Testing errors and observed data

Prottoy A. Akbar

Principles of Empirical Analysis (ECON-A3000)
Lecture 6

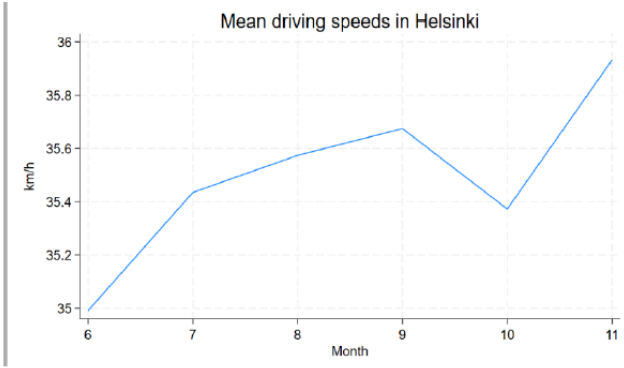
Today: Common mistakes, MTO, and observed data

- Things that often go wrong in statistical reasoning
- You will understand the following concepts:
  1. **false positives vs negatives** (a.k.a. type I and II errors)
  2. statistical power
  3. publication bias, file-drawer effect and p-hacking
  4. multiple hypothesis problem
  5. pre-registration and replication files
- A large-scale randomized controlled trial (RCT) with people
- An application
  - using data outside of experiments

Homework 1 feedback

- Grades should be out soon.
- Main feedback: lots of scope for partial credit for trying (even if incorrect). So, show your process.
- Question 5 asked you to explore the data and share an interesting finding.
  - Next slides: some examples from your responses.

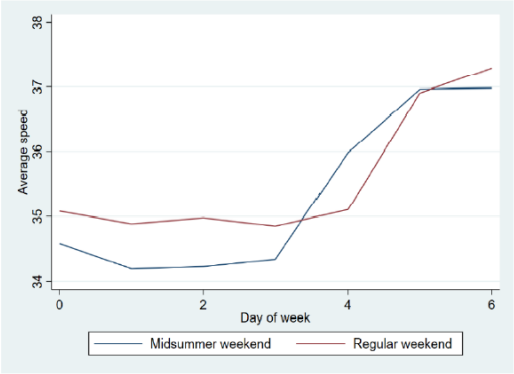
Homework 1 q.5 example (Teemu Virta)



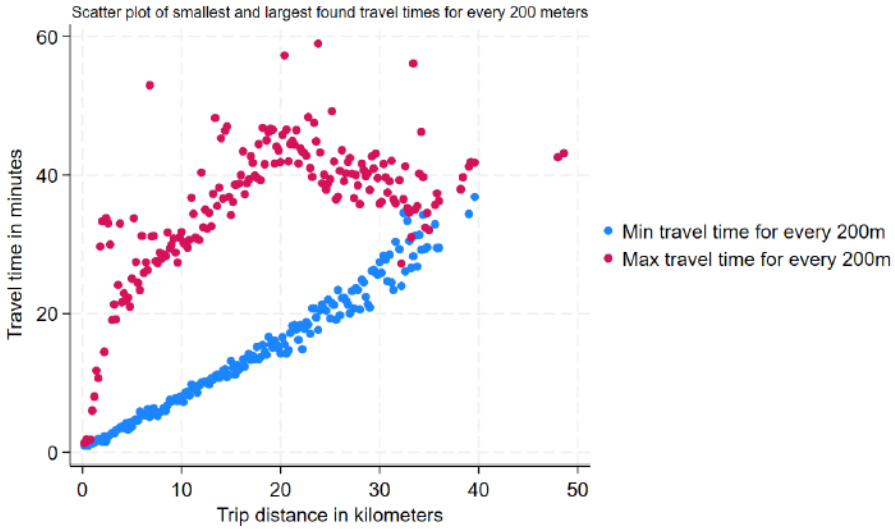
Mean driving speed by month. Sadly we only have data from half a year but it is interesting that we see an uptick in November, even though it is a wintermonth and driving conditions are probably worse.s

Homework 1 q.5 example (Joona Suominen)

5. The graph shows average speed on different days of the week for the dates 17 June to 27 June of 2019 (blue) during which (Saturday 22 June or day 5) the Finnish festival of Midsummer was held. In red is the average of average speeds for different days of the week on all other days in the sample. I was trying to see if average traffic speed differed significantly during Midsummer but the data do not seem to indicate it.



Homework 1 q.5 example (Ismo Laine)



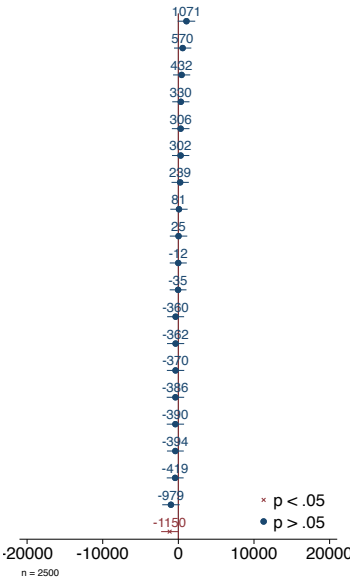
Shortly: Higher correlation with min than max travel times.

Testing errors (recap)

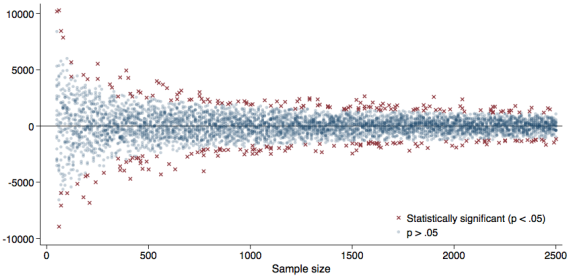
		Reality	
		Effect	No effect
Result of an experiment	Effect	True positive	**False positive**
	No efect	**False negative**	True negative

- False positive: Claiming an effect when it does not exist
  - also known as "type I error" or "acceptance error"
- **False negative**: Not finding an effect when it does exist
  - a.k.a. "type II error" or "rejection error"
- **Power**: the probability of finding an effect when it exists

False positives with larger samples (recap)



- 20 simulations with $n = 2500$
  - even less spectacular false positive
  - and still tighter confidence intervals (CI average width is 2,300 euros)
- More simulations
  - 20 rounds for 50,60,....,2500 observations
  - 0–5 false positives per round
  - overall 5.2% of simulations false positive



Take-aways from the first simulation (recap)

- The likelihood of a false positive does not vary with sample size
  - by definition, depends only on the p-value required for calling the esimate statistically significant (significance level)
- Small samples lead to large point estimates for false positives
  - small sample $\rightarrow$ wide CI $\rightarrow$ only large estimates significant
  - thus false positives from small samples may cause more damage
    - policy mistakes more likely if the effects are believed to be large
    - sadly, few people understand the dangers of underpowered studies
  - results from small samples sometimes get huge media attention
    - unfortunately, editors and referees of scientific journals may also like spectacular and statistically signficant results

Publication bias, file-drawer effect and p-hacking

- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments

Publication bias, file-drawer effect and p-hacking

- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments
- The problem is that we may get to see only the "significant" ones
  - **publication bias**: academic journals may be more likely to publish statistically significant results than insignificant "imprecise zeros"

Publication bias, file-drawer effect and p-hacking

- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments
- The problem is that we may get to see only the "significant" ones
  - **publication bias**: academic journals may be more likely to publish statistically significant results than insignificant "imprecise zeros"
  - **file-drawer effect**: researchers never finish papers with statistically insignificant results, because they would not be published anyways
    - less likely in large RCTs (funding agencies require to publish something)

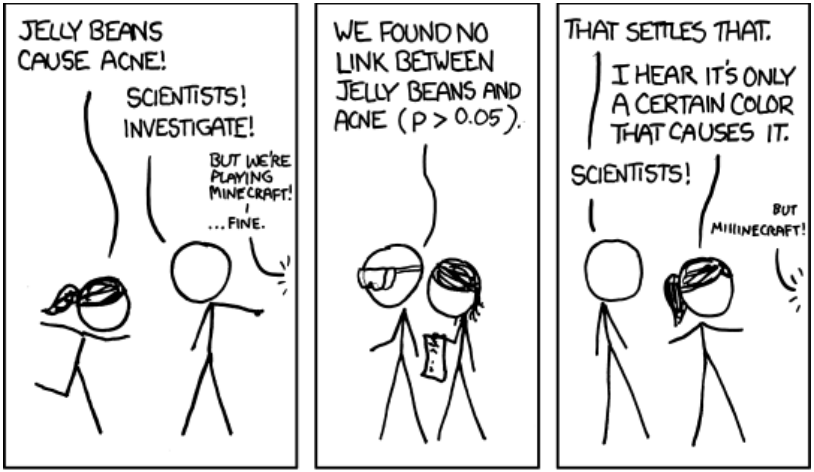
Publication bias, file-drawer effect and p-hacking

- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments
- The problem is that we may get to see only the "significant" ones
  - **publication bias**: academic journals may be more likely to publish statistically significant results than insignificant "imprecise zeros"
  - **file-drawer effect**: researchers never finish papers with statistically insignificant results, because they would not be published anyways
    - less likely in large RCTs (funding agencies require to publish something)
  - **p-hacking**: researcher reports only a specification with $p < .05$

Publication bias, file-drawer effect and p-hacking

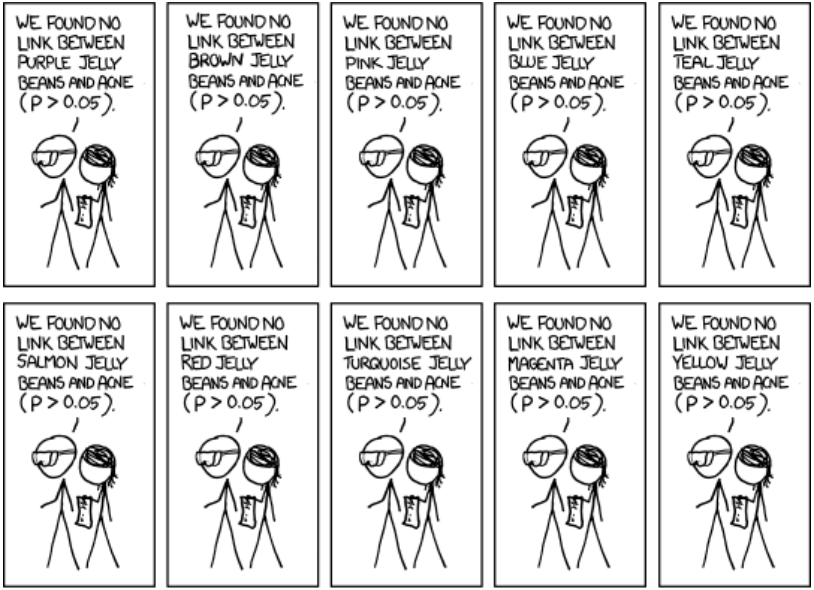
- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments
- The problem is that we may get to see only the "significant" ones
  - **publication bias**: academic journals may be more likely to publish statistically significant results than insignificant "imprecise zeros"
  - **file-drawer effect**: researchers never finish papers with statistically insignificant results, because they would not be published anyways
    - less likely in large RCTs (funding agencies require to publish something)
  - **p-hacking**: researcher reports only a specification with $p < .05$
- No-one needs to be nefarious for these problems to arise
  - people who fabricate results rarely want to be researchers
  - but: honest researchers may "follow the data" into wrong conclusions

Multiple comparisons problem

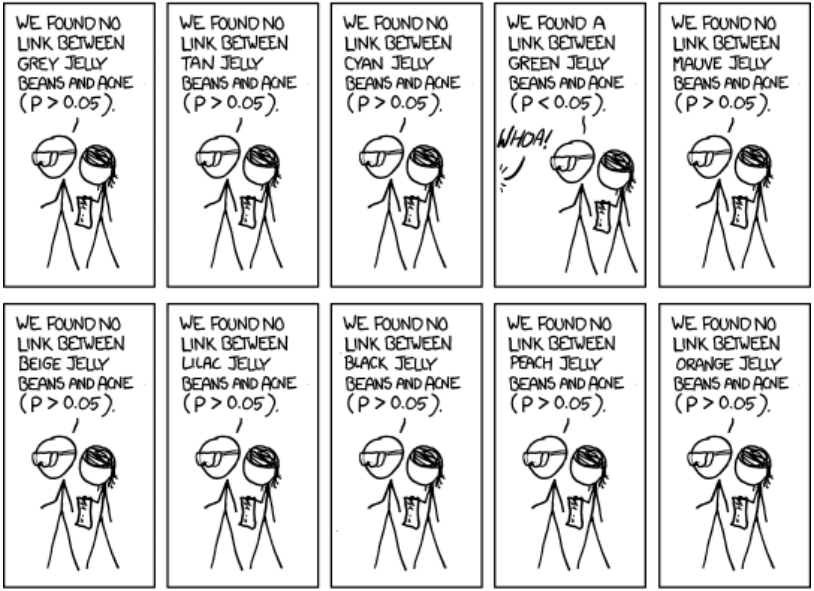


xkcd 882

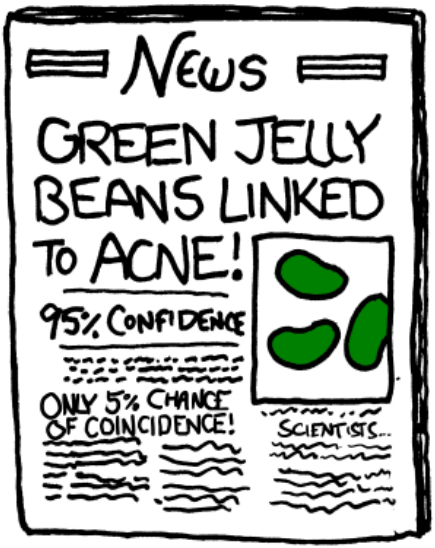
Multiple comparisons problem



Multiple comparisons problem



Multiple comparisons problem



- **Multiple comparisons problem** occurs when many comparisons are performed, but this is not taken into account in hypothesis testing
- A *human* error that can happen even with the best intentions
  - "the Garden of Forking Paths"
  - can take also other forms (e.g. subsample analysis)
- Tests taking into account the number of comparisons exist
  - you'll learn some of them in the more advanced courses

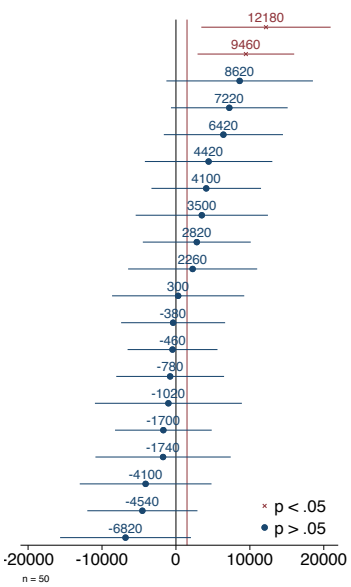
Approaches for human mistakes

- Pre-registration of randomized control trials (RCTs)
  - researchers can "tie their hands" by documenting their primary outcomes and specifications before seeing the data
  - long tradition in medicine; now also required in economics
- Replication files
  - top economics journals require researchers to post their code and data (or details about accessing the data) of published papers
  - allows other researchers to analyze the robustness of the results
- Running larger experiments

False negatives

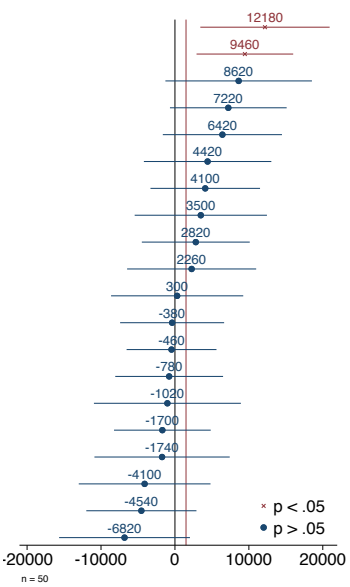
- Statistical error of not detecting an effect when it exists
  - getting $p > .05$ when there is an effect
- Let's demonstrate this with another simulation
  - identical to the one before except that now the treatment increase annual income of the treated by 1,500 euros

False negatives in small samples



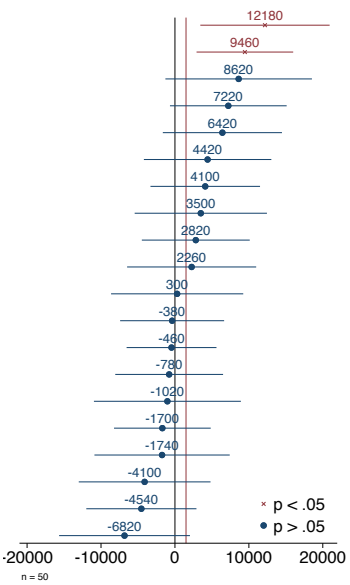
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control

False negatives in small samples



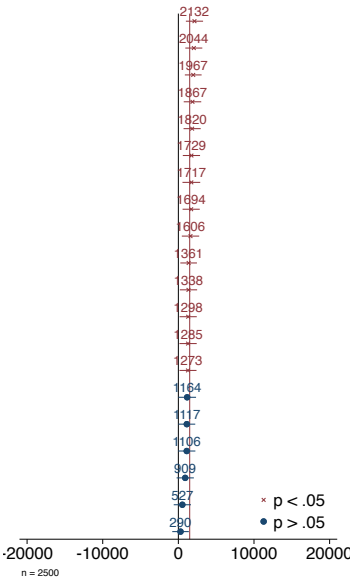
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control
- 2 out of 20 is statistically significant
  - but they are also severely wrong in the sense of being 6–8 times larger than the truth!

False negatives in small samples



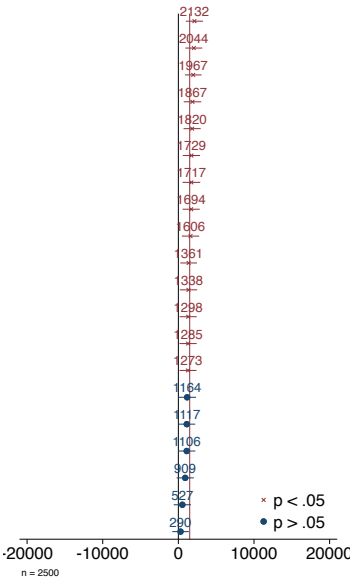
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control
- 2 out of 20 is statistically significant
  - but they are also severely wrong in the sense of being 6–8 times larger than the truth!
- 18 out of 20 are false negatives
  - 5 some of them are larger with the wrong sign than the true effect!
- Take-away: these estimates contain very little information

False negatives with larger samples

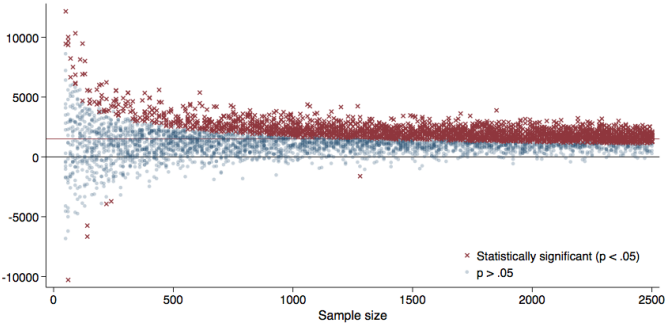


- 20 simulations with $n = 2500$
  - 12 out of 20 statistically significant
  - all relatively close to to the truth

False negatives with larger samples



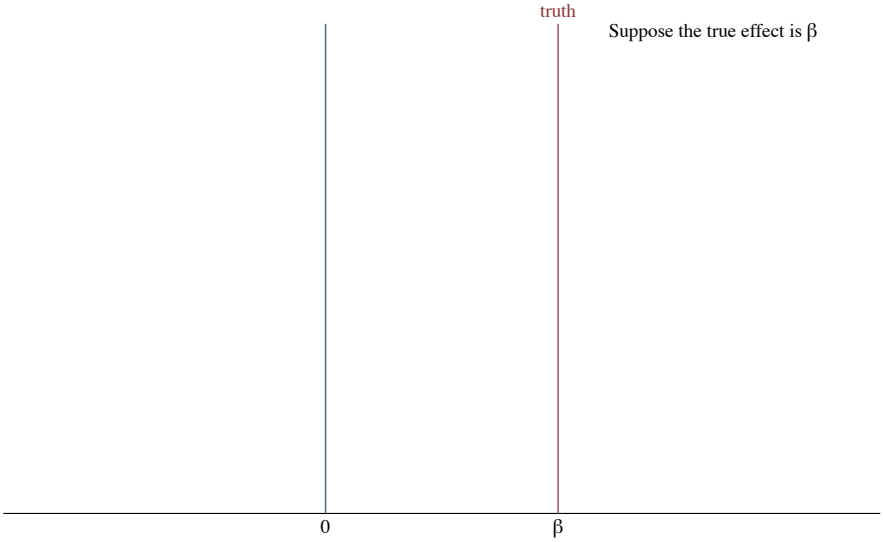
- 20 simulations with $n = 2500$
  - 12 out of 20 statistically significant
  - all relatively close to to the truth
- More simulations
  - 20 rounds for 50,60,....,2500 observations
  - as n increases, share of false negatives and wild point estimates decrease



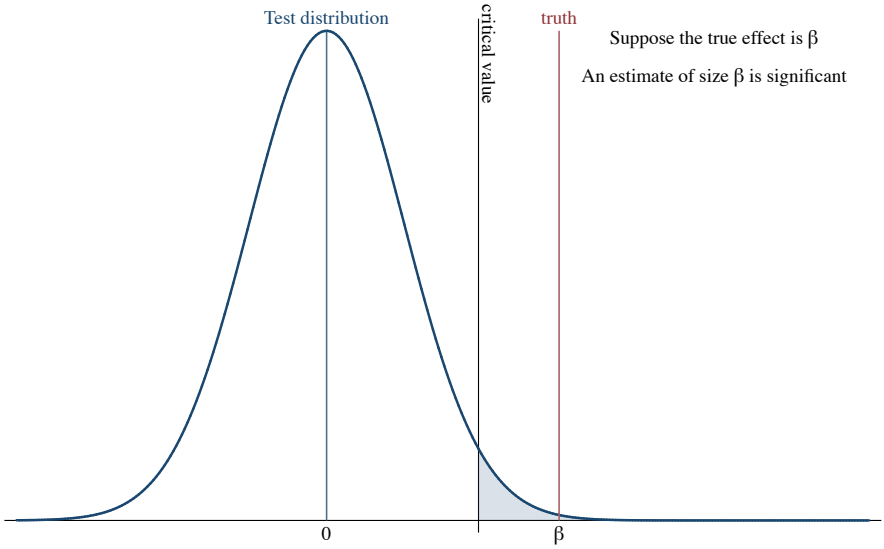
Power

- **Power** $= Pr(\text{reject } H_0 | H_1 \text{ is true})$
  - in our context: how likely are we to conclude that a treatment has an impact, when it truly has an impact
- Power depends on
  - true effect size
  - sample size
  - variability of the outcome variable
  - statistical significance level
- Next: a graphical illustration of power

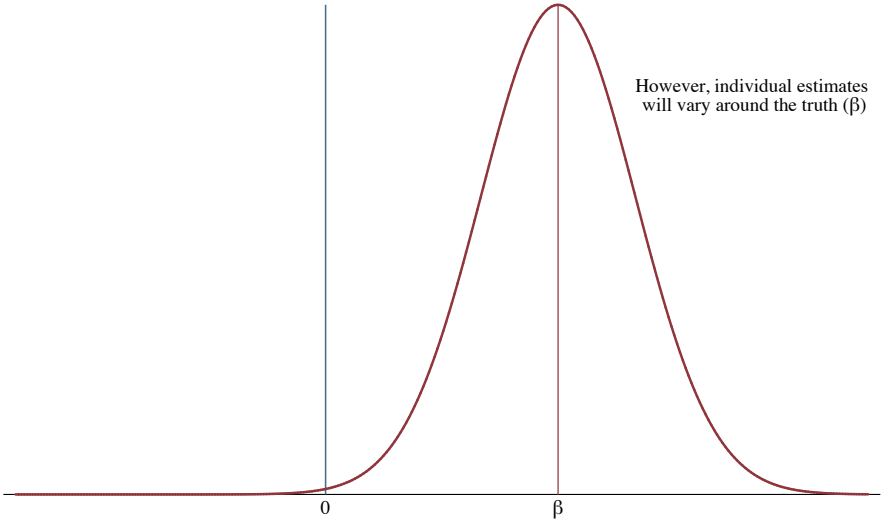
Power



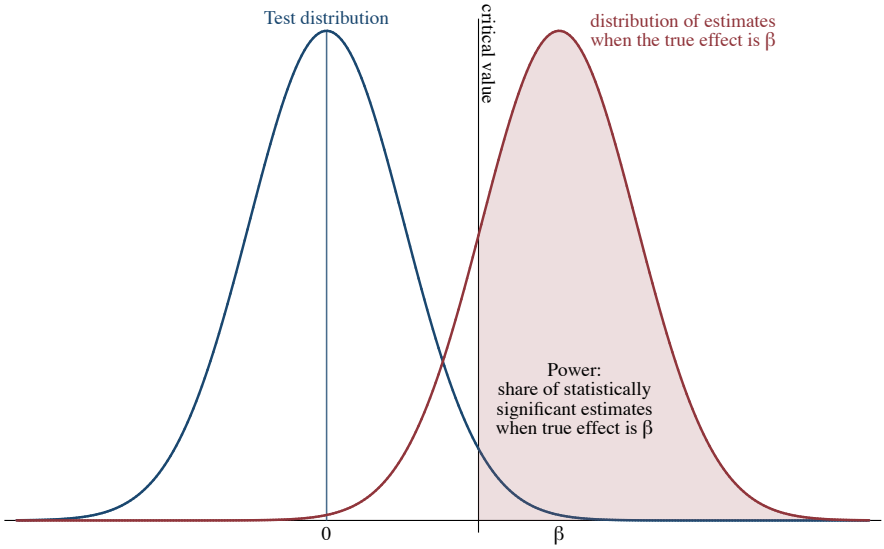
Power


Test distribution
critical value
truth
Suppose the true effect is β
An estimate of size β is significant
0
β

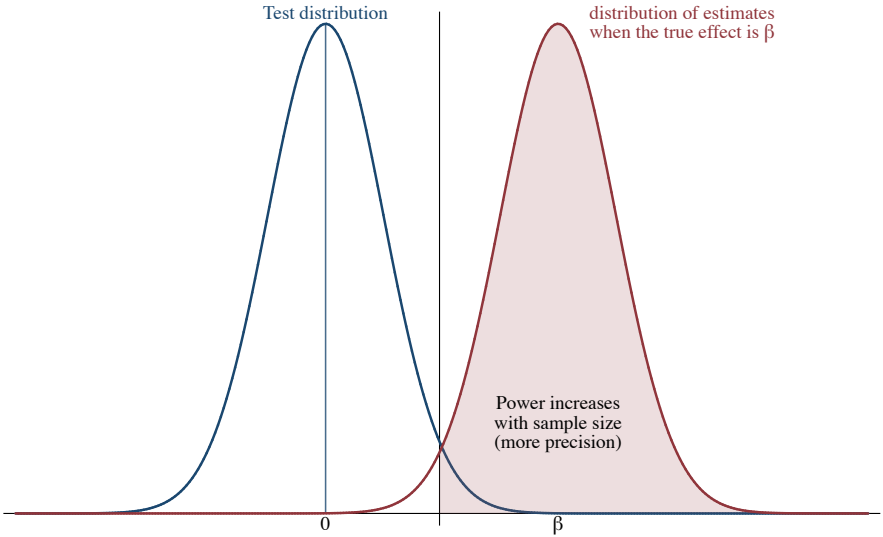
Power


However, individual estimates
will vary around the truth (β)
0
β

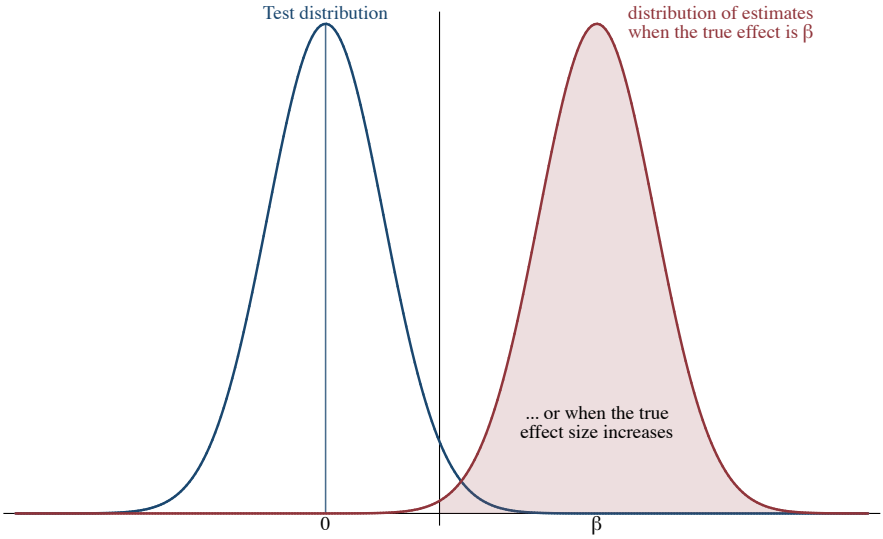
Power


Test distribution
critical value
distribution of estimates
when the true effect is β
Power:
share of statistically
significant estimates
when true effect is β
0
β

Power


Test distribution
distribution of estimates
when the true effect is β
Power increases
with sample size
(more precision)
0
β

Power


Test distribution
distribution of estimates
when the true effect is β
... or when the true
effect size increases
0
β

Question of power

1. How large would the true effect need to be in order for us to have sufficient power?

Question of power

1. How large would the true effect need to be in order for us to have sufficient power?
2. How large an experiment do we need, in order to be able to detect an effect of a certain size?

Summary

- We discussed two kinds of errors
  - statistical: well-defined properties of statistical tests
  - human: messy reality of how people (mis)use/interpret statistics

Summary

- We discussed two kinds of errors
  - statistical: well-defined properties of statistical tests
  - human: messy reality of how people (mis)use/interpret statistics
- Key concepts to understand
  - false negative vs false positive, statistical power

Summary

- We discussed two kinds of errors
  - statistical: well-defined properties of statistical tests
  - human: messy reality of how people (mis)use/interpret statistics
- Key concepts to understand
  - false negative vs false positive, statistical power
- Ways to avoid human errors
  - being alert and suspicious (particularly regarding your own results)
  - tying one's hands: pre-registration, replication, machine learning...

Moving to Opportunity

Moving to Opportunity

- One of the most famous social experiments of all time
  - target group: households with children living in high-poverty public housing projects (primarily minority, single mother families)
  - implemented in 1994-98 in Baltimore, Boston, Chicago, LA, New York

Moving to Opportunity

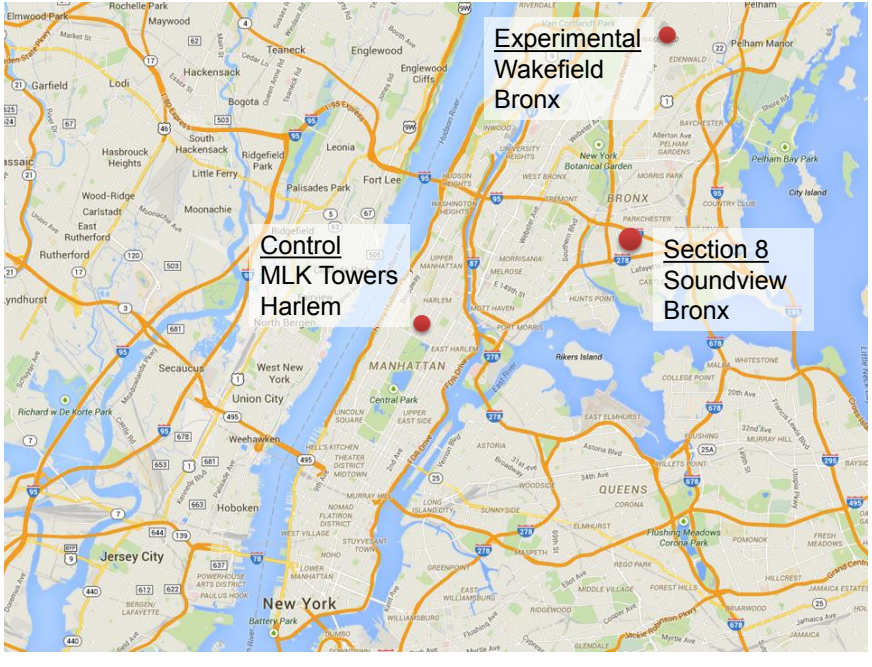
- One of the most famous social experiments of all time
  - target group: households with children living in high-poverty public housing projects (primarily minority, single mother families)
  - implemented in 1994-98 in Baltimore, Boston, Chicago, LA, New York
- Random assignment of 4,600 families into three groups:
  - control: not offered a voucher, stayed in public housing
  - section 8: offered conventional housing vouchers, no restrictions
  - experimental: offered housing vouchers to low-poverty neighborhoods

Moving to Opportunity

- One of the most famous social experiments of all time
  - target group: households with children living in high-poverty public housing projects (primarily minority, single mother families)
  - implemented in 1994-98 in Baltimore, Boston, Chicago, LA, New York
- Random assignment of 4,600 families into three groups:
  - control: not offered a voucher, stayed in public housing
  - section 8: offered conventional housing vouchers, no restrictions
  - experimental: offered housing vouchers to low-poverty neighborhoods
- Many families chose not to use the voucher they were offered
  - 48% of experimental group used voucher
  - 66% of Section 8 group used voucher

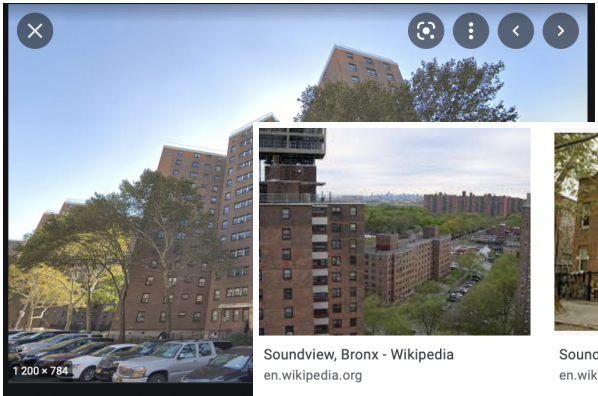
The MTO parts of these slides draw heavily from lecture 3 of Raj Chetty's excellent course Using Big Data to Solve Economic and Social Problems. I'm also borrowing quite a bit from Tuukka Saarimaa's (also excellent) Urban Economics course.

Common MTO Residential Locations in New York



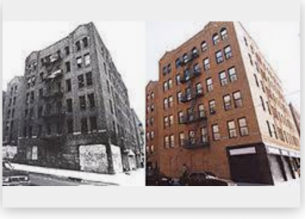
First results from a Google image search, Jan 28th, 2022

First results from a Google image search, Jan 28th, 2022



Soundview, Bronx - Wikipedia
en.wikipedia.org

Soundview, Bronx - Wikipedia
en.wikipedia.org



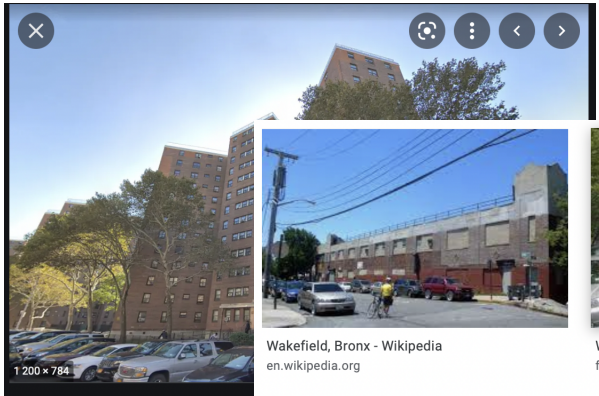
Soundview, Bronx - Wikipedia
en.wikipedia.org

Bronx dad, 26, shot to death as he sat with ...
nydailynews.com

Father of three shot dead outside Bronx housi...
nydailynews.com

Soundview Projects in New Yor...
virtualglobetrotting.com

First results from a Google image search, Jan 28th, 2022



Wakefield, Bronx - Wikipedia
en.wikipedia.org

WAKEFIELD, Bronx - Forgotten New Y...
forgotten-ny.com

Living in Wakefield, the Bronx - Slide Show ...
nytimes.com

Wakefield–241st Street station - Wikip...
en.wikipedia.org

4K60 Walking NYC's Northernmost Neighborhood : ...
youtube.com

WAKEFIELD, Bronx - Forgotten New York
forgotten-ny.com

Early research

MEAN EFFECT SIZES FOR SUMMARY MEASURES OF OUTCOMES[a]

	All Adults
	E – C (i)
Economic self-sufficiency	0.017 (0.031)
Absence of physical health problems	0.012 (0.024)
Absence of mental health problems	0.079* (0.030)
Absence of risky behavior	
Education	
Overall	0.036 (0.020)

[a] E – C denotes experimental – control

Robust standard errors adjusted for household clustering are in parentheses; * = p-value < 0.05.

Kling, Liebman, Katz (2007): Experimental Analysis of Neighborhood Effects.

Early research

MEAN EFFECT SIZES FOR SUMMARY MEASURES OF OUTCOMES[a]

	All Adults	
	E – C (i)	S – C (ii)
Economic self-sufficiency	0.017 (0.031)	0.037 (0.033)
Absence of physical health problems	0.012 (0.024)	0.019 (0.026)
Absence of mental health problems	0.079* (0.030)	0.029 (0.033)
Absence of risky behavior		
Education		
Overall	0.036 (0.020)	0.028 (0.022)

[a]E – C denotes experimental – control; S – C denotes Section 8 – control. Estimates are the intent-to-treat mean effect sizes,

Robust standard errors adjusted for household clustering are in parentheses; * = p-value < 0.05.

Kling, Liebman, Katz (2007): Experimental Analysis of Neighborhood Effects.

MEAN EFFECT SIZES FOR SUMMARY MEASURES OF OUTCOMES[a]

	All Adults		All Youth		Female Youth		Male Youth		M – F Youth	
	E – C	S – C	E – C	S – C	E – C	S – C	E – C	S – C	E – C	S–C
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)
Economic self-sufficiency	0.017	0.037								
	(0.031)	(0.033)								
Absence of physical health problems	0.012	0.019	−0.038	−0.020	0.025	0.077	−0.112*	−0.114	−0.138	−0.192*
	(0.024)	(0.026)	(0.038)	(0.040)	(0.053)	(0.055)	(0.053)	(0.061)	(0.076)	(0.084)
Absence of mental health problems	0.079*	0.029	0.102	0.138*	0.267*	0.192*	−0.052	0.054	−0.319*	−0.138
	(0.030)	(0.033)	(0.053)	(0.056)	(0.062)	(0.067)	(0.080)	(0.092)	(0.101)	(0.113)
Absence of risky behavior			−0.023	−0.039	0.142*	0.129*	−0.181*	−0.208*	−0.323*	−0.337*
			(0.043)	(0.050)	(0.053)	(0.059)	(0.062)	(0.071)	(0.080)	(0.092)
Education			0.050	0.028	0.138*	0.056	−0.053	−0.001	−0.191*	−0.057
			(0.041)	(0.047)	(0.065)	(0.068)	(0.047)	(0.060)	(0.080)	(0.090)
Overall	0.036	0.028	0.018	0.018	0.136*	0.109*	−0.099*	−0.078*	−0.235*	−0.187*
	(0.020)	(0.022)	(0.025)	(0.026)	(0.034)	(0.034)	(0.031)	(0.037)	(0.047)	(0.051)

[a] E – C denotes experimental − control; S – C denotes Section 8 − control. Estimates are the intent-to-treat mean effect sizes, from Equation (1), fully interacted with gender in columns (v)–(x) as described in the text. The estimated equations all include site indicators and the baseline covariates listed in Appendix A with those in Table A1 included for adults and those in Tables A1 and A2 included for youth. M – F Youth is male − female difference. Adult economic self-sufficiency: + adult not employed and not on TANF + employed + 2001 earnings − on TANF − 2001 government income. Adult mental health: − distress index − depression symptoms − worrying + calmness + sleep. Adult physical health: − self-reported health fair/poor − asthma attack past year − obesity − hypertension − trouble carrying/climbing. Adult overall includes 15 measures in self-sufficiency, physical health, and mental health. Youth physical health: − self-reported health fair/poor − asthma attack past year − obesity − nonsports injury past year. Youth mental health: − distress index − depression symptoms − anxiety symptoms. Youth risky behavior: − marijuana past 30 days − smoking past 30 days − alcohol past 30 days − ever pregnant or gotten someone pregnant. Youth education: + graduated high school or still in school + in school or working + WJ-R broad reading score + WJ-R broad math score. Youth overall includes 15 measures in physical health, mental health, risky behavior, and education. Sample sizes in the E, S, and C groups are 1,453, 993, and 1,080 for adults and 749, 510, and 548 for youth ages 15–20 on 12/31/2001. Robust standard errors adjusted for household clustering are in parentheses; * = p-value < 0.05.

Kling, Liebman, Katz (2007): Experimental Analysis of Neighborhood Effects.

Early research

- Making sense of the previous table
  - outcomes: indices that aggregate information over multiple measures
    - for example, the index of economic self-sufficiency includes five measures of employment, earnings, and public assistance
  - each index has mean 0 and standard deviation 1

Early research

- Making sense of the previous table
  - outcomes: indices that aggregate information over multiple measures
    - for example, the index of economic self-sufficiency includes five measures of employment, earnings, and public assistance
  - each index has mean 0 and standard deviation 1
- Impacts of being offered an experimental voucher (4–7 years later)
  - no effects on adult economic self-sufficiency or physical health

Early research

- Making sense of the previous table
  - outcomes: indices that aggregate information over multiple measures
    - for example, the index of economic self-sufficiency includes five measures of employment, earnings, and public assistance
  - each index has mean 0 and standard deviation 1
- Impacts of being offered an experimental voucher (4–7 years later)
  - no effects on adult economic self-sufficiency or physical health
  - improved mental health for adults
  - positive effect on teenage girls
  - negative effect on teenage boys

The most recent results

- Chetty, Hendren, Katz (2016) focus on those moving as children
  - group 1: younger than 13 (average 8.2) at assignment
  - group 2: 13-18 years old (average 15.1) at assignment

The most recent results

- Chetty, Hendren, Katz (2016) focus on those moving as children
  - group 1: younger than 13 (average 8.2) at assignment
  - group 2: 13-18 years old (average 15.1) at assignment
- MTO data linked to 1996–2012 federal income tax returns
  - 4,604 households and 15,892 individuals
    - primary focus on 8,603 children born in or before 1991
  - about 85% of children matched
    - match rates do not differ significantly across treatment groups
    - baseline covariates balanced across treatment groups in matched data

The most recent results

- Chetty, Hendren, Katz (2016) focus on those moving as children
  - group 1: younger than 13 (average 8.2) at assignment
  - group 2: 13-18 years old (average 15.1) at assignment
- MTO data linked to 1996–2012 federal income tax returns
  - 4,604 households and 15,892 individuals
    - primary focus on 8,603 children born in or before 1991
  - about 85% of children matched
    - match rates do not differ significantly across treatment groups
    - baseline covariates balanced across treatment groups in matched data
- Using administrative data (tax records) is quite new in the US
  - earlier work based typically on survey data
  - in the Nordic countries, we have a long tradition (and much better infrastructure) for using administrative data in research

Compliance

- Often only part of the treatment group actually gets the treament
  - e.g. only 48% of those randomized into the experimental group in MTO chose to use the voucher (column 1 of the previous slide)
  - similarly, 66% of the section 8 group used the voucher

Compliance

- Often only part of the treatment group actually gets the treament
  - e.g. only 48% of those randomized into the experimental group in MTO chose to use the voucher (column 1 of the previous slide)
  - similarly, 66% of the section 8 group used the voucher

- Compliance *choice* is potentially affected by potential outcomes
  - e.g. those expecting to benefit the least becoming never-takers

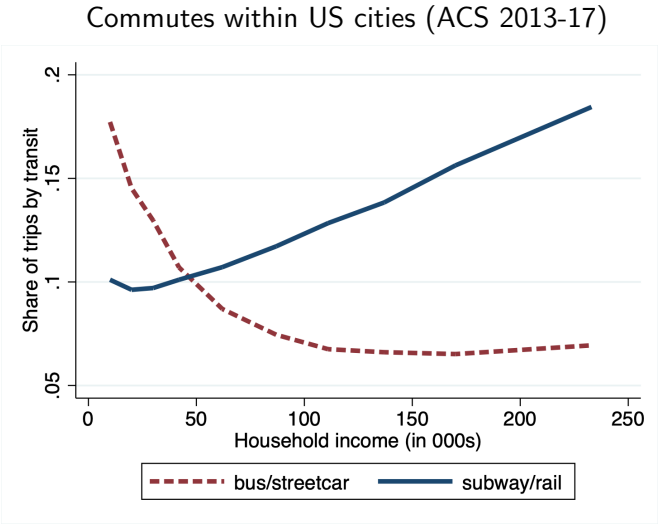
  $\rightarrow$ comparing those who actually gets the treatment to the entire control group is not a valid comparison

The data we observe

What we typically observe

- in the absence of experimental settings
- is one state of the world in equilibrium
- We can still learn from choices people have made in this equilibrium!
  - with the same tools we have been learning for experimental settings.
- Rest of today: an application

Public transit ridership in US cities

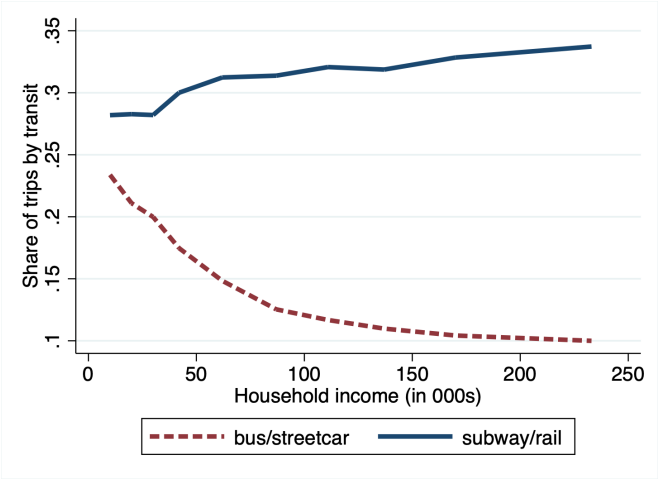


- Low-income commuters ride bus more
- High-income commuters ride subway/rail more
- Why? When typically no difference in fares between bus and rail transit?
- Why do low-income commuters appear to ride bus more than rail transit?

Akbar (2023): Public Transit Access and Income Segregation.

Public transit ridership in US cities

Commutes within US cities with both road and rail transit ridership $> 5\%$ (ACS 2013-17)

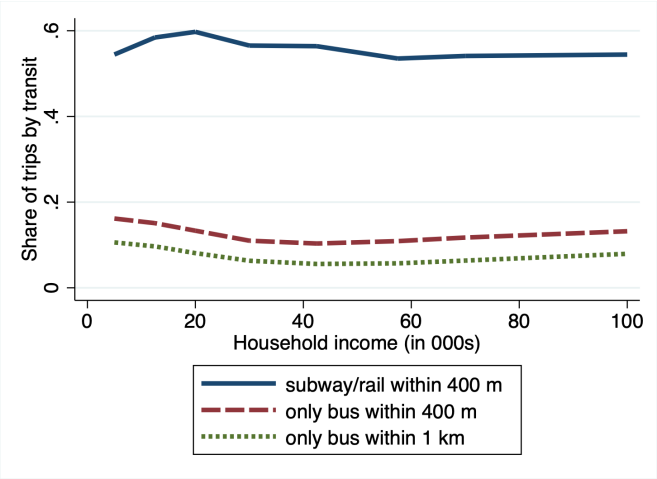


- Many cities with no rail transit. They are typically lower-income cities.
- Figure makes more sense when we focus on cities with both road and rail transit.
- But still bus ridership is decreasing and rail ridership is increasing with income!

Akbar (2023): Public Transit Access and Income Segregation.

Public transit ridership in US cities

Commutes within US cities by proximity to transit stop (ACS 2013-17)

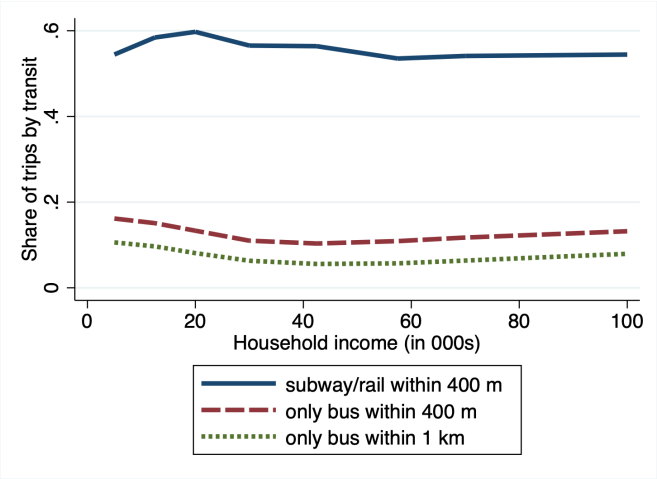


- More transit ridership closer to transit stops, especially if also close to rail transit stop.

Akbar (2023): Public Transit Access and Income Segregation.

Public transit ridership in US cities

Commutes within US cities by proximity to transit stop (ACS 2013-17)

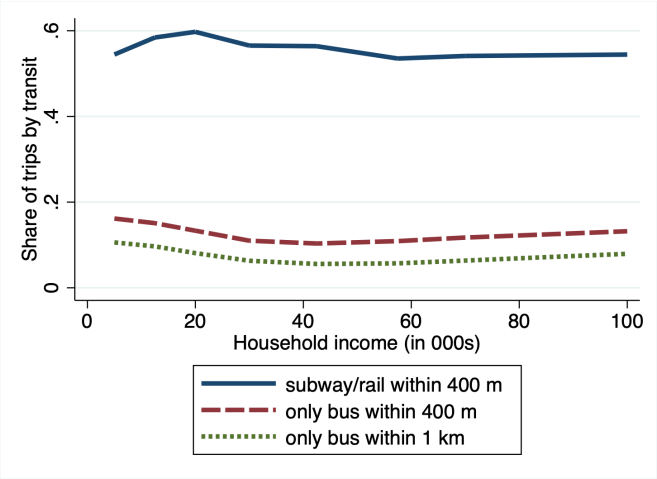


- More transit ridership closer to transit stops, especially if also close to rail transit stop.
- But conditional on proximity to stop, no notable difference in ridership by income!

Akbar (2023): Public Transit Access and Income Segregation.

Public transit ridership in US cities

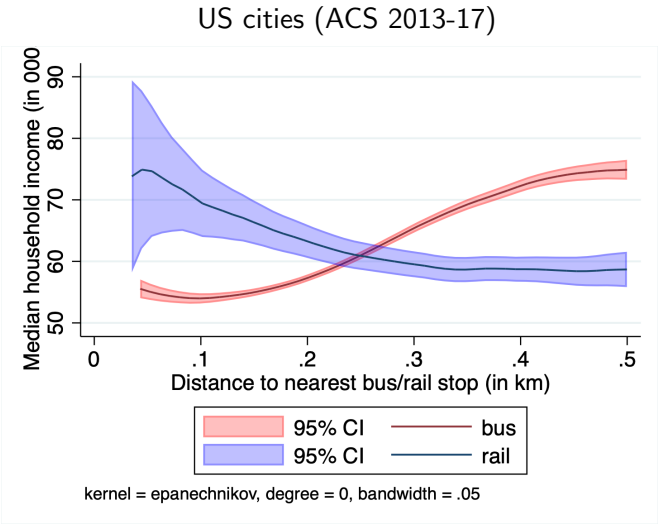
Commutes within US cities by proximity to transit stop (ACS 2013-17)



- More transit ridership closer to transit stops, especially if also close to rail transit stop.
- But conditional on proximity to stop, no notable difference in ridership by income!
- So, why the difference unconditional on proximity?

Akbar (2023): Public Transit Access and Income Segregation.

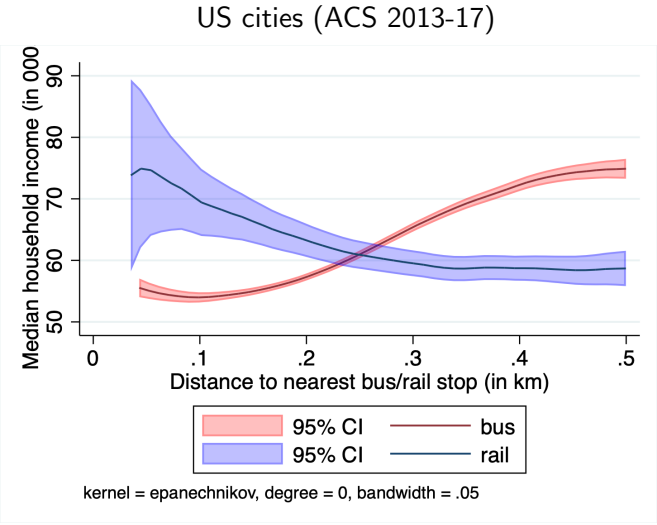
Median household incomes by proximity to transit stops



Unpublished ongoing work for Akbar (2023): Public Transit Access and Income Segregation.

- Higher-income households reside closer to rail transit stops.
- Lower-income households reside closer to bus transit stops.

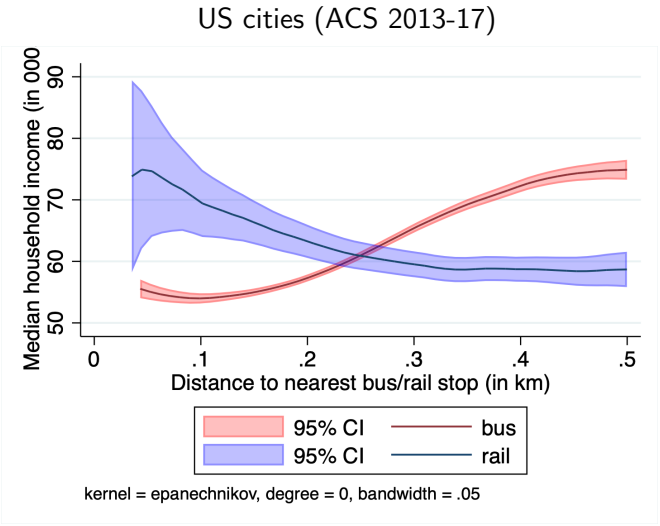
Median household incomes by proximity to transit stops



Unpublished ongoing work for Akbar (2023): Public Transit Access and Income Segregation.

- Higher-income households reside closer to rail transit stops.
- Lower-income households reside closer to bus transit stops.
- But why?

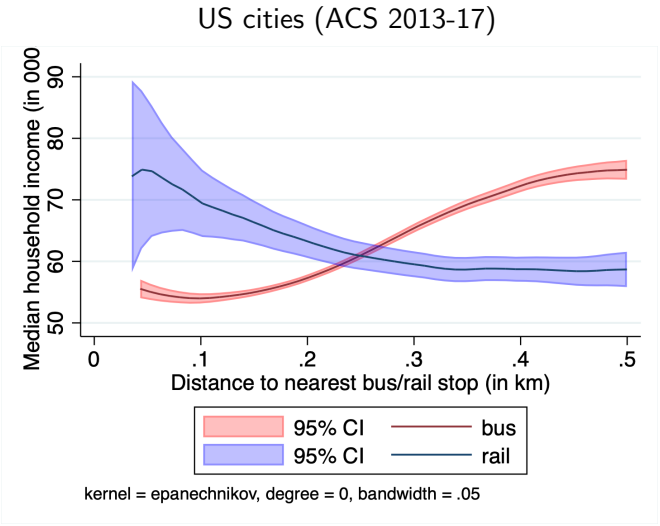
Median household incomes by proximity to transit stops



Unpublished ongoing work for Akbar (2023): Public Transit Access and Income Segregation.

- Higher-income households reside closer to rail transit stops.
- Lower-income households reside closer to bus transit stops.
- But why?
  - more expensive to reside near rail transit stops?

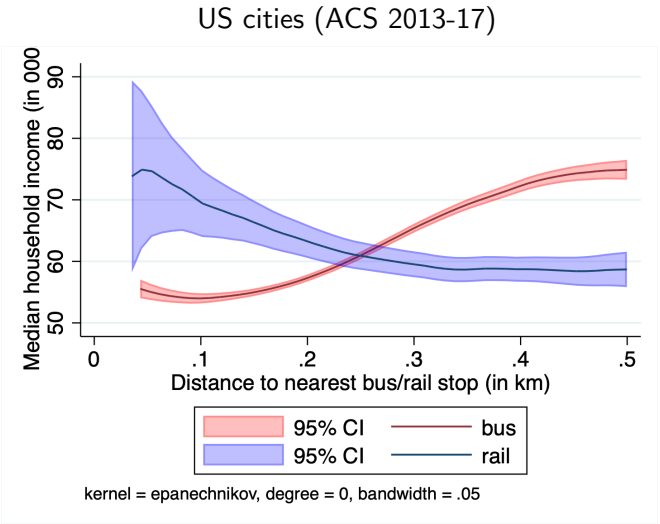
Median household incomes by proximity to transit stops



Unpublished ongoing work for Akbar (2023): Public Transit Access and Income Segregation.

- Higher-income households reside closer to rail transit stops.
- Lower-income households reside closer to bus transit stops.
- But why?
  - more expensive to reside near rail transit stops?
  - higher income households more willing to pay the higher housing costs?

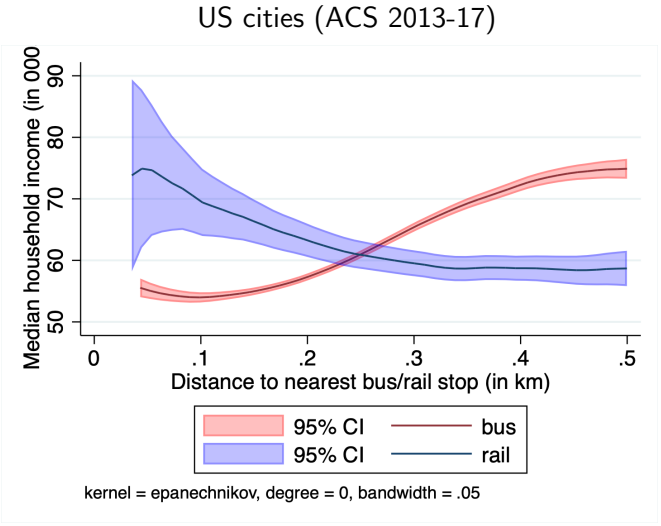
Median household incomes by proximity to transit stops



Unpublished ongoing work for Akbar (2023): Public Transit Access and Income Segregation.

- Higher-income households reside closer to rail transit stops.
- Lower-income households reside closer to bus transit stops.
- But why?
  - more expensive to reside near rail transit stops?
  - higher income households more willing to pay the higher housing costs?
  - rail transit operation targeted at high income neighborhoods?

Median household incomes by proximity to transit stops



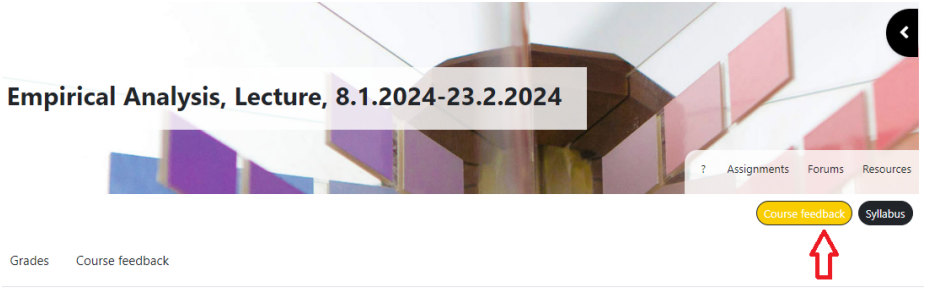
Unpublished ongoing work for Akbar (2023): Public Transit Access and Income Segregation.

- Higher-income households reside closer to rail transit stops.
- Lower-income households reside closer to bus transit stops.
- But why?
  - more expensive to reside near rail transit stops?
  - higher income households more willing to pay the higher housing costs?
  - rail transit operation targeted at high income neighborhoods?
  - Need more analytical structure for causal attribution.

Homework 3

Worksheet 4: Mid-period course feedback

Please complete the **Course feedback survey on MyCourses**.



We want the response sample to be representative of your population. So, extra incentive:

- 50% extra (on this worksheet) to everyone if $> 90\%$ response rate.

Enjoy the rest of the course!