

EC 1152 - Using Big Data to Solve Economic and Social Problems

Review Session #2
TF: Diana Goldemberg

Prof: Raj Chetty
Harvard University
Spring 2019

Outline

- Project #1
 - Stata crash workshop
- Absolute Mobility and conditional probability
 - Also check the handout
- Causal Effects
 - Randomized Experiments
 - (teaser) Quasi-Experiments

Logistics

- Office Hours:
 - Wed 4.30-6.30, Barker 103
- Any outstanding issues with sectioning?
- Remember, can always submit questions & anonymous feedback before sections using this Google form [<https://goo.gl/forms/RAOQFBIj6SXdFOZJ3>]
- Find this prez and files at: https://github.com/dianagold/Ec1152_diana
Or at GoogleDrive: <http://bit.ly/ec1152drive>

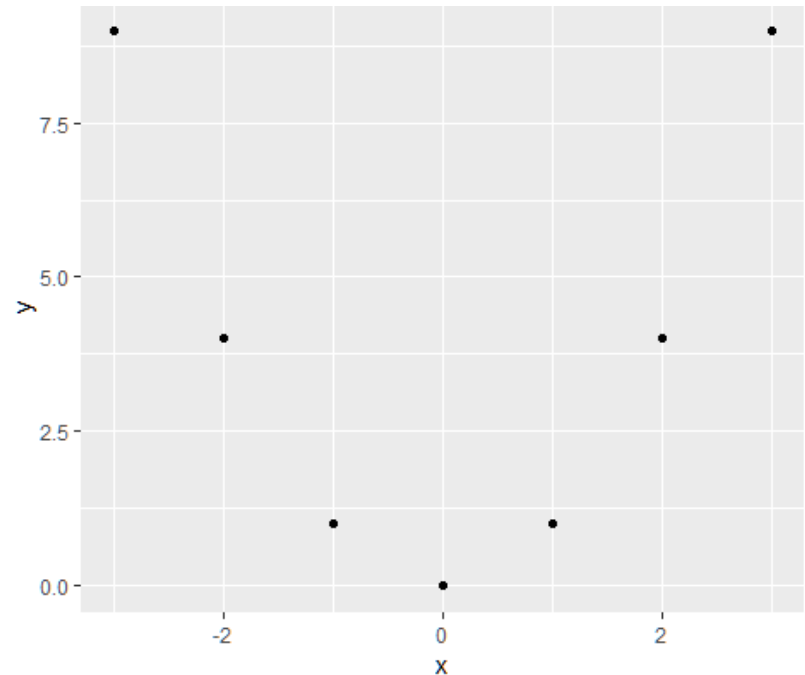
Project #1

- First project is due Thursday, 2/21 on Canvas
- Some clarifications (see also the Canvas discussion)
 - Your 4-6 page narrative should address all of the main questions from the project document, but you do **not** need to include every figure you create in the document.
 - For example, when creating scatterplots and tables of mobility by race, choose a couple to discuss in the narrative, and just include the rest of them in your code/logs so that we can see you generated them.
 - You do not need to number each question as you answer them—we will be able to tell. Please structure your response as an essay, but make sure to answer all of the questions.

Project #1

- Correlation coefficients measure **linear** correlation:

X	Y
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9



- Are X and Y related? What do you think the correlation coefficient is?
- Takeaway:** Zero correlation does not mean X and Y are unrelated!
- Covariates = Other variables that are (potentially) related to a variable of interest, typically included as an independent variable in a regression.

Stata hands-on demo

- *Stata will be used in section*
- *But you're very welcomed to follow the Jupyter notebook for Python and the hints at the end of the Assignment for the R commands*
- *All files at: https://github.com/dianagold/Ec1152_diana*

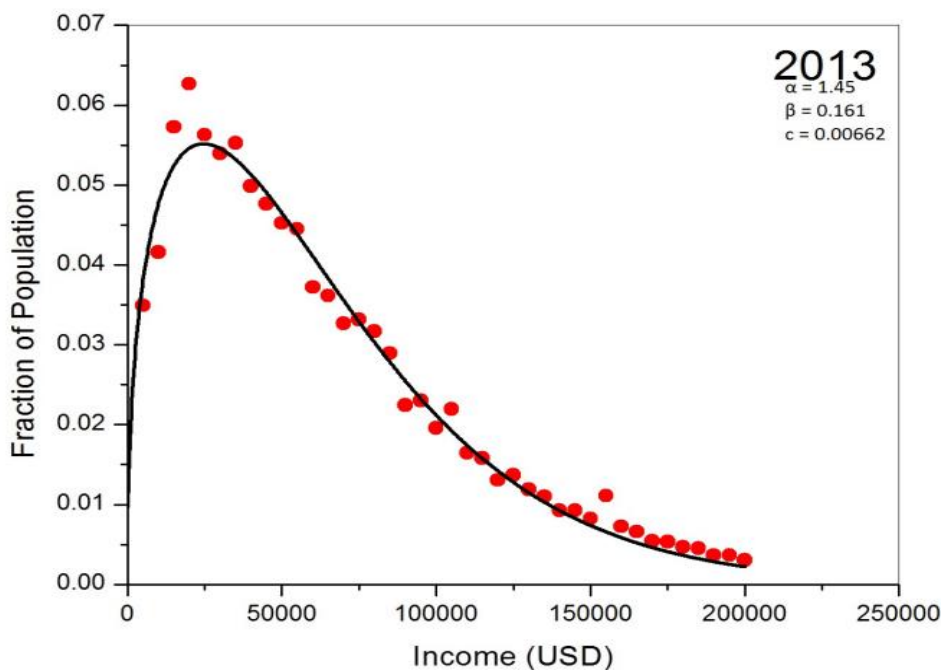
Absolute Mobility

Absolute Mobility

- Chetty et al (2017) **absolute mobility**: What fraction of children have a higher standard of living than their parents did?
- Three pieces of information needed:
 - Average parent income at age 30 at each percentile rank of the income distribution.
 - Average child income at age 30 at each percentile rank of the income distribution.
 - Joint distribution (“copula”) of parent and child income ranks
- This is enough information to calculate absolute mobility!

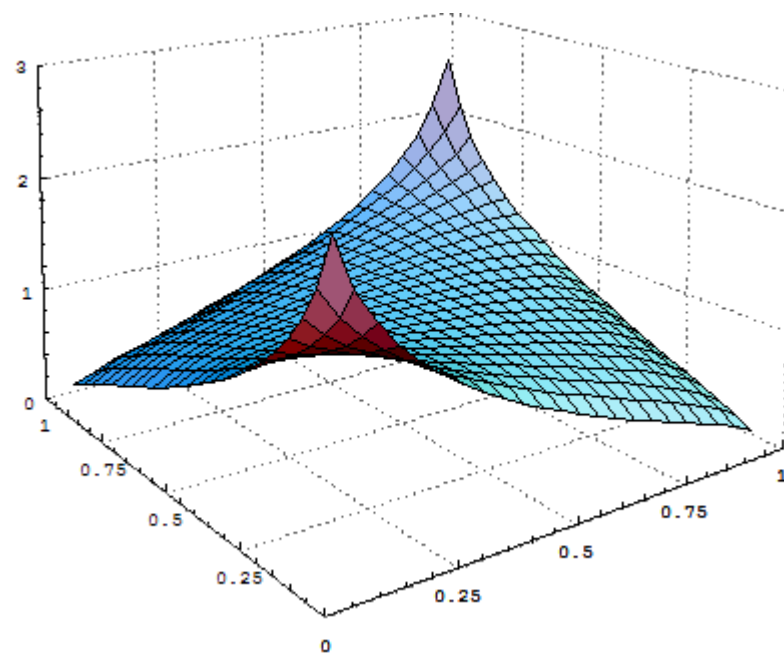
Wait, What's a Joint Distribution?

Last Week: probability distribution functions (PDFs) in one variable



Graph shows probability $X = \text{some value}$, e.g. $P(X = 5)$

This Week: **joint distribution** functions in two variables

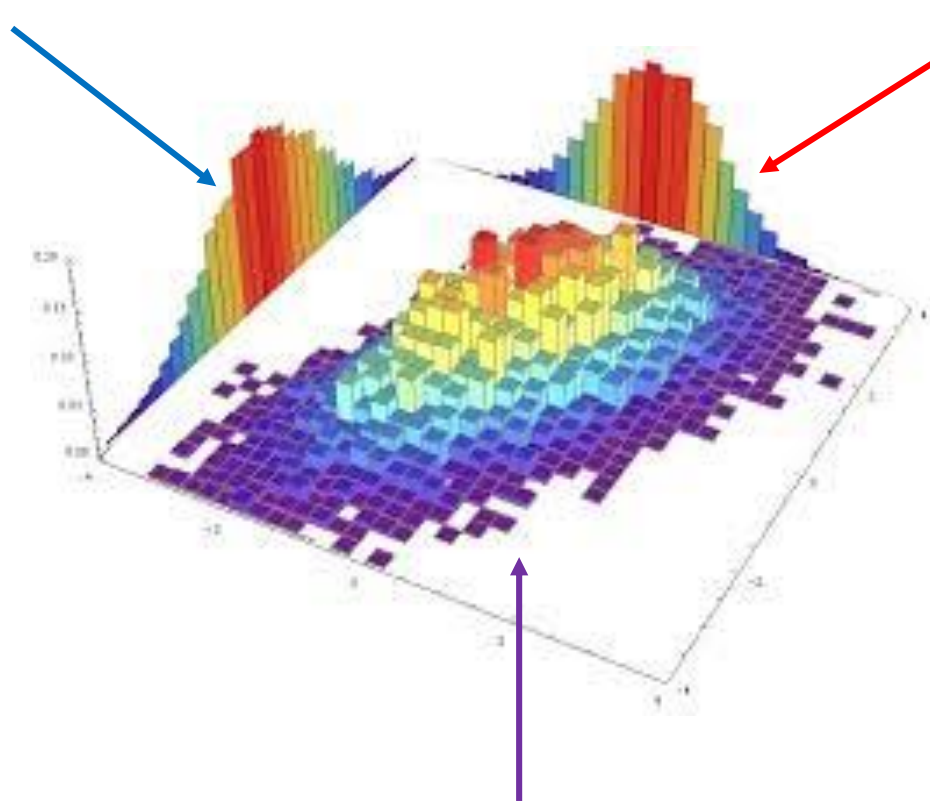


Graph shows probability $X = \text{some value}$ **AND** $Y = \text{some value}$, e.g. $P(X=5, Y=10)$

Wait, What's a Joint Distribution?

Marginal distribution: $P(X = x)$

Marginal distribution: $P(Y = y)$



Joint distribution: $P(X = x, Y = y)$

Wait, What's a Joint Distribution?

- May also be interested in the probability of X occurring, given that Y has occurred.
 - For example, probability a kid ends above the 80th percentile of income given that their parent was below the 20th percentile.
- We call this **conditional probability**, and we can write it as:
 - $\Pr(K_rank > 80 \mid P_rank < 20)$
 - “The probability that Kid rank > 80 given that Parent rank < 20”
 - “The probability that Kid rank > 80 conditional Parent rank < 20”
- This is related to the **joint probability**:
 - $$\Pr(K_rank > 80 \mid P_rank < 20) = \frac{\Pr(K_r > 80 \text{ and } P_r < 20)}{\Pr(P_r < 20)}$$

Absolute Mobility: Example Calculation

- Suppose you knew that child income had the following quintiles:

Rank	20	40	60	80
Income	20, 514	38,008	62,734	94,563

- And similarly, for parent income:

Rank	20	40	60	80
Income	26,764	43,290	58,235	76,847

- The joint rank distribution for Milwaukee, WI

		Parent Rank				
		< 20	20 - 40	40 - 60	60 - 80	> 80
Child Rank	< 20	.058	.037	.029	.028	.024
	20 - 40	.046	.036	.035	.041	.031
	40 - 60	.025	.031	.041	.056	.045
	60 - 80	.013	.024	.041	.066	.065
	80 - 100	.007	.017	.035	.072	.097

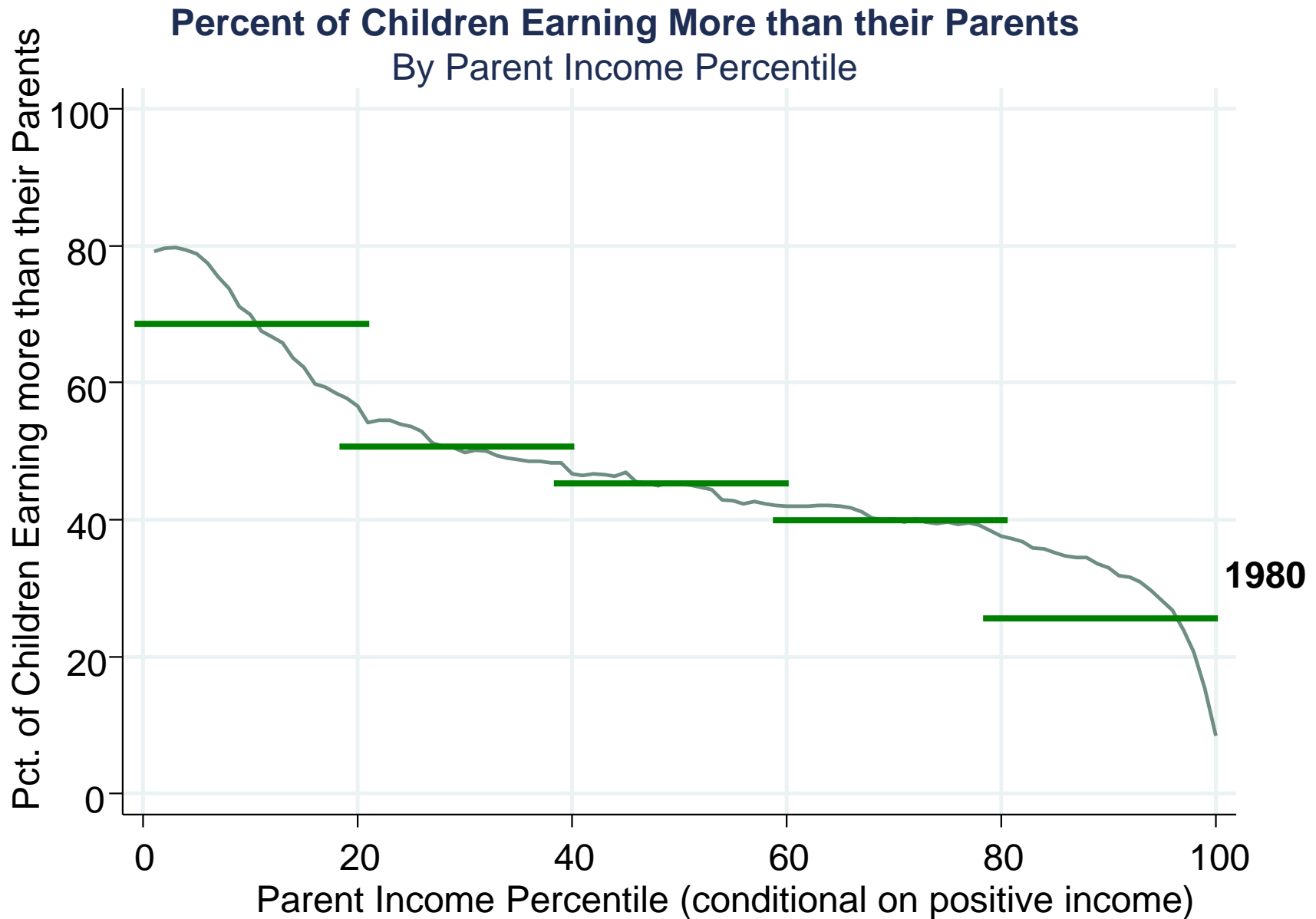
Absolute Mobility: Example Calculation

- Can fill in every box this way:

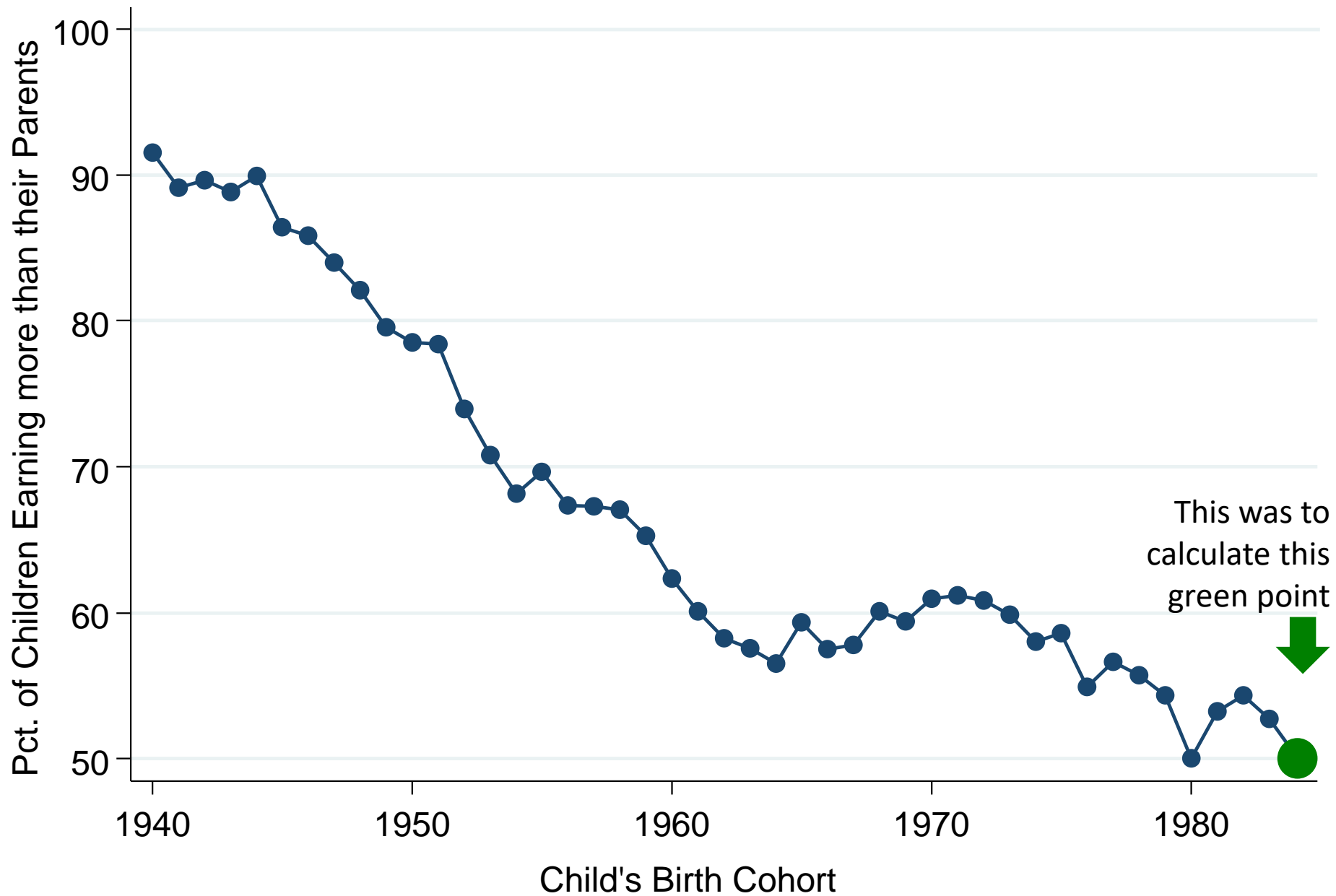
		Parent Rank				
		< 20	20 - 40	40 - 60	60 - 80	> 80
Child Rank	< 20	.058	.037	.029	.028	.024
	20 - 40	.046	.036	.035	.041	.031
	40 - 60	.025	.031	.041	.056	.045
	60 - 80	.013	.024	.041	.066	.065
	80 - 100	.007	.017	.035	.072	.097

- Can you calculate exact absolute mobility?
- What about a lower bound? Absolute mobility is at least...
 - Add up green boxes = .234
- What about an upper bound? Absolute mobility is at most...
 - Add up green and yellow boxes = .73
- Big range! Real paper uses percentiles to make it smaller.

Absolute Mobility: Example Calculation



Mean Rates of Absolute Mobility by Cohort




Causal Effects

Causal Effects

- Last week: correlation is not causation.
- This week: What is causation?

cau·sa·tion

/kô'zāSH(ə)n/ 

noun

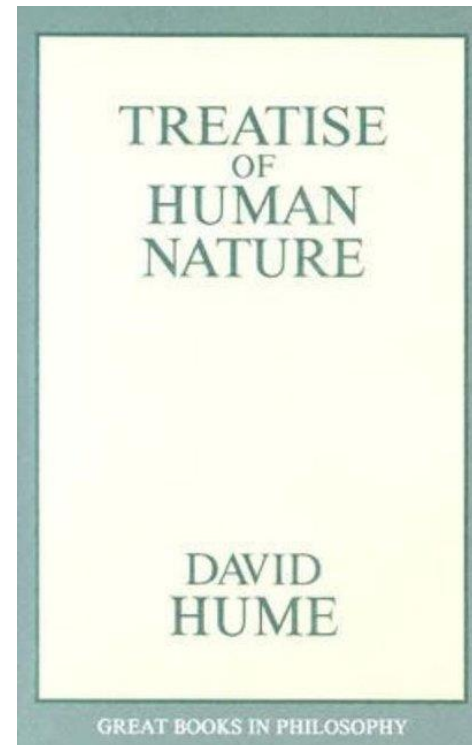
the action of causing something.

"investigating the role of nitrate in the causation of cancer"

- the relationship between cause and effect; causality.

plural noun: **causations**

- Philosophically difficult question!



Causal Effects

In this class, and in economics and social science more generally, causal effects contrast the **factual** outcome and its **counterfactual**, often intimately linked to (real or hypothetical) experiments.

- “If Jane would have gotten one more year of education, how would her wages be different?”
- “If the Fed were to raise interest rates, how would unemployment change relative to if the Fed did not raise interest rates?”
- “If a child were to grow up in Minneapolis instead of Atlanta, how much more likely is it that she would become an inventor?”

Causal Effects: Counterfactual

Notice in all the hypotheticals the idea of situations or people being “otherwise identical”. The **counterfactual** represents the state of the world in the absence of the policy/program you want to evaluate.

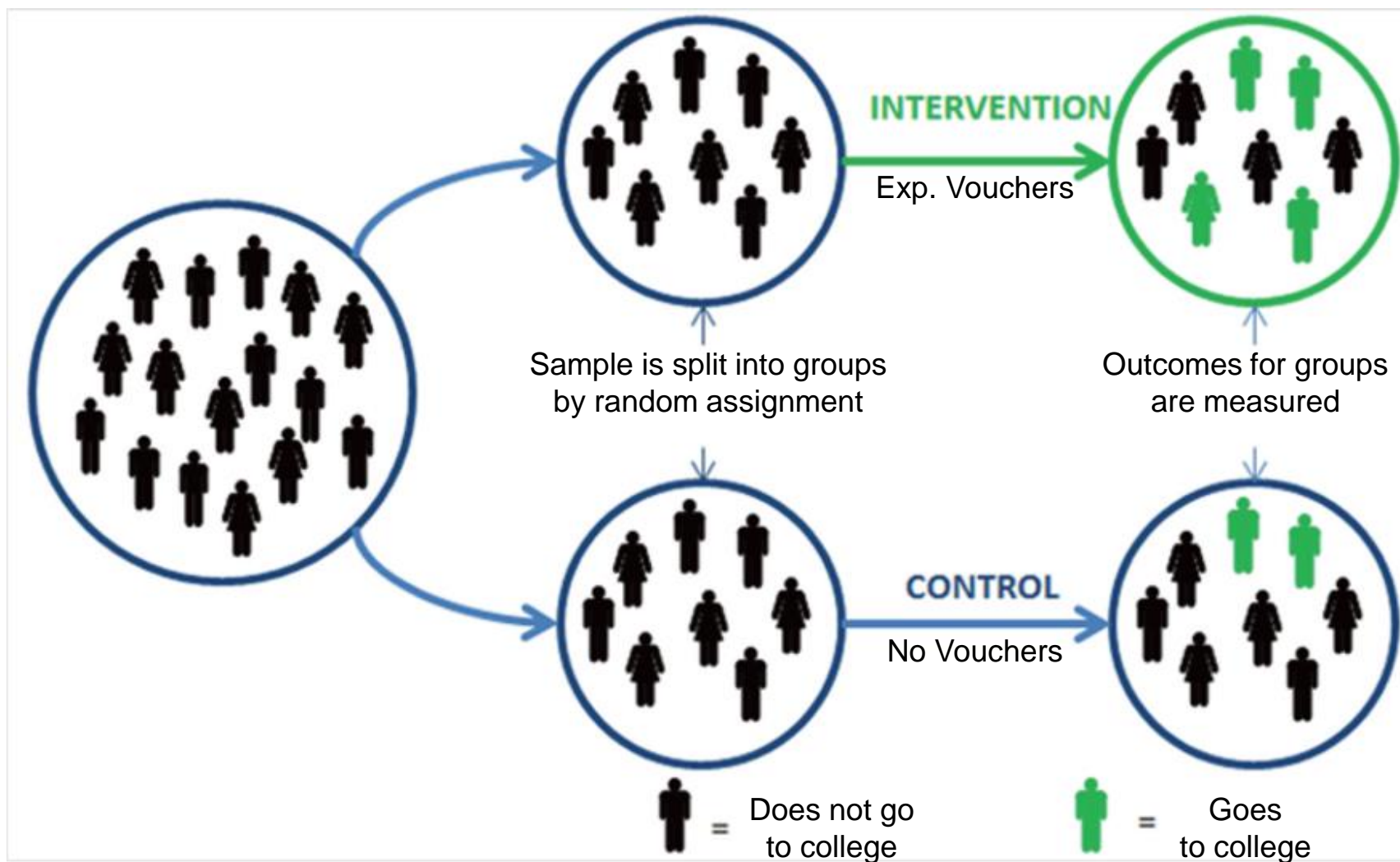
- Jane vs. Jane with one more year of education
- The US economy vs. the US economy with slightly higher interest rates

Problem: counterfactuals can not be directly observed (can’t compare a person to themselves)

Solution: we need to “mimic” or construct a credible counterfactual

- What’s wrong with comparing participants and non-participants?
- Better ways to achieve this?

Causal Effects: Randomized Experiments



Note: Moving to Opportunity had two treatment (intervention) arms and one control, this is a simplification.

Causal Effects: Randomized Experiments

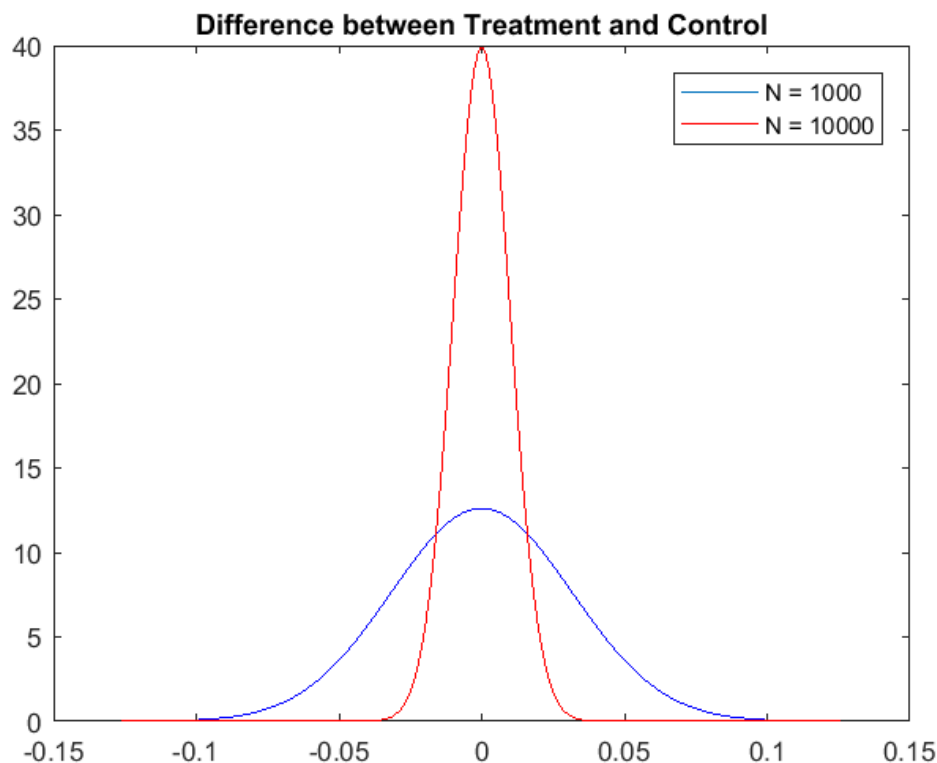
Randomized Experiments are the “Gold Standard” for creating comparable/“otherwise identical” groups. Why?

- **Key Point 1:** Random assignment ensures that, at the outset of the experiment, members of the groups (treatment and control) **do not differ systematically**.
- **Key Point 2:** Differences between the groups, which are solely due to chance, decrease with sample size.
- **Key Point 3:** Thus, any difference that subsequently arises between them can be **attributed to the intervention** rather than to other factors.

Causal Effects: Randomized Experiments

Illustrating **Key Point 2**

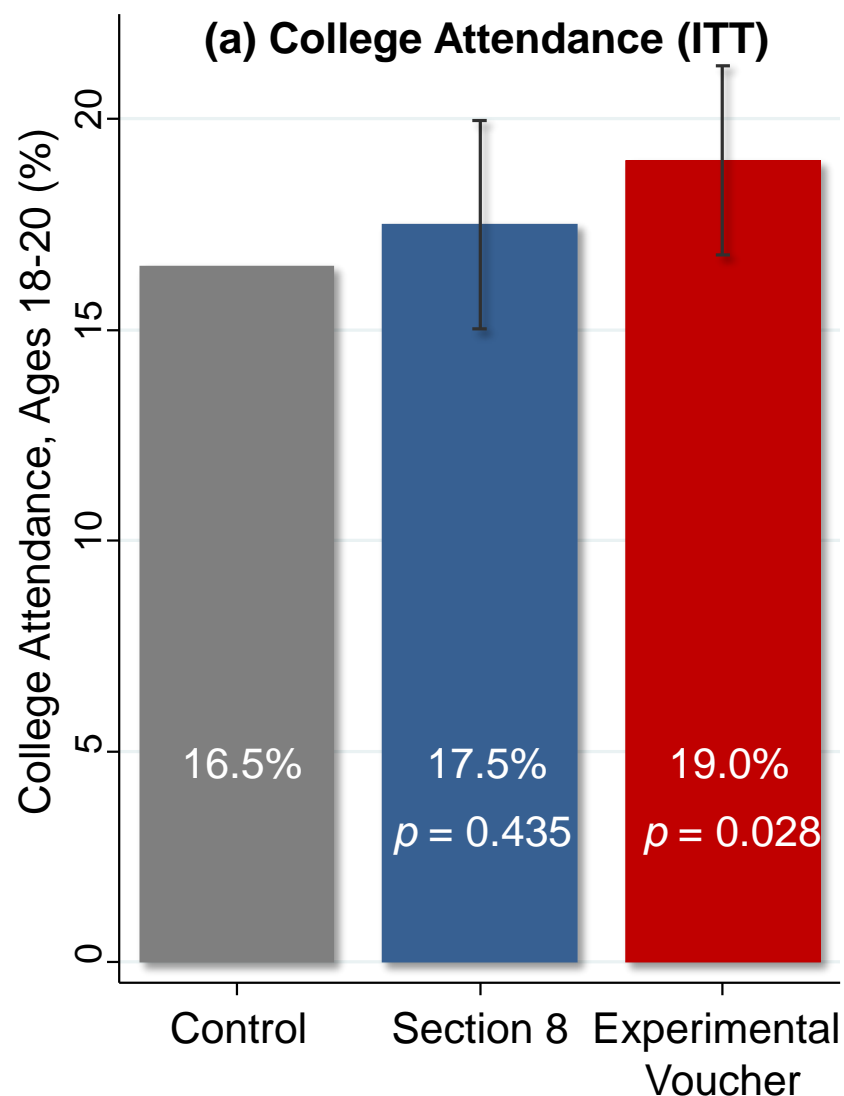
- Population of 50% female, 50% male. Suppose you randomly assign N people, half to treatment, half to control.
- Below is the (approximate) distribution of the difference between the fraction of the treatment group that is male and the fraction of the control group that is male, for $N = 1,000$ vs. $N = 10,000$:



Causal Effects: Randomized Experiments

Illustrating **Key Point 3**

Impacts of MTO on Children Below
Age 13 at Random Assignment



Randomized Experiments: Limitations

- What are some limitations of randomized experiments? Consider the example of attempting to randomly assign people to move to opportunity.
 - Non-compliance: I don't use my voucher
 - **Attrition**: I stop talking to the researchers/disappear.
 - Hawthorne effects: I know I'm in the treatment group and want to do really well.
 - John Henry effects: I know I'm in the control group and want to do really well (maybe because I want to show the researcher that my neighborhood should be more respected).
 - **Small samples/Too expensive**
 - **Non-scalability**: Can't move everyone to opportunity, and can't be sure what would happen if we did.

Randomized Experiments: Limitations

- Last Two Slides, Summarized:
 - Randomized experiments are the Gold Standard from estimating causal effects.
 - Randomized experiments can be expensive, non-generalizable, and feature many pitfalls.
- So what can we do?
 - Big data helps solve/eliminate attrition: just use administrative records.
 - Quasi-experimental methods (stay tuned!!!)