HANDOUT 2 – CAUSALITY AND VALIDITY

AGENDA

- Causality and counterfactuals
- Randomized experiments vs. observational studies
  - Example 1: Tennessee STAR
  - Example 2: California School districts
- Internal and external validity
- Vocabulary
- Takeaways

BIBLIOGRAPHY FOR TODAY´S CLASS

- Mosteller, M. (1995) (*)
- Stock and Watson, 1.2, 3.4, 9.1, 9.4 (**)
- Angrist and Pischke, 1 (pp. 39-46)

CAUSALITY

Econometrics is a set of statistical techniques for exploring empirical relationships between variables.

- "Empirical" means "based on data"

Relationships between variables means either:

- **Correlation** (or association):
  - If a variable X is correlated (or associated) with a variable Y, we shall see both variables "moving together" in the data
  - In X and Y are correlated (or associated), knowing the value or trajectory of X helps in *predicting* the value of Y

- **Causation**:
  - If X is causally related to Y, changing the value of X would lead to a change in the value of Y

Consider a plan to reduce Cambridge schools' class sizes by 20%. Why is it important whether the relationship between class size and student outcomes is causal or correlational?

Causality means that changes in a specific factor X lead to changes in an outcome Y:

- Y may be a function of many factors other than X
- We want to measure any changes in Y that are directly attributable to a change in X, not to these other factors
- Econometrics makes a lot of emphasis on isolating the impacts of X on Y from that of other variables (**identification strategy**)

We need to think in the **counterfactual**:

- What would have happened to Y in the absence of an intervention (if X did not occur)?
- We need to build a parallel world

Think about counterfactuals for these examples:

- Do industrial emissions X rise the temperature of a planet Y?

- Does getting an MPP from HKS X affect your job prospects Y?

Inferring causality is hard because we do not see counterfactuals! Instead, we "mimic" a counterfactual using statistical methods and data

RANDOMIZED EXPERIMENTS VS. OBSERVATIONAL STUDIES

Conceptually, randomized experiments are the "gold standard", the ideal method for estimating the causal effect of a "treatment"

- Randomized controlled trials (RCTs), randomized trials, random assignment studies, social experiments, etc.

In RCTs we compare two groups that are similar, except for the "treatment":

- A sample of people is drawn from the population
- People are randomly assigned to treatment or control group
- Treatment is administered to the treatment group
- Control group is not offered treatment, aiming at mimicking the counterfactual (what would have happened to the treated in the absence of treatment?)

What if you do not have a randomized experiment?

Observational data / Observational study:

- Researchers analyze data from a situation where they had no control
- Much of this course will be spent learning how to extract useful causal information from observation studies

As a starter – continuing with the class size sample – we will analyze two observational studies on the question of whether Cambridge schools should reduce its schools' class size:

- Experiment: Tennessee STAR
- Observational data: California School districts

TENNESSEE STAR EXPERIMENT

Design:

Execution:

- Below are mean characteristics of the two groups, with p-values from a t-test of the null hypothesis that the two means are equal:

	Small	Regular	Difference	p-value
Free Lunch (%)	47.2	48.5	-1.3	0.325
Mean (%)	51.5	51.3	0.2	0.883
African-American (%)	31.1	32.5	-1.4	0.302

- Did randomization "work"?

- Then explore the effects of class size on test scores in two steps:
  - **Estimation:** What is the difference in mean scores between the groups?

  - **Test hypothesis:** Are these means are significantly different?

- We want to test the null hypothesis that the two groups have the same mean test scores, or:
  - Ho: μS = μR ("Small" vs. "Regular").

- To test this, we need the t-statistic comparing the difference in sample means to the standard error of that difference:

$$t = \frac{\bar{Y}_S - \bar{Y}_R}{SE(\bar{Y}_S - \bar{Y}_R)} \qquad SE(\bar{Y}_S - \bar{Y}_R) = \sqrt{\frac{(\sigma_{Y_S})^2}{N_S} + \frac{(\sigma_{Y_R})^2}{N_R}}$$

- What is (in words) the p-value derived from calculated t-statistic?

EXPERIMENTAL DATA: TENNESSEE STAR EXPERIMENT

Stata output on test scores in regular (=0) and small (=1) classes:

small class in K	Summary of testscr Mean	Std. Dev.	Freq.
0	918.20133	72.214225	4048
1	931.94189	76.358633	1738
Total	922.32872	73.746597	5786

- What is the difference in mean test scores between groups?

- What is the t-statistic to test the hypothesis that the two groups have the same mean score?

- Can we reject the null hypothesis? Is the difference statistically significant at the 5% level? What does that mean (in words)?

Recap of confidence intervals: Instead of checking whether our sample is consistent with a specific value of the mean, we can construct the set of all values of u that are consistent with:

a) The null hypothesis
b) The observed mean

The confidence interval will contain the average value of the mean:

- 90% of the time: using sample mean value minus/plus 1.645 times the standard deviation
- 95% of the time: using sample mean value minus/plus 1.960 times the standard deviation
- 99% of the time: using sample mean value minus/plus 2.575 times the standard deviation

- What are the confidence intervals for the mean difference under the null hypothesis?

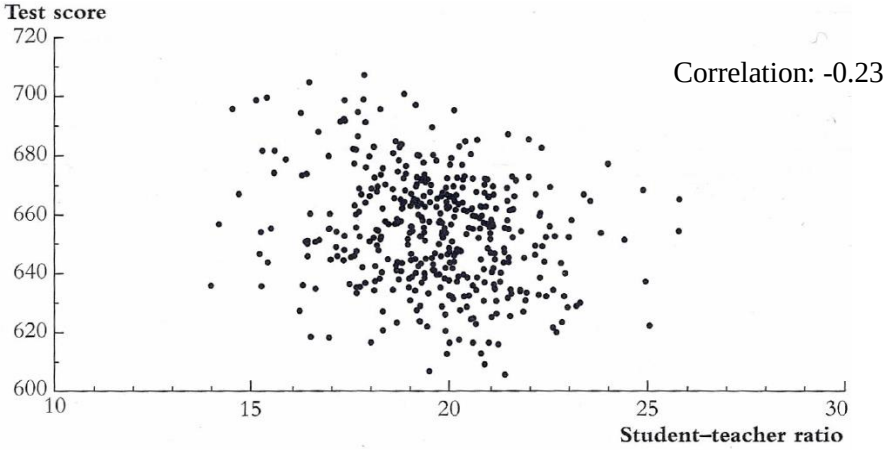
Confidence interval	t-value	Lower bound limit	Upper bound limit
90%	1.645		
95%	1.960		
99%	2.575		

- What are the confidence intervals for the mean difference under the alternative hypothesis?

Confidence interval	t-value	Lower bound limit	Upper bound limit
90%	1.645		
95%	1.960		
99%	2.575		

OBSERVATIONAL DATA: CLASS SIZES IN CALIFORNIA

- Consider an observational study from California:
  - Sample consists of 420 California School districts
  - For each district, we have a mean test score that will be our outcome Y and a measure of class size (student-teacher ratio or STR) that will be our "explanatory" variable X

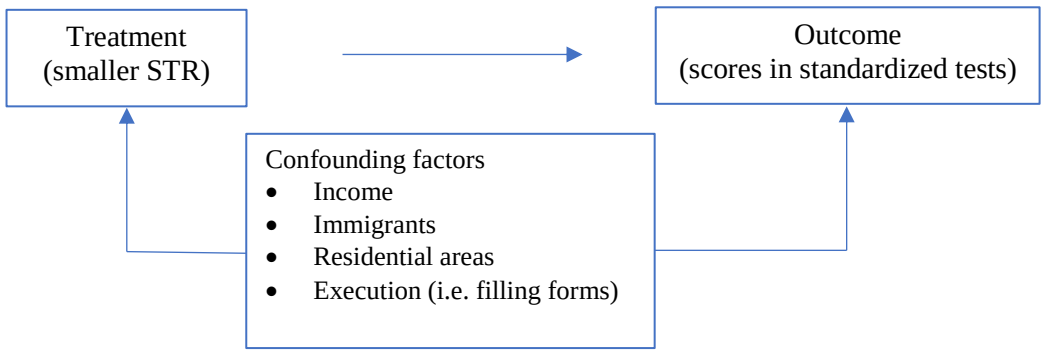


- What can we do with observational data?
  - We could arbitrarily define small (STR<20), regular (STR>=20).
  - Use estimation and hypothesis testing to see if districts with small classes had different scores than those with regular classes.
  - Stata output on test scores in regular (=0) and small (=1) classes:

small	Summary of testscr Mean	Std. Dev.	Freq.
0	649.97885	17.853364	182
1	657.35126	19.358012	238
Total	654.15655	19.053348	420

- Class size in California:
  - What is the difference in mean test scores between groups?
  - What is the t-statistic to test the hypothesis that both groups have the same mean score?
  - What is the confidence internal (95% level) for mean difference?
  - Can we reject the null hypothesis? Is the difference statistically significant at the 5% level? What does that mean (in words)?

- Straightforward comparison of means is naive:



- There might be differences – other than treatment – between these groups that may affect outcomes, but also the possibility of receiving treatment.
- Class sizes in California: What is wrong with that?
    - Below are mean characteristics of the two groups, with p-values from a t-test of the null hypothesis that the two means equal:

	Small	Regular	Difference	p-value
% receiving free lunch (low income)	41.6	48.7	-7.1	0.001
% English as a second language	12.5	20.0	-7.5	<0.001
Average income (in 000' of $)	16.3	14.0	2.3	<0.001

    - What do p-values tell you about these two groups of districts?

III – INTERNAL AND EXTERNAL VALIDITY

- Validity of research is established following a set of criteria that we will study in this class
- No valid/invalid research, rather shades of validity

Internal Validity:	A study is internally valid if statistical inference on causal effects are valid for the *population and settings studied*. **Answers:** Does the relationship capture the causal effect of interest for the population represented in the sample?
External Validity:	A study is externally valid if inferences made on causal effects for the *population and settings* studied can generalized to other populations and settings (*the population of interest*). **Answers:** Are findings generalizable to other places, times, people? Do other empirical studies yield similar results?

How would you describe the **internal validity** of:

- The Tennessee START experiment?

- The California observational study?

Threats to external validity:

- Differences in population:
    - Differences in baseline characteristics between the population studied and the population of interest.
    - Study is out of date (*"that used to be that way, but is not anymore"*)
- Difference in the setting of the study: *Yes, those populations are similar, but the conditions prevailing in the place studied are totally different than those prevailing in the place of interest*.
    - Difference in the institutional environment.
    - Difference in law.
    - Difference in the physical environment where study was carried.

How would you describe the **external validity** of:

- The Tennessee START experiment?

- The California observational study?

So, let's run experiments?

- Experiments are rare in social science

  … they are *expensive*: STAR costed US$2.5 million per year x 4 years = US$12,0 million

  … they are often perceived as *unethical* . Parents consider it unfair if their child is randomly put in a large class which may hinder his career

  … they are (almost) never ideal, which limits *internal validity*

  … they are (often) not representative, which puts *external validity* under question

IV – TAKEAWAYS

- Distinguishing correlation from causation requires careful identification of the counterfactual
- **Randomized experiments** are ideal to do this, but the specific context and execution that spur high **internal validity might restrict external validity**
- In **observational studies**, difference in mean outcomes between groups is usually not a good estimate of **causal effect** because there are other differences (other than treatment) between groups affecting outcome (scores)
- Econometrics provides statistical techniques that try to isolate the causal effect of X on Y when we have observational data:
  - Bivariate regression does not do this.
  - Multiple regression does, to some extent.
  - Quasi-experimental techniques do even better, if done right.

Vocabulary from Randomized Experiments:

- Control group: the group that does not receive the treatment or intervention in an experiment
- Treatment group: the group that receives the treatment or intervention in an experiment
- Counterfactual: the counterfactual measures what would have happened to beneficiaries in the absence of the intervention
- Average treatment effect: a measure used to compare treatments, measures the difference in mean (average) outcomes between units assigned to treatment and units assigned to control
- Internal validity: a statistical analysis is internally valid if the statistical inferences about causal affects are valid for the population being studied
- External validity: the analysis is externally valid if its inferences and conclusions can be generalized from the population and setting to other population and settings
- Causal effect: the expected effect of a given intervention or treatment as measured in an ideal randomized controlled experiment
- Identification strategy: the manner in which a research uses observational data to approximate a real experiment