Can you conclude "handing in a test early *causes* a student to have a different mean test score" (statement 1b)? NO. You cannot conclude this. This conceptual jump is not rationally justified. But is it a good question? YES

We indeed wish to know causal mechanisms in the world e.g. a drug company wants to know if it's drug *causes* a higher prob of cancer remission, Amazon wants to know if offering you a coupon *causes* a prob. of sale to increase, the Federal Reserve wants to know if lowering interest rates *causes* inflation to decrease, etc. Most scientific quesitons are worded this way. So we need a way to answer them.

Scenario 2. A factory has a batch of wire on a spool. They cut n_1 short pieces and n_2 long pieces. They are interested in proving that the electrical resistance is different in long and short pieces. They measure resistances y_1,1,..., y_1,n_1 independent of y_2,1, ..., y_2,n_2 and they run a Wald test. Let's say they reject H_0 and conclude "the mean resistances between the short and long wires is different" (Statement 2a).

But... can you conclude that "a different wire length *causes* a different resistance" (Statement 2b)?

Likely 2b is true. What's the difference between 1b and 2b?

The distinguishing difference is the "treatment variable" (the categorical difference between the two populations which is full time vs reduced exam time in Scenario 1 and short vs long length in Scenario 2) is "confounded" (confused, obfuscated, hidden) in Scenario 1 but not Scenario 2. The confounding is due to other variables being different in the two populations. In Scenario 1... student IQ, amount of time studied, etc. These confounding variables influence both the cause (treatment variable) and the effect (the response, the test score). If there is no confounding variable, the b statement is justified.

Some vocab. In this class, we will consider treatment variables consisting of two categories (AKA two "arms") called the treatment (T) and the control (C). T is usually the situation whose effect on the phenomenon $\vec{y} = [y_1 \dots y_n]^T$ (AKA "response", "endpoint" or "outcome") you wish to investigate and C is usually the "business as usual case". In a clinical trial the T is usually the "pill" and the C is usually the "placebo". Let the treatment variable be denoted $\vec{w} = [w_1 \dots w_n]^T$ where w_i = 1 means the ith subject got T and w_i = 0 means the ith subject got C. n_T is the number of subjects that got T and n_C is the number of subjects that got C and n = n_T + n_C is the total sample size.

FYI: more than two arms or non categorical treatment variables are also important but we just don't have time...

The Rubin Causal Model / Potential Outcomes Framework: every subject i has a potential outcome for T, y_T,i and potential outcome for C, y_C,i. Before the subject is "assigned" (given, allocated) to T or C, both are observable. But after they are assigned only one is observable. Every subject has a unit-level causal treatment effect, $y_{T,i} - y_{C,i}$, which is not measurable since only one of the two quantities can be measured. This is called "the fundamental problem of causal inference".

How do we learn anything? We make assumptions. Let's assume a constant, additive treatment effect theta:

$$y_{T,i} = y_{C,i} + \theta$$

We can't measure theta. It's the target of inference. We need an estimator. So first define Y_T as the DGP that produces y_T,i's and Y_C as the DGP that produced y_C,i's both are probably not identically distributed but we will assume independence (look up SUTVA online for a related concept). I know that:

$$\theta = E[Y_T] - E[Y_C] \overset{?}{\approx} \bar{y}_T - \bar{y}_C = \frac{1}{n_T} \sum_{\{i:\, w_i = 1\}} y_{T,i} - \frac{1}{n_C} \sum_{\{i:\, w_i = 0\}} y_{C,i}$$

Can't we use the "naive estimate" above? NO. Why? Just as we said before. The subjects in T are not necessarily the "same" as subjects in C. There may be a confounding variable. Assume one confounder and denote its value for subject i as x_i (e.g. let x_i be the number of hours studied in Scenario 1). Assume the confounder is linear in the response:

$$y_{C,i} = \beta x_i + e_i$$

beta is scaling parameter (another unknown quantity but a nuisance parameter) and e_i is a random component you don't understand (AKA "error" or "noise"). Putting these two equations together we get:

$$y_{T,i} = \theta + \beta x_i + e_i$$

Our naive estimate under these assumptions is:

$$\bar{y}_T - \bar{y}_C = \frac{1}{n_T} \sum_{\{i \text{ s.t. } w_i = 1\}} \theta + \beta x_i + e_i - \frac{1}{n_C} \sum_{\{i \text{ s.t. } w_i = 0\}} \beta x_i + e_i$$

$$= \left(\theta + \beta \bar{x}_T[\vec{w}] + \bar{e}_T[\vec{w}]\right) - \left(\beta \bar{x}_C[\vec{w}] + \bar{e}_C[\vec{w}]\right)$$

$$= \theta + \beta(\bar{x}_T - \bar{x}_C)[\vec{w}] - (\bar{e}_T - \bar{e}_C)[\vec{w}]$$

If you assume the e_i's are the result of a sum of a bunch of random stuff and there's no difference between the DGP's in the T and C groups, then the CLT says:

$$\bar{\varepsilon}_T - \bar{\varepsilon}_C \overset{\cdot}{\sim} N\left(0, \sigma_\varepsilon^2 \left(\frac{1}{n_T} + \frac{1}{n_C}\right)\right) \Rightarrow \bar{e}_T - \bar{e}_C \approx 0 \quad \forall \vec{w}$$

$$\Rightarrow E[\bar{y}_T - \bar{y}_C] = \theta + \beta(\bar{x}_T - \bar{x}_C)[\vec{w}]$$

Is this estimator biased? YES! It's biased by

$$\bar{y}_T - \bar{y}_C = \underbrace{\theta}_{>0} + \underbrace{\beta(\bar{x}_T - \bar{x}_C)}_{<0 \;\; >0} + \underbrace{(\bar{e}_T - \bar{e}_C)}_{\approx 0} \quad \text{for assignment } \vec{w}$$

In scenario 1... theta is the causal effect of handing in the exam early (negative).... but xbarT-xbarC is the average study time in the hand-in-exam-early group minus the average study time in the hand-in-exam-on-time group and xbarT > xbarC they studied more! So it's positive! And more positive than theta is negative. ebarT-ebarC is a nuisance and close to zero.

The reason the estimate is biased is due to "selection bias" which is bias in the w_i's. Students who studied more are more likely give themselves the T assignment (hand in exam early).

How to fix it? You can use "linear regression" to isolate and remove the effect of the confounder x or "matching" (we won't discuss these). But if there's a confounder you didn't think of... you're hosed.

To almost certainly get around this, you can't merely be an observer of the situation (scenario 1 is called an "observational study"), you need to be able to assign T/C's (AKA manipulate) to the values w_1, ..., w_n. Sometimes this is unethical. This ability to manipulate is called an "experimental study".

The gold standard to estimate causality is the "randomized experiment" (Fisher, 1925) meaning w_i's are picked randomly.

There are many ways to do this assignment ("experimental design"). If T/C comes from a 50-50 coin flip, it's called the "Bernoulli trial". I work on designs that lower the MSE of the estimator theta.

The design where n_T = n_C = n/2 and all such assignments are equally likely, $p(\vec{w}) = 1 / \binom{n}{n/2}$ is called "the completely randomized design. You can show:

$$E_{\vec{w}}\left[E_\varepsilon\left[\bar{y}_T - \bar{y}_C\right]\right] = \theta \quad \text{all } \vec{w}\text{'s}$$

this is the essential reason why we use randomized experiments.

You're never doing more than one experiment; you always have just one assignment w_1, ..., w_n. So how is this any different??

$$\bar{y}_T - \bar{y}_C = \theta + \underbrace{\beta(\bar{x}_T - \bar{x}_C)[\vec{w}]}_{\approx 0} + \underbrace{(\bar{e}_T - \bar{e}_C)[\vec{w}]}_{\approx 0}$$

getting this even closer to zero is what I work on.