

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
- Propensity Score Reweighting

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>

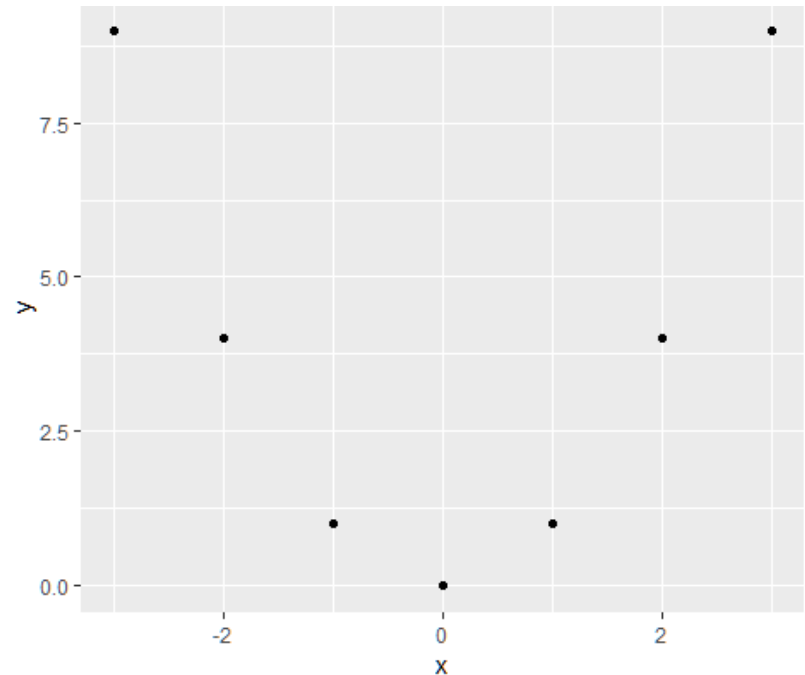
Project #1

- 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*

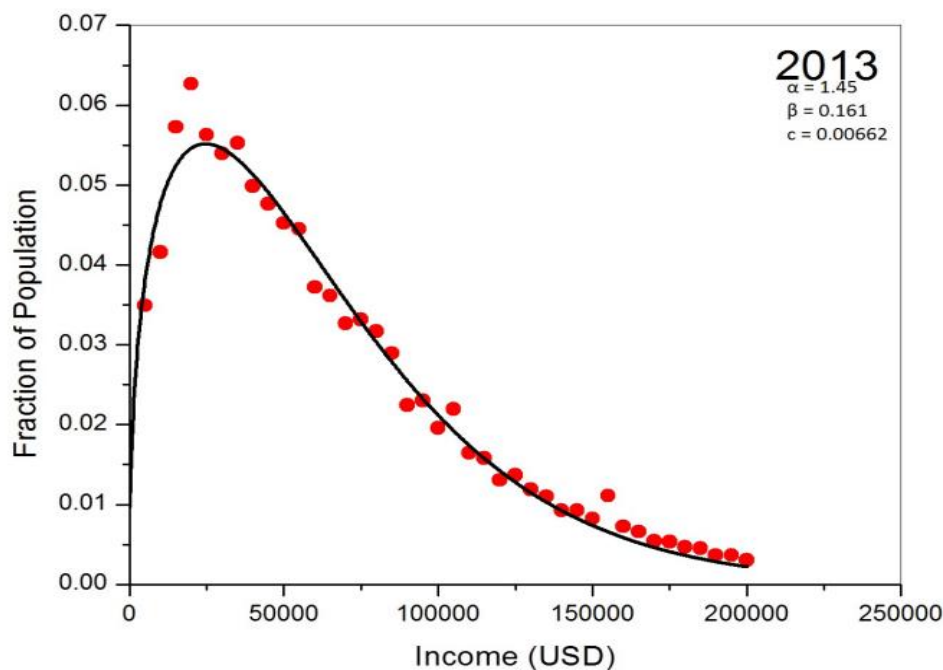
Absolute Mobility

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!

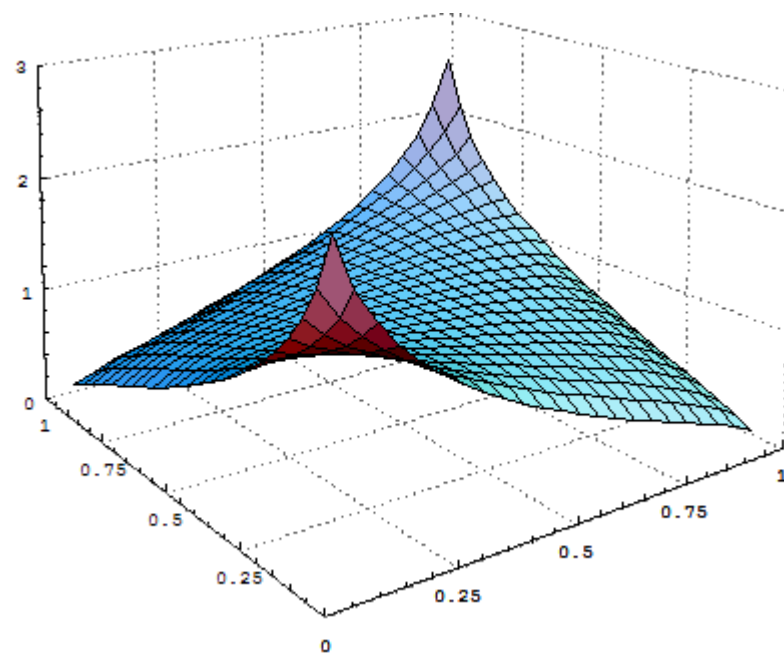
Wait, What's a Joint Distribution?

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

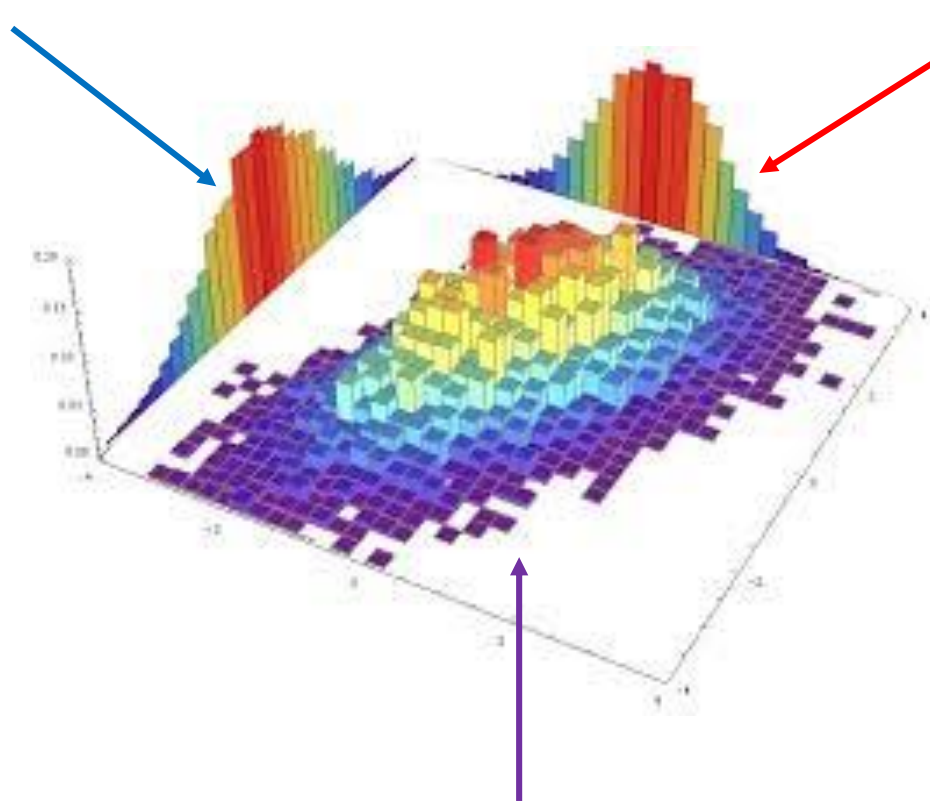


Graph shows probability $X = \text{some value}$ **AND** $Y = \text{some value}$, e.g. $P(X=5, Y=10)$

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

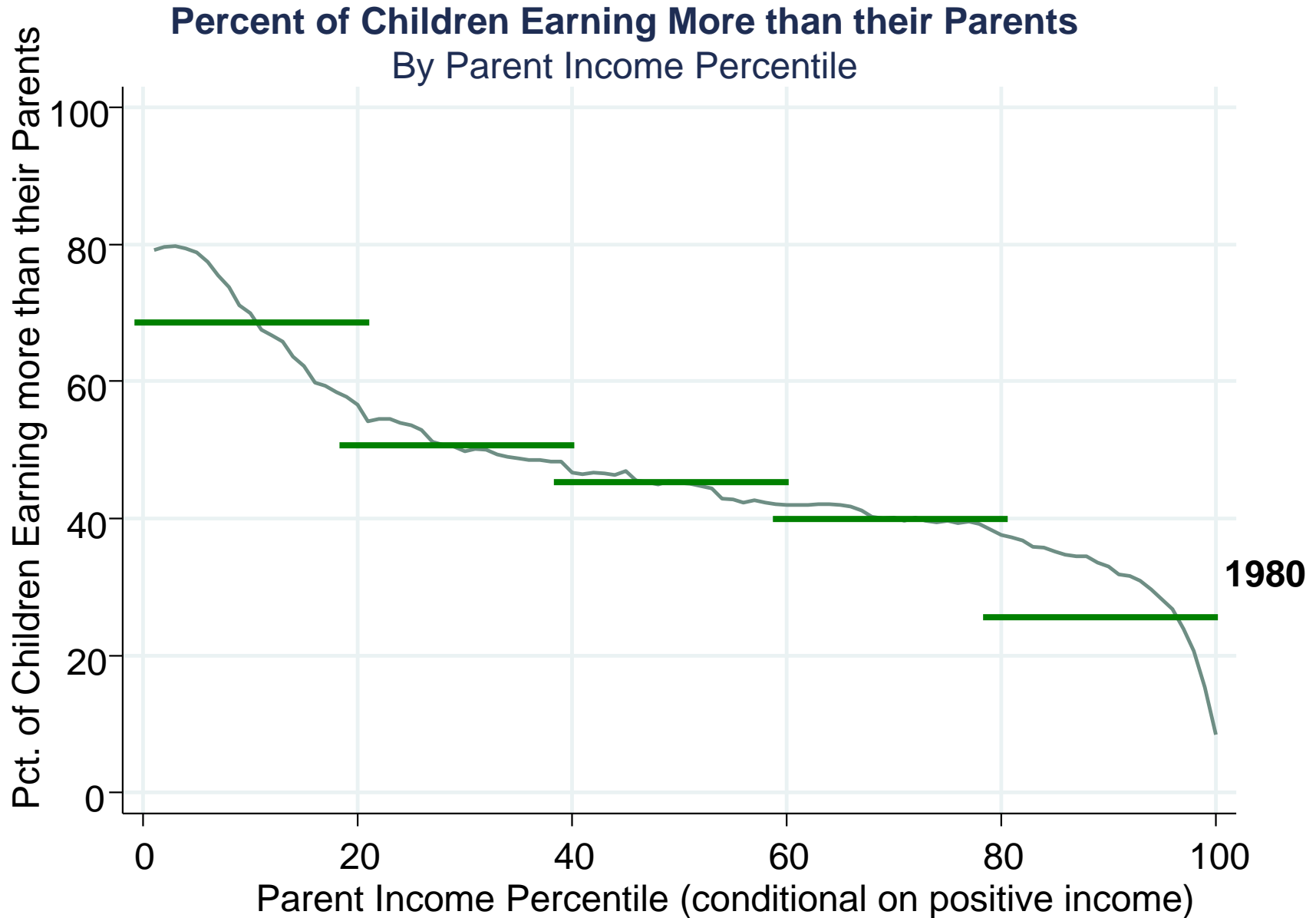
Absolute Mobility: Example Calculation

- Can fill in every box this way:

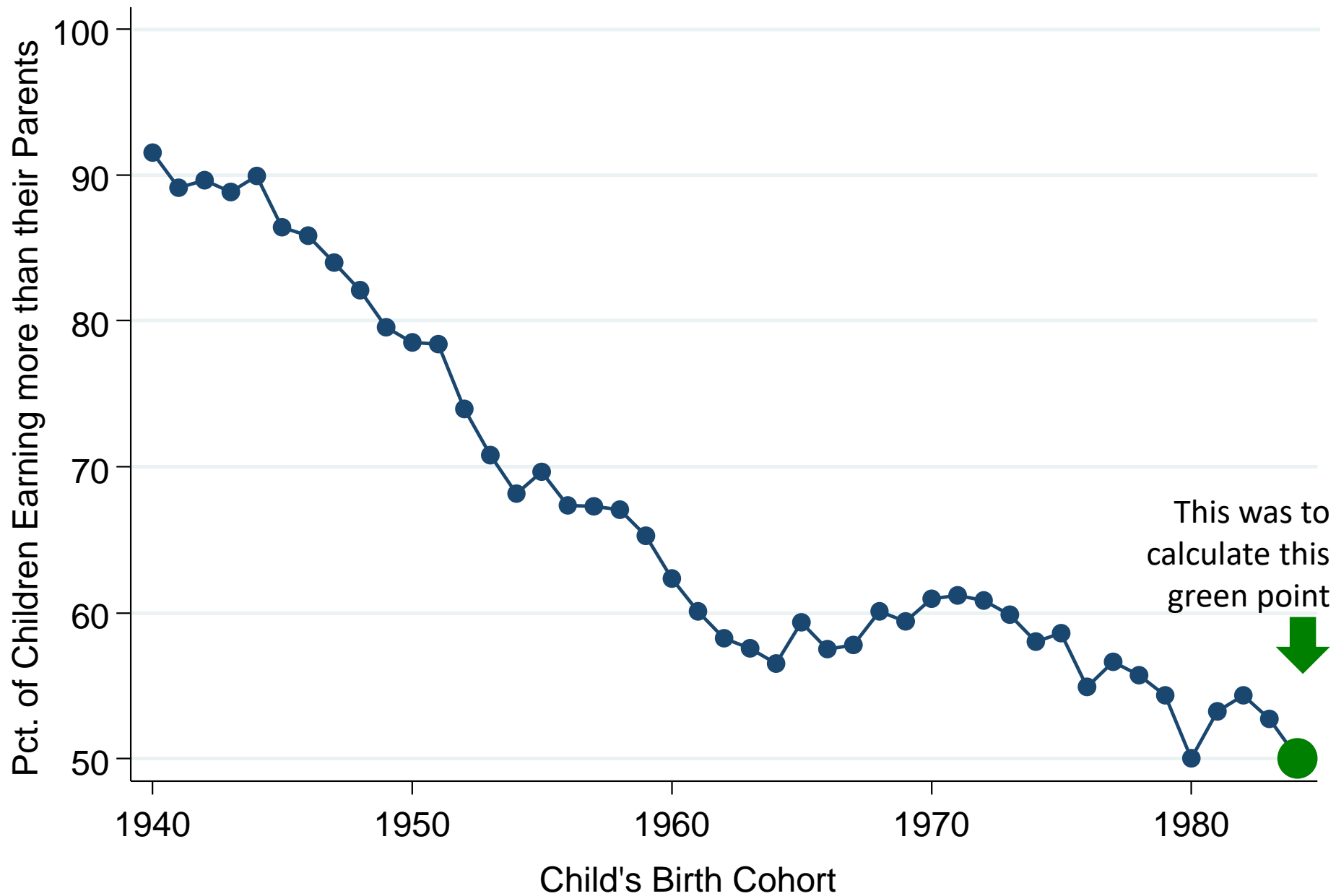
		Parent Rank				
		< 20	20 - 40	40 - 60	60 - 80	> 80
Child Rank	< 20	.058	.037	.029	.028	.024
	20 - 40	.046	.036	.035	.041	.031
	40 - 60	.025	.031	.041	.056	.045
	60 - 80	.013	.024	.041	.066	.065
	80 - 100	.007	.017	.035	.072	.097

- 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



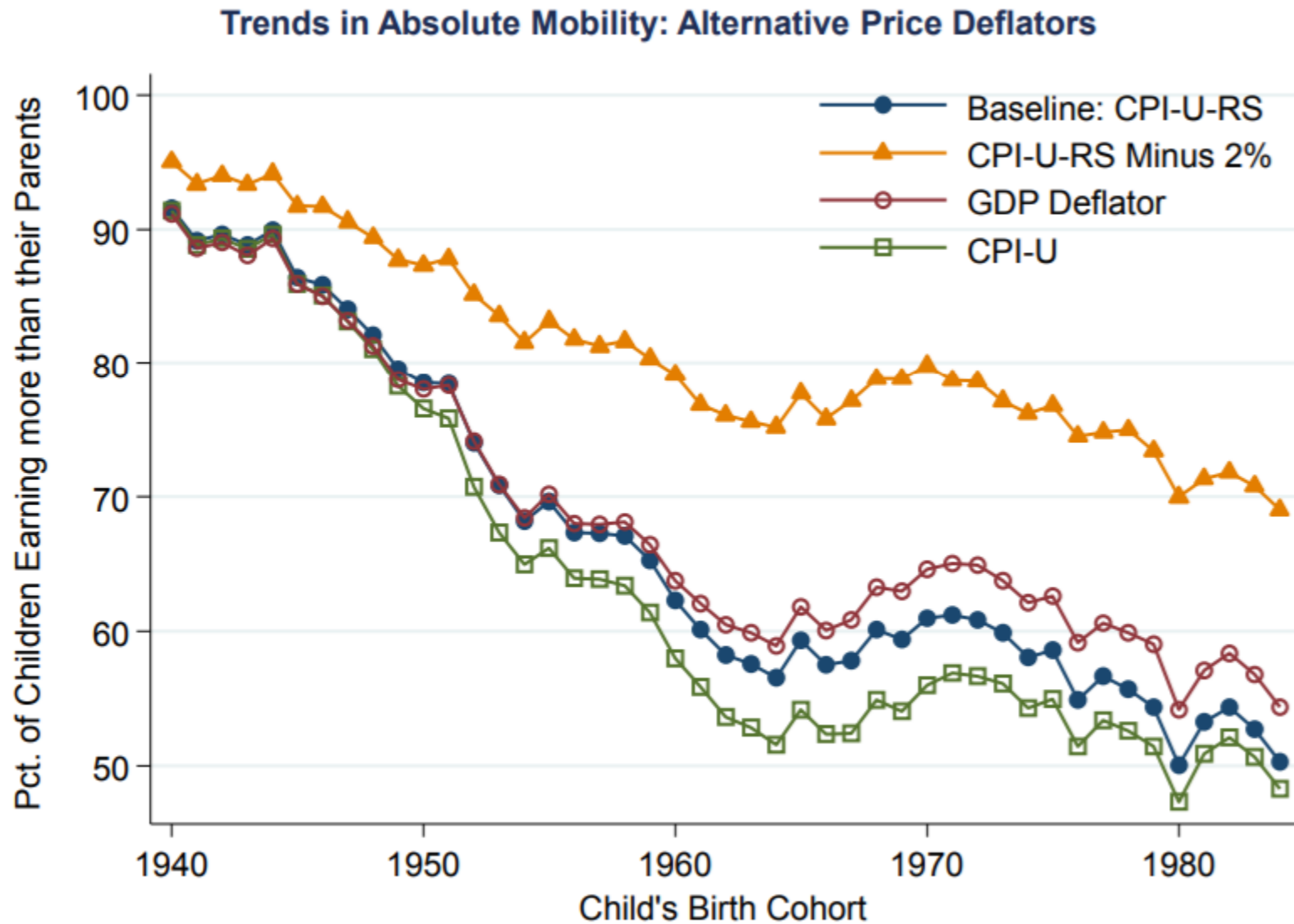
Mean Rates of Absolute Mobility by Cohort



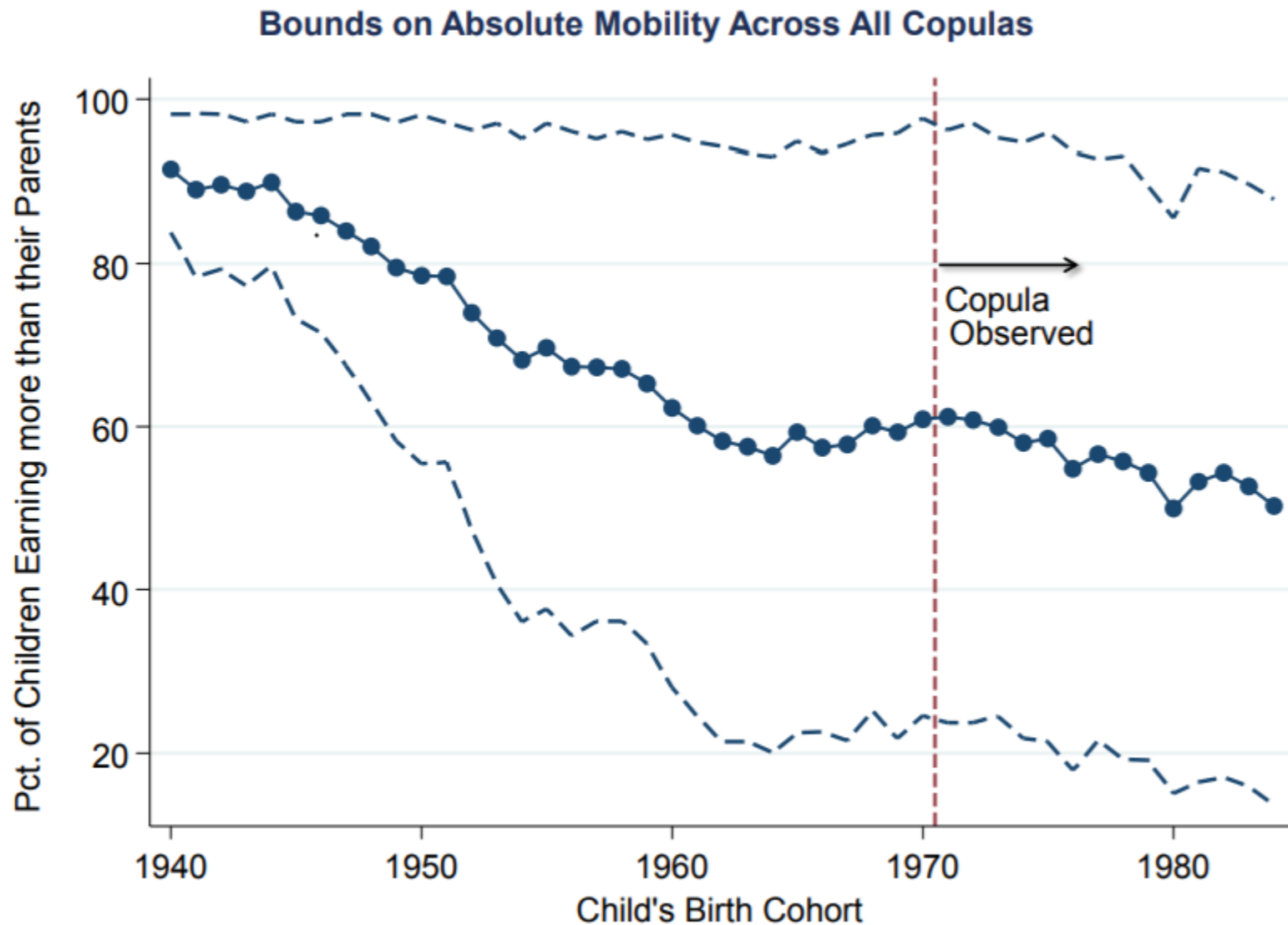
Absolute Mobility: Potential Issues?

- What are some possible issues threatening the validity of absolute mobility estimates?
- Some debate over the proper inflation adjustment
 - Solution: Present results under a variety of inflation adjustments, general pattern is fairly robust.
- Do not observe the parent/child rank joint distribution “copula” before 1971.
 - Solution A: Assume stability, i.e. that it was the same then as it is now.
 - Solution B: Derive bounds on the measure of mobility under different possible joint distributions. Once back to 1940-1950, it doesn't matter much.

Absolute Mobility: Potential Issues?

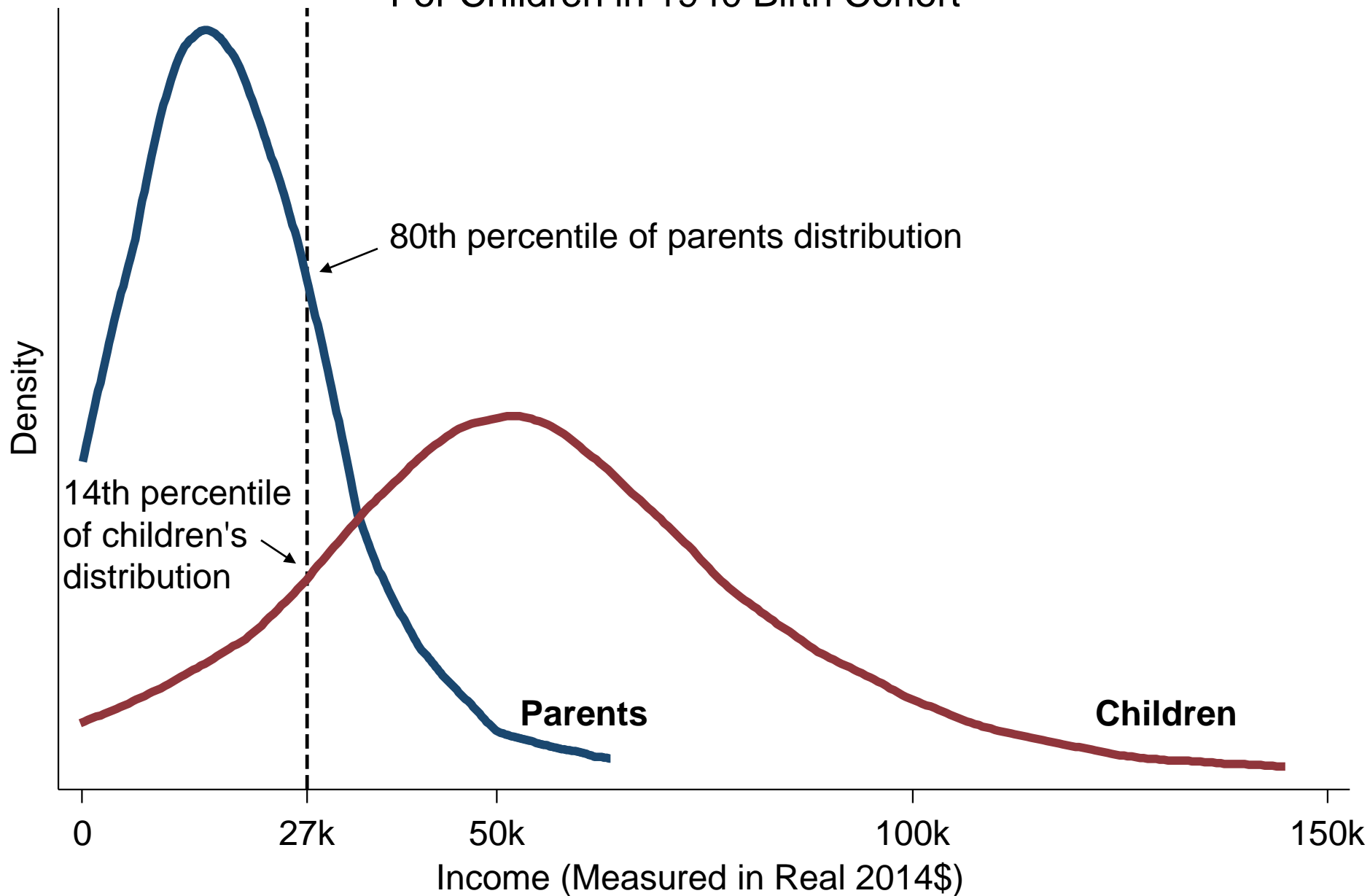


Absolute Mobility: Potential Issues?



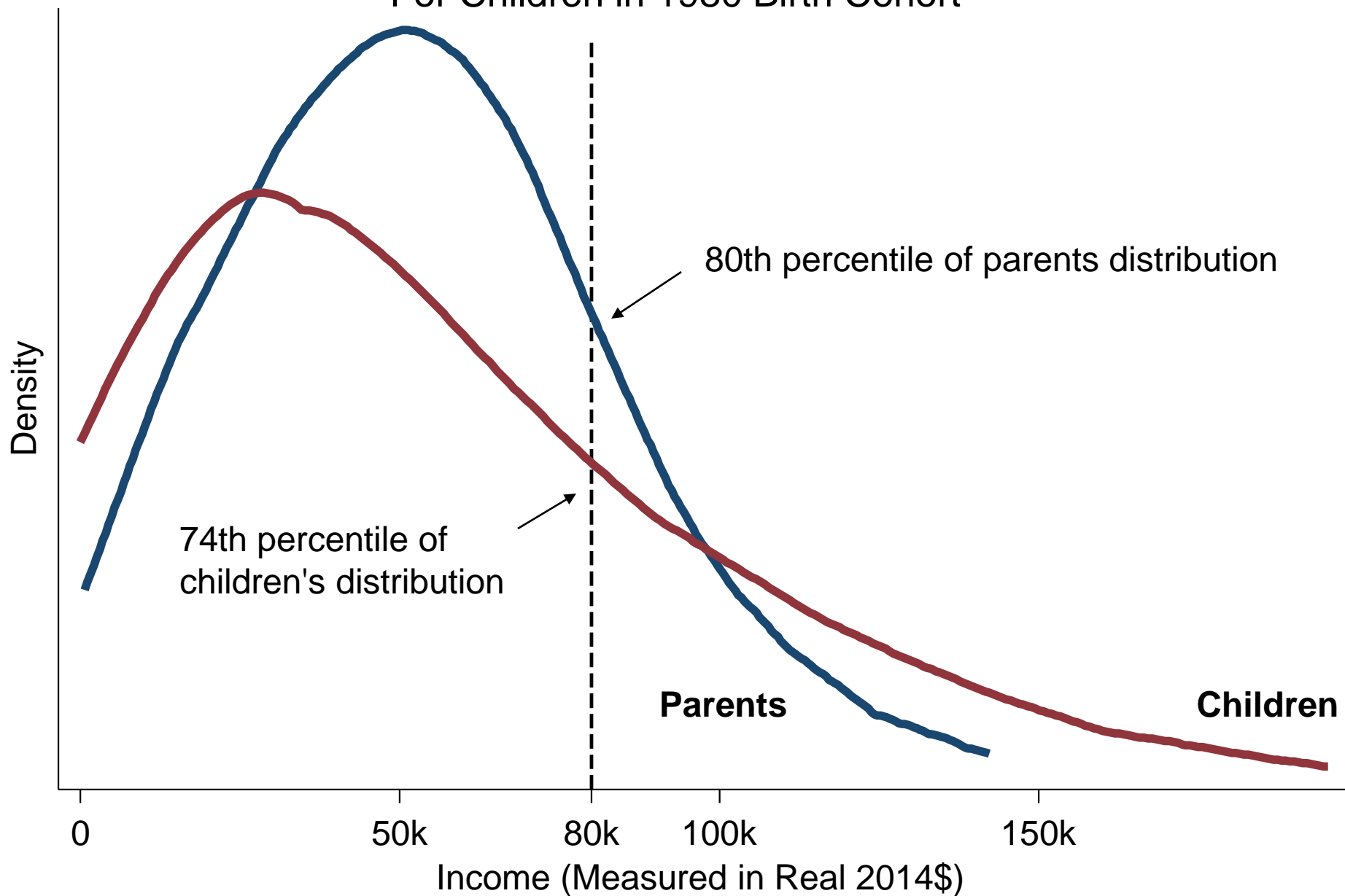
Household Income Distributions of Parents and Children at Age 30

For Children in 1940 Birth Cohort

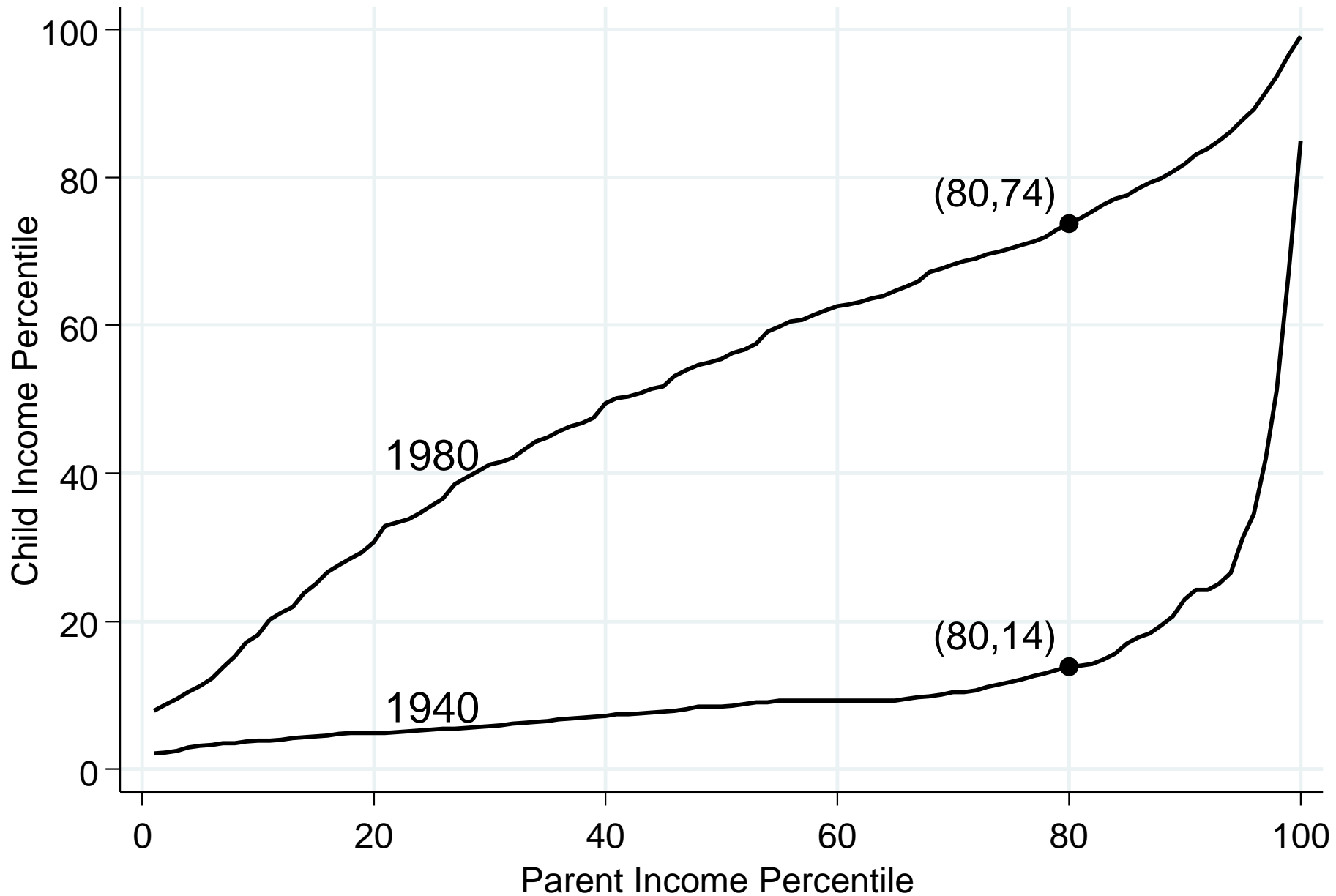


Household Income Distributions of Parents and Children at Age 30

For Children in 1980 Birth Cohort



Child Rank Required to Earn More Than Parents




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**

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: Partial Equilibrium vs. General

- Think precisely about what causal effect you are interested in, in terms of the corresponding hypothetical experiment.

“What is the impact of moving to opportunity on future wages?”



“If we take one person, and move them from a low-opportunity neighborhood to a high-opportunity neighborhood, how will their future wage path be different than if we had not moved them?”

Partial Equilibrium Effects

“If we move everyone in Seattle from a low-opportunity neighborhood to a high-opportunity neighborhood, how will the distribution of wages be different than if we had not moved them?”

General Equilibrium Effects

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.

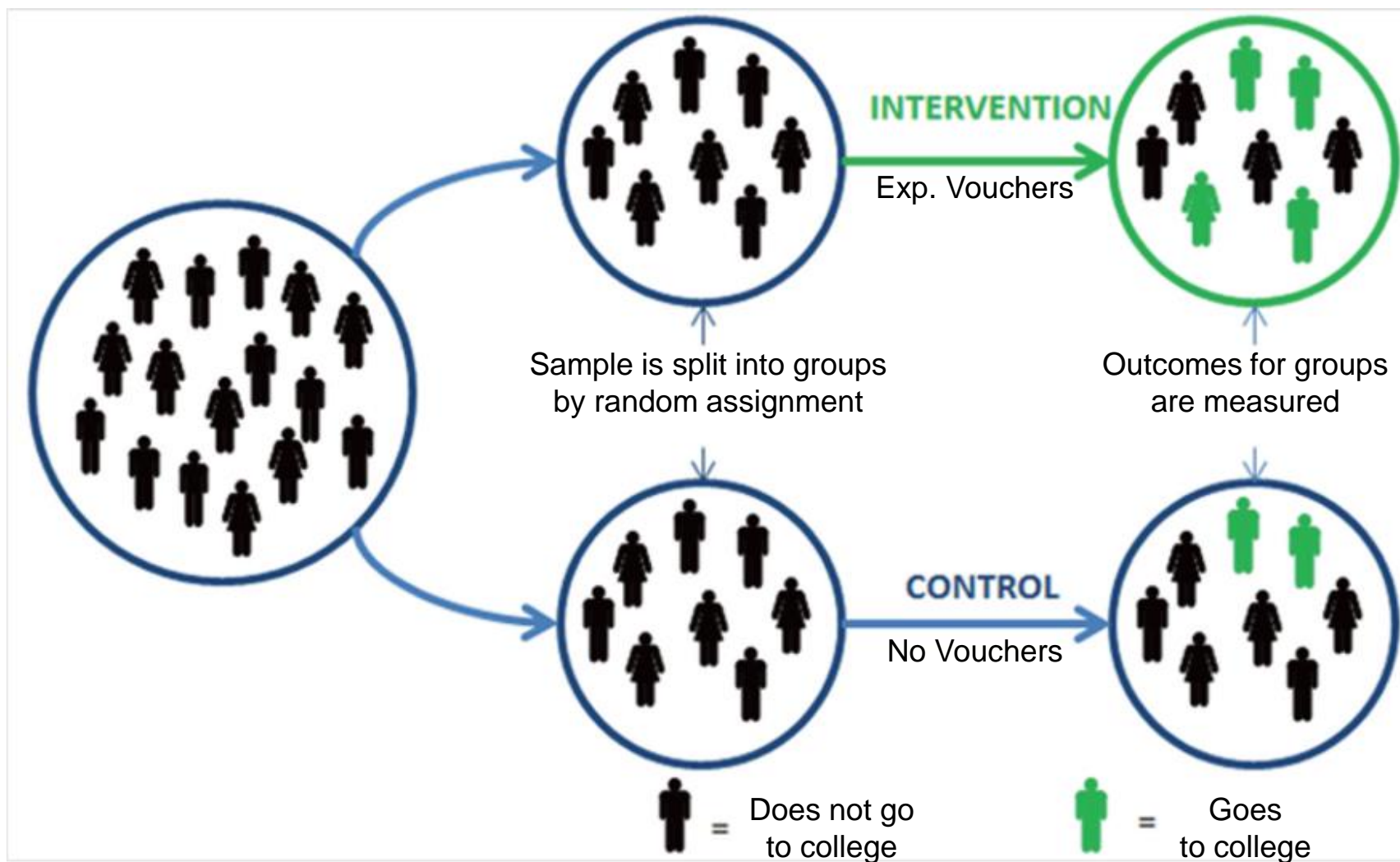
- Jane vs. Jane with one more year of education
- The US economy vs. the US economy with slightly higher interest rates

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. $N = 10,000$:

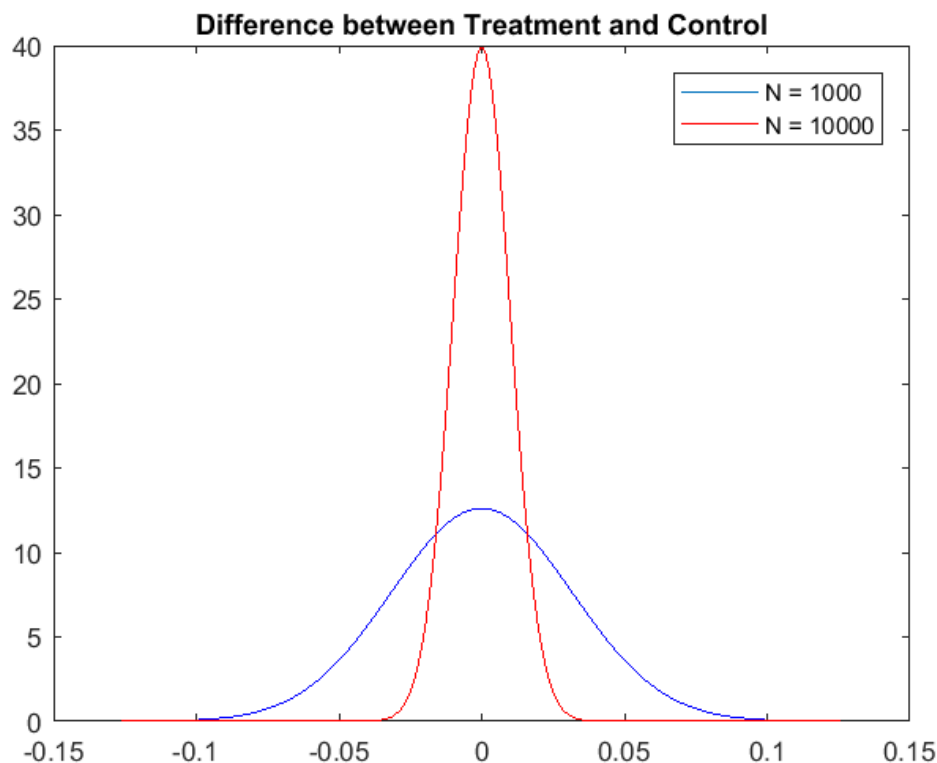


TABLE 1
Summary Statistics and Balance Tests for Children in MTO-Tax Data Linked Sample


	< Age 13 at Random Assignment			Age 13-18 at Random Assignment		
	Control Grp. Mean (1)	Exp. vs. Control (2)	Sec 8. vs. Control (3)	Control Grp. Mean (4)	Exp. vs. Control (5)	Sec 8. vs. Control (6)
Linked to tax data (%)	86.4	-0.8 (1.4)	-0.4 (1.5)	83.8	1.5 (2.0)	-0.1 (2.2)
Child's age at random assignment	8.2	-0.1 (0.1)	-0.0 (0.1)	15.1	0.1 (0.1)	-0.1 (0.1)
Household Head Completed High School (%)	34.3	4.2 ⁺ (2.4)	0.4 (2.6)	29.5	5.0 (3.1)	0.7 (3.3)
Household Head Employed (%)	23.8	1.0 (2.1)	-2.2 (2.2)	25.3	3.0 (2.9)	-0.4 (3.0)
Household Head gets AFDC/TANF (%)	79.5	0.6 (1.9)	1.8 (2.0)	75.0	-0.8 (2.9)	-1.0 (3.0)
Household Head never married (%)	65.1	-4.3 ⁺ (2.3)	-3.1 (2.6)	53.0	-3.1 (3.2)	-6.3 ⁺ (3.4)
Household Head had teenage birth (%)	28.6	-0.9 (2.2)	-0.3 (2.5)	29.1	-3.6 (2.9)	-2.5 (3.2)
N. of Children in Linked MTO-Tax Data	1613	1969	1427	686	959	686

[there were
more lines
in here]

Notes: This table presents summary statistics and balance tests for match rates and a subset of variables collected prior to randomization; Appendix Table 1a replicates this table for all 52 control variables we use in our analysis. The estimates in the first row (fraction linked to tax data) are based on all children in the MTO data who were born in or before 1991. The estimates in the remaining rows use the subset of these observations successfully linked to the tax data. Columns 1-3 include children below age 13 at random assignment; Columns 4-6 include those above age 13 at random assignment. Columns 1 and 4 show the control group mean for each variable. Columns 2 and 5 report the difference between the experimental voucher and control group, which we estimate using an OLS regression (weighted to adjust for differences in sampling probabilities across sites and over time) of each variable on indicators for being assigned to the experimental voucher group, the section 8 voucher group, as well as indicators for randomization site. Columns 3 and 6 report the coefficient for being assigned to the section 8 group from the same regression. The estimates in Columns 2-3 and 5-6 are obtained from separate regressions. Standard errors, reported in parentheses, are clustered by family (* = $p < 0.10$, * = $p < 0.05$, ** = $p < 0.01$). The final row lists the number of individuals in the control, experimental, and section 8 groups in the linked MTO-Tax data sample.

MTO: Estimating Treatment Effects

- Regression specifications:

$$y_i = \alpha + \beta_E^{ITT} Exp_i + \beta_S^{ITT} S8_i + s_i \delta_s + \epsilon_i$$


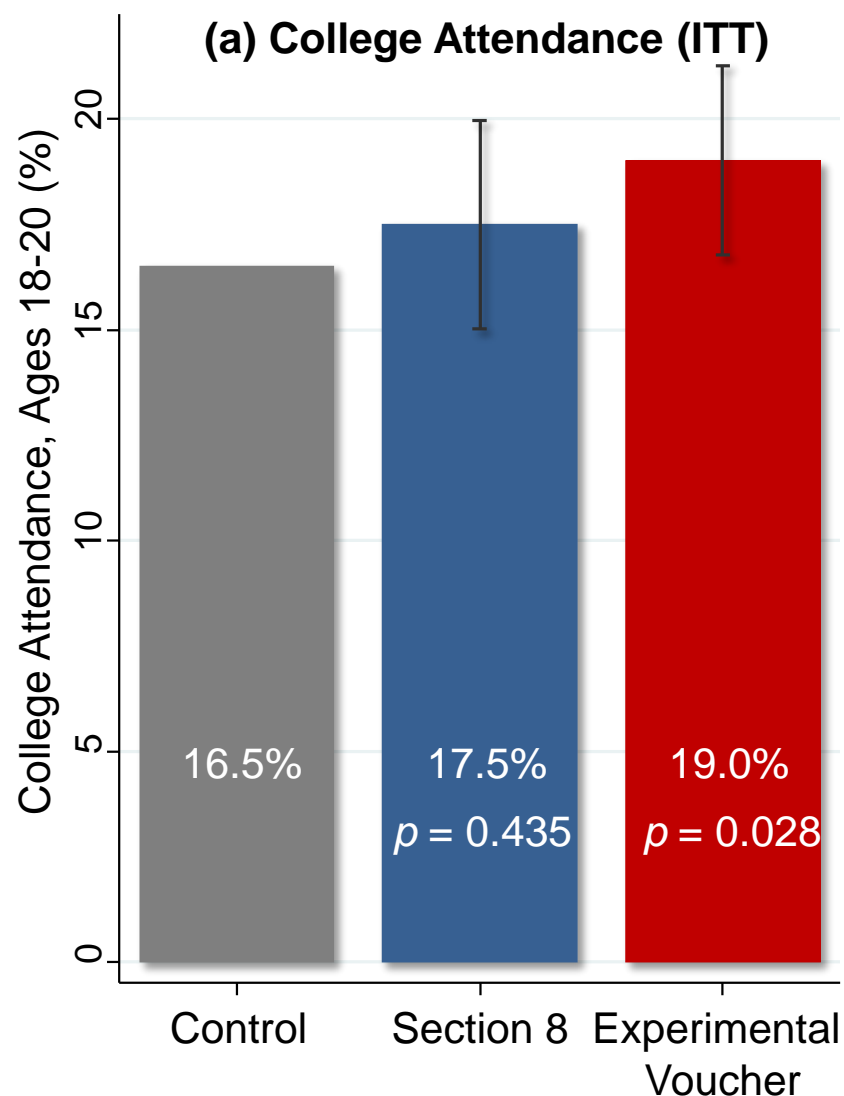
The diagram illustrates the components of the regression equation. Two blue arrows originate from the text 'Treatment Indicators' and point to the variables Exp_i and $S8_i$. Another blue arrow originates from the text 'Site Indicators' and points to the term $s_i \delta_s$.

- These intent-to-treat (ITT) estimates identify effect of being *offered* a voucher to move through MTO
- From ITT to treatment-on-treated (TOT) estimates, needs to take into account voucher take-up (for young children: 48% for Exp and 66% for S8)

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.

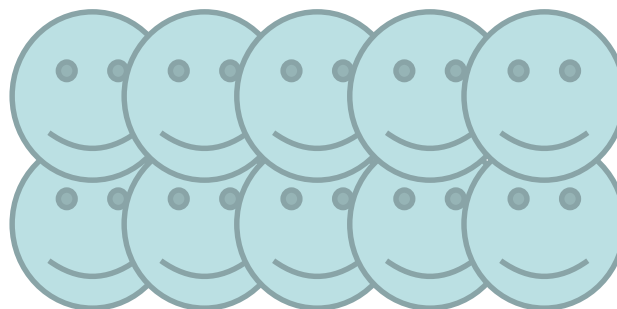
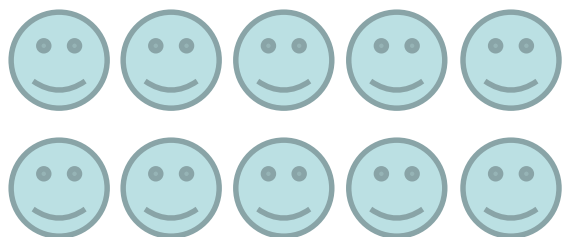
Randomized Experiments: Limitations

- 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!!!)

Propensity Score Reweighting

Propensity Score Reweighting

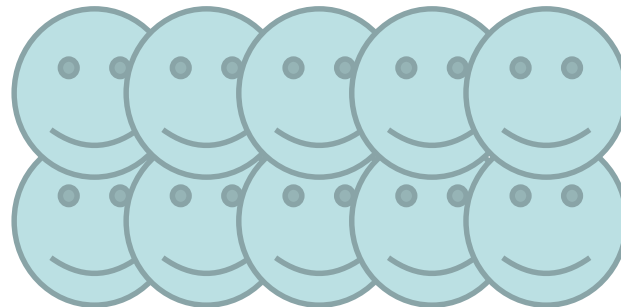
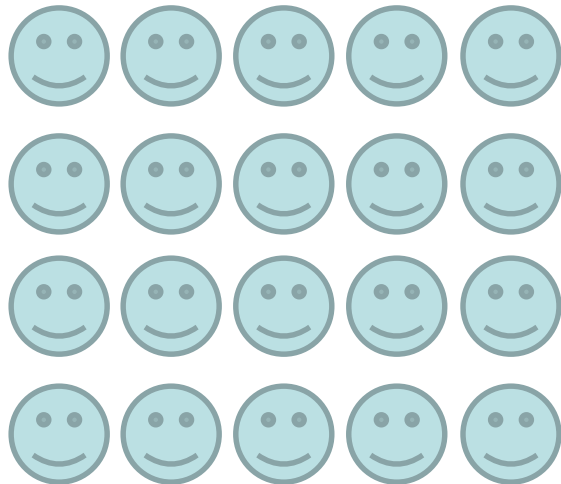
- One tool from econometrics that can help when we don't have a random experiment.
- Example: Suppose people are either tall or short, and wages are a function only of your height
 - Tall people make 2 dollars a day
 - Short people make 1 dollar per day
- Initially we have 10 tall people and 10 short people working.



- Average Income = $[(10)*1 + (10)*2]/20 = \mathbf{1.50 \text{ per day}}$

Propensity Score Reweighting

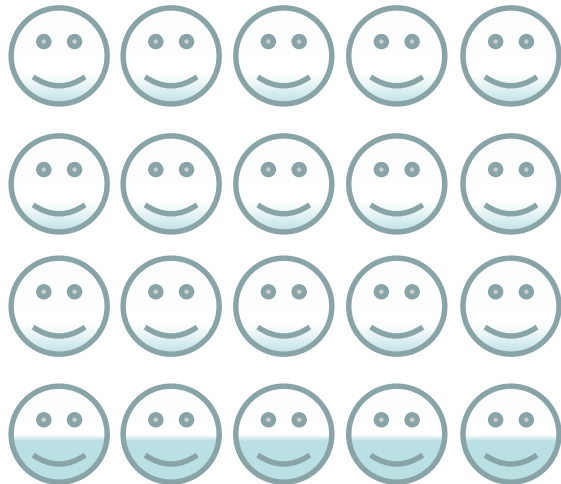
- Improve Earned Income Tax Credit
 - Tall people make 2.10 dollars a day
 - Short people make 1.20 dollar per day
- More short people enter the workforce: Now have 20 short people and 10 tall people.



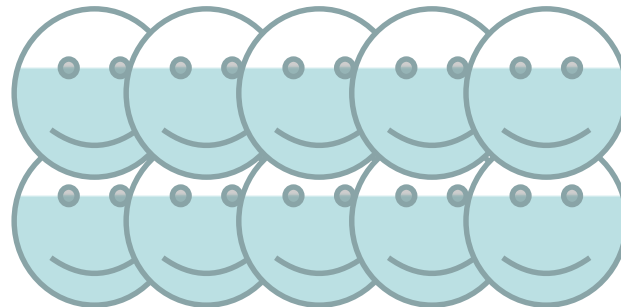
- Average Income = $[1.20(20) + 2.10(10)]/30 = \mathbf{1.50 \text{ per day}}$
 - Is the program a failure? Why or why not?

Propensity Score Reweighting

- **Key Point:** The group before the change and the group after the change were not “otherwise identical.” They had different distributions of height!
- **Propensity Score Reweighting Approach:** Weight the people after the change such that the distribution of height after the change matches the distribution of height before the change.



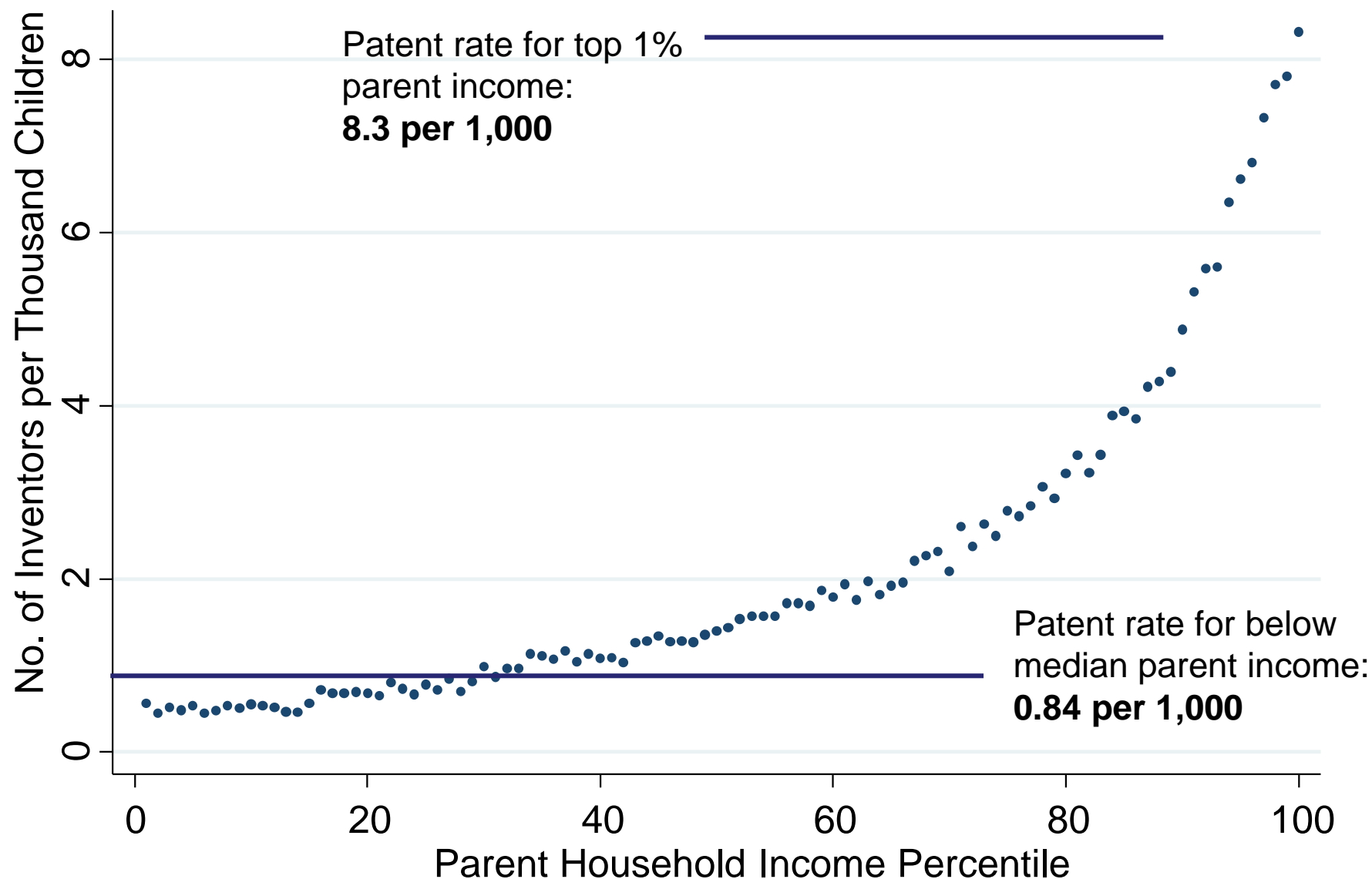
Optimal weights: 1/3 weight on short people,
2/3 weight on tall.



- Re-weighted Average Income = $[1.20(20) * \underline{.333} + 2.10(10) * \underline{.667}] / 30$
= **1.65 per day**

Who Becomes an Inventor? Propensity Score Reweighting

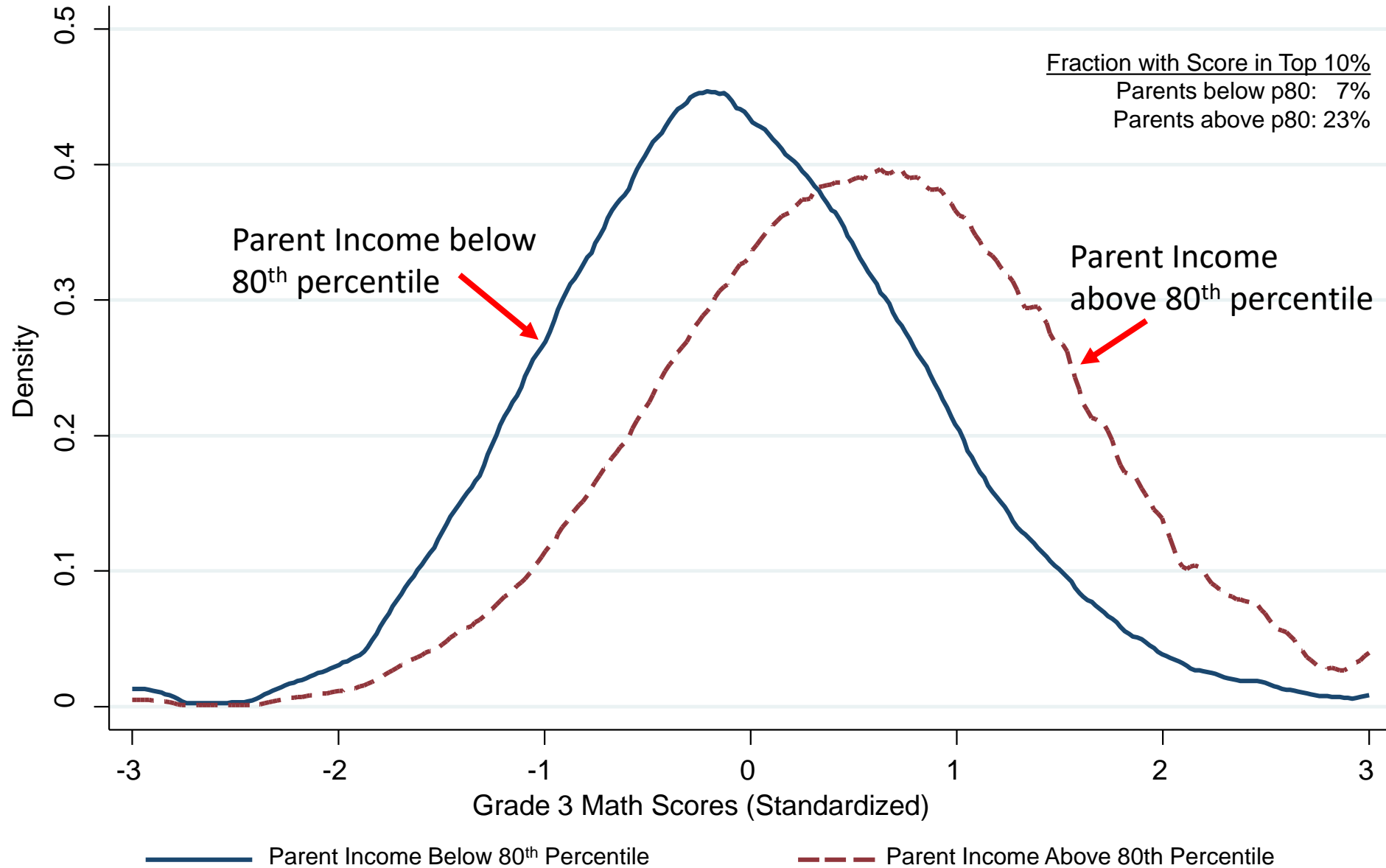
- Relationship between becoming inventor and parent income?



Who Becomes an Inventor? Propensity Score Reweighting

- Bell, Chetty, Jaravel, Petkova, and van Reenen (2016) link data on patents and NYC test scores to IRS income data
- Children of rich parents are much more likely to become inventors than the children of poor parents
- But poor children also have lower 3rd grade math test scores (ability)
- How much of the gap in patent rates can be explained by ability?
- Define weights w_i that equalize test score dist'n for rich and poor
- Compare reweighted means, medians, or the entire distribution of outcome variable across the two groups

Distribution of Math Test Scores in 3rd Grade for Children of Low vs. High Income Parents



Who Becomes an Inventor? Propensity Score Reweighting

- First, regress a rich parents indicator on observables:

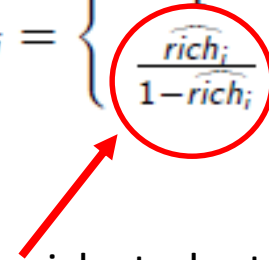
$$rich_i = \alpha + \beta \text{testscore}_i + u_i$$

where testscore_i is a vector of indicators for each test score percentile

- Use estimates $\hat{\alpha}$ and $\hat{\beta}$ to form predicted probabilities (or “propensity score”):

$$\widehat{rich}_i = \hat{\alpha} + \hat{\beta} \text{testscore}_i$$

- If covariates continuous, use probit or logit to force \widehat{rich}_i to be in $[0,1]$
- Using these predicted probabilities as sampling weights will make test scores match across the two groups:

$$w_i = \begin{cases} 1 & \text{if parents' income in top 20 percent} \\ \frac{\widehat{rich}_i}{1 - \widehat{rich}_i} & \text{otherwise} \end{cases}$$


- Weight on non-rich students increases in how much your test score “looks like” that of a rich person. (Upweight high-scoring, low-income students)

Who Becomes an Inventor? Propensity Score Reweighting

What Fraction of the Gap in Patenting by Parent Income
is Explained by Differences in Test Scores?

	Patent Rate (per 1000 Individuals)	Gap Relative to Above p80 Group
Above 80 th Pctile.	1.93	
Below 80 th Pctile.	0.52	1.41
Below 80 th Pctile. (Reweighting Scores)	0.95	0.97 (= 1.93 – 0.95)
% of gap accounted for by 3 rd grade scores		31.2% (s.e. = 6.8%)

Patent Rates by Race and Ethnicity

