr(D = 1) < 1$.

Then average **observed** outcome in treatment group $d \in \{0,1\}$ recovers average **potential** outcome for treatment d

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

Hence, average **predictive** effect recovers the average **treatment** effect

$$\pi := E[Y|D = 1] - E[Y|D = 0]$$
$$= E\left[Y^{(1)}\right] - E\left[Y^{(0)}\right] =: \delta$$

Statistical Inference

What can we do with finite samples

- Assume we have access to n samples (Y_i, D_i)
- Drawn i.i.d. from the distribution of the random variables (Y, D)
- Want estimate $\hat{\delta}$ of $\delta = E[Y|D=1] - E[Y|D=0]$ (identification)
- Denote with $E_n[\cdot]$ the empirical average $E_n[X] = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Estimate $\hat{\delta}$ by using empirical averages for each sub-population

$$\hat{\theta}_d := \frac{E_n[Y\ 1\{D=d\}]}{E_n[1\{D=d\}]} = \frac{1}{\#\{i: D_i = d\}} \sum_{i:D_i=d} Y_i$$

$$\hat{\delta} = \hat{\theta}_1 - \hat{\theta}_0$$

How accurate is the estimate?

Under mild regularity conditions

$$\sqrt{n}\{\hat{\theta}_d - \theta_d\}_{d\in\{0,1\}} \overset{a}{\sim} N(0, V)$$

where

$$V = \begin{pmatrix} \dfrac{\mathrm{Var}(Y|D=0)}{P(D=0)} & 0 \\ 0 & \dfrac{\mathrm{Var}(Y|D=1)}{P(D=1)} \end{pmatrix}$$

Hence

$$\sqrt{n}\left(\hat{\delta} - \delta\right) \overset{a}{\sim} N(0, V_{11} + V_{22})$$

$X \overset{a}{\sim} Y$ means that as $n \to \infty$:

$$\sup_{R\in\mathcal{R}} |P(X \in R) - P(Y \in R)| \approx 0$$

where $\mathcal{R}$ set of all hyper-rectangles

Proof Sketch

Trivially we can write $\theta_d = E\left[Y^{(d)}\right]\frac{E_n[1(D=d)]}{E_n[1(D=d)]}$

$$\hat{\theta}_d - \theta_d = \frac{E_n\left[Y^{(d)}\,1(D=d)\right]}{E_n[1(D=d)]} - \theta_d = \frac{E_n\left[\left(Y^{(d)} - E\left[Y^{(d)}\right]\right)1(D=d)\right]}{E_n[1(D=d)]}$$

By Law of Large Numbers (LLN)

$$\hat{\theta}_d - \theta_d \approx \frac{E_n\left[\left(Y^{(d)} - E\left[Y^{(d)}\right]\right)1(D=d)\right]}{P(D=d)}$$

Difference is average of the i.i.d. mean zero r.v.s: $\frac{\left(Y_i^{(d)} - E\left[Y^{(d)}\right]\right)1(D_i=d)}{P(D=d)}$

With zero covariance and variance: $\frac{E\left[\left(Y_i^{(d)} - E\left[Y^{(d)}\right]\right)^2 1(D_i=d)\right]}{P(D=d)^2} = \frac{Var(Y|D=d)}{P(D=d)}$

Statement follows by Central Limit Theorem (CLT)

Variance estimate

- Same statement also holds with consistent estimate of variance

$$\hat{V} = \begin{pmatrix} \frac{\mathrm{Var_n}(Y|D=0)}{E_n[1-D]} & 0 \\ 0 & \frac{\mathrm{Var_n}(Y|D=1)}{E_n[D]} \end{pmatrix}$$

Confidence Interval

- $X \overset{a}{\sim} Y \equiv \sup_{[\ell,u]} |P(X \in [\ell,u]) - P(Y \in [\ell,u])| \approx 0$

- If we consider $[\ell,u]$ the $\left(\frac{\alpha}{2}, 1-\frac{\alpha}{2}\right)$ quantile of $N\left(0,\hat{V}\right)$ then

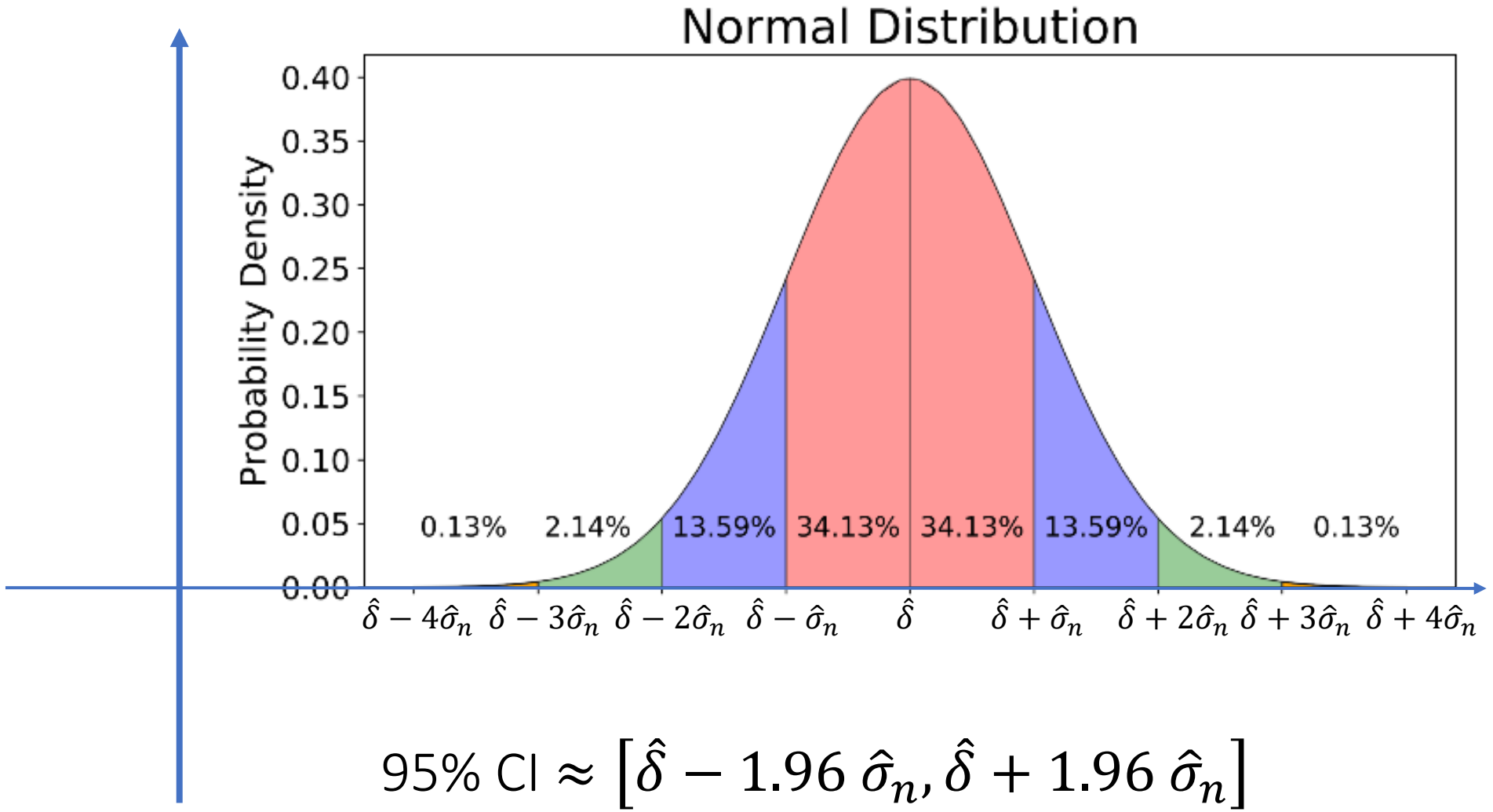
$$P\left(\sqrt{n}\left(\hat{\delta}-\delta\right) \in [\ell,u]\right) \approx 1-\alpha$$

Standard error

- Equivalently, let z_α the α quantile of $N(0,1)$ and $\hat{\sigma}_n = \sqrt{\hat{V}/n}$ then:

$$P\left(\delta \in \left[\hat{\delta} - z_{1-\frac{\alpha}{2}}\hat{\sigma}_n, \hat{\delta} + z_{1-\frac{\alpha}{2}}\hat{\sigma}_n\right]\right) \approx 1-\alpha$$

Confidence Interval



$$95\%\ \text{CI} \approx \left[\hat{\delta} - 1.96\ \hat{\sigma}_n, \hat{\delta} + 1.96\ \hat{\sigma}_n\right]$$

Let's try it out!

Relative effect

- Many times (e.g. vaccine trials) we are interested in relative effect

$$RE = \frac{E\left[Y^{(1)} - Y^{(0)}\right]}{E\left[Y^{(0)}\right]} = \frac{\theta_1 - \theta_0}{\theta_0}$$

- We can construct a plug-in estimate

$$\widehat{RE} = \frac{\hat{\theta}_1 - \hat{\theta}_0}{\hat{\theta}_0} = \frac{\hat{\theta}_1}{\hat{\theta}_0} - 1$$

- By delta method with $G = \left(-\theta_1/\theta_0^2, 1/\theta_0\right)$

$$\sqrt{n}\left(\widehat{RE} - RE\right) \overset{a}{\sim} N\left(0, \frac{\theta_1^2 V_{11}}{\theta_0^4} + \frac{V_{22}}{\theta_0^2}\right)$$

Delta method: for any function f, with $G = \nabla f(\theta)$

$$\sqrt{n}\left(f\left(\hat{\theta}\right) - f(\theta)\right) \approx G\sqrt{n}\left(\hat{\theta} - \theta\right) \overset{a}{\sim} N(0, G'VG)$$

Example: Pfizer Vaccine

Efficacy Endpoint Subgroup	BNT162b2 N[a]=19965 Cases n1[b] Surveillance Time[c] (n2[d])	Placebo N[a]=20172 Cases n1[b] Surveillance Time[c] (n2[d])	Vaccine Efficacy % (95% CI)[e]
Overall	9 2.332 (18559)	169 2.345 (18708)	94.6 (89.6, 97.6)
Age group (years)			
16 to 17	0 0.003 (58)	1 0.003 (61)	100.0 (-3969.9, 100.0)
18 to 64	8 1.799 (14443)	149 1.811 (14566)	94.6 (89.1, 97.7)
65 to 74	1 0.424 (3239)	14 0.423 (3255)	92.9 (53.2, 99.8)
≥75	0 0.106 (805)	5 0.109 (812)	100.0 (-12.1, 100.0)

Approximate Confidence Interval

- Outcomes are binary $y \in \{0,1\}$
- Distribution of outcome for each d is Bernoulli with success p_d
- For each subpopulation, mean outcome $\hat{p}_d = \frac{\text{Cases}_d}{N_d}$ is estimate of p_d

$$\text{Vaccine Efficacy } (VE) = -\widehat{RE} = \frac{\hat{p}_0 - \hat{p}_1}{\hat{p}_0}$$

- An estimate of the variance of y is $\hat{p}_d(1 - \hat{p}_d)$
- 95% confidence interval can be derived by delta method
- Alternative: approximate bootstrap; simulate $\hat{p}_d + N\left(0, \hat{V}_d/n\right)$

Let's try it out!

Pre-Treatment Covariates

Precision and Heterogeneity

Pre-Treatment Covariates

- Assume we have covariates W that correspond to variables determined prior to treatment assignment (e.g. age, income groups)
- How can we use them?

- Heterogeneity: how does the effect vary with these covariates

- Formalized by the Conditional Average Treatment Effect (CATE)

$$\delta(W) \coloneqq E\left[Y^{(1)} - Y^{(0)} \middle| W\right]$$

Identification of CATE under Random Assignment

Suppose that treatment is randomly assigned (i.e. RCT)

$$\left(Y^{(d)}, W\right) \perp\!\!\!\perp D$$

and $0 < \Pr(D = 1) < 1.$

Then conditional **observed** outcome in treatment group $d \in \{0,1\}$ recovers conditional **potential** outcome for treatment d

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

Hence, conditional **predictive** effect recovers the CATE

$$\begin{aligned}\pi(W) &:= E[Y|D = 1, W] - E[Y|D = 0, W] \\ &= E\left[Y^{(1)}|W\right] - E\left[Y^{(0)}|W\right] =: \delta(W)\end{aligned}$$

If we only care about ATE are co-variates useful?

Co-variates for Sanity Check

- Since treatment is supposed to be independent of co-variates W

$$W|D = 1 \sim W|D = 0$$

- For instance, $E[W|D = 1] = E[W|D = 0] = E[W]$
- D does not predict any covariate
- Equivalently

$$D|W \sim D$$

- D is not predictable by any covariate
- We can test all these conditions on the samples to uncover violations of random assignment
- These are typically referred to as co-variate balance tests

Co-variates for Precision

- Even if we are only interested on ATE covariates can be valuable for precision
- Suppose variance of y is large but can be explained largely by W
- Then we can use W to remove all the explained variation from y
- Then perform our ATE analysis on the remnant variation
- This is oftentimes performed in practice via ordinary linear regression of y on the vector $(1, D, W)$ (after centering W, i.e. $E[W] = 0$)

Is this consistent?

- Suppose that the conditional expectation function (CEF) of the outcome is indeed linear, with $(D, 1, W)$

$$E[\, Y \mid D, W \,] = D\alpha + \alpha_0 + W'\beta$$

- Then note that

$$E[Y(0)] = E\big[E[Y|D = 0, W]\big] = \alpha_0$$
$$E[Y(1)] = E\big[E[Y|D = 1, W]\big] = \alpha + \alpha_0$$

- Baseline outcome is coefficient associated with the intercept 1
- Average effect is coefficient associated with treatment D
- Next lecture: this does not require the linear CEF assumption

Analysis of Variance (ANOVA)

Sneak peek into some material from next lecture

Variance of Estimate

- The OLS theory that we will cover in the next section yields

$$\sqrt{n}(\hat{\alpha} - \alpha) \overset{\text{a}}{\sim} N(0, V_\alpha) \qquad V_\alpha = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]^2}$$

- ϵ is residual outcome:

$$\epsilon = y - D\alpha - \alpha_0 - W'\beta \qquad E[\epsilon | D, W] = 0 \text{ (by Linear CEF)}$$

- $\widetilde{D}$ is residual treatment (removing whatever is linearly predictable from $(1, W)$)

$$\widetilde{D} = D - E[D]$$

Variance without adjustment

- OLS theory that we will cover in the next section yields

$$V_\alpha = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]^2}$$

- ϵ is residual outcome: $\epsilon = y - D\alpha - \alpha_0 - W'\beta$ with $E[\epsilon|D, W] = 0$
- Two means estimate is equivalent to OLS without W. OLS theory gives variance V_α of same form but with residual:

$$\bar{\epsilon} = y - D\alpha - \alpha_0 = W'\beta + \epsilon$$

$$\begin{aligned} E\left[\bar{\epsilon}^2 \widetilde{D}^2\right] &= E\left[(W'\beta + \epsilon)^2 \widetilde{D}^2\right] \\ &= E\left[(W'\beta)^2 \widetilde{D}^2\right] + E\left[\epsilon^2 \widetilde{D}^2\right] + 2E\left[\beta' W \epsilon \widetilde{D}^2\right] \\ &= E\left[(W'\beta)^2 \widetilde{D}^2\right] + E\left[\epsilon^2 \widetilde{D}^2\right] + 2E\left[\beta' W E[\,\epsilon \mid D, X\,] \widetilde{D}^2\right] \\ &= E\left[(W'\beta)^2 \widetilde{D}^2\right] + E\left[\epsilon^2 \widetilde{D}^2\right] \end{aligned}$$

$$\bar{V}_\alpha \geq V_\alpha$$

Variance of OLS estimate with extra co-variates (adjusted) is weakly smaller than two-means estimate (un-adjusted)

Heteroskedasticity Robust Variance

- Variance formula

$$V_\alpha = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]^2}$$

- is valid even when the linear CEF assumption is violated
- Inference is asymptotically valid!
- Important to note that this formula is known as the “heteroskedasticity robust variance formula” (HC0)
- Many software packages make the simplification that the residual ϵ is independent of D, W, leading to $V_\alpha = E\left[\epsilon^2\right]/E\left[\widetilde{D}^2\right]$. This is incorrect in most cases!

Precision Beyond Linear CEF

- The precision statement invoked the property that the residual of the OLS regression of y on D, X is mean zero conditional on D, X
- If linear CEF is violated, then all we know is the orthogonality property

$$E\left[\epsilon \begin{pmatrix} D \\ 1 \\ W \end{pmatrix}\right] = 0 \quad \left[\text{FOC of: } \min_{\alpha,\alpha_0,\beta} E\left[(y - D\alpha - \alpha_0 - W'\beta)^2\right]\right]$$

- This is not sufficient to argue that the cross-term vanishes

$$E\left[\beta' W \epsilon \widetilde{D}^2\right] = E\left[\beta' E[\, W\epsilon \mid D\,]\widetilde{D}^2\right]$$

- Note that we only need that:

$$E[\, W\epsilon \mid D\,] = 0$$

Let's try it out!

OLS with Interactive Terms

- It is advisable that instead of running OLS of y on $D, 1, W$ we also include interaction terms, i.e. y on $D, 1, W, DW$ [Lin'13]
- In the absence of any model assumptions, the coefficient of D and of the intercept, recover the ATE and the mean baseline outcome
- These interactive terms enforce the residual ϵ of OLS to satisfy the stronger orthogonality property with $X = (1, W)$

$$E\left[\epsilon \begin{pmatrix} X \\ DX \end{pmatrix}\right] = 0$$

- $E[\epsilon DX] = 0 \Rightarrow E[\,\epsilon X \mid D = 1\,] = 0 \Rightarrow E[\,\epsilon X \mid D = 0\,] = 0$
- Interaction term in $\bar{V}_\alpha$ is zero and we get that $\bar{V}_\alpha \geq V_\alpha$ without assumptions
- OLS with interactive terms always has weakly smaller variance than two means estimate!

Let's try it out!

Example: Heterogeneous Effects

- Suppose that:

ATE

effect modifier

$$E[\,y \mid D, X\,] = D\alpha + \alpha_0 + DW'\gamma + W'\beta$$

- What OLS is estimating is the solution to:

$$E\left[\left(y - D\tilde{\alpha} - \tilde{\alpha}_0 - W'\tilde{\beta}\right)\begin{pmatrix} D \\ X \end{pmatrix}\right] = 0$$

$$E\left[\left(D\alpha + \alpha_0 + DW'\gamma + W'\beta - D\tilde{\alpha} - \tilde{\alpha}_0 - W'\tilde{\beta}\right)\begin{pmatrix} D \\ X \end{pmatrix}\right] = 0$$

- Since $E[W] = 0$ and $W \perp D$: $\alpha = \tilde{\alpha}, \alpha_0 = \tilde{\alpha}_0$
- But $\tilde{\beta} = \beta + E[D]\gamma$
- Residual of OLS is $\epsilon = \widetilde{D}W'\gamma + \nu$ with $E[\nu|D, W] = 0$
- Then interaction term is:

$$\mathrm{E}\left[\beta' W \epsilon \widetilde{D}^2\right] = E\left[\widetilde{D}^3\right]\beta' E[WW']\gamma \neq 0$$

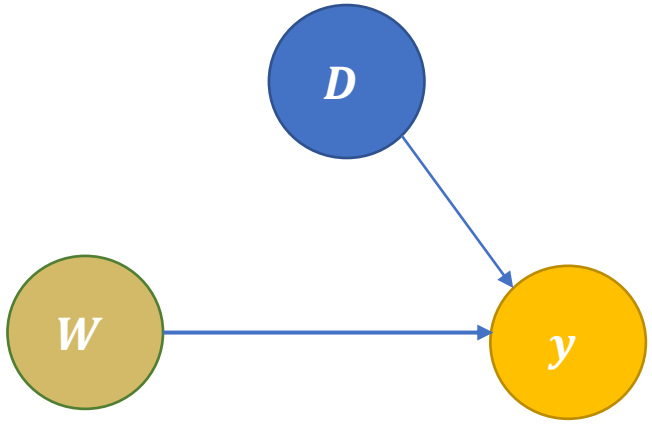
Even if you only care about ATE, if you have p covariates and $p \ll n$ run OLS with interactive terms!

Guaranteed improved precision, plus can uncover potential dimensions of heterogeneity

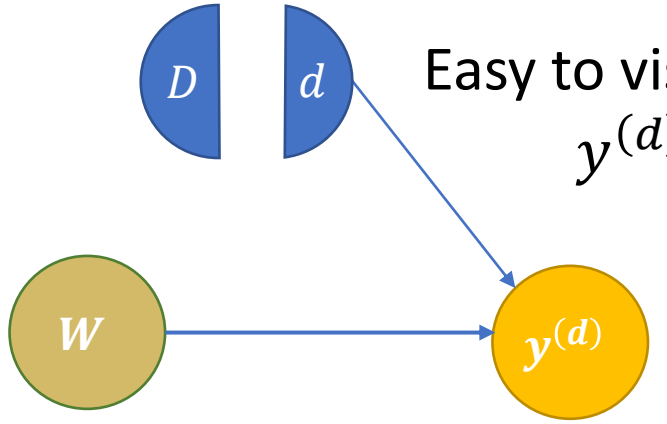
RCTs and Causal Diagrams

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 $y^{(d)} \perp D$

Single World Intervention Graph (SWIG)

Limitations of RCTs

Externalities, Stability and Equilibrium Effects

- Stable Unit Treatment Value Assumption (SUTVA)
- Implicit in our notation the potential outcome of a unit depends only on its own treatment
- This can be violated due to what is known as spillover effects, externalities or general equilibrium effects
- In vaccine example: if large fraction of population is treated, then the effect of vaccinating an additional unity changes, due to herd immunity
- In labor: if we make an intervention that incentivizes a large fraction of population to attend college, the college-wage premium will likely decrease

Ethical, Practical and Generalizability Concerns

- **Ethical Concerns:** Many potential trial would correctly be judged unethical by a human subject trial review board; Key principles [78 Belmont report]: (i) respect for persons, (ii) beneficence, (iii) justice
- **Practical Concerns:** Practically infeasible due to high cost (e.g. expensive treatment, expensive data collection, low signal regime, long-term outcomes)
- **Generalizability:** Local population used for RCT might not generalize to broader population