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

Review Session #1
TF: Diana Goldemberg

Prof: Raj Chetty
Harvard University
Spring 2019

Disclaimer: multiple TFs contributed to those slides. All mistakes are mine.

Logistics

- I'm Diana.
- Take 2 min to fill out this survey please [bit.ly/ec1152d006]
  Find this prez at: https://github.com/dianagold/Ec1152_diana
- We'll meet every Thursdays @ 4.30-5.30pm (Sever 208)
  - Introductory level, no previous Stats background
  - Focus on intuition and applications
- Office Hours:
  - Wednesdays @ 4.30-6.30pm (Barker 103)
  - I'm also available by appointment and after sections.
- Expectations:
  - Email (**diana_goldemberg@g.harvard.edu**) response times: within 24 hours M-F; 48 hours on the weekend
  - Google form to submit questions before section

Outline

- Level the playing field: Summary Stats & Inference
- Intergenerational Mobility main graph
- Backdrop on Regression Analysis
- Stata demo: regression on bowling alleys
- Correlation is not causation (!)

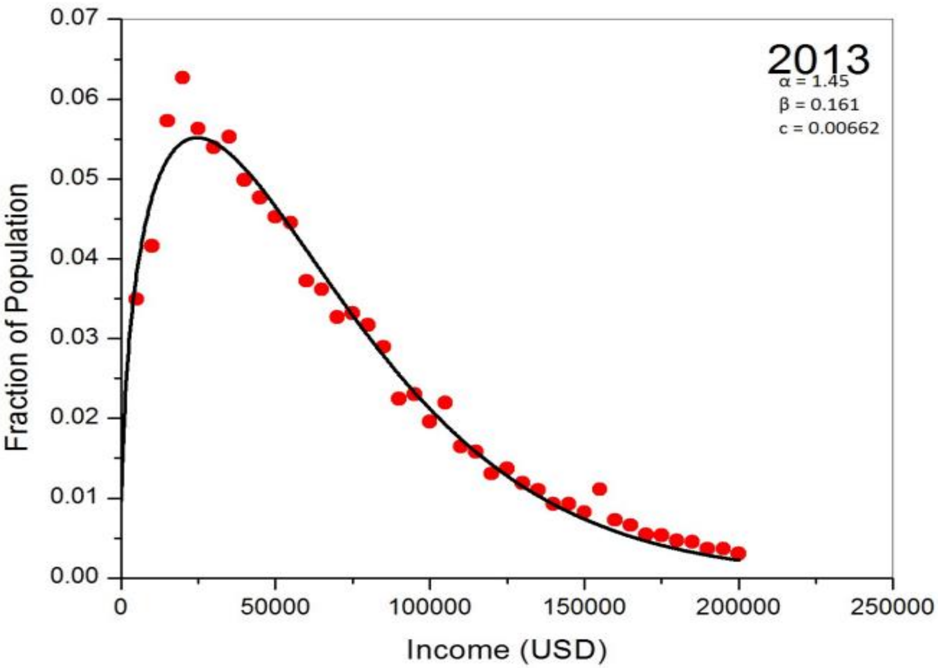
Summary Statistics

- Suppose you have access to all of Professor Chetty's data
    - That means you know everyone's income in the USA in 2017

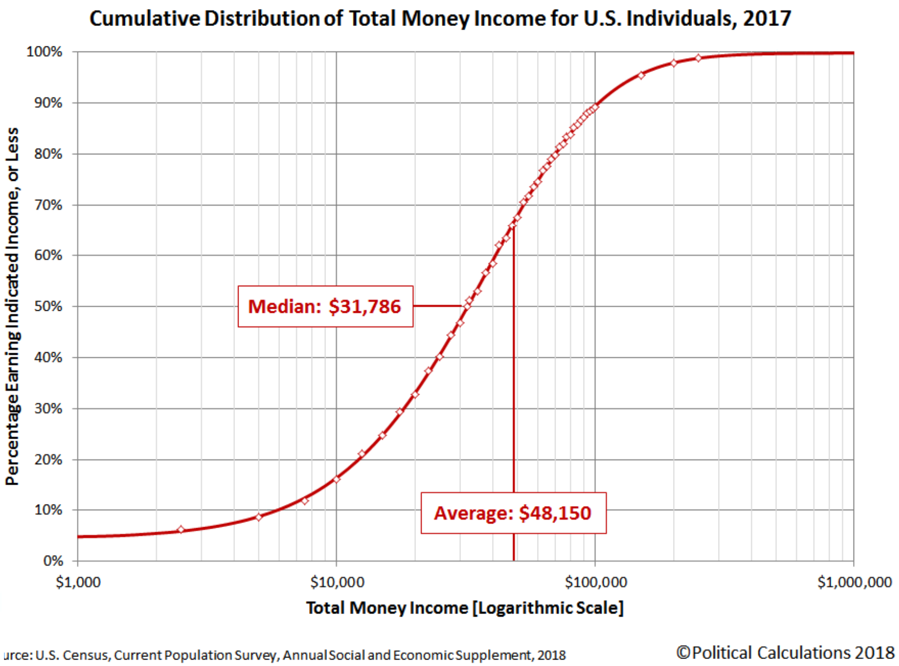
- In 2017, the mean U.S. Individual Income was $48,150.
    - How many individuals made close to $48,150 (say within $1000)?
    - How many individuals made more than $48,150?

- What pieces of information related to your data might you want to know...
    - To understand the "center" of the distribution?
    - To understand the "dispersion" of the distribution?
    - To visualize your distribution?

Summary Statistics



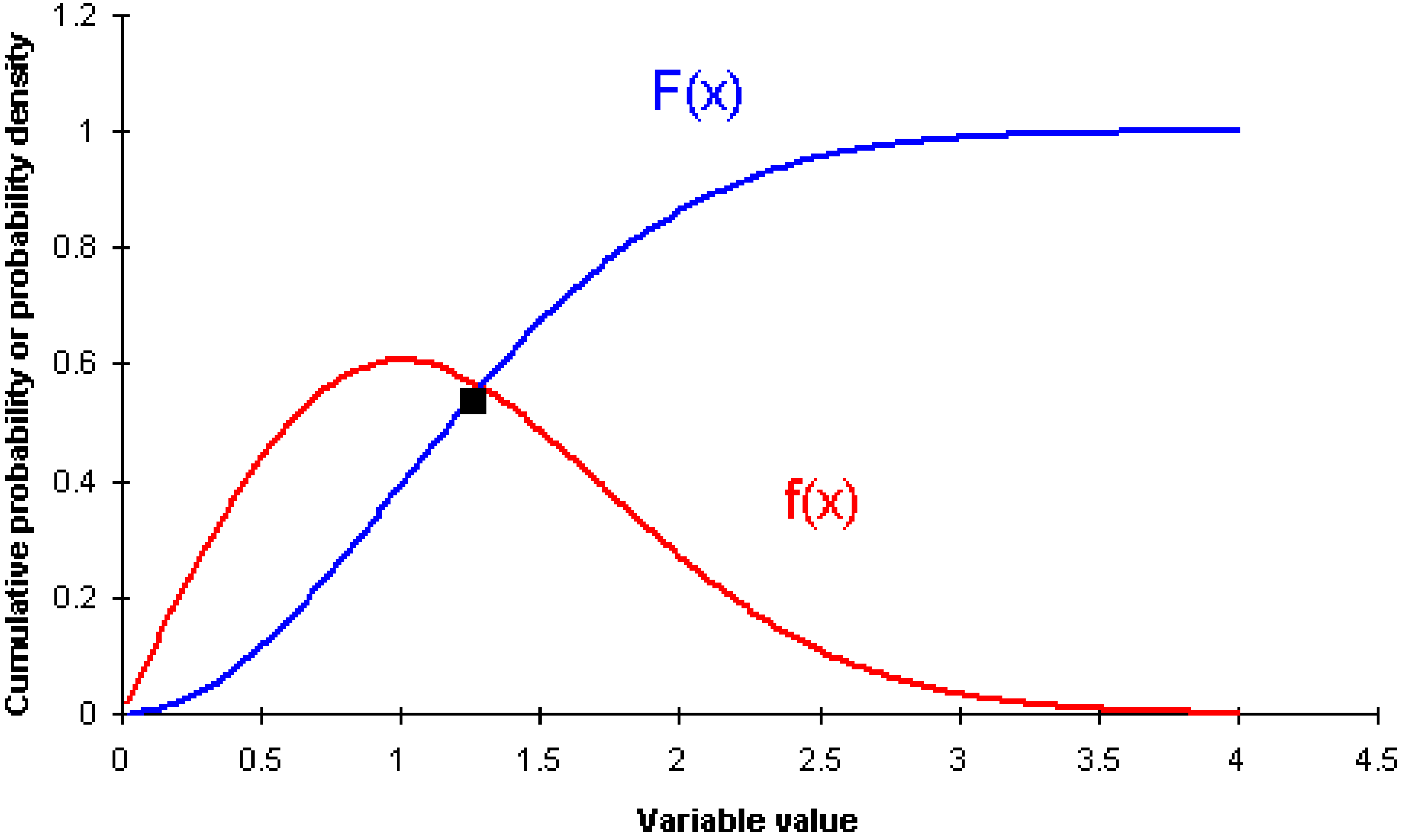
Probability Distribution Function



Cumulative Distribution Function

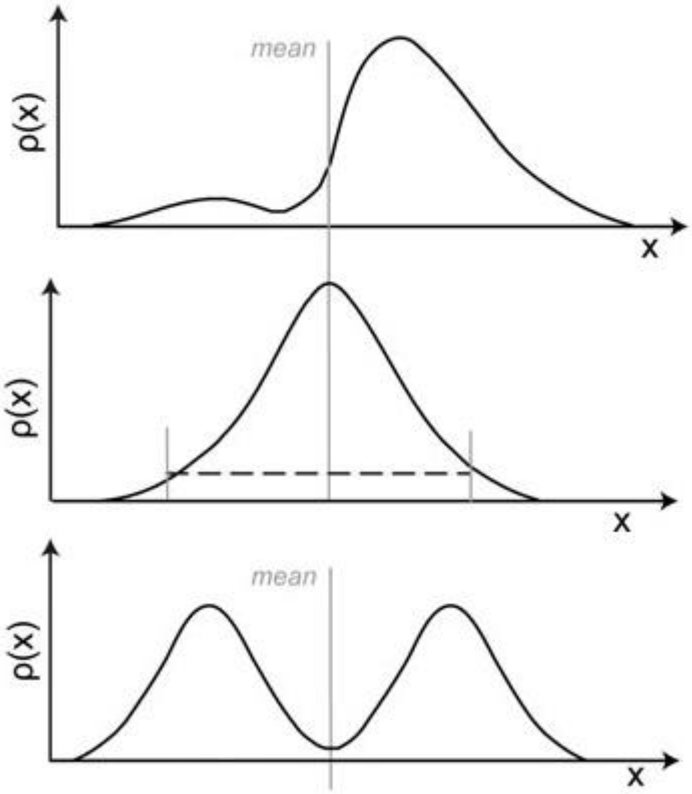
- A "low income" person in 2017 (aka: 25^{th} percentile) earns up to…?

Summary Statistics



Summary Statistics

- Three distributions with the same mean, and same variance / standard error



- Always important to visualize when possible!

Summary Statistics: Takeaways I

- Summary statistics and visualization are a good place to start when facing a new dataset.
- There is no one summary statistic that tells you everything you need to know.
- Common measures of centrality:
  - Mean: What is the "center of mass" of the data? If all the income were divided equally, how much would everyone receive?
  - Median: What is the "typical" value of the data? For what income level do half of people make more, and half of people make less? The 50th percentile.
  - Mode: What is the most common value of the data? If you had to guess the exact amount that a randomly chosen individual makes, what would be the best guess?

Summary Statistics: Takeaways II

- Common measures of dispersion:
  - Variance: Mean of squared deviations from the mean.
  - Standard Deviation: Square Root of the Variance. Has nice statistical properties for certain distributions (as does variance).
  - Interquartile range: What is the difference between the 75th percentile and the 25th percentile in your data? How spread out is the "middle half" of your data?
- Visualizing a single variable:
  - Probability distribution function (PDF): Easiest to think of this as visualizing relative frequency of your data. Higher point in PDF means a value is more common in your data.
  - Cumulative distribution function (CDF): For each value in your data, plots what fraction of your data is less than that value. Always starts at zero and rises to one.

Population, Sample and Inference

Suppose that:

- I only like **BLUE** m&m's
- Yesterday I opened one pack of each, finding

4 / 15 = **26.8%**

4 / 20 = **20.0%**

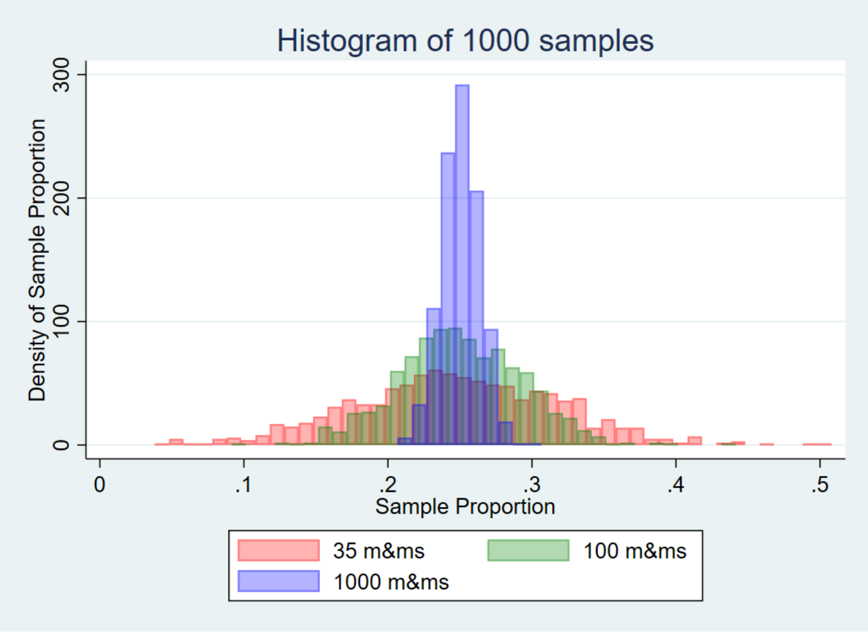
- Should I only buy Peanut m&m's from now on, trusting that they have a bigger share of **BLUE** m&m's?

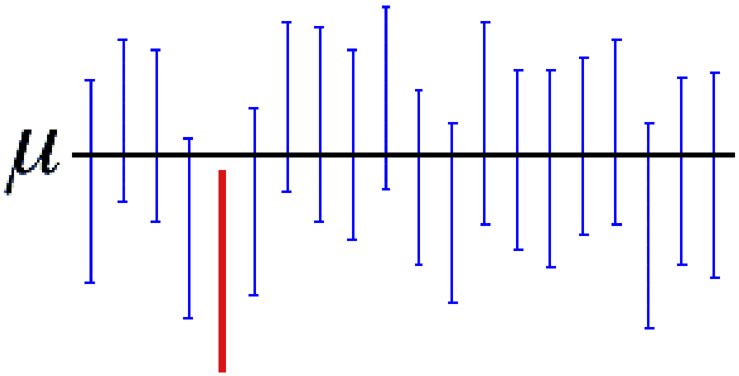
- According to Mars, **BLUE** m&m's represent **25.0%** of their production in the NJ plant (but changes across their plants!), and does not varying between fillings (m/p/pb)

=> combining them: **8** / 35 = **22.9%**

Population, Sample and Inference

- **Confidence intervals**:
  - My sample: I estimate the true value that Mars uses as [**8.9%, 36.8%**] with 95% confidence, using my 35 sample, that is the **22.9%** plus or minus 1.96*standard errors
  - Reality: a new sample of 35 m&m's will have [**10.7%, 39.3%**] with 95% confidence, that is the **25.0%** plus or minus 1.96*standard deviations





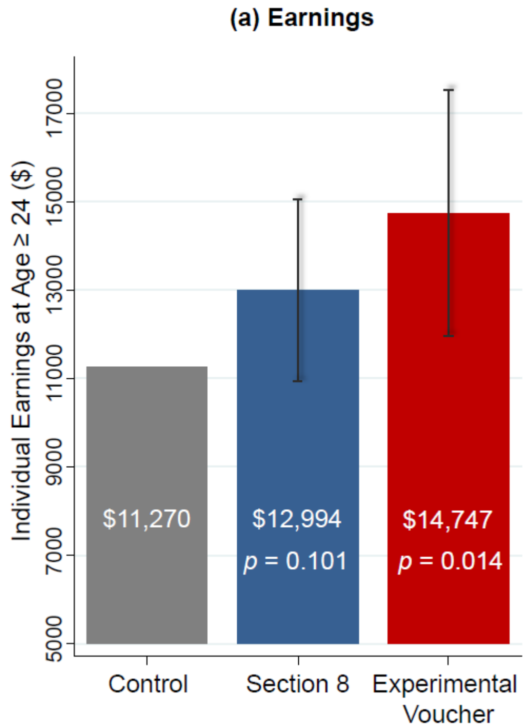
A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

Population, Sample and Inference

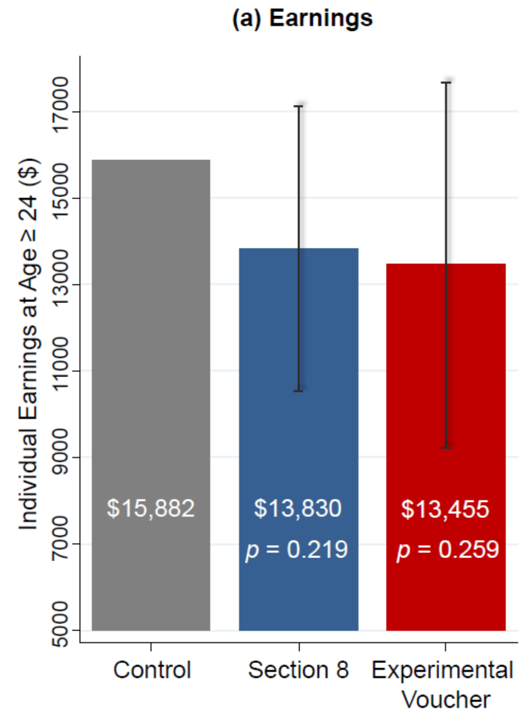
- **P-values**: a friendly answer to testing hypothesis
  - Peanut and peanut butter m&m's have the same distribution of blues. *p = 0.65*
  - The m&m's I ate follow the stated distribution of blues by Mars. *p=0.78*
  - Translate the chance that your hypothesis is true and you observed your result.
    Low p => reject hypothesis [stars]; High p => cannot reject hypothesis

Bringing it back to Chetty's lecture on MTO…

Impacts of MTO on Children Below 13



Impacts of MTO on Children Age 13-18



Inference: Takeaways

Statistical inference is the theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling.

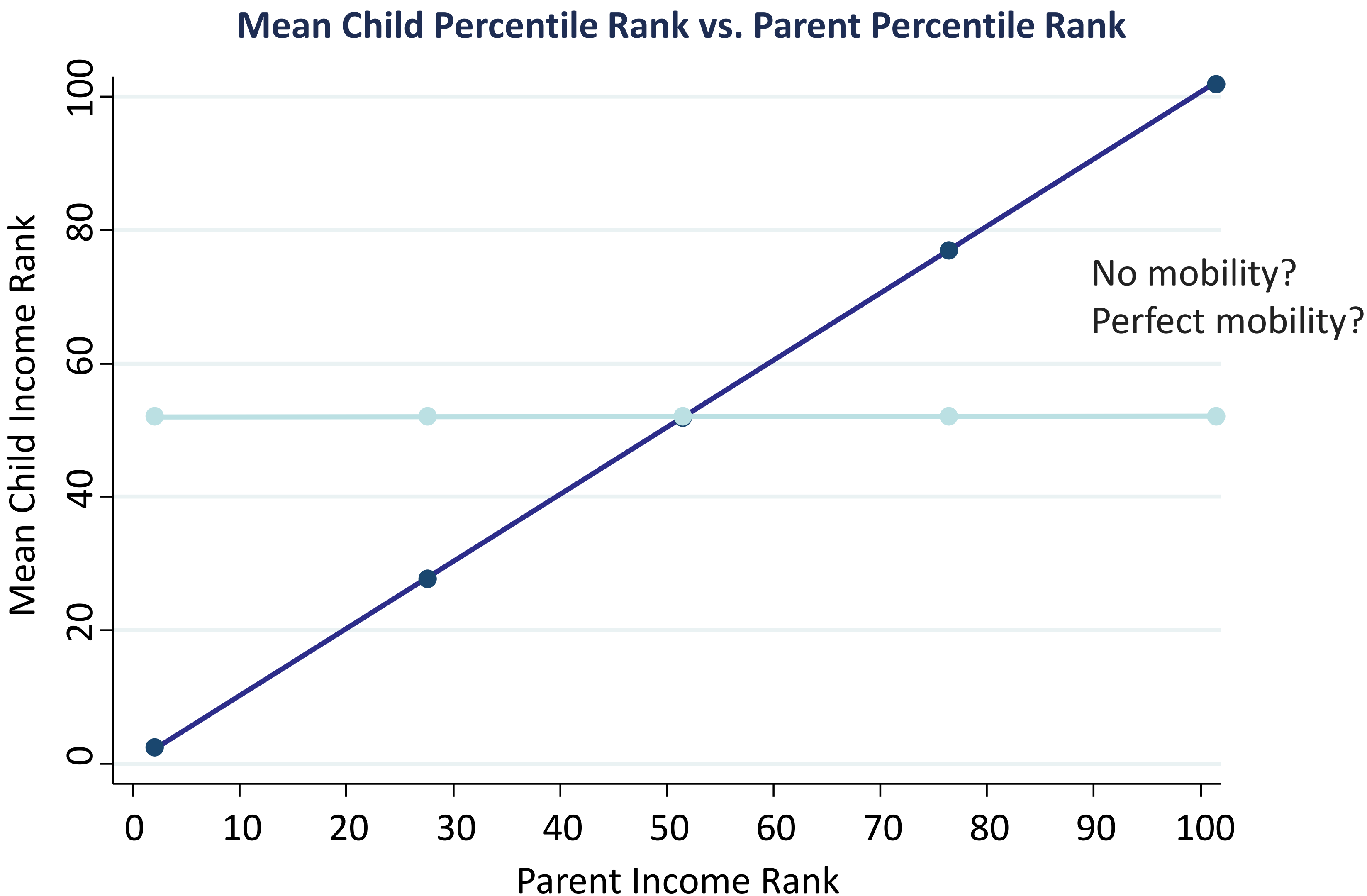
- Randomness exists. Results on a sample of data will not always match the population value.

- To deal with this, we calculate new statistics:
  - Standard errors tells us how far we might expect the sample mean to be from the true population mean.
  - Confidence intervals provide a net that we can use to try to "catch" the population mean with a pre-specified level of certainty.
  - P-values are the most friendly answer to hypothesis testing. Usually translates “what is the probability that we would observe such an extreme result by pure chance?”
    - Typical significance levels: p-value below 0.1, 0.05, 0.01? [stars]

- All of these values are given by statistical software, but you need to know which ‘questions’ to ask. P-value is your friend, always look for the p-value and the test it is addressing!!!

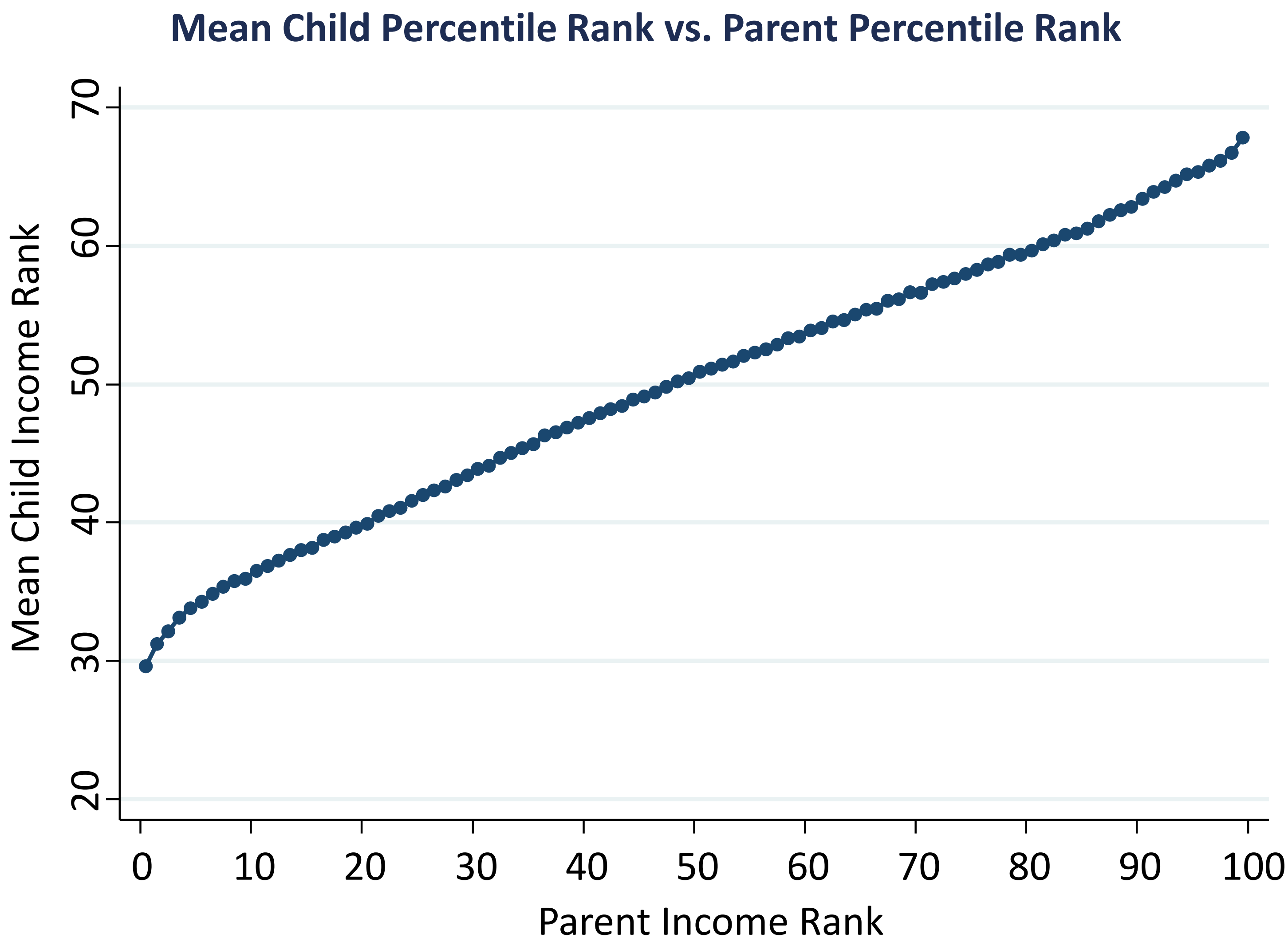
Intergenerational Mobility

- Main graph?



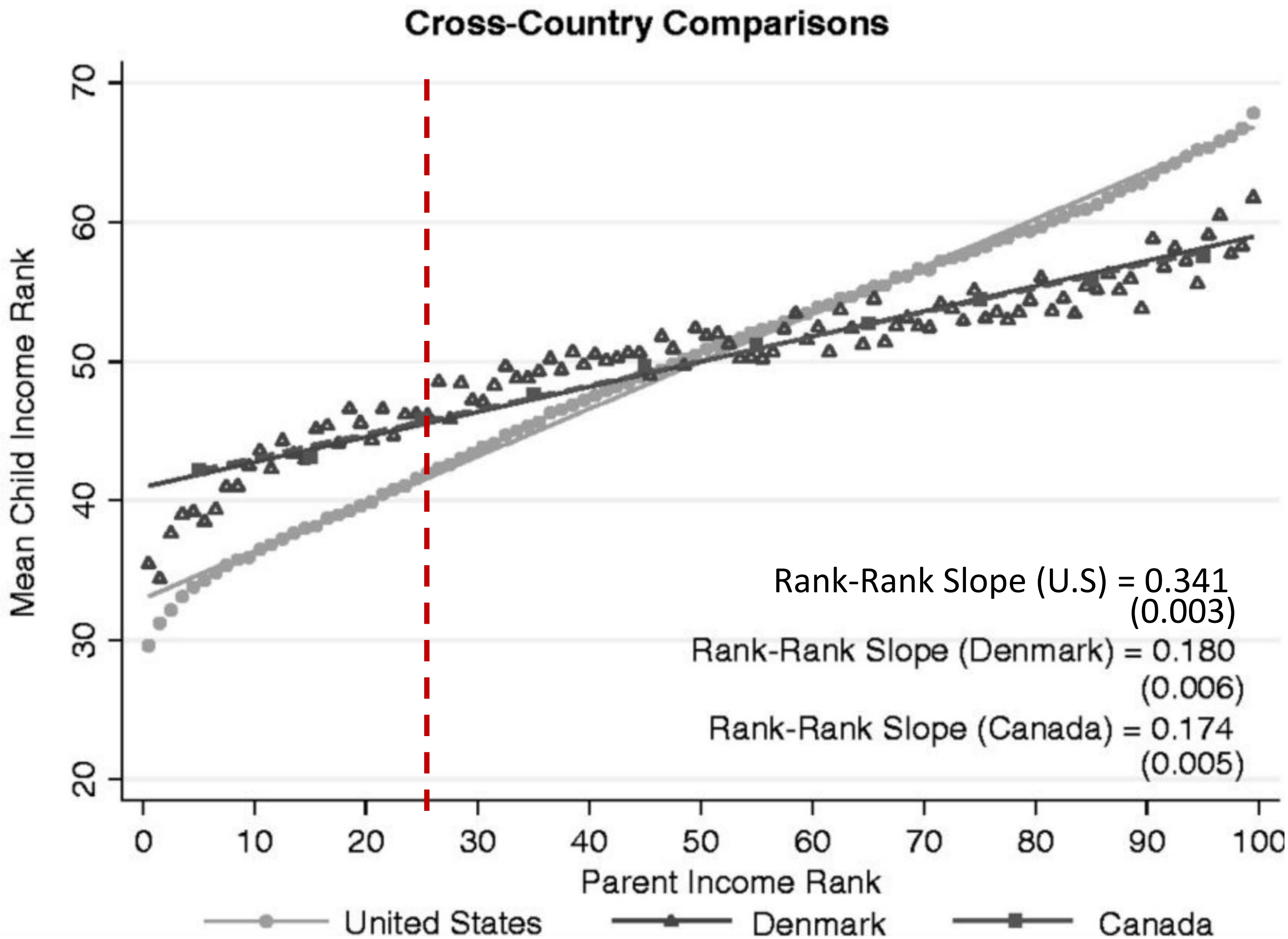
Intergenerational Mobility

- Think of some measures that translate mobility
  [this is the Figure I.A, Chetty et al 2018a]



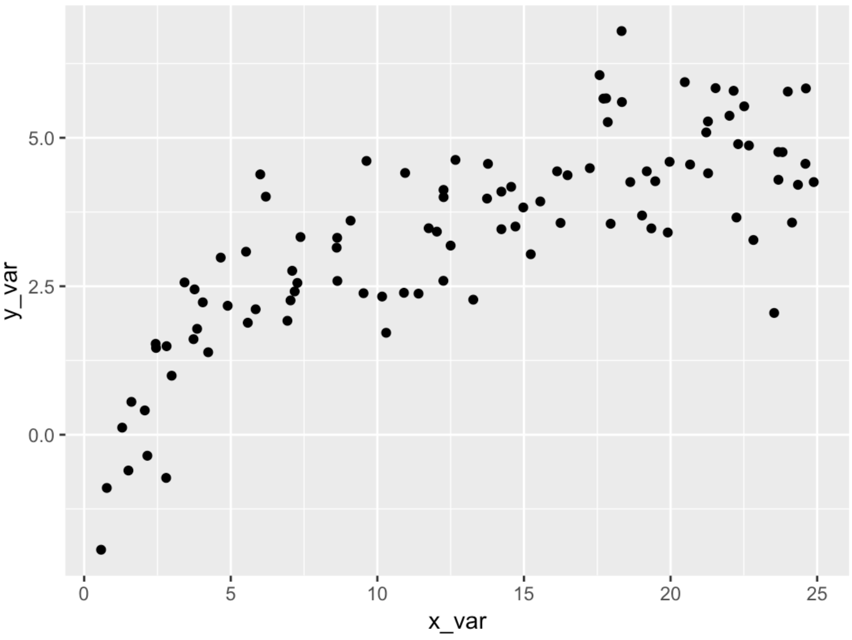
Intergenerational Mobility

- In simple terms: how well do kids from poor parents do?

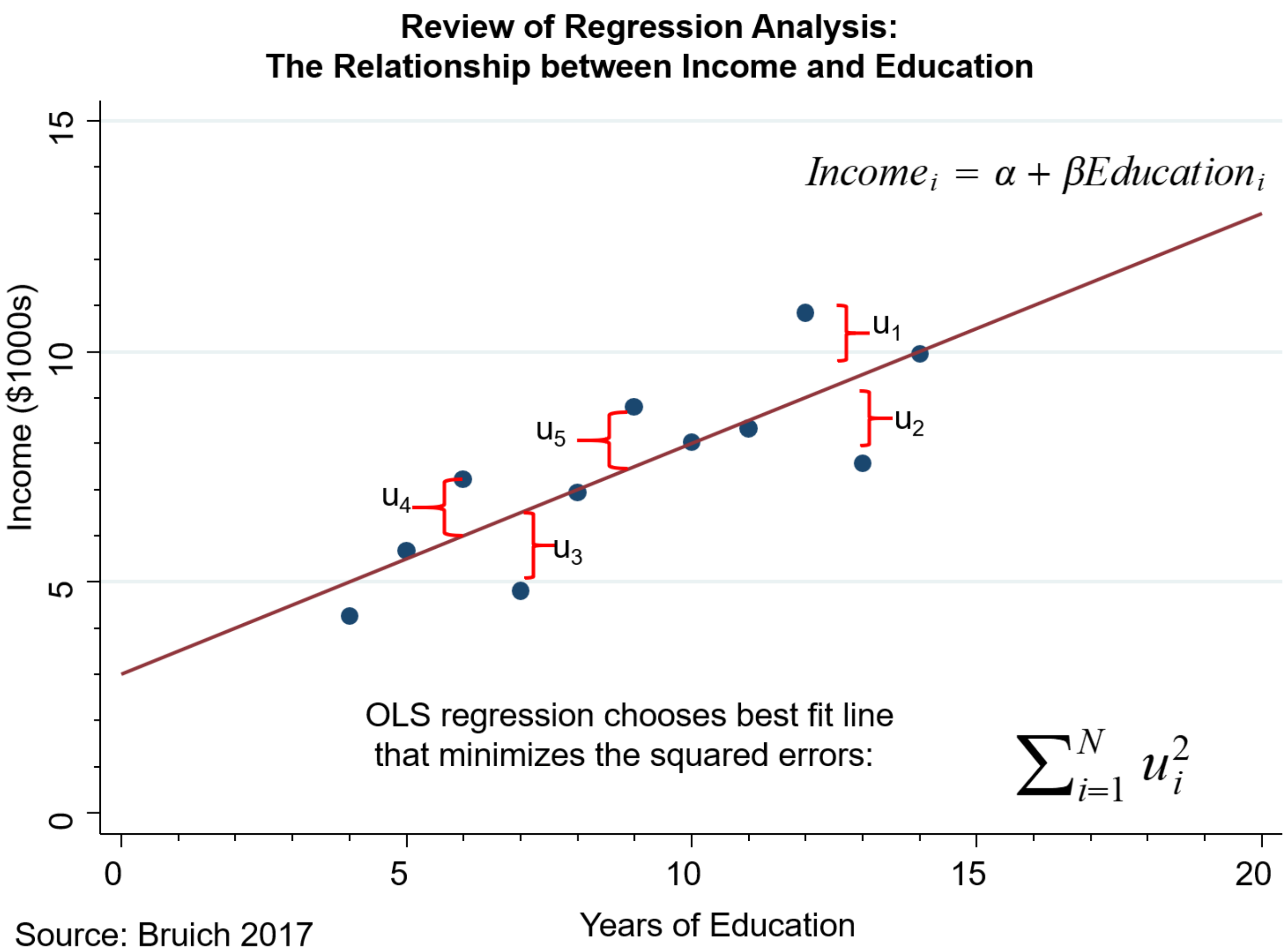


Backdrop on Regression Analysis

- On the previous slides, we took a series of points from our data, and we drew a line through them
    - How did we even do that?
- Suppose you have data on years of education and income for a group of people. How would you try to fit a line through that data?



Backdrop on Regression Analysis



- What's an interpretation of α and of β?
- Does this line give results for the population or for a sample? To what consequences?

Regression Analysis: Common Output

```
. reg e_rank_b bowl_per_capita, robust

Linear regression                               Number of obs     =        586
                                                F(1, 584)         =     339.47
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4124
                                                Root MSE          =     3.9697
```

e_rank_b	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
bowl_per_capita	12.04453	.6537132	18.42	0.000	10.76061	13.32844
_cons	39.28227	.2571105	152.78	0.000	38.77729	39.78724

Where is the regression slope? Intercept?

How precise are those estimates, or: where are their standard errors?

What is the probability that you would get this result (this slope estimate) even if the true population coefficient was <u>zero</u>? (This is the p-value, your best friend!!!)

Note that a p-value smaller than 5% means that the 95% CI will not include zero!

Regression Analysis: Standardization

- How can we compare the strength of relationship between different variables on a "level playing field?"

- For example, how could we tell if average years of education in a district or fraction of people married in a district is more closely associated with income in that district?

  - Key point: we need a **standardized** measure of correlation

  - It turns out, we can run a very simple regression to get a **correlation coefficient** that:
    - Is always between –1 and 1.
    - Is –1 if variables are perfectly linearly related in a negative way
    - Is 1 if variables are perfectly linearly related in a positive way
    - Is 0 if variables are not at all linearly related

Regression Analysis: Standardization

Suppose you have variables X_i, Y_i. Construct X_i* and Y_i* as follows:

$$Y_i^* = \frac{Y_i - \text{Mean}(Y_i)}{\text{Std\_Deviation}(Y_i)} \qquad X_i^* = \frac{X_i - \text{Mean}(X_i)}{\text{Std\_Deviation}(X_i)}$$

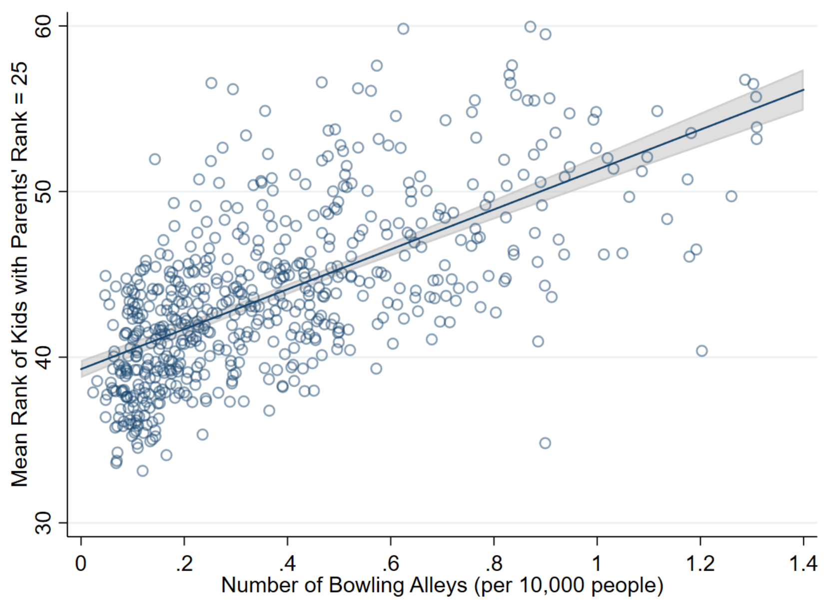
Then you can compute the least squares regression equation:

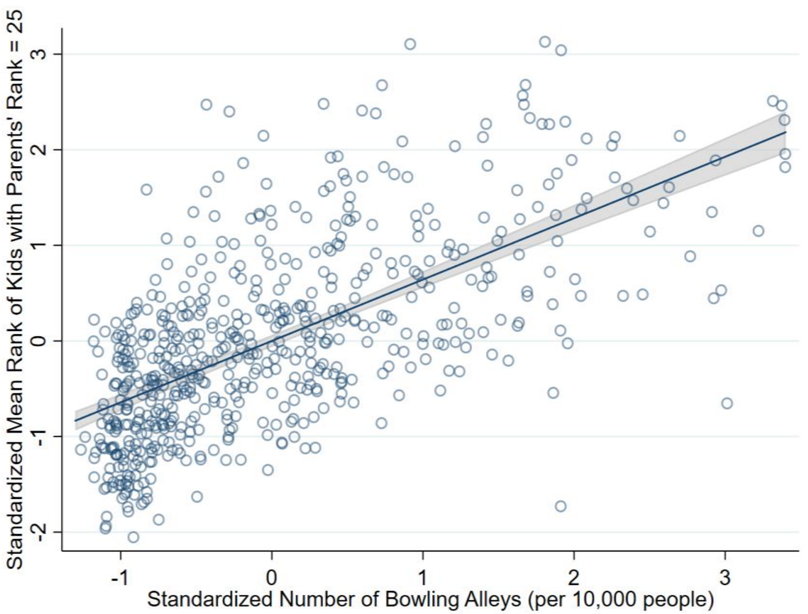
$$Y_i^* = a + r X_i^*$$

Fact: When you compute the above regression, you will always find that:

- The intercept a = 0
- The slope r is a correlation coefficient of the type we described.
- If you square r, **("R squared")** you get a number between 0 and 1 that is equal to the fraction of variation in Y explained by a linear regression on X.

Regression Analysis: Standardization





Unstandardized slope (left) tells you that how much of an increase in mean rank of kids with parent rank = 25 is associated with one additional bowling alley per 10,000 people.

Standardized slope (left) tells you **how correlated** mean rank of kids with parent rank = 25 and bowling alleys per 10000 people are, on a scale of -1 to 1.

Regression Analysis: Takeaways

- Regression analysis allows us to fit a line to data in a systematic way.

- In this class, we will begin with "Ordinary Least Squares" regression (OLS).
  - OLS minimizes sum of the squared errors between data points & the fitted line.

- Nice features of OLS and other regression techniques:

  - The slope of the line often has a natural interpretation.
    - "One more year of education is associated with B in increased earnings"

  - When data is noisy, OLS allows you to focus in on the trends and patterns..

  - In certain circumstances, OLS constitutes our "best guess" at the Y value when all we know is X. "Knowing only a person's education is X, I'd guess their earnings are Y on average.“

- Regression coefficient estimates come with their standard errors, which are needed to test hypothesis (inference and p-values!)

Stata hands-on demo

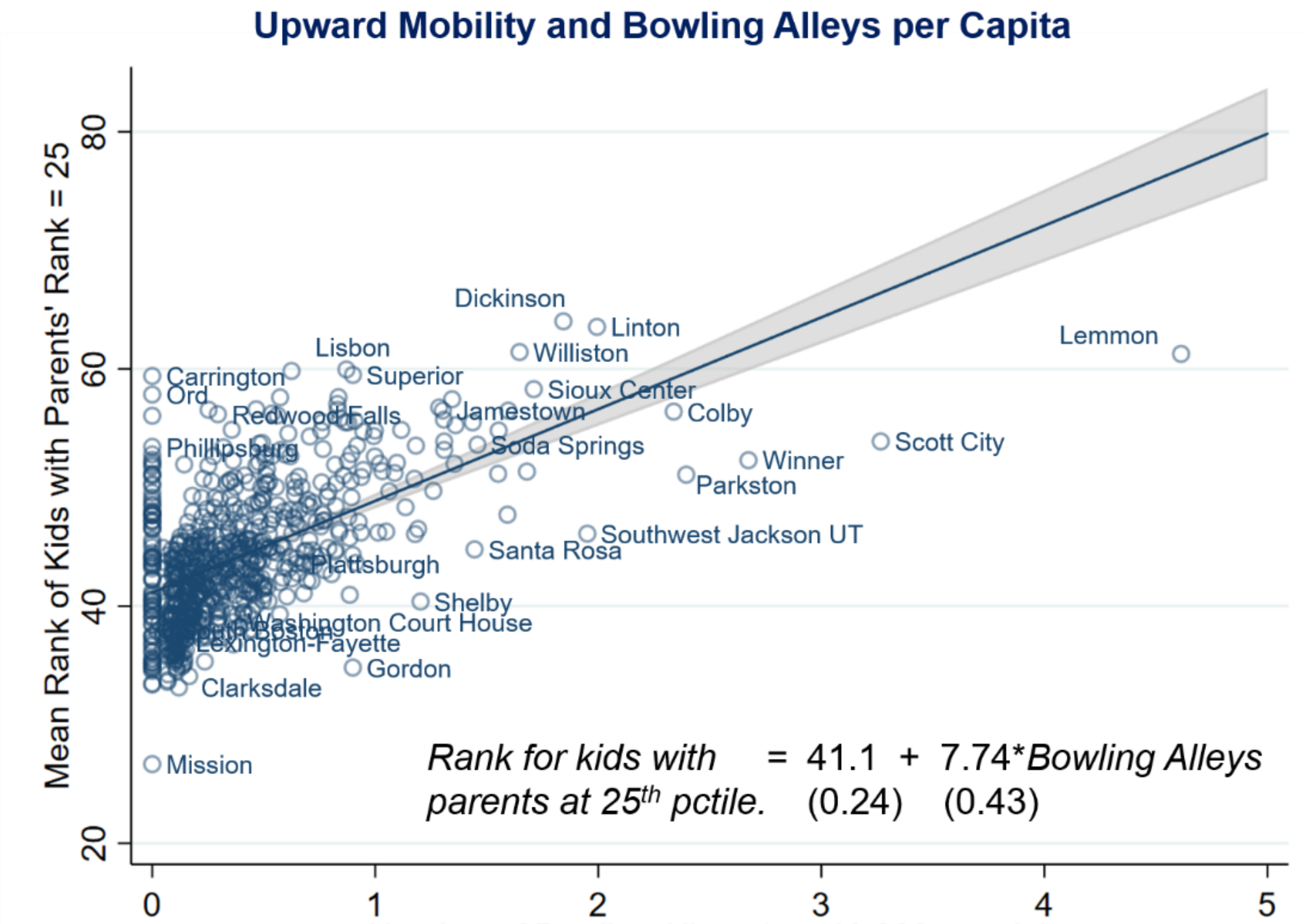
- *Stata will be used in section*
- *But you're very welcomed to follow the Jupyter notebooks for:*
  - *R*
  - *Python*
- *All files at:* https://github.com/dianagold/Ec1152_diana

Stata demo

Required files at: https://github.com/dianagold/Ec1152_diana

- If you have Stata in your computer, you may want to do it along
  - How to install & hints: https://canvas.harvard.edu/courses/19323
  - Optional workshop: Monday at 5:30 pm in Emerson Hall 105
- Why are we using Stata?
  - The most popular software used by economists for applied econometrics
  - Works for "big data": up to 20 billion observations and 32 thousand variables (contingent on RAM)
- Upward mobility (Y) as a linear regression of Bowling Alleys per capita (X)
- Tasks:
  - Get means and stdevs
  - Standardize Y and X
  - Use OLS to estimate correlation coefficients

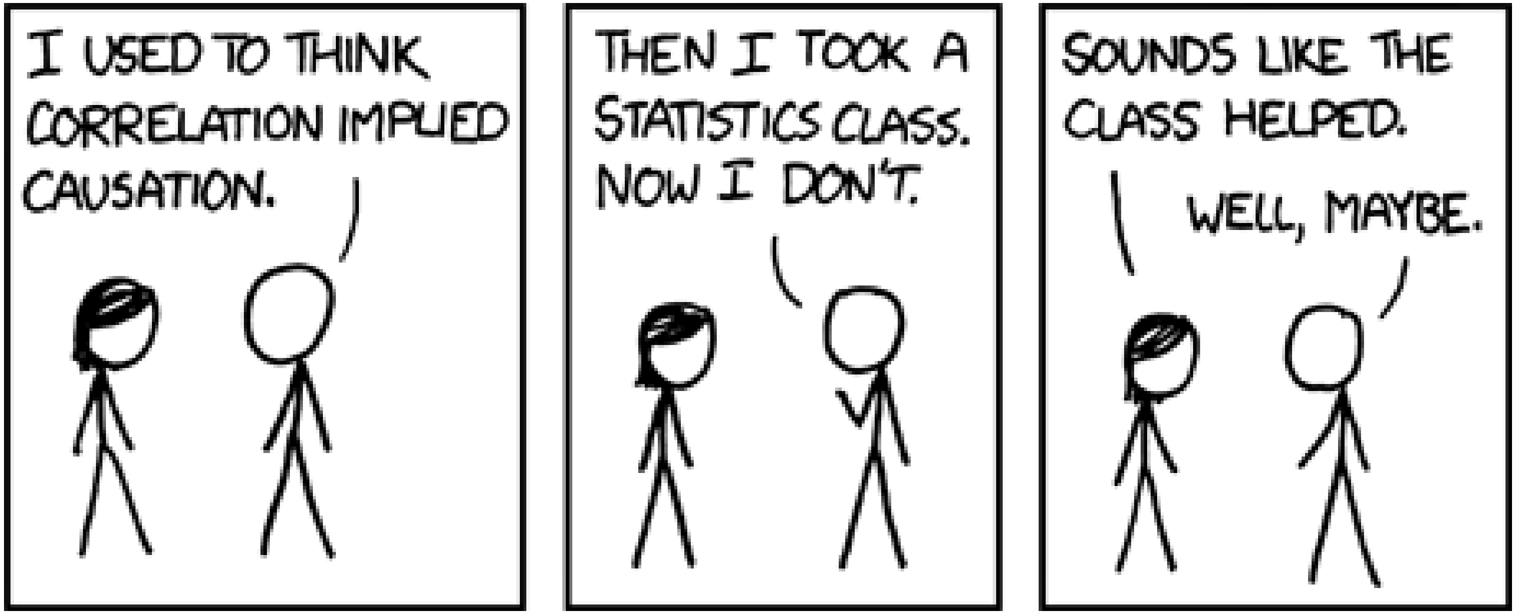
Correlation is not causation!



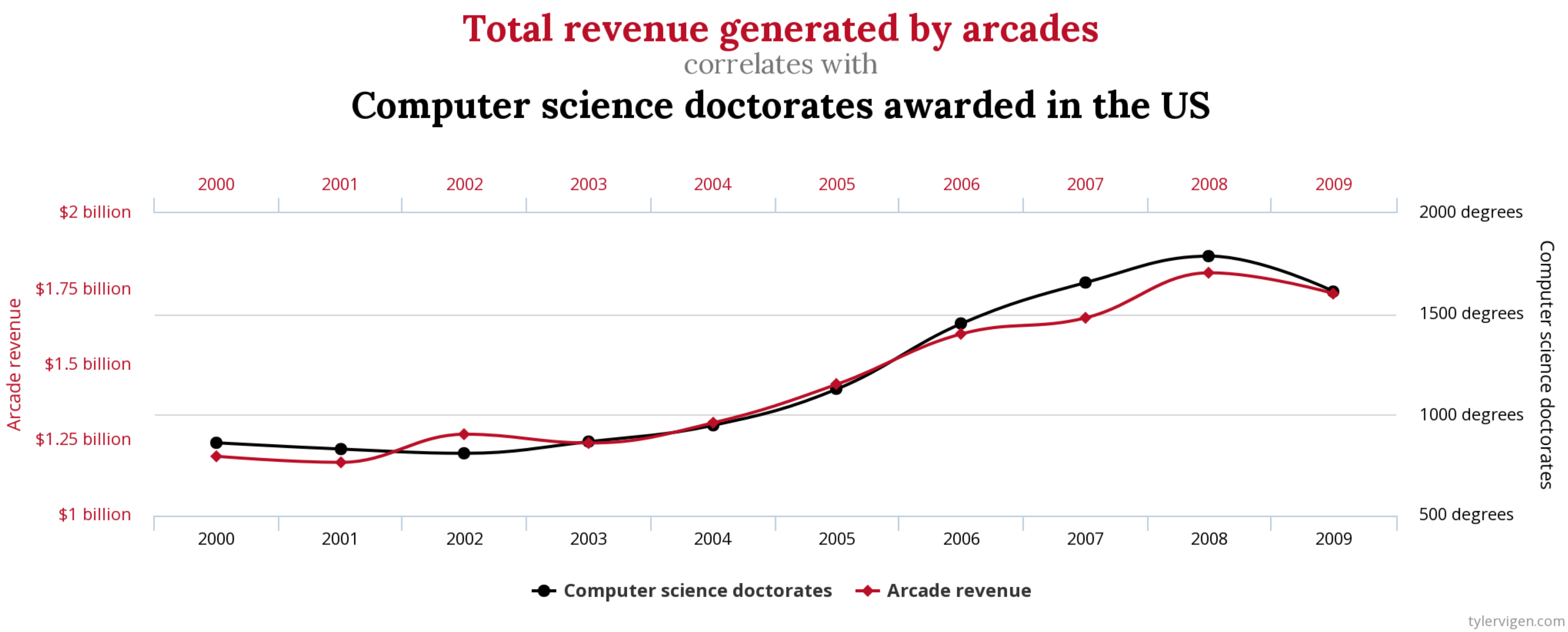
Source: Bruich 2018

Correlation is not causation!

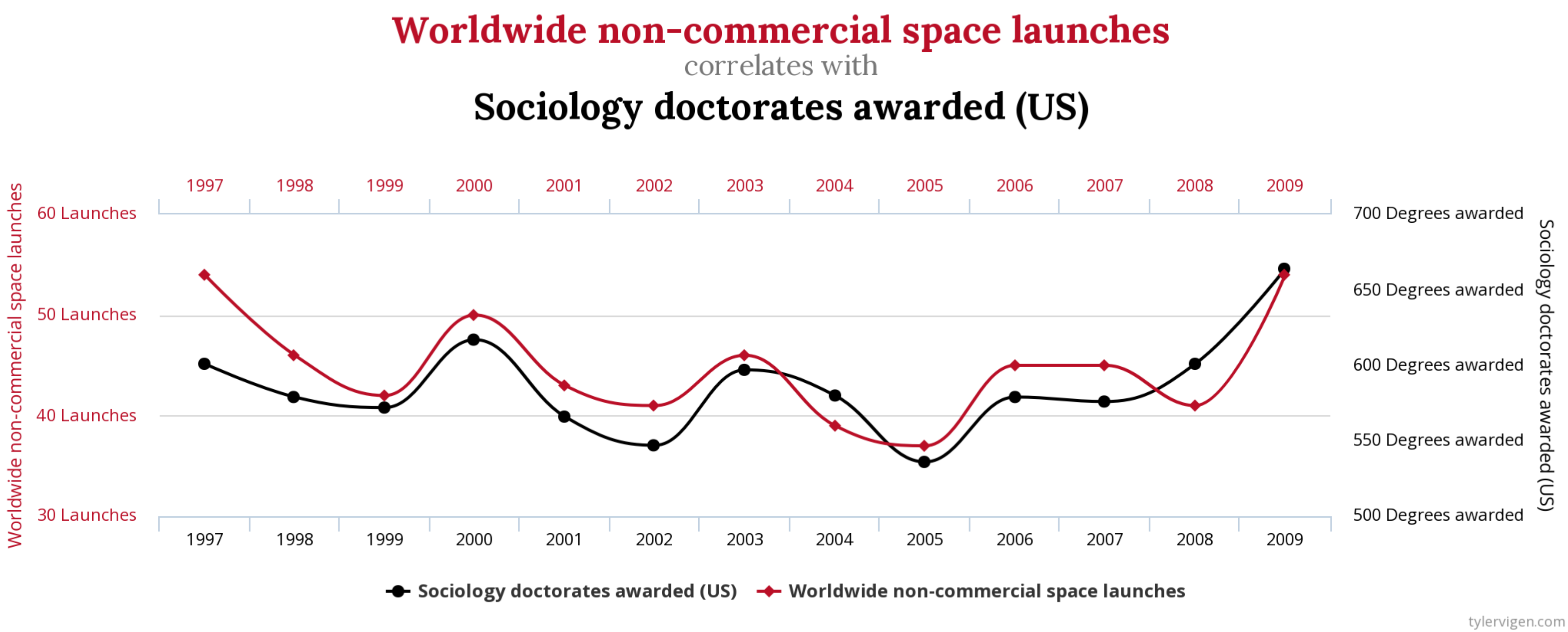
- Have you seen this meme before?
- Have you take a Stats class before?
- Correlation or causation?



Correlation is not causation! Examples



Correlation is not causation! Examples



Correlation is not causation! Examples