Correlation or causation?

Clement de Chaisemartin, UCSB

Roadmap

1. Defining what we are looking for: potential outcomes and treatment effects.
2. Omitted variable bias.
3. No omitted variable bias in Randomized Controlled Trials.
4. Statistical tests in Randomized Controlled Trials.
5. Application: the effect of health insurance.
6. Other methods to measure causal effects than RCTs.
7. The impact evaluation industry

The effect of a treatment on an outcome.

- In these lectures, we assume we consider a population of n people.
- We want to know the effect of a **treatment variable** D_i on an **outcome variable** Y_i**.**
- E.g.: what is the effect of having access to health insurance (D_i) on health (Y_i)?
- E.g.: what is the effect of number of years of schooling (D_i) on wages (Y_i)?
- Often, the treatments we are interested in are binary variables (having / not having health insurance), but sometimes they are not (years of schooling).
- In this section of the lectures, we focus on the case where the treatment is binary.
- First, we are going to define precisely what we would like to measure, in the context of an example.

Your mission

- Measure if 1st grade students in Kenya perform better in tracked schools than in non-tracked schools.
- You will consider schools with 2 1st grade classes, classes A and B.
- Students take a test at the start of the year.
- Tracked schools: students below the school median in that test sent to classroom A, students above the median sent to classroom B. Classroom A weaker, classroom B stronger.
- Non-tracked schools: classrooms A and B have the same numbers of strong and weak students.
- **Treatment D_i: binary variable equal to 1 if student i is in tracked school. Outcome Y_i: student i's test score in end of first grade.**

Big picture economics question

- How are education and learning produced?
- One key input to the education production function=teacher.
- Another key input: classmates.
  - Maybe strong students help weaker ones: peer effects.
  - Maybe weaker students do not dare to ask questions because feel they would slow down class too much: censoring effect.
- Tracking:
  - Probably increases teachers' effectiveness: easier to teach a homogeneous class than a very heterogeneous one.
  - Probably reduces peer effects: weak students are all in the same classroom, and not strong student in that classroom => do not benefit from help of strong students.
  - Probably reduces censoring effect: now weak students less shy to ask questions when not too strong students in their class.
- Effect of tracking on test scores? Unclear, depends on relative strength of those effects. => we need to study this question empirically.

Potential outcomes...

- We now introduce the concept of potential outcomes, to define precisely the effect of tracking on test scores.
- You want to measure the effect of studying in a tracked versus a non-tracked school on the test score that Sharon, a Kenyan 1st grade student, will achieve in the end of 1st grade.
- To measure that effect, you need to measure:
  - the 1st grade test score Sharon will obtain if she studies in a tracked school: $y_{\text{Sharon}}(1)$.
  - the 1st grade test score Sharon will obtain if she studies in a non-tracked school: $y_{\text{Sharon}}(0)$.
- $y_{\text{Sharon}}(1)$ and $y_{\text{Sharon}}(0)$ are the two **potential** test scores that Sharon will obtain if she studies in a tracked school and if she studies in a non-tracked school.

… And treatment effects

- $y_{\text{Sharon}}(1)$ and $y_{\text{Sharon}}(0)$ are the two **potential** test scores that Sharon will obtain if she studies in a tracked school and if she studies in a non-tracked school.
- The effect of studying in a tracked school on Sharon's end of 1st grade test score is the difference between these two potential outcomes: $y_{\text{Sharon}}(1) - y_{\text{Sharon}}(0)$.
- We compare the performance of the same person, Sharon, with and without tracking, to isolate the effect of tracking.
- Cetebus paribus analysis (everything else equal).
- We say that tracking is the **treatment** we are interested in. That's because impact evaluation literature comes from medicine, were they study effects of medical treatments.

Sharon slightly benefits from tracking

- Assume that $y_{\text{Sharon}}(0) = 0.85$ and $y_{\text{Sharon}}(1) = 0.90$.
- $y_{\text{Sharon}}(0) = 0.85$ means that Sharon will get 85% of the correct answers in the end of 1st year test if she studies in non-tracked school.
- $y_{\text{Sharon}}(1) = 0.90$ means that Sharon will get 90% of the correct answers in the end of 1st year test if she studies in a tracked school.
- The effect of studying in a tracked school on Sharon's test score is $y_{\text{Sharon}}(1) - y_{\text{Sharon}}(0) = 0.90 - 0.85 = 0.05$.
- Tracking increases her proportion of correct answers in the end of first year test by 0.05.
- Small effect. Sharon = strong student. Does well irrespective of tracking. With tracking, teacher can challenge her a bit more (only strong students in her class) => improves a bit but not huge difference.
- Is it plausible to assume that for all students, the effect of tracking is to increase their % of correct answers by 0.05? Discuss this question with your neighbor for 2mns.

iClicker time

- Is it plausible to assume that for all students, the effect of tracking is to increase their % of correct answers by 0.05, like for Sharon?
- A) Yes, this is plausible.
- B) No, this is not plausible.

Mercy strongly benefits from tracking

- Now consider another student, Mercy.
- Assume $y_{\text{Mercy}}(0) = 0.30$ and $y_{\text{Mercy}}(1) = 0.60$.
- Mercy will get 30% of the correct answers in the end of 1st year test if she studies in a non-tracked school, and will get 60% of correct answers if she studies in a tracked school.
- Mercy is a weaker student than Sharon. In a non-tracked school, lags behind her peers, cannot keep up with pace of instruction. Does not dare to tell teacher when does not understand.
- In tracked school, Mercy feels more comfortable to ask questions, and the teacher can tailor the instruction to meet the needs of weaker students.
- Therefore, Mercy would strongly benefit from tracking: $y_{\text{Mercy}}(1) - y_{\text{Mercy}}(0) = 0.60 - 0.30 = 0.30$: tracking increases her proportion of correct answers by 0.30.

Can we compute the effect of tracking on Sharon and Mercy?

- In this hypothetical example, we have assumed we know the potential test scores of Sharon and Mercy, without and with tracking.
- In practice, can you observe at the same time $y_{\text{Sharon}}(0)$ and $y_{\text{Sharon}}(1)$? Therefore, can you compute $y_{\text{Sharon}}(1) - y_{\text{Sharon}}(0)$? Discuss this question with your neighbor for 1mn.

iClicker time

- In practice, can you observe at the same time $y_{\text{Sharon}}(0)$ and $y_{\text{Sharon}}(1)$? Therefore, can you compute $y_{\text{Sharon}}(1) - y_{\text{Sharon}}(0)$?
- A) Yes
- B) No

Fundamental problem of causal inference

- Sharon has two potential end of 1st grade test scores:
  - $y_{\text{Sharon}}(0)$: score if studies in non-tracked school.
  - $y_{\text{Sharon}}(1)$: score she obtains if studies in tracked school.
- Can you observe both $y_{\text{Sharon}}(0)$ and $y_{\text{Sharon}}(1)$?
- No! Either Sharon goes to a non-tracked school, in which case her end of first grade test score Y_i is $y_{\text{Sharon}}(0)$…
- … Or Sharon goes to a tracked school, in which case her end of first grade test score Y_i is $y_{\text{Sharon}}(1)$.
- Sharon cannot go both to tracked and non-tracked school, so we cannot observe both what happens to her if she goes to tracked school, and what happens to her if she goes to non-tracked school.
- **We have $\boldsymbol{Y_i = D_i y_i(1) + (1 - D_i) y_i(0)}$:**
  - for students that go to tracked schools (D_i=1), the test score that we observe for them in the end of the year is $y_i(1)$
  - for students that do not go to tracked schools (D_i=0), the test score that we observe for them in the end of the year is $y_i(0)$.
- **Fundamental problem of causal inference:** we can never observe the same person with and without the treatment. **Therefore, we can never compute the effect of a treatment on a specific person.**

Frost's "Road not taken": poetic version of fundamental problem of causal inference.

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;
Then took the other, as just as fair,
[…]
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

Frost cannot know both his potential life where he travels road A, and his potential life where he travels road B.

Here, he claims that travelling road A rather than B made his life better. Actually he cannot know.

Let's try to learn average effect of treatment among students that go to tracked schools.

- n: number of 1st grade students in Kenya. $n_1 < n$: number of students that go to tracked schools. $n - n_1$: number of students in non-tracked schools.
- For every i included between 1 and n:
  - D_i equal to 1 if student i goes to a tracked school, 0 otherwise.
  - $y_i(0)$: potential end of first grade test score of student i if goes to a non-tracked school, $y_i(1)$: her test score if goes to a tracked school.
- Average effect of going to tracked school on scores of students that go a tracked school is

$$\frac{1}{n_1} \sum_{i:D_i=1} \big(y_i(1) - y_i(0)\big)$$

Average treatment effect on the treated (ATT).

- P3Sum: $ATT = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$.
- **In these lectures, the parameter we will try to learn is ATT.**

ATT = average of $y_i(1)$ - average of $y_i(0)$ for students who go to tracked schools.

- $ATT = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$
- ATT: average of $y_i(1)$ for students who go in tracked schools, minus average of $y_i(0)$ for the same students.
- If we have a data set where we observe Y_i, the end of 1st grade test scores of all Kenyan students, we can compute one these two averages, not the other one. Which is the one we can compute, which is the one we cannot compute? Discuss this question with your neighbor for 2 minutes.

iClicker time

- $ATT = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$
- If we have a data set where we observe Y_i, the end of 1st grade test scores of all Kenyan students, we can compute one these two averages, not the other one. Which is the one we can compute, which is the one we cannot compute?
- A) We can compute $\frac{1}{n_1}\sum_{i:D_i=1} y_i(1)$ but not $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$
- B) We can compute $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$ but not $\frac{1}{n_1}\sum_{i:D_i=1} y_i(1)$

ATT = average $y_i(1)$ - average $y_i(0)$ for students in tracked schools, but we only observe average $y_i(1)$.

- $ATT = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$
- If we have data set with end of 1st grade test scores of all Kenyan students, we can compute 1 of these 2 averages, not the other. Which is the one we can compute, which is the one we cannot compute?
- We can compute $\frac{1}{n_1}\sum_{i:D_i=1} y_i(1)$: for students that go to tracked schools, $Y_i = y_i(1)$: the test score they obtain in the end of 1st grade is $y_i(1)$, their test score when they go to a tracked school. Average test score of those n_1 students gives us $\frac{1}{n_1}\sum_{i:D_i=1} y_i(1)$.
- We cannot compute $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$: average test scores that the students who went to tracked schools would have obtained if had gone to non-tracked schools.
- **We need to find an estimator of $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$.**

What you need to remember

- We want to know effect of binary **treatment** D_i on **outcome** Y_i of person i.
- E.g.: we want to measure effect of tracking on the test score of person i.
- Measuring treatment effect on outcome of unit i requires measuring the **potential outcomes** of that person with and without the treatment, $y_i(1)$ and $y_i(0)$, to be able to compute $y_i(1) - y_i(0)$, **effect of treatment** on outcome of person i.
- We cannot do this, **fundamental problem of causal inference**.
- **We have $\boldsymbol{Y_i = D_i y_i(1) + (1 - D_i) y_i(0)}$:**
  - Either person i receives treatment, and then we only observe her potential outcome with treatment $y_i(1)$.
  - Or person does not receive treatment, and then we only observe her potential outcome without treatment $y_i(0)$.
- **We cannot measure the effect of the treatment on a specific individual.**
- But maybe we can measure **average effect of the treatment on the treated,** $\text{ATT} = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$.
- We can compute $\frac{1}{n_1}\sum_{i:D_i=1} y_i(1)$: average Y_i of the treated people.
- We cannot compute $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$: average outcome without treatment of the treated people. **Unobserved.**
- **We need to find an estimator of $\boldsymbol{\frac{1}{n_1}\sum_{i:D_i=1} y_i(0)}$.**

Roadmap

1. Defining what we are looking for: potential outcomes and treatment effects.
2. Omitted variable bias.
3. No omitted variable bias in Randomized Controlled Trials.
4. Statistical tests in Randomized Controlled Trials.
5. Application: the effect of health insurance.
6. Other methods to measure causal effects than RCTs.
7. The impact evaluation industry

The effect of health insurance on health.

- In this section, we use other example to illustrate discussion: what is effect of having health insurance on health, and health expenditures?
- Treatment D_i: having access to health insurance, outcome Y_i: health, health expenditures.
- Background: US is one of the only advanced economies where large fraction of population still does not have health insurance. In Western Europe, free health insurance for almost everyone.
- Economists disagree on consequences of granting health insurance.
- Insurance reduces the cost of healthcare for patients => should increase expenditures if demand elastic. But maybe then people can pay for preventive care (e.g. get their cholesterol checked) and then avoid diseases with very large costs (e.g. heart attacks) => maybe actually insurance reduces health expenditures.
- Then, maybe increase in health expenditures => you get access to better care => better health. Or maybe insurance => you know you will not need to pay for your health care => you start adopting risky health behaviors (smoking, drinking): moral hazard => your health deteriorates.
- Conflicting economic theories give opposite answers to same question: we need to study this question empirically.
- Aside. Health expenditures = 10k USD per person and per year in the US, against 5k in Western Europe, but life expectancy much lower in the US. Life expectancy in the US in 2016: 79.3 years old, same as Cuba, below Chile or Greece, countries with much lower income / head.

What if we ran an OLS regression of health on whether people are insured?

- 2015 National Health Interview Survey (NHIS): representative survey of American population.
- Respondents asked to rate overall health on scale from 1 to 5.
- Let Y_i be health of respondent i in 2015 and let D_i be binary variable equal to 1 if i had health insurance in 2015 and 0 if did not.
- We could run an OLS regression of Y_i on a constant and D_i, and use $\hat{\beta}_1$, the coefficient of health insurance variable as our measure of the effect of health insurance on health.
- $\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i-\bar{D})(Y_i-\bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(D_i-\bar{D})^2}$. Numerator: covariance between insurance and health, measures whether health and insurance move together or in opposite directions => maybe $\hat{\beta}_1$ good measure of effect of insurance on health.
- We run regression and find $\hat{\beta}_1 = 0.31$. Also, $\hat{\beta}_1$ statistically significant at 5% level.
- Can we conclude from the fact that $\hat{\beta}_1$ is positive and significant that being insured improves health?

iClicker time

- Using 2015 NHIS data, we run an OLS regression of Y_i (2015 health) on a constant and D_i (insurance in 2015), and find $\hat{\beta}_1 = 0.31$. Also, $\hat{\beta}_1$ statistically significant at 5% level. Can we conclude from this that being insured improves health?
- A) Yes
- B) No

No, due to omitted variable bias.

- $y_i(0)$: health of respondent i in 2015 if uninsured in 2015.
- $y_i(1)$: health of respondent i in 2015 if insured in 2015.
- To simplify, we momentarily assume that $y_i(1) - y_i(0) = \rho$, for some number ρ.
- We assume that effect of health insurance on health is the same for every respondent in the NHIS survey. **Constant treatment effect assumption.**
- Unrealistic, but makes the derivation of omitted variable bias formula simpler (formula still holds without that assumption).
- Under constant treatment effect assumption,

$$ATT = \frac{1}{n_1}\sum_{i:D_i=1}\big(y_i(1) - y_i(0)\big) = \frac{1}{n_1}\sum_{i:D_i=1}\rho = \rho.$$

- Under the constant treatment effect assumption, what we are trying to learn is ρ.

Two useful formulas before we derive omitted variable bias formula.

- Let Y_i be actual health of respondent i in 2015.
- $Y_i = (1 - D_i)y_i(0) + D_i y_i(1)$.
- Under the constant treatment effect assumption,

$\boldsymbol{Y_i} = (1 - D_i)y_i(0) + D_i y_i(1) = y_i(0) - D_i y_i(0) + D_i y_i(1) = y_i(0) + \big(y_i(1) - y_i(0)\big)D_i = \boldsymbol{y_i(0) + \rho D_i}$.

- Therefore,

$\overline{\boldsymbol{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n}(y_i(0) + \rho D_i) = \frac{1}{n}\sum_{i=1}^{n} y_i(0) + \rho\frac{1}{n}\sum_{i=1}^{n} D_i = \boldsymbol{\overline{y}(0) + \rho\overline{D}}$.

The omitted variable bias formula (1/2)

- $Y_i = y_i(0) + \rho D_i$ and $\bar{Y} = \bar{y}(0) + \rho \bar{D}$, $\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$.
- Plugging $Y_i = y_i(0) + \rho D_i$ and $\bar{Y} = \bar{y}(0) + \rho \bar{D}$ into formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) + \rho D_i - \bar{y}(0) - \rho \bar{D})}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}[(D_i - \bar{D})(y_i(0) - \bar{y}(0)) + (D_i - \bar{D})(\rho D_i - \rho \bar{D})]}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0)) + \frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})\rho(D_i - \bar{D})}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0)) + \rho\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$$

$$= \rho + \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0))}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$$

The omitted variable formula (2/2)

- $\hat{\beta}_1 = \rho + \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0))}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$
- $\hat{\beta}_1 =$ effect of health insurance of health + covariance between respondents' health without insurance ($y_i(0)$) and whether they are insured or not (D_i)/ variance of insurance variable.
- **General result: in regression of an outcome on a constant and a treatment, the coefficient of the treatment is equal to the effect of the treatment + covariance between outcome without treatment ($y_i(0)$) and treatment (D_i) / variance of treatment.**
- This second term is called the omitted variable bias term.
- Therefore, $\hat{\beta}_1 = \rho$, the effect of health insurance on health, if and only if 0 covariance between respondents' health without insurance ($y_i(0)$) and whether they are insured or not (D_i).
- Otherwise $\hat{\beta}_1 \neq \rho$, meaning that coefficient of D_i in regression of Y_i on a constant and D_i is not equal to the effect of D_i on Y_i.
- Do you think that whether someone gets insured or not is uncorrelated with her health without insurance?

iClicker time

- It is plausible that whether someone gets insured is
- A) positively correlated with the health of that person without insurance
- B) uncorrelated with the health of that person without insurance
- C) negatively correlated with the health of that person without insurance

Getting insured is likely to be positively correlated with health without insurance

- People getting insured are people rich enough to pay for insurance. Richer people also tend to be more educated. In NHIS survey, insured have more education and a higher income than uninsured.

	Insured	Uninsured
Average Years of Education	14.31	11.56
Average Family Income	106,467	46,656

- More educated people tend to smoke less and exercise more. So being insured may be positively correlated with $y_i(0)$, health without insurance: even without insurance insured people would smoke less and exercise more so they would be in better health.
- $\hat{\beta}_1 = \rho + \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0))}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$, $\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0)) > 0$, so $\hat{\beta}_1 > \rho$: $\hat{\beta}_1$ overestimates effect of health insurance on health.
- Maybe insured healthier not because insured, but because smoke less & exercise more. Smoking & exercising = **omitted variables**.

Omitted variable with binary treatment (1/2)

- When treatment binary, omitted variable formula becomes simpler, and we can derive it without assuming constant treatment effect.
- $\hat{\beta}_1$ is coeff. of D_i, binary variable, in reg. of Y_i on a constant and D_i.

$$\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:D_i=1} Y_i - \frac{1}{n-n_1}\sum_{i:D_i=0} Y_i$$

$$= \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$$

$$= \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$$

$$= ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0).$$

- $\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$, $\mathrm{ATT} = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$, so $\hat{\beta}_1$ uses $\frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$, average health of uninsured, to estimate $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$ average health of insured without insurance.

Omitted variable with binary treatment (2/2)

- When treatment binary, omitted variable formula becomes simpler:

$$\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0).$$

- $\hat{\beta}_1$= average effect of being insured on health among insured people (ATT)+ difference between average health without insurance ($y_i(0)$) of insured ($D_i = 1$) and uninsured ($D_i = 0$) people.
- **General result: in regression of outcome on constant and binary treatment, coefficient of treatment equal to average effect of treatment among treated people (ATT)+ difference between average $y_i(0)$ of treated ($D_i = 1$) and untreated ($D_i = 0$) people.**
- Second term is omitted variable bias term in previous formula. When treatment binary, sometimes called selection bias term.
- Therefore, $\hat{\beta}_1 = ATT$, if and only if insured and uninsured would have same average health in 2015 if uninsured in 2015. Otherwise $\hat{\beta}_1 \neq ATT$.
- Do you think that insured and uninsured would have same average health in 2015 if insured had remained uninsured?

iClicker time

- Do you think that insured and uninsured have same average health without insurance?
- A) Yes
- B) No

Insured would probably be in better health than uninsured even if they had not gotten insurance.

- In the NHIS survey, it is indeed the case that insured respondents have more education and a higher income than uninsured respondents.

	Insured	Uninsured
Average Years of Education	14.31	11.56
Average Family Income	106,467	46,656

- More educated people tend to smoke less and exercise more. So even if they had not been insured in 2015, the average health of insured people in 2015 would probably have been higher than average health of uninsured people: $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0) > 0$.
- $\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$ and $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0) > 0$, so $\hat{\beta}_1 > ATT$: $\hat{\beta}_1$ overestimates the average effect of health insurance on the health of insured people.
- Maybe insured are in better health not because insured, but because smoke less & exercise more. Smoking & exercising = **omitted variables**.

The attribution problem

- With binary treatment,

$\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:D_i=1} Y_i - \frac{1}{n-n_1}\sum_{i:D_i=0} Y_i$

- Good measure of the effect of D_i on Y_i only if the only difference between people with $D_i = 1$ and people with $D_i = 0$ is that the first group got the treatment and not the other one.
- In the health insurance example, this is not the case. Many other differences between insured and uninsured: insured richer, more educated, smoke less, exercise more.
- Therefore, we cannot know whether the difference between the average health of the two groups comes from the fact one group is insured and not the other one, or from the fact insured are richer, smoke less, etc. => Maybe insured would have been in better health even if uninsured, maybe ≠ in health not due to insurance.
- **Attribution problem**: if people with $D_i = 1$ and $D_i = 0$ differ on many characteristics (treatment + other characteristic), you cannot know which characteristic generates difference in their outcome.

Definition of omitted variables

- Omitted variables are variables that are not included in the regression, that are correlated with D_i, and that have an effect on $y_i(0)$.
- E.g.: in the regression of health on a constant and insurance, the smoking status of each individual is an omitted variable. Smoking is correlated with insurance (uninsured people tend to smoke more), and smoking has an effect on $y_i(0)$, health without insurance.
- **When we have an omitted variable in a regression, the coefficient of D_i in that regression does not measure the causal effect of D_i on Y_i.**

Correlation is not causation!

- $\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$ measures the correlation between health and insurance.
- Measures whether Y_i and D_i move in the same or opposite directions: do people who are insured tend to be in better health than people who are not insured.
- Health and insurance move in same direction, but maybe not because insurance has positive effect on health, maybe because insured people exercise more and smoke less than uninsured. **Smoking and exercise are omitted variables in this regression.**
- Correlation is not causation! The fact that having insurance is positively correlated with health does not mean that having insurance has a positive effect on health.
- Assume that in the 2015 NHIS data set, there is another binary variable X_i equal to 1 if respondent i smokes and to 0 otherwise. How could you use X_i to form a better measure of the effect of health insurance on health?

iClicker time

- Assume that in the 2015 NHIS data set, there is another binary variable X_i equal to 1 if respondent i smokes and to 0 otherwise. Which of the following would be a better measure of the effect of health insurance on health than $\hat{\beta}_1$?
- A) $\hat{\alpha}$, the coefficient of X_i in a regression of Y_i on X_i.
- B) $\hat{\gamma}_1$, the coefficient of D_i in a regression of Y_i on a constant, D_i, and X_i.
- C) The average value of X_i.

$\hat{\gamma}_1$!

- $\hat{\gamma}_1$ = coeff. of D_i in a regression of Y_i on a constant, D_i, and X_i. D_i and X_i binary, so follows from slides on multivariate regression that

$$\hat{\gamma}_1 = w\left(\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i\right) + (1-w)\left(\frac{1}{n_{11}}\sum_{i:D_i=1,\ X_i=1} Y_i - \frac{1}{n_{01}}\sum_{i:D_i=0,\ X_i=1} Y_i\right)$$

- $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$: difference between average health of insured and uninsured that don't smoke ($X_i = 0$).
- Therefore, this difference cannot come from the fact insured smoke more than uninsured: both groups do not smoke.
- $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$: difference between average health of insured and uninsured that smoke ($X_i = 1$).
- Therefore, this difference cannot come from the fact insured smoke more than uninsured: both groups smoke.
- $\hat{\gamma}_1$ compares health of insured and uninsured, controlling for smoking status. **$\hat{\gamma}_1$ shuts down omitted variable bias in $\hat{\beta}_1$ coming from difference in smoking rates between insured and uninsured.**
- We run regression and find $\hat{\gamma}_1 = 0.23$. Also, $\hat{\gamma}_1$ statistically significant at 5% level. Can we conclude that being insured improves health?

iClicker time

- Using 2015 NHIS data, we run an OLS regression of Y_i (2015 health) on a constant, D_i (insurance in 2015), and X_i (whether person smokes). We find $\hat{\gamma}_1 = 0.23$. Also, $\hat{\gamma}_1$ statistically significant at 5% level. Can we conclude from this that being insured improves health?
- A) Yes
- B) No

No, due to omitted variable bias.

- $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$: diff. between average health of insured and uninsured people that do not smoke ($X_i = 0$).
- This diff. may not come from the effect of insurance on health, may just come from the fact that insured non smokers exercise more and are richer than uninsured non smokers.
- $\frac{1}{n_{10}}\sum_{i:D_i=1,\ X_i=0} Y_i - \frac{1}{n_{00}}\sum_{i:D_i=0,\ X_i=0} Y_i$: diff. between average health of insured and uninsured people that smoke ($X_i = 1$).
- This diff. may not come from the effect of insurance on health, may just come from the fact that insured smokers exercise more and are richer than uninsured smokers.
- **Controlling for smoking status reduces but does not solve the omitted variable bias problem.**
- There are still omitted variables in that regression. For instance, exercise is not included in the regression, it is correlated with insurance (uninsured people exercise less), and it has an effect on $y_i(0)$, health without insurance.

Couldn't we control for all omitted variables?

- Couldn't we just include all variables correlated with D_i and that have an effect on $y_i(0)$ in regression? Then, we would no longer have any omitted variable.
- Issue: often, there are variables correlated with D_i, that have an effect on $y_i(0)$, but that are not included in our data set, so we cannot include them in our regression!
- In the insurance and health example, "motivation to stay in good health" is probably correlated with insurance: people who are the most motivated to stay in good health will purchase insurance, people less motivated won't.
- "motivation to stay in good health" also has an effect on $y_i(0)$: the higher your motivation to be in good health, the better your health habits, and the better your health.
- However, something like "motivation to stay in good health" is very hard to measure => we rarely have that variable in our data set and we cannot include it in the regression.

The effect of education on wages

- Data set of homework 3. Representative sample of 14086 wage earners.
- We regress their ln(wage) Y_i on a constant and their years of schooling D_i.
- $\hat{\beta}_1$ positive and significant.

```
Linear regression                               Number of obs     =     14,086
                                                F(1, 14084)       =    2209.70
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1474
                                                Root MSE          =     .79406
```

ln_weekly_wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
years_of_schooling	.1153305	.0024535	47.01	0.000	.1105214	.1201396
_cons	4.929266	.0338734	145.52	0.000	4.86287	4.995663

- Can we conclude from that regression that schooling has a positive effect on individuals' wages, meaning that thanks to their years of schooling, people manage to get higher wages than $y_i(0)$, the wage they would have obtained without any schooling?

iClicker time

- We regress ln(wage) on a constant and years of schooling. $\hat{\beta}_1$ positive and significant. Can we conclude from that regression that schooling has a positive effect on individuals' wages, meaning that thanks to their years of schooling, people manage to get higher wages than $y_i(0)$, the wage they would have obtained without any schooling?
- A) Yes
- B) No, because the R-squared of the regression is low.
- C) No, because there are omitted variables in this regression.

No because of omitted variable bias.

- $\hat{\beta}_1$ compares the average wage of people whose years of schooling differ by one.
- People with more schooling tend to come from better-off families.
- Maybe the fact that people with more schooling earn more money does not come from the fact that they completed more years of schooling, but just comes from the fact their parents are better-off and could help them get a better job. Even without that extra year of schooling, would still have obtained better earnings than people with one year of schooling less.
- People with more schooling tend have a higher IQ than people with less schooling.
- Maybe the fact that people with more schooling earn more money does not come from the fact that they completed more years of schooling, but just comes from the fact their IQ is higher. Even without that extra year of schooling, would still have obtained better earnings than people with one year of schooling less, just because their IQ is higher.
- Parents' earnings and IQ are omitted variables in this regression.
- **The R2 of a regression has nothing to do with whether the regression measures a causal effect or just a correlation.**

$\hat{\beta}_1$ overestimates the effect of schooling on wages.

- Omitted variable bias formula: $\hat{\beta}_1 = \rho + \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0))}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$.
- Presumably, $\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0)) > 0$.
- Positive correlation between years of schooling, and $y_i(0)$: people with more schooling than average $(D_i - \bar{D}) > 0$ probably would earn more money than the average without any schooling $y_i(0) - \bar{y}(0)$, e.g. because smarter.
- Therefore, $\hat{\beta}_1 > \rho$. $\hat{\beta}_1$ overestimates the effect of schooling on wages.

The effect of attending an independent school on student's achievement.

- Quote from a report by the National Association of Independent Schools: "NAIS worked with Gallup for several years to investigate the life outcomes and well-being of graduates of independent schools. Gallup's analysis found that a higher percentage of NAIS graduates than public school graduates enrolled in college immediately after high school (85 percent of NAIS graduates compared to 69 percent of public school graduates)."
- These figures come from a regression of Y_i (whether student i goes to college or not) on a constant and D_i (whether student i was in public or independent school).
- $\hat{\beta}_1 = 0.85 - 0.69 = 0.16.$
- Can we conclude from this regression that independent schools have a positive effect on students' chances of going to college?

iClicker time

- We regress college attendance of student i on a constant and whether student i went to an independent school or a public school. $\hat{\beta}_1$ positive and significant. Can we conclude from this regression that independent schools have a positive effect on students' chances of going to college?
- A) Yes
- B) No

No because of omitted variable bias.

- $\hat{\beta}_1$ compares the college attendance rate of students in independent and public schools.
- Students in independent schools tend to come from better-off families => therefore, more likely that their families can pay for private tutors for them.
- Maybe the fact that students in independent schools more likely to go to college does not come from the fact that independent schools prepare them better for college, but just comes from the fact their parents are better-off and can pay for a private tutor for them. Even if had gone to a public school, would still have been more likely to go to college thanks to that private tutor.
- Parents' earnings is an omitted variable in this regression.
- Because attending an independent school is binary, we can use the omitted variable bias formula for a binary treatment:

$$\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$$

- Likely that $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0) > 0$: even if they had gone to a public school, students going to independent schools would have been more likely to attend college than students going to public schools. Therefore, $\hat{\beta}_1 > ATT$.

What you need to remember (1/2)

- We would like to compute $ATT = \frac{1}{n_1}\sum_{i:D_i=1}(y_i(1) - y_i(0))$: average effect of treatment among treated people.
- Idea: reg. Y_i on constant and D_i, use $\hat{\beta}_1$, coeff. of D_i, as estimator of ATT.
- Omitted variable bias formula: if $y_i(1) - y_i(0) = \rho$ (constant treatment effect), $ATT = \rho$ but $\hat{\beta}_1 = \rho + \frac{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})(y_i(0) - \bar{y}(0))}{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2}$.
- $\hat{\beta}_1 = ATT$ iif 0 covariance between $y_i(0)$ and D_i. Otherwise $\hat{\beta}_1 \neq ATT$: coeff. of D_i in reg. of Y_i on constant and D_i not equal to effect of D_i on Y_i.
- When D_i binary, simpler omitted variable formula (holds even if treatment effect not constant): $\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$.
- Therefore, $\hat{\beta}_1 = ATT$ iif treated and untreated people have same average outcome without treatment. Otherwise $\hat{\beta}_1 \neq ATT$.
- Treated and untreated people will have same average outcome without treatment only if look very similar (same income, same education…).
- If two groups look different on important variables, cannot know if difference in their average outcome comes from treatment or from those variables. **Attribution problem.**
- E.g.: are insured people healthier than uninsured because insured, or because smoke less?

What you need to remember (2/2)

- You need to know how to use the omitted variable formula to assess whether it is more likely that $\hat{\beta}_1 > \rho$, or $\hat{\beta}_1 < \rho$ ($\hat{\beta}_1 > ATT$, or $\hat{\beta}_1 < ATT$ when treatment binary).
- Omitted variable: variable not included in regression, correlated with D_i, and has an effect on $y_i(0)$.
- E.g.: in the regression of health on a constant and insurance, smoking is omitted variable.
- When we have an omitted variable in a regression, the coefficient of D_i in that regression does not measure the causal effect of D_i on Y_i, only measures the correlation between D_i and Y_i, but correlation is not causation.
- Instead of running regression of Y_i on constant and D_i, we can get a better measure of the effect of D_i on Y_i by running regression of Y_i on constant, D_i, and some variables correlated with D_i and that have effect on $y_i(0)$.
- Issue: often, we cannot include all the omitted variables in our regression, because there are some that we cannot even measure: e.g. "motivation to be in good health" in health insurance example.
- Misleading claims where people confuse correlation and causation very pervasive, especially in political discourse. You need to know how to debunk such claims for final.

Roadmap

1. Defining what we are looking for: potential outcomes and treatment effects.
2. Omitted variable bias.
3. No omitted variable bias in Randomized Controlled Trials.
4. Statistical tests in Randomized Controlled Trials.
5. Application: the effect of health insurance.
6. Other methods to measure causal effects than RCTs.
7. The impact evaluation industry

To avoid attribution problem, we need to form balanced treatment & control groups.

- You are mayor of town where 4 adults do not have health insurance. 2 healthy (orange). 2 unhealthy (blue). You do not observe who is healthy/unhealthy (confidential info).

- You have budget to give insurance to 2 people (**treatment group**). 2 people will remain uninsured (**control group**). You want to study effect of insurance on consumption of care, health... Idea: compare in 1 year outcomes in treatment & control groups.
- You worry that treatment and control groups might not bear same % of healthy people. If so, you cannot know whether treatment-control comparison picks effect of health insurance, or the fact initial health of 2 groups was ≠. **Omitted variable.**
- Could you ensure that 2 people who receive health insurance bear same % of healthy people as those who do not receive insurance?

iClicker time

- To avoid attribution problem, could you ensure that the 2 people who receive health insurance bear same % of people healthy as 2 people who do not receive health insurance?
- A) Yes
- B) No

No!

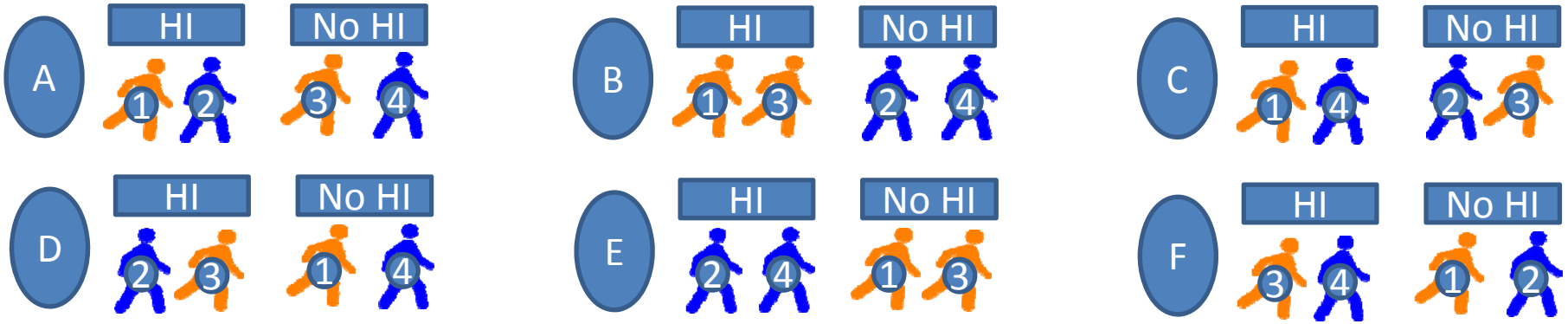
- You are the mayor of a town where 4 adults do not have health insurance. 2 are healthy (orange). 2 are unhealthy(blue). You do not observe who is in good/unhealthy (confidential info).

- You have the budget to give health insurance to 2 people (treatment group). 2 people will remain uninsured (control group).
- To avoid attribution problem, could you ensure that the 2 people who receive health insurance bear the same % of people healthy as 2 people who do not receive insurance?
- No. If you observed that 1 and 3 are healthy while 2 and 4 are unhealthy, you could give health insurance to, say 1 and 4, while 2 and 3 remain insured. Thus treatment and control groups would be balanced: they would both bear 50% of healthy people. But the issue is that you do not observe who is healthy/unhealthy => you cannot do that.

What if we randomly choose two people that receive health insurance? Preliminary question.

- 4 people, 2 orange (healthy), 2 blue (unhealthy). You can give health insurance (HI) to 2.

- If randomly choose those 2 people, 6 possible outcomes, and probability that each outcome gets selected is 1/6:



- Let H_T denote the % of healthy people in the treatment group.
- Let H_C denote the % of healthy people in the control group.
- H_T and H_C: random variables. Depend on lottery outcome realized.
- What is the value of H_T if outcome A gets selected? What is the value of H_T if outcome E gets selected? Discuss this question with your neighbor during two minutes.

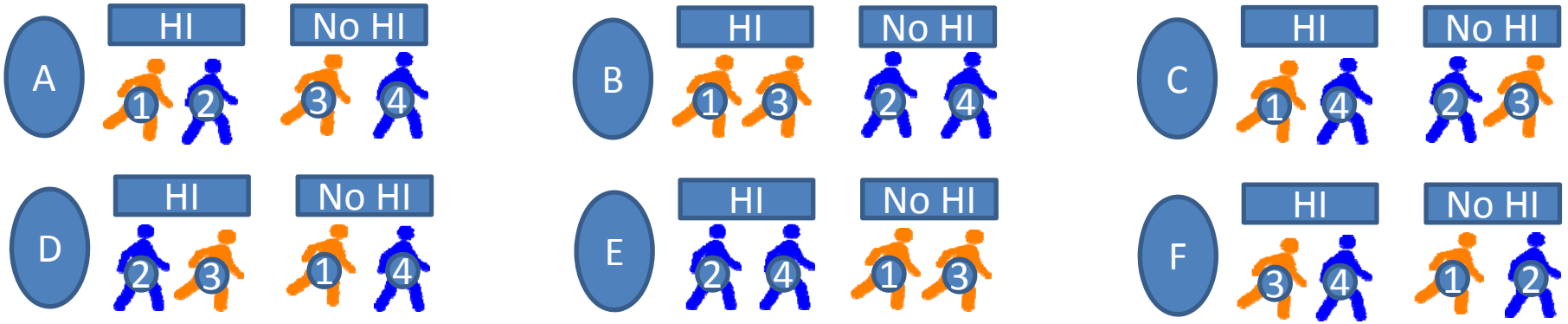
iClicker time

- What is the value of H_T if outcome A gets selected? What is the value of H_T if outcome E gets selected?
- A) If A gets selected $H_T = 0.5$, while if E gets selected $H_T = 0$
- B) If A gets selected $H_T = 1$, while if E gets selected $H_T = 0.5$.

Answer to preliminary question.

- 4 people, 2 orange (healthy), 2 blue (unhealthy). You can give health insurance (HI) to 2.

- If randomly choose those 2 people, 6 possible outcomes, and probability that each outcome gets selected is 1/6:

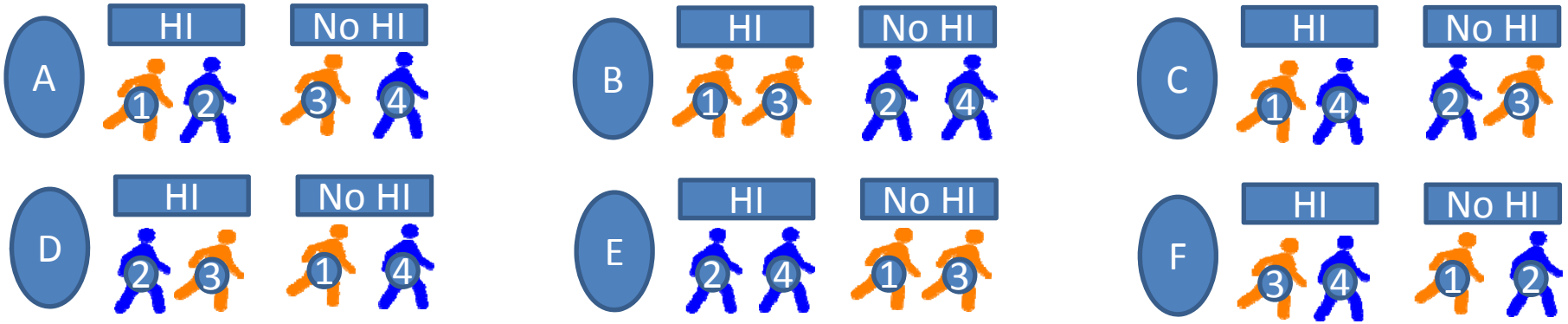


- What is value of H_T if A selected? What is the value of H_T if E selected?
- If A selected, $H_T = 0.5$: 1 healthy person / 2 in treatment group.
- If E selected, $H_T = 0$: 0 healthy person / 2 in treatment group.

What if we randomly choose the two people that receive health insurance?

- 4 people, 2 orange (healthy), 2 blue (unhealthy). You can give health insurance (HI) to 2.

- If randomly choose those 2 people, 6 possible outcomes, and probability that each outcome gets selected is 1/6:



- H_T: % healthy people in treatment group. H_C: % healthy people in control group.
- What is the expectation of H_T? What is the expectation of H_C?

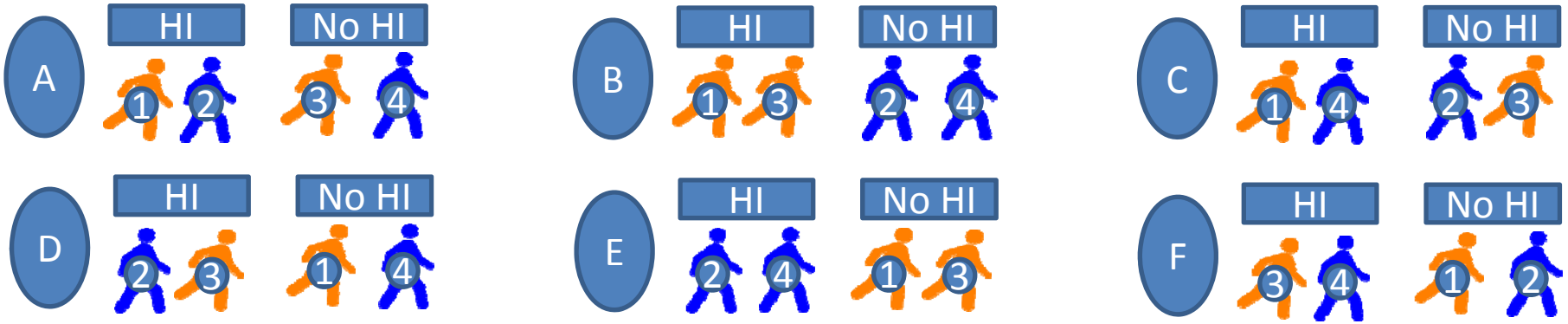
iClicker time

- What is the expectation of H_T? What is the expectation of H_C?
- A) $\mathrm{E}(H_T) = 0.4$ and $\mathrm{E}(H_C) = 0.6$.
- B) $\mathrm{E}(H_T) = 0.5$ and $\mathrm{E}(H_C) = 0.5$.
- C) $\mathrm{E}(H_T) = 0.6$ and $\mathrm{E}(H_C) = 0.4$.

On average, the lottery creates balanced groups!

- 4 people, 2 orange (healthy), 2 blue (unhealthy). You can give health insurance (HI) to 2.

- If randomly choose those 2 people, 6 possible outcomes, and probability that each outcome gets selected is 1/6:

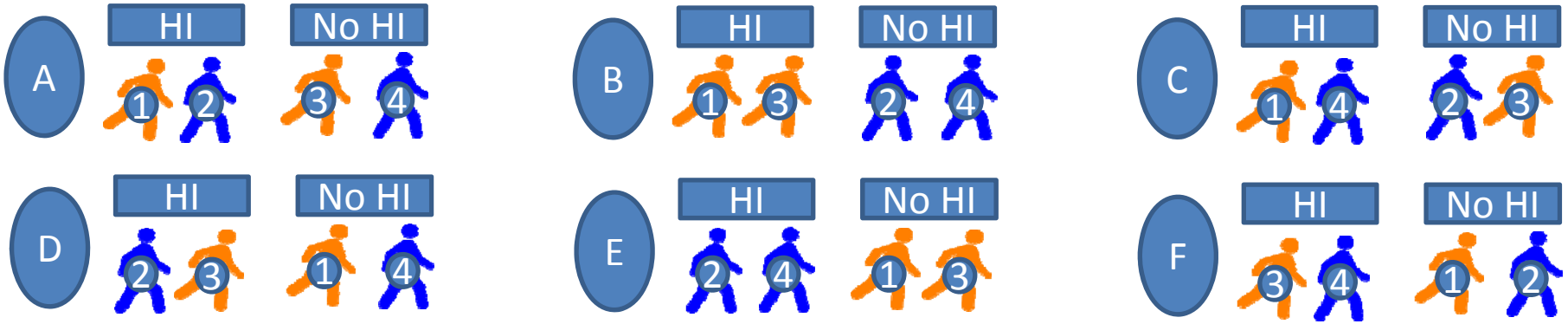


- H_T: % healthy people in treatment group. H_C: % healthy people in control group.
- What is the expectation of H_T? What is the expectation of H_C?
- $E(H_T) = 1/6 \times 0.5+1/6 \times 1+1/6 \times 0.5+1/6 \times 0.5+1/6 \times 0+1/6 \times 0.5=0.5.$
- $E(H_C) = 1/6 \times 0.5+1/6 \times 0+1/6 \times 0.5+1/6 \times 0.5+1/6 \times 1+1/6 \times 0.5=0.5.$
- **On average both groups have 50% healthy people. On average, 2 groups are balanced!**

Will the lottery always produce balanced groups?

- 4 people, 2 orange (healthy), 2 blue (unhealthy). You can give health insurance (HI) to 2.

- If randomly choose those 2 people, 6 possible outcomes, and probability that each outcome gets selected is 1/6:



- H_T: % healthy people in treatment group. H_C: % healthy people in control group.
- What is probability that $H_T = H_C$, meaning that the groups are balanced? What is probability that $H_T \neq H_C$, meaning that the groups are not balanced? Discuss this question with your neighbor during two minutes.

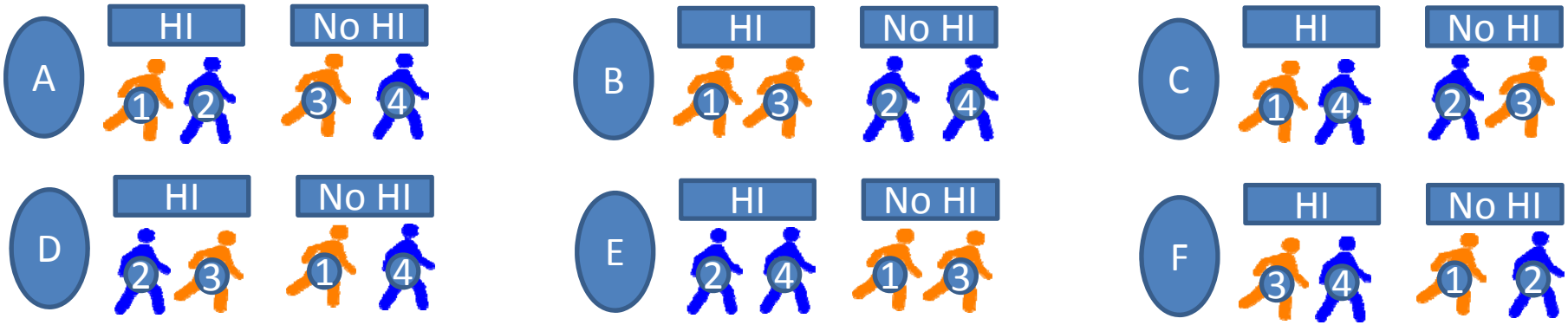
iClicker time

- What is probability that $H_T = H_C$, meaning that the groups are balanced? What is probability that $H_T \neq H_C$, meaning that the groups are not balanced?
- A) The probability that $H_T = H_C$ is equal to 1 while the probability that $H_T \neq H_C$ is equal to 0.
- B) The probability that $H_T = H_C$ is equal to 0.5 while the probability that $H_T \neq H_C$ is equal to 0.5.
- C) The probability that $H_T = H_C$ is equal to 2/3 while the probability that $H_T \neq H_C$ is equal to 1/3.

No, lotteries can produce imbalanced groups.

- 4 people, 2 orange (healthy), 2 blue (unhealthy). You can give health insurance (HI) to 2.

If randomly choose those 2 people, 6 possible outcomes, and probability that each outcome gets selected is 1/6:



- H_T: % healthy people in treatment group. H_C: % healthy people in control.
- What is probability that $H_T = H_C$? Probability that $H_T \neq H_C$?
- If outcome A, C, D, or F gets selected, $H_T = H_C = 0.5$: groups perfectly balanced. This has 1/6+1/6+1/6+1/6=2/3 probability of happening.
- If B or E gets selected, groups are perfectly imbalanced. E.g.: if B gets selected, $H_T = 1$ and $H_C = 0$. All people who receive HI healthy, all people who do not receive HI unhealthy. B or E has 1/6+1/6=1/3 probability of happening.

Reminder: combinations

- Assume you have n people, and you want to draw k people out of those n. The number of different subsets of the n people you can draw is

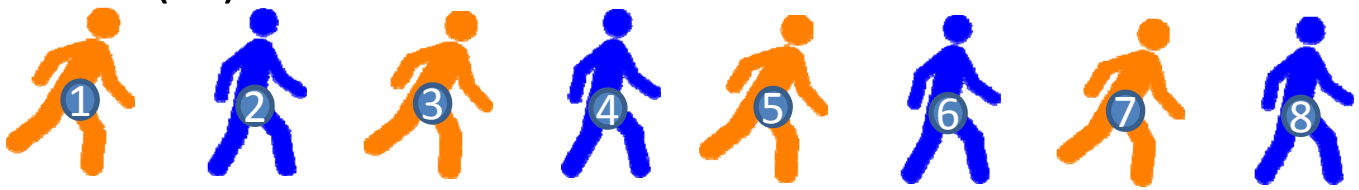
$$\binom{n}{k} = \frac{n!}{(n-k)!k!},$$

Where for any integer j, $j! = j \times (j-1) \times \cdots \times 1$.

- When you randomly assign n_1 units out of n to treatment group, there are $\binom{n}{n_1}$ different treatment groups you can create.
- Example: if you have 3 people (a, b, and c) and you randomly assign 2 of them to treatment group, the different treatment groups you can create are (a,b), (a,c), and (b,c). Three different treatment groups.
- $n = 3, n_1 = 2$, therefore $n - n_1 = 1$.
- $\binom{3}{2} = \frac{3!}{2! \times 1!} = \frac{3 \times 2 \times 1}{2 \times 1 \times 1} = 3$. The formula works.

But probability that lottery creates very imbalanced groups diminishes when size of lottery grows…

- 8 people, 4 orange (healthy), 4 blue (unhealthy). You can give health insurance (HI) to 4.



- If randomly choose those 4 people, follows from previous slide that number of possible draws is $\binom{8}{4} = \frac{8\times7\times6\times5}{4\times3\times2} = 70$ (draw 4 people out of 8).
- For only 2 lottery draws, groups are perfectly imbalanced:
  - 1-3-5-7 receive HI and 2-4-6-8 do not receive it
  - 2-4-6-8 receive HI and 1-3-5-7 do not receive it.
- => Chances that lottery produces perfectly imbalanced groups is 2/70 against 1/3 when we randomly assigned 2 people out of 4 to treatment, as in previous slides.
- **Larger lottery has lower proba. of creating imbalanced groups.**

…And vanishes when size of lottery goes to infinity.

- n people, $n/2$ orange (healthy), $n/2$ blue (unhealthy). You give HI to n_1 people.

- If randomly choose those n_1 people, proba that HI and control groups both bear 50% of healthy people goes to 1 when $n_1 \to +\infty$.
- **With infinity of people, lottery will create perfectly balanced groups.**
- => if number of people in lottery large, high chances that produces almost perfectly balanced groups.
- There may be small diffs. between your groups: you randomly send a bit more healthy people to treatment group than to control group. But differences unlikely to be large.
- When you toss 1000 times a fair coin, possible that you get 505 heads while you would expect 500: 2.4% chances that this happens. But extremely unlikely that you get 550 heads: 0.02% chances.
- Here same thing:
  - You have 2000 people, 1000 healthy and 1000 unhealthy.
  - You randomly give HI to 1000 of those 2000 people.
  - Maybe your lottery sends 505 healthy people to treatment group and 495 to control group.
  - But almost impossible that your lottery sends 550 healthy people to treatment group and 450 to control group.

Randomization creates groups that are balanced on all dimensions

- If number of people in health insurance (HI) lottery large enough, HI and control groups will be balanced on all dimensions.
- Balanced on dimensions you can observe (gender, age…) but also on dimensions you cannot observe (health, motivation…).
- Magic of randomization: even if you do not observe who is healthy/unhealthy, if your sample is large enough you can be highly confident that randomization will create groups where the % of healthy/unhealthy people is very similar.
- **Randomization solves attribution problem**: the only difference between your treatment and your control group is that the treatment group receives the treatment.
- Any difference between the average outcomes of the two groups must come from the effect of the treatment, not from something else.

The Oregon health insurance experiment.

- Medicaid: free health insurance for people with limited resources.
- Tight eligibility criteria: large % of US population remains uninsured.
- Oregon wants to expand Medicaid to adults not otherwise eligible for public insurance, who are Oregon residents, have been without health insurance for six months, have income below federal poverty level (FPL), and have assets below $2,000.
- 74,922 individuals, and Oregon only has money to insure 29,834.
- => assign random number to each of 74,922 individuals.
- 29,834 individuals with lowest numbers get enrolled in Medicaid (treatment group). Others not enrolled (control group).
- Treatment group gets health insurance, control group does not.
- How can you check that lottery indeed creates balanced groups? Discuss this question with your neighbor during two minutes.

iClicker time

- How can you check that the lottery indeed creates balanced groups?
- A) By counting the number of people in the treatment and in the control group, and checking that these two numbers are close to each other.
- B) By comparing the socio-demographic characteristics of the members of the two groups and by checking that they are similar.

By running balancing checks!

- Groups have similar demographics. Very small differences. That's because the lottery has many people.

	Insured by lottery	Uninsured by lottery
% Female	55.7%	55.0%
% English first language	92.2%	92.4%
% live in rural area	77.3%	77.5%
Average of average income in ZIP code of residence	39,265 USD	39,310 USD
Average Year of Birth	1968.00	1968.16
Number of people	29,834	45,088

Randomization creates balanced groups, choice does not.

- Demographics of people insured / uninsured by lottery very similar.
- Demographics of people insured / uninsured by choice very different.

	Insured by lottery	Uninsured by lottery
% Female	55.7%	55.0%
% English first language	92.2%	92.4%
% live in urban area	77.3%	77.5%
Average of average income in ZIP code of residence	39,265 USD	39,310 USD
Average Year of Birth	1968.00	1968.16
Number of people	29,834	45,088

	Insured by choice	Uninsured by choice
Average Years of Education	14.31	11.56
Average Family Income	106,467	46,656
Number of people	8,114	1,281

Is $\hat{\beta}_1$ good estimator of treatment effect in an RCT?

- n people eligible for a treatment (e.g.: HI). We can give treatment to $n_1 < n$ people.
- RCT: randomly choose n_1 people that get treatment.
- D_i: binary variable equal to 1 if person i randomly selected to get the treatment, and to 0 otherwise.
- Some time after lottery, we observe Y_i, outcome variable we are interested in for each person included in lottery (e.g.: the health of each person in the Oregon health experiment 1 year after lottery).
- $Y_i = D_i y_i(1) + (1 - D_i) y_i(0)$
- We want to estimate $ATT = \frac{1}{n_1} \sum_{i:D_i=1} (y_i(1) - y_i(0))$: average effect of treatment among treated people.
- We cannot compute ATT, as we do not observe $\frac{1}{n_1} \sum_{i:D_i=1} y_i(0)$.
- Instead: reg. Y_i on constant and D_i, use $\hat{\beta}_1$, coeff. of D_i as estimator of ATT.
- In a RCT, should we expect $\hat{\beta}_1$ to be close to ATT?

iClicker time

- In a RCT, we regress outcome Y_i on D_i: binary variable equal to 1 if person i randomly selected to get the treatment. Should we expect $\hat{\beta}_1$ to be close to ATT?
- A) Yes
- B) No

Yes! No omitted variable bias.

- Because treatment binary, we can use the omitted variable bias formula for a binary treatment:

$$\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$$

- Treatment randomly assigned so $\frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$ should be close to $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$.
- Average $y_i(0)$ of untreated people should be close to average $y_i(0)$ of treated people. The two groups are formed randomly => should be pretty similar.
- For instance, balancing checks in the Oregon Health experiment showed that average of all variables (average year of birth, average income in ZIP code of residence, etc.) were very close in the two groups. Therefore, the average of $y_i(0)$ should also be very close in the two groups!
- Thus, $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$ should be small…
- … so $\hat{\beta}_1$ should be close to ATT!
- In sessions, you will show that $\hat{\beta}_1$ unbiased estimator of ATT.

What you need to remember

- A treatment and control group created by a lottery are balanced on average across all possible lottery draws.
- For some draws, the groups might still be unbalanced.
- But if many people participate in the lottery, the number of lottery draws where the two groups are very unbalanced becomes very small relative to the total number of possible lottery draws.
- => with very high probability, the two groups will be almost perfectly balanced on every characteristic, both on characteristics you can observe (income, age) and on characteristics you cannot observe (motivation…).
- You can check this: you can compare the social, demographic, psychological, etc. characteristics of your treatment and your control groups at the time of the lottery. You should find that they are very similar. That's what we found in the Oregon Health Experiment example.
- **Therefore, randomization solves attribution problem.** At the time of the lottery, no difference between treatment and control group. If differences start emerging after treatment group receives treatment, must be due to treatment, not to something else.
- **In a RCT**, to measure effect of treatment, you can just reg. outcome Y_i on constant and treatment D_i. No omitted variable bias, so $\hat{\beta}_1$, coeff. of D_i, should be close to ATT. $\hat{\beta}_1$ unbiased estimator of ATT.
- **Only true in RCT: if treatment not randomly assigned and treated people choose to get treated, it is very likely that there will be omitted variable bias!**

Roadmap

1. Defining what we are looking for: potential outcomes and treatment effects.
2. Omitted variable bias.
3. No omitted variable bias in Randomized Controlled Trials.
4. Statistical tests in Randomized Controlled Trials.
5. Application: the effect of health insurance.
6. Other methods to measure causal effects than RCTs.
7. The impact evaluation industry

Can we conclude that $ATT \neq 0$ whenever $\hat{\beta}_1 \neq 0$?

- $ATT = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i(1) - \frac{1}{n_1}\sum_{i=1}^{n_1} y_i(0).$
- We cannot compute ATT but we use $\hat{\beta}_1$ to estimate it.
- We have

$$\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$$

- Can we conclude that $ATT \neq 0$ whenever $\hat{\beta}_1 \neq 0$?

iClicker time

- Can we conclude that $ATT \neq 0$ whenever $\hat{\beta}_1 \neq 0$?
- A) Yes
- B) No

No

- $\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$
- One can have $\hat{\beta}_1 \neq 0$ while $ATT = 0$, if $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0) \neq 0$.
- $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$: difference between the average of $y_i(0)$ in our randomly formed treatment and control groups.
- In Oregon health experiment, difference between, say, the average year of birth of our two groups is close to 0 but not exactly equal to 0.
- => likely that $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$ close, but not exactly equal, to 0.

Intuition for statistical tests in RCTs

- $\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$
- If $\hat{\beta}_1 > 0$ but close to 0, maybe $ATT = 0$, but out of bad luck we formed treatment group where average of $y_i(0)$ a little bit larger than in control group.
- On the other hand, if $\hat{\beta}_1 > 0$ and far from 0, unlikely that $ATT = 0$. Having $\hat{\beta}_1 > 0$ and far from 0 and $ATT = 0$ means that $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0) > 0$ and far from 0: we formed treatment group where average of $y_i(0)$ much larger than in control group. Very unlikely to happen.
- If $\hat{\beta}_1 < 0$ but close to 0, maybe $ATT = 0$, but out of bad luck we formed treatment group where average of $y_i(0)$ little lower than in control group.
- On the other hand, if $\hat{\beta}_1 < 0$ and far from 0, unlikely that $ATT = 0$.

An example

- In the Oregon health experiment, Y_i is a binary variable equal to 1 for individuals who say they are in good health one year after the lottery, and to 0 for people who say they are not in good health.
- You reg. Y_i on a constant and D_i, a binary variable equal to 1 for individuals randomly assigned to receive health insurance, and to 0 for other individuals.
- Assume that you find $\hat{\beta}_1 = 0.05$, meaning that % of people who say they are healthy is 5 points higher in treatment than in control group. Is it plausible that $ATT = 0$? Hint: $ATT = 0$ and $\hat{\beta}_1 = 0.05$ imply what value of $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$? In the balancing checks, we compared average of 3 binary variables (female, English first language, live in urban area) in treatment and control groups. Did we find very large differences?

iClicker time

- In the Oregon health experiment, Y_i is a binary variable equal to 1 for individuals who say they are in good health one year after the lottery, and to 0 for people who say they are not in good health.
- You reg. Y_i on a constant and D_i, a binary variable equal to 1 for individuals randomly assigned to receive health insurance, and to 0 for other individuals.
- Assume that you find $\hat{\beta}_1 = 0.05$, meaning that % of people who say they are healthy is 5 points higher in treatment than in control group. Is it plausible that $ATT = 0$?
- A) Yes
- B) No

No!

- In Oregon health experiment, Y_i is equal to 1 for individuals who say they are in good health one year after the lottery, and to 0 for people who say they are not. You reg. Y_i on constant and D_i, equal to 1 for individuals randomly assigned to receive health insurance, and to 0 for others.
- Assume that $\hat{\beta}_1 = 0.05$, plausible that $ATT = 0$?
- $\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$
- $\hat{\beta}_1 = 0.05$ & $ATT = 0$ => $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0) = 0.05$:

% of people in good health without insurance 5 points higher in treatment than control group.

- Balancing checks: % females 0.7 points higher in treatment than control, % of people with English as 1st language 0.2 points lower in treatment than control, % people living in city 0.2 points lower in treatment than control.
- % of people in good health without insurance very unlikely to be 5 percentage points higher in treatment than control: 7 times higher than highest difference on other percentages!
- Very unlikely that lottery creates groups with such large difference of average $y_i(0)$.

Central limit theorem (CLT) in RCTs

- $ATT = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$
- Let $\bar{y}(1) = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1)$ and $\bar{y}(0) = \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$.
- $\hat{\beta}_1 = \bar{y}(1)$-$\bar{y}(0)$. Difference between the two averages.
- Let $\hat{V}(\bar{y}(1)) = \frac{\frac{1}{n_1}\sum_{i:D_i=1}(y_i(1)-\bar{y}(1))^2}{n_1}$. Estimator of variance of $\bar{y}(1)$ (cf. polling).
- When outcome binary, $\hat{V}(\bar{y}(1)) = \frac{\bar{y}(1)(1-\bar{y}(1))}{n_1}$.
- Let $\hat{V}(\bar{y}(0)) = \frac{\frac{1}{n-n_1}\sum_{i:D_i=0}(y_i(0)-\bar{y}(0))^2}{n-n_1}$. Estimator of variance of $\bar{y}(0)$ (cf. polling). When outcome binary, $\hat{V}(\bar{y}(0)) = \frac{\bar{y}(0)(1-\bar{y}(0))}{n-n_1}$.
- Let $V(\hat{\beta}_1)$=$\hat{V}(\bar{y}(1))$+$\hat{V}(\bar{y}(0))$.
- **If n larger than 100, $\frac{\hat{\beta}_1 - ATT}{\sqrt{V(\hat{\beta}_1)}}$ approximately follows N(0,1) distribution.**

A 5% level t-test that $ATT = 0$

- $ATT = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1) - \frac{1}{n_1}\sum_{i:D_i=1} y_i(0)$
- Let $\bar{y}(1) = \frac{1}{n_1}\sum_{i:D_i=1} y_i(1)$ and $\bar{y}(0) = \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$.
- $\hat{\beta}_1 = \bar{y}(1)$-$\bar{y}(0)$. Difference between the two averages.
- Let $\hat{V}(\bar{y}(1)) = \frac{\frac{1}{n_1}\sum_{i:D_i=1}(y_i(1)-\bar{y}(1))^2}{n_1}$.
- Let $\hat{V}(\bar{y}(0)) = \frac{\frac{1}{n-n_1}\sum_{i:D_i=0}(y_i(0)-\bar{y}(0))^2}{n-n_1}$.
- Let $V(\hat{\beta}_1)$=$\hat{V}(\bar{y}(1))$+$\hat{V}(\bar{y}(0))$.
- 5%-level test of $ATT = 0$:

Reject $ATT = 0$ if $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} > 1.96$ or $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} < -1.96$.

Otherwise, do not reject $ATT = 0$.

If $ATT = 0$, we only have 5% chances of wrongly rejecting $ATT = 0$.

- $\frac{\widehat{\beta}_1 - ATT}{\sqrt{V(\widehat{\beta}_1)}}$ approximately follows N(0,1) distribution.
- If $ATT = 0$, $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}}$ approximately follows N(0,1) distribution.
- We reject $ATT = 0$ if $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}} > 1.96$ or $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}} < -1.96$.
- A N(0,1) variable has 5% chances of being above 1.96 or below -1.96 => 5% chances of wrongly rejecting $ATT = 0$.
- Similarly, one can construct a 10% level test of $ATT = 0$ by comparing $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}}$ to 1.64 and -1.64.
- Similarly, one can construct a 1% level test of $ATT = 0$ by comparing $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}}$ to 2.57 and -2.57.

A 95% confidence interval for ATT

- 95% CI for ATT: $\left[\hat{\beta}_1 - 1.96\sqrt{V(\hat{\beta}_1)}, \hat{\beta}_1 + 1.96\sqrt{V(\hat{\beta}_1)}\right]$.
- When you randomly assign n_1 units out of n to treatment group, there are $\binom{n}{n_1}$ different treatment groups you can create.
- $\hat{\beta}_1 = \bar{y}(1)$-$\bar{y}(0)$. Difference between average outcome of units in the treatment and in the control group => $\hat{\beta}_1$ depends on who assigned to treatment and control group => $\binom{n}{n_1}$ possible values of $\hat{\beta}_1$, and $\binom{n}{n_1}$ possible values of the confidence interval.
- For 95% of those values,

$$ATT \in \left[\hat{\beta}_1 - 1.96\sqrt{V(\hat{\beta}_1)}, \hat{\beta}_1 + 1.96\sqrt{V(\hat{\beta}_1)}\right].$$

- 90% CI: $\left[\hat{\beta}_1 - 1.64\sqrt{V(\hat{\beta}_1)}, \hat{\beta}_1 + 1.64\sqrt{V(\hat{\beta}_1)}\right]$.
- 99% CI: $\left[\hat{\beta}_1 - 2.57\sqrt{V(\hat{\beta}_1)}, \hat{\beta}_1 + 2.57\sqrt{V(\hat{\beta}_1)}\right]$.

What you need to remember

- In RCT, one cannot conclude that $ATT \neq 0$ whenever $\hat{\beta}_1 \neq 0$.
- $\hat{\beta}_1 = ATT + \frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$, so one can have $\hat{\beta}_1 \neq 0$ while $ATT = 0$, if $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0) \neq 0$.
- However, $\frac{1}{n_1}\sum_{i:D_i=1} y_i(0) - \frac{1}{n-n_1}\sum_{i:D_i=0} y_i(0)$ cannot be too large because groups have been formed randomly => one can conclude that $ATT \neq 0$ when $\hat{\beta}_1$ is "far enough" from 0.
- Formally: 5%-level test of $ATT = 0$: Reject $ATT = 0$ if $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} > 1.96$ or $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} < -1.96$. Otherwise, do not reject $ATT = 0$.
- $V(\hat{\beta}_1)$ defined in the slides, you should know how to compute it when outcome is binary.
- 95% CI for ATT: $\left[\hat{\beta}_1 - 1.96\sqrt{V(\hat{\beta}_1)}, \hat{\beta}_1 + 1.96\sqrt{V(\hat{\beta}_1)}\right]$. For 95% of the treatment and control groups we can form, ATT belongs to CI.

Roadmap

1. Defining what we are looking for: potential outcomes and treatment effects.
2. Omitted variable bias.
3. No omitted variable bias in Randomized Controlled Trials.
4. Statistical tests in Randomized Controlled Trials.
5. Application: the effect of health insurance.
6. Other methods to measure causal effects than RCTs.
7. The impact evaluation industry

Findings from Oregon experiment, 1 year after lottery

	Uninsured by lottery	Insured by lottery	$\hat{\beta}_1$	$\sqrt{V(\hat{\beta}_1)}$
Number of hospital admissions	0.067	0.088	0.021	0.0074
Outpatient visits last 6 months	0.574	0.786	0.212	0.025
Cholesterol ever checked	0.625	0.739	0.114	0.026
Mammogram within last 12 months (women >40)	0.298	0.485	0.187	0.040
Spent > 30% income in health expenses	0.055	0.010	-0.045	0.019
Any medical debt	0.568	0.435	-0.133	0.045
Screens positive depression	0.300	0.208	-0.092	0.040
Health is good (reported)	0.548	0.681	0.133	0.026
Alive	0.987	1.00	0.013	0.027
High blood pressure	0.163	0.150	-0.013	0.014
High cholesterol	0.141	0.117	-0.024	0.027

Findings from Oregon experiment, 1 year after lottery

- For "Number of hospital admission", can you reject at the 5% level the null hypothesis $ATT = 0$?

	Uninsured by lottery	Insured by lottery	$\hat{\beta}_1$	$\sqrt{V(\hat{\beta}_1)}$
Number of hospital admissions	0.067	0.088	0.021	0.0074

iClicker time

- For “Number of hospital admission”, can you reject at the 5% level the null hypothesis $ATT = 0$?
- A) Yes
- B) No

Yes!

- For "Number of hospital admission", can you reject at the 5% level the null hypothesis $ATT = 0$?
- We reject $ATT = 0$ at the 5% level if $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}} > 1.96$ or $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}} < -1.96$.
- For the number of hospital admissions, $\hat{\beta}_1 = 0.021$ and $\sqrt{V(\hat{\beta}_1)}$=0.007, therefore $\frac{\widehat{ATT}}{\sqrt{\hat{V}(\widehat{ATT})}} = 3$. We reject $ATT = 0$ at the 5% level.
- Intuition: the difference between average hospital admissions in the insured and uninsured groups is too large to come only from the fact that out of bad luck, the insured group has a higher propensity to go to the hospital than the uninsured group. Insurance must have an effect.
- => insurance increases the number of visits to the hospital.
- Strictly positive elasticity of hospital admissions to price.

Findings from Oregon experiment, 1 year after lottery

- For how many of the 10 other outcomes, can you reject at the 5% level the null hypothesis $ATT = 0$? Discuss with your neighbor during 2 minutes.

	Uninsured by lottery	Insured by lottery	$\hat{\beta}_1$	$\sqrt{V(\hat{\beta}_1)}$
Outpatient visits last 6 months	0.574	0.786	0.212	0.025
Cholesterol ever checked	0.625	0.739	0.114	0.026
Mammogram within last 12 months (women >40)	0.298	0.485	0.187	0.040
Spent > 30% income in health expenses	0.055	0.010	-0.045	0.019
Any medical debt	0.568	0.435	-0.133	0.045
Screens positive depression	0.300	0.208	-0.092	0.040
Health is good (reported)	0.548	0.681	0.133	0.026
Alive	0.987	1.00	0.013	0.027
High blood pressure	0.163	0.150	-0.013	0.014
High cholesterol	0.141	0.117	-0.024	0.027

iClicker time

- For how many of the 10 other outcomes, can you reject at the 5% level the null hypothesis $ATT = 0$?
- A) For 6 of them
- B) For 7 of them
- C) For 8 of them

Yes for 7 of them…

- For the 10 other outcomes, can you reject at the 5% level the null hypothesis ATT=0?
- If you compute $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}}$ for the 10 other outcomes in the table, you will find that this quantity is greater than 1.96 or lower than -1.96 for:
  - Outpatient visits last 6 months
  - Cholesterol ever checked
  - Mammogram within last 12 months (women >40)
  - Spent > 30% income in health expenses
  - Any medical debt
  - Screens positive for depression
  - Health is good (reported)
- For all these outcomes, you can reject ATT=0.
- Difference between treatment and control group large => very unlikely that these differences are only due to chance, very likely that due to the effect of health insurance.

… But no for 3 of them.

- For the 10 other outcomes, can you reject at the 5% level the null hypothesis ATT=0?
- If you compute $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}}$ for the 10 other outcomes in the table, you will find that this quantity is included between -1.96 and 1.96 for:
  - Alive
  - High blood pressure
  - High cholesterol
- For all these outcomes, you cannot reject ATT=0.
- Difference between treatment and control group not so large => possible that comes from small differences between treatment and control groups created by chance, not from the effect of health insurance.

Summarizing the results

- When you give comprehensive health insurance for free to people:
- Health expenditures increase: **health expenditures elastic to price.**
- In particular, they consume more preventive care (cholesterol checked, mammograms…).
- They are also **less likely to have medical debt** and to face catastrophic health expenses (>30% income).
- Maybe because of this, their **mental health improves** (less likely to screen for depression), and they report that their health is better.
- On the other hand, **their physical health does not improve**: not more likely to be alive after one year, not less likely to have hypertension or high cholesterol.
- Keep in mind that those results are measured 1 year after people received health insurance. Will effects on physical health appear in the longer run? The fact insured people consume more preventive care suggests that might happen but we need to wait for longer-term results to know.

What you need to remember

- You need to know how to interpret results from a RCT.
- $\hat{\beta}_1$: difference between mean outcome in treatment and control group.
- We reject $ATT = 0$ at 5% level if:

$$\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} > 1.96 \text{ or} \frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} < -1.96.$$

- If outcome is not binary, I will give you $V(\hat{\beta}_1)$ or $\sqrt{V(\hat{\beta}_1)}$ so you can perform the test.

- If outcome is binary, I may not give $V(\hat{\beta}_1)$ or $\sqrt{V(\hat{\beta}_1)}$, I may just give you $\bar{y}(1)$ and $\bar{y}(0)$, and then you need to use $\hat{V}(\bar{y}(1)) = \frac{\bar{y}(1)(1-\bar{y}(1))}{n_1}$, $\hat{V}(\bar{y}(0)) = \frac{\bar{y}(0)(1-\bar{y}(0))}{n-n_1}$, and $V(\hat{\beta}_1)$=$\hat{V}(\bar{y}(1))$+$\hat{V}(\bar{y}(0))$ to run the test.

Roadmap

1. Defining what we are looking for: potential outcomes and treatment effects.
2. Omitted variable bias.
3. No omitted variable bias in Randomized Controlled Trials.
4. Statistical tests in Randomized Controlled Trials.
5. Application: the effect of health insurance.
6. Other methods to measure causal effects than RCTs.
7. The impact evaluation industry

Not all questions can be answered by RCTs

- Is it bad for employment to increase the minimum wage?
- Neoclassical economists say yes: if expensive to hire someone, then maybe companies prefer to buy machines to do the job, => capital substituted to labor, so high minimum wage reduces employment.
- Neo-Keynesian economists say no: if workers feel wage fair, maybe they become more motivated and productive => the company makes more profits and can hire even more workers.
- 2 conflicting theories => we need to study this question empirically.
- You cannot run randomized experiment: randomly set the minimum wage at 5$/hour in some companies and at 15$/hour in other companies, and look whether those where minimum wage = 15$ start hiring less or firing more workers. There can be only 1 minimum wage!
- There are other econometric techniques to answer questions that cannot be answered through RCTs.
- **If you want to learn those techniques, you should take 140B!**

Roadmap

1. Defining what we are looking for: potential outcomes and treatment effects.
2. Omitted variable bias.
3. No omitted variable bias in Randomized Controlled Trials.
4. Statistical tests in Randomized Controlled Trials.
5. Application: the effect of health insurance.
6. Other methods to measure causal effects than RCTs.
7. The impact evaluation industry

The impact evaluation industry

- The following consulting firms run RCTs for federal or local government in the US. They hire econ undergrads.
  - Mathematica Policy Research: https://www.mathematica-mpr.com/
  - MDRC: http://www.mdrc.org/
  - Many others.
- The following research institutions help researchers conduct RCTs in developing countries. They also hire econ undergrads.
  - J-Pal: https://www.povertyactionlab.org/careers
  - Innovation for poverty action: http://www.poverty-action.org/
- Many firms run RCTs with you everyday, to determine how they can maximize their revenues:
  - E.g., Google: whenever you make a search on Google, you are part of a RCT. Is it better to put the sponsored links left or right of the screen? What will maximize the chances that people click on those sponsored links? A random group of users gets their sponsored links on left of screen, another group gets them on right, and see who clicks the most!
  - Facebook, Capital one, etc.