# Introduction to Data

Matti Sarvimäki

Principles of Empirical Analysis
Lecture 1

# This course

- Aims to *complement* MS-A0503
  - you do need to learn everything from MS-A0503!
  - ... but you need more to make sense of real-world data

# This course

- Aims to *complement* MS-A0503
  - you do need to learn everything from MS-A0503!
  - ... but you need more to make sense of real-world data
- Here, we learn to ask questions such as
  - How was this data collected? How reliable is it?
  - Is that a large number? What should we compare it to?
  - How precise are the estimates? Do we have enough power?
  - Does this particular correlation imply causation?
  - What are the identifying assumptions of this research design?

# This course

- Aims to *complement* MS-A0503
    - you do need to learn everything from MS-A0503!
    - ... but you need more to make sense of real-world data
- Here, we learn to ask questions such as
    - How was this data collected? How reliable is it?
    - Is that a large number? What should we compare it to?
    - How precise are the estimates? Do we have enough power?
    - Does this particular correlation imply causation?
    - What are the identifying assumptions of this research design?
- This course is about ways to learn about the world through observation and experimentation. It is not a math course
    - but: meaningful empirical inquiry is impossible without the math
    - we'll learn a few concepts and some notation along the way, but most of the math will be covered at MS-A0503

# Why this course?

- Economics is increasingly empirical
  - empirical work dominates some subfields
  - important for all branches of economics

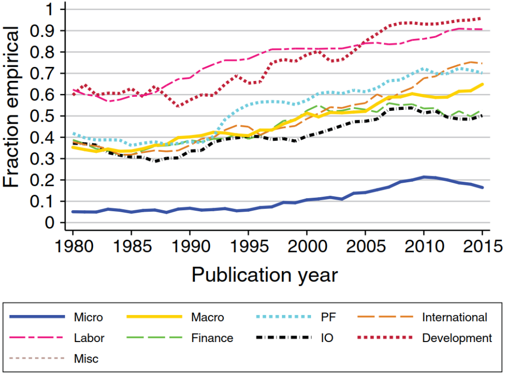



FIGURE 4. WEIGHTED FRACTION EMPIRICAL BY FIELD

*Source:* Angrist, J., Azoulay, P., Ellison, G., Hill R. and S. Lu 2017. Economic research evolves: Fields and styles. American Economic Review, Papers and Proceedings, 107, 5, 293-297.

# Why this course?

- Economics is increasingly empirical
  - empirical work dominates some subfields
  - important for all branches of economics
- Theory and empirics are complements
  - theories need to be tested and quantified
  - empirical findings need to interpreted

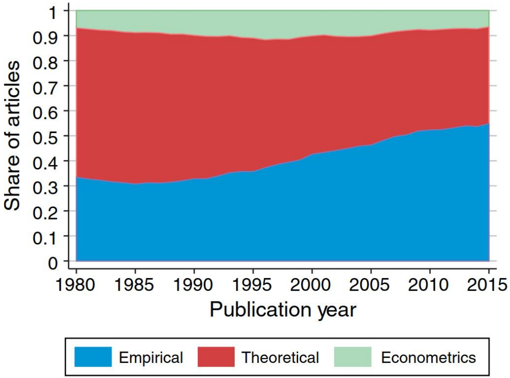
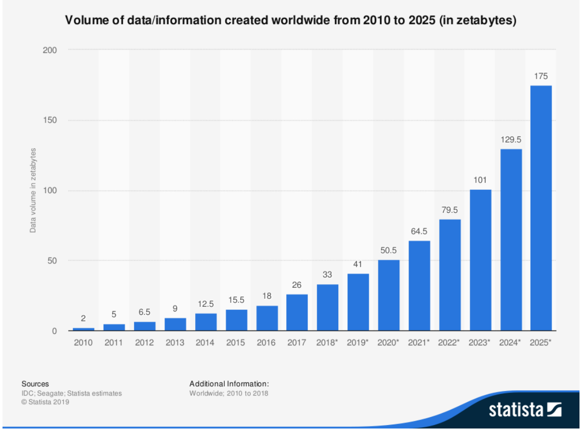



FIGURE 6. WEIGHTED PUBLICATIONS BY STYLE

*Source:* Angrist, J., Azoulay, P., Ellison, G., Hill R. and S. Lu 2017. Economic research evolves: Fields and styles. American Economic Review, Papers and Proceedings, 107, 5, 293-297.

# Why this course?

- Economics is increasingly empirical
  - empirical work dominates some subfields
  - important for all branches of economics
- Theory and empirics are complements
  - theories need to be tested and quantified
  - empirical findings need to interpreted
- New opportunities are constantly emerging due to
  - more (digital) data becoming available
  - improvements in computing power
- But, old mistakes are still being made
  - more data is wonderful, but not a cure-all




*Source:* Statista.com. 1 zetabite = 1 billion terabytes = $1000^7$ bytes.

# Types of empirical research

- Three complementary approaches
  1. Descriptive: summarizing data, establishing facts
  2. Causal: how X *affects* Y?
  3. Prediction: how X *predicts* Y?

# Types of empirical research

- Three complementary approaches
  1. Descriptive: summarizing data, establishing facts
  2. Causal: how X *affects* Y?
  3. Prediction: how X *predicts* Y?
- Example: ice cream and forest fires
  - descriptive: strong correlation between the two
  - causal: banning ice cream would probably not reduce forest fires
  - prediction: if all we observed was ice cream sales, we probably should use it for preparing for forest fires

# Types of empirical research

- This course is about descriptive and causal work
  - reflects the focus of most economics research
  - data science more focused on prediction

# Types of empirical research

- This course is about descriptive and causal work
  - reflects the focus of most economics research
  - data science more focused on prediction
- Lucas critique and "black box" prediction models
  - data is generated by agents who make optimal decisions given policy $\rightarrow$ if policy changes, they will make different decisions
  - thus predicting the impact of policy changes by naively extrapolating from historical data may be misleading

# Types of empirical research

- This course is about descriptive and causal work
  - reflects the focus of most economics research
  - data science more focused on prediction
- Lucas critique and "black box" prediction models
  - data is generated by agents who make optimal decisions given policy $\rightarrow$ if policy changes, they will make different decisions
  - thus predicting the impact of policy changes by naively extrapolating from historical data may be misleading
- This is not to say that prediction is useless
  - black box prediction is a powerful tool for some tasks
  - prediction tools used as inputs in descriptive and causal analysis
  - prediction often follows from causal inference and quantitative models
  - the intersection of econometrics, statistics and machine learning is one of the most existing methodological frontiers at the moment

# Practicalities

- Lectures: Matti Sarvimäki, Tuukka Saarimaa
  - weeks 1–3: Mondays 14.15–15.45, Wednesdays 13.15–14.45
  - weeks 4–6: Mondays 14.15–15.45, Thursdays 9.15–10.45
  - office hours: upon request
- Exercises: Riku Buri
  - weeks 1–3: Thursdays 9.15–10.45
  - weeks 4–6: Wednesdays 13.15–14.45
  - return your exercises before each session
    - first deadline: Jan 21st
    - submit written answers and code through Mycourses
    - you can download Stata from download.aalto.fi; other statistical programs such as R are also fine
  - we encourage you to work together with the exercises
  - ... but everyone has to return unique answers and code

# Practicalities

- Exam: Feb 23rd (retake April 26th)
    - online, open books
    - details announced later
- Materials
    - lectures and exercise (required)
        - everything available at course website
        - lectures will be recorded
    - helpful textbooks (optional)
        - Angrist and Pischke: Mastering 'Metrics, any edition
        - Cunningham Causal Inference: The Mixtape
    - helpful videos: Mastering Econometrics
        - partly used as pre-class assignment material
- Grading: 50% exercises and pre-class assignments, 50% exam
    - max 5% extra for class activity

# Course outline and learning objectives

- Data and measurement
  1. introduction, data
  2. samples and descriptive statistics
  3. more descriptive statistics
- Experimental methods
  1. causality and research designs
  2. statistical significance
  3. statistical power
  4. noncompliance
- Quasi-experimental methods
  1. observational data and quasi-experiments
  2. difference-in-difference (DiD)
  3. regression discontinuity design (RDD)
  4. regression and matching
- Structural methods

# Course outline and learning objectives

- Data and measurement
  1. introduction, data
  2. samples and descriptive statistics
  3. more descriptive statistics
- Experimental methods
  1. causality and research designs
  2. statistical significance
  3. statistical power
  4. noncompliance
- Quasi-experimental methods
  1. observational data and quasi-experiments
  2. difference-in-difference (DiD)
  3. regression discontinuity design (RDD)
  4. regression and matching
- Structural methods

- After this course you
  1. **can critically evaluate the data needs and the data used for a given empirical question**. (today)
  2. can describe and interpret data in a meaningful way, linking it to questions of interest and to economic theory.
  3. understand the concept of causality and the basic principles of randomized control trials.
  4. can interpret and critically evaluate the economic and statistical significance of the results of an empirical analysis.
  5. are familiar with the basics of the most important methods for estimating causal effects using data from quasi-natural experiments.

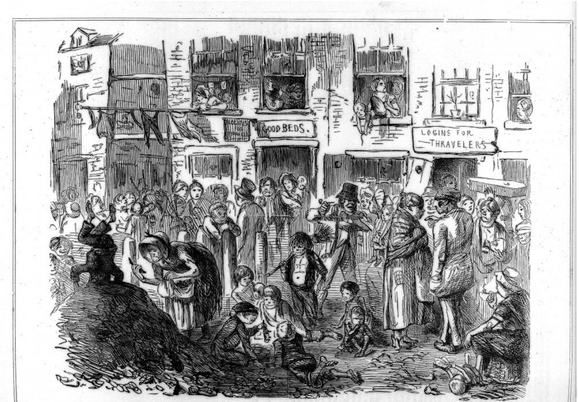
# Today

- The power and sources of data
  - data-driven decision making is in vogue
  - on balance, this is real progress ... but bad data and bad analysis can also lead to spectacular mistakes
- Aim: Learning to ask two types of questions
  - How was this data collected? How reliable is it?
  - Is that a large number? What should we compare it to?

# The power and sources of data

# Cholera in Victorian London

- Cholera arrived in London in 1831
  - "The combination of scary symptoms and fear of the unknown seized the public's imagination and chlolera was characterised as a foreign epidemic (it was commonly known as Asiatic cholera), which was 'invading' the nation."

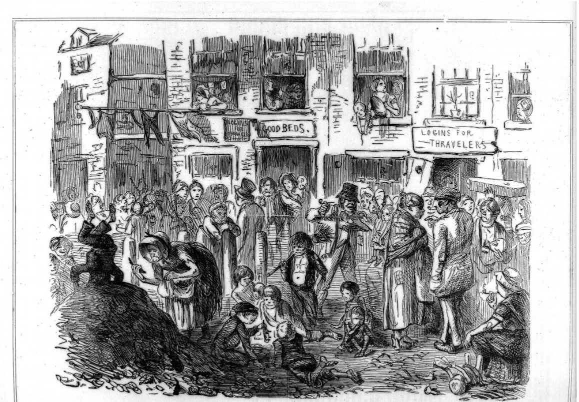


A COURT FOR KING CHOLERA.

*Source:* Science Museum

# Cholera in Victorian London

- Cholera arrived in London in 1831
  - "The combination of scary symptoms and fear of the unknown seized the public's imagination and chlolera was characterised as a foreign epidemic (it was commonly known as Asiatic cholera), which was 'invading' the nation."
- Competing theories of cholera's causes
  - miasmas: particles in the air from decaying matter ("smell is disease")
  - germs: unknown germ transmitted by individuals ingesting water
- Both consistent with London's extremely bad sanitation conditions at the time




A COURT FOR KING CHOLERA.

*Source:* Science Museum

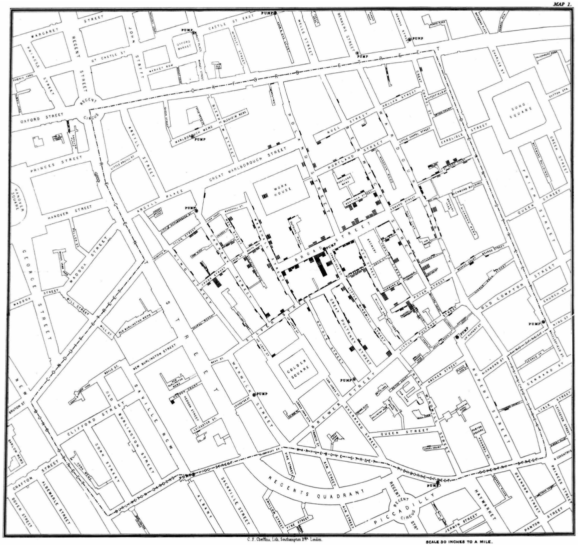
# John Snow and the 1854 Broad Street Outbreak

- A particularly severe outbreak occurred in 1854 near Broad Street in Soho
  - 127 people died in three days
- John Snow identified the source as the public water pump on Broad Street and convinced authorities to disable it by removing its handle
  - initially: talking to local residents


John Snow memorial on Broadwick Street, Soho, London.

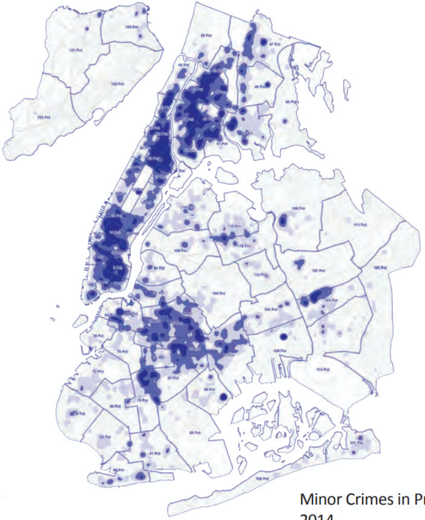
# John Snow and the 1854 Broad Street Outbreak

- A particularly severe outbreak occurred in 1854 near Broad Street in Soho
  - 127 people died in three days
- John Snow identified the source as the public water pump on Broad Street and convinced authorities to disable it by removing its handle
  - initially: talking to local residents
  - later: map showing how cholera cases were clustered around this water pump
- This is just one example of how systematic data collection revolutionized medicine and public health



Original map by John Snow showing the clusters of cholera cases in the London epidemic of 1854. *Source:* Wikipedia.

- Let's discuss your pre-class case
    - what was Jack Maple's big idea?



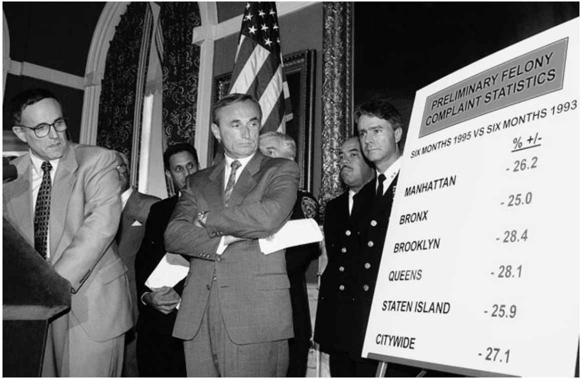
Minor Crimes in Progress- 911
2014

Figure 11

*Source:* urbanomnibus.net, 20 June 2018.

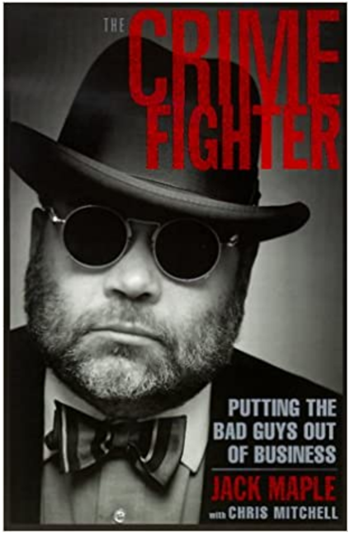
# Jack Maple and NYC Policing

- Let's discuss your pre-class case
  - what was Jack Maple's big idea?
  - why was it such a big success?




Mayor Rudy Giuliani and Commissioner William Bratton discussing crime statistics. *Source:* The Activist History Review

- Let's discuss your pre-class case
  - what was Jack Maple's big idea?
  - why was it such a big success?
  - do we actually know it was a success?




*Source:* Amazon

- Let's discuss your pre-class case
  - what was Jack Maple's big idea?
  - why was it such a big success?
  - do we actually know it was a success?
- CompStat used as a management tool
  - subjecting police commanders to scrutiny gives them incentives to perform better


"Valdez defended his work, and that of the detectives from the 40th Precinct, who're standing behind him. "We're up in robberies, assaults, burglaries and grand larcenies, but we're also up in arrests for those index crimes," said Valdez, his voice cracking. The brass didn't let up, on Valdez or a troop of other precinct commanders who appear at the podium to parry questions about crime on their turf." *Source:* NBC News, 17 April 2016.

# Jack Maple and NYC Policing

- Let's discuss your pre-class case
  - what was Jack Maple's big idea?
  - why was it such a big success?
  - do we actually know it was a success?
- CompStat used as a management tool
  - subjecting police commanders to scrutiny gives them incentives to perform better … and to fabricate data
    - underreporting of serious crime
    - unnecessary arrests to meet targets

The New York Times

## *Retired Officers Raise Questions on Crime Data*

46

By William K. Rashbaum

Feb. 6, 2010

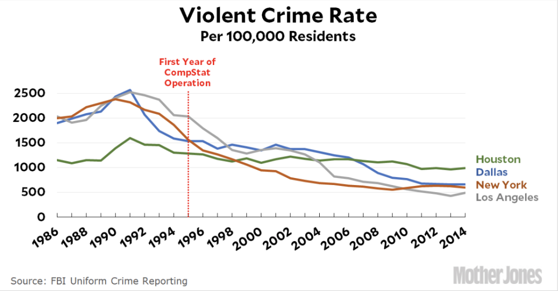
More than a hundred retired New York Police Department captains and higher-ranking officers said in a survey that the intense pressure to produce annual crime reductions led some supervisors and precinct commanders to manipulate crime statistics, according to two criminologists studying the department.

*Source:* New York Times, 6 Feb 2010.

# Jack Maple and NYC Policing

- Let's discuss your pre-class case
  - what was Jack Maple's big idea?
  - why was it such a big success?
  - do we actually know it was a success?
- CompStat used as a management tool
  - subjecting police commanders to scrutiny gives them incentives to perform better … and to fabricate data
    - underreporting of serious crime
    - unnecessary arrests to meet targets
- Unclear if CompStat affected crime
  - even if we believe the crime statistics
  - we will talk a lot about causal inference from lecture 4 onwards

Crime began dropping in New York years before CompStat was up and running. Crime dropped in Los Angeles and Toronto and Stockholm at the same time, even though they didn't have CompStat. And crime dropped everywhere in New York City, regardless of how strongly various areas were targeted by CompStat.



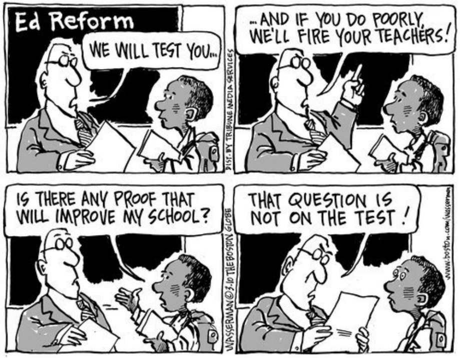

*Source:* MotherJones, 2 March 2018.

# The Point: Know your data!

- Every serious data analysis starts with a question:
  Where is this data coming from?
  - how and why was it collected?
  - who are included, who are excluded?
  - how exactly are variables defined?
  - what are the incentives of the data publisher?
  - what are the incentives of the respondents?

# The Point: Know your data!

- Every serious data analysis starts with a question: Where is this data coming from?
    - how and why was it collected?
    - who are included, who are excluded?
    - how exactly are variables defined?
    - what are the incentives of the data publisher?
    - what are the incentives of the respondents?




*Source:* Kentucky School News and Commentary

# The Point: Know your data!

- Every serious data analysis starts with a question: Where is this data coming from?
  - how and why was it collected?
  - who are included, who are excluded?
  - how exactly are variables defined?
  - what are the incentives of the data publisher?
  - what are the incentives of the respondents?
- There is a lot of good *and* bad data out there!
  - data does not need to be perfect, but you need to understand each dataset's strengths and limitations
  - true regardless of whether you are doing the analysis or consuming other people's analysis

The New York Times

***Atlanta Educators Convicted in School Cheating Scandal***

Donald Bullock, a former Atlanta testing coordinator, was led to a holding cell after his conviction. A judge ordered most of the educators jailed immediately. Pool photo by Kent D. Johnson

*Source:* New York Times, 1 April 2015.

# Data sources

- Bedrock: national statistical offices and alike
  - permanent, standardized surveys
    - e.g. census, labor force surveys
  - administrative register data
    - e.g. tax register, population register

Census enumerators in the 19th century UK (top) and 2020 US (bottom).

# Data sources

- Bedrock: national statistical offices and alike
  - permanent, standardized surveys
    - e.g. census, labor force surveys
  - administrative register data
    - e.g. tax register, population register
- Other sources
  - private proprietary data
    - e.g. tech companies, grocery store chains
    - sometimes shared with researchers
  - one-off surveys, digitalized archival records, scraped text, satellite data...

The New York Times Magazine

**How Companies Learn Your Secrets**

570

Antonio Bolfo/Reportage for The New York Times

By Charles Duhigg

Feb. 16, 2012

Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that? "

Pole has a master's degree in statistics and another in economics, and has been obsessed with the intersection of data and human behavior most of his life. His parents were teachers in North

*Source:* New York Times Magazine, 16 Feb 2012.

# Example 1: Deaths of Despair

# Example 1: Deaths of Despair

- In 2015, Anne Case and Angus Deaton published one of the most discussed papers of the past years
  - "This paper documents a marked increase in the all-cause mortality of middle-aged white non-Hispanic men and women in the United States between 1999 and 2013. This change reversed decades of progress in mortality and was unique to the United States [...]

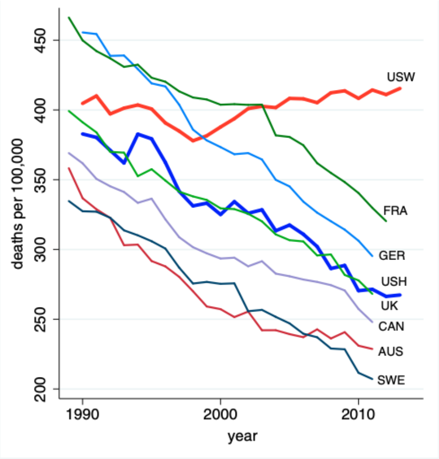



**Fig. 1.** All-cause mortality, ages 45–54 for US White non-Hispanics (USW), US Hispanics (USH), and six comparison countries: France (FRA), Germany (GER), the United Kingdom (UK), Canada (CAN), Australia (AUS), and Sweden (SWE).

*Source:* Case and Deaton (2015). PNAS

# Example 1: Deaths of Despair

- In 2015, Anne Case and Angus Deaton published one of the most discussed papers of the past years
  - "This paper documents a marked increase in the all-cause mortality of middle-aged white non-Hispanic men and women in the United States between 1999 and 2013. This change reversed decades of progress in mortality and was unique to the United States [...]
  - [...] largely accounted for by increasing death rates from drug and alcohol poisonings, suicide, and chronic liver diseases and cirrhosis."

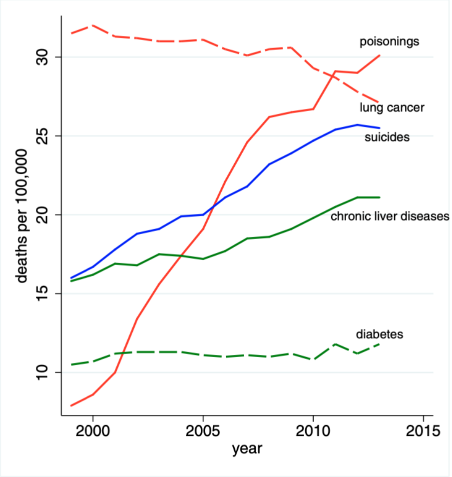
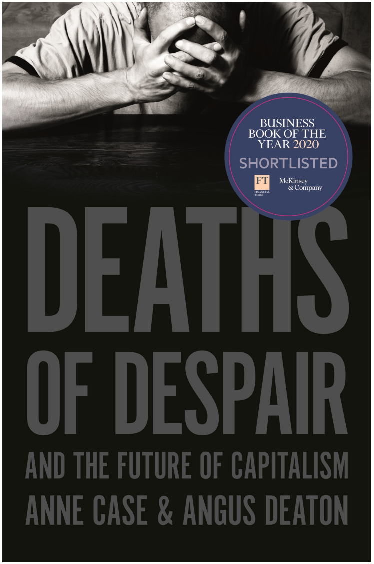



**Fig. 2.** Mortality by cause, white non-Hispanics ages 45–54.

*Source:* Case and Deaton (2015). PNAS

# Example 1: Deaths of Despair

- In 2015, Anne Case and Angus Deaton published one of the most discussed papers of the past years
  - "This paper documents a marked increase in the all-cause mortality of middle-aged white non-Hispanic men and women in the United States between 1999 and 2013. This change reversed decades of progress in mortality and was unique to the United States [...]
  - [...] largely accounted for by increasing death rates from drug and alcohol poisonings, suicide, and chronic liver diseases and cirrhosis."
- Much further research and hypothesis followed
  - "[...] a groundbreaking account of how the flaws in capitalism are fatal for America's working class."
- Question: How do we know mortality is rising?




*Source:* Princeton University Press.

# Example 1: Deaths of Despair

## Materials and Methods

### Mortality Data.

We assembled data on all-cause and cause-specific mortality from the CDC Wonder Compressed and Detailed Mortality files as well as from individual death records from 1989 to 2013. For population by ethnicity and educational status, we extracted data from American Community Surveys and, before 2000, from Current Population Surveys. International data on mortality were taken from the Human Mortality Database www.mortality.org; these are not separated by race and ethnicity. Specific causes of death are constructed for 1999–2013 using International Statistical Classification of Diseases and Related Health Problems 10$^{th}$ Revision (ICD10) codes: alcoholic liver diseases and cirrhosis (ICD10 K70, K73-74), suicide (X60-84, Y87.0), and poisonings (X40-45, Y10-15, Y45, 47, 49). Poisonings are accidental and intent-undetermined deaths from alcohol poisoning and overdoses of prescription and illegal drugs.

*Source:* Case and Deaton (2015). PNAS

- Answer: Governments systematically collect (high-quality) population data

# Example 1: Deaths of Despair

- Case and Deaton (2020) illustrate the magnitude of this trend as following
  - between 1999–2017, additional 600,000 mid-life Americans died in comparison to what would have happened if "progress had gone on as expected".
  - life-expectancy at birth for white Americans fell between 2013 and 2017

# Example 1: Deaths of Despair

- Case and Deaton (2020) illustrate the magnitude of this trend as following
  - between 1999–2017, additional 600,000 mid-life Americans died in comparison to what would have happened if "progress had gone on as expected".
  - life-expectancy at birth for white Americans fell between 2013 and 2017
- Are these large numbers?

# Example 1: Deaths of Despair

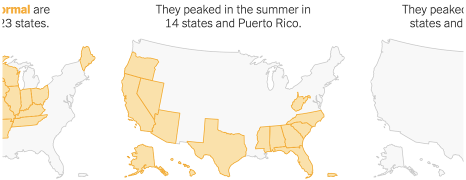
- Case and Deaton (2020) illustrate the magnitude of this trend as following
  - between 1999–2017, additional 600,000 mid-life Americans died in comparison to what would have happened if "progress had gone on as expected".
  - life-expectancy at birth for white Americans fell between 2013 and 2017
- Are these large numbers?
  - roughly 670,000 Americans have died to HIV/AIDS since the early 1980s
  - life-expectancy has never fell in the US for three years in a row since 1933 (start of full data)
    - in a subset of states for which earlier data is available, the only precedent is for 1915–1918

# Example 1: Deaths of Despair

- Case and Deaton (2020) illustrate the magnitude of this trend as following
  - between 1999–2017, additional 600,000 mid-life Americans died in comparison to what would have happened if "progress had gone on as expected".
  - life-expectancy at birth for white Americans fell between 2013 and 2017
- Are these large numbers?
  - roughly 670,000 Americans have died to HIV/AIDS since the early 1980s
  - life-expectancy has never fell in the US for three years in a row since 1933 (start of full data)
    - in a subset of states for which earlier data is available, the only precedent is for 1915–1918

## True Pandemic Toll in the U.S. Reaches 377,000

By Josh Katz, Denise Lu and Margot Sanger-Katz Updated Dec. 16, 2020




Deaths in every state of the country are higher than they would be in a normal year, according to an analysis of estimates from the Center for Disease Control and Prevention.

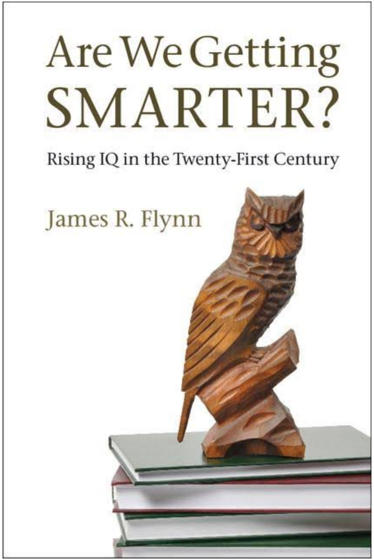
The data show how the coronavirus pandemic, which is peaking in many states, is bringing with it unusual patterns of death, higher than the official totals of deaths that have been directly linked to the virus.

*Source:* New York Times. Dec. 15, 2020.

# Example 2: The Flynn Effect

# Example 2: The Flynn Effect

- IQ tests need to be occasionally renormed to have mean 100, standard deviation 15
    - test groups tend to get higher scores with the old than with the new test → average IQ has increased at a rate of 0.2 SD per decade since the 1950s!
    - although this progress seem to have now stalled




*Source:* Cambridge University Press.

# Example 2: The Flynn Effect

- IQ tests need to be occasionally renormed to have mean 100, standard deviation 15
  - test groups tend to get higher scores with the old than with the new test $\rightarrow$ average IQ has increased at a rate of 0.2 SD per decade since the 1950s!
  - although this progress seem to have now stalled
- Is that 0.2 SD per decade a large number?

# Example 2: The Flynn Effect

- IQ tests need to be occasionally renormed to have mean 100, standard deviation 15
    - test groups tend to get higher scores with the old than with the new test → average IQ has increased at a rate of 0.2 SD per decade since the 1950s!
    - although this progress seem to have now stalled
- Is that 0.2 SD per decade a large number?
    - implies that in 1932, average IQ would have been about 80 and 37% of population below 75

# Example 2: The Flynn Effect

- IQ tests need to be occasionally renormed to have mean 100, standard deviation 15
  - test groups tend to get higher scores with the old than with the new test → average IQ has increased at a rate of 0.2 SD per decade since the 1950s!
  - although this progress seem to have now stalled
- Is that 0.2 SD per decade a large number?
  - implies that in 1932, average IQ would have been about 80 and 37% of population below 75
- Well, is *that* a small number?

# Example 2: The Flynn Effect

- IQ tests need to be occasionally renormed to have mean 100, standard deviation 15
  - test groups tend to get higher scores with the old than with the new test → average IQ has increased at a rate of 0.2 SD per decade since the 1950s!
  - although this progress seem to have now stalled
- Is that 0.2 SD per decade a large number?
  - implies that in 1932, average IQ would have been about 80 and 37% of population below 75
- Well, is *that* a small number?
  - low IQ often seen as disability, e.g., no-one with IQ below 75 will get the death penalty in the US

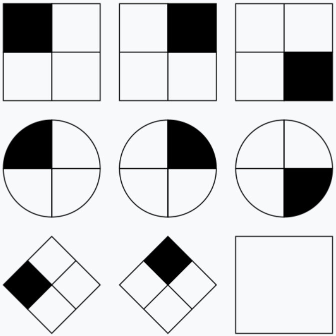
**Levine and Marks 1928 IQ classification**[60][61]

| IQ Range ("ratio IQ") | IQ Classification |
|---|---|
| 175 and over | Precocious |
| 150–174 | Very superior |
| 125–149 | Superior |
| 115–124 | Very bright |
| 105–114 | Bright |
| 95–104 | Average |
| 85–94 | Dull |
| 75–84 | Borderline |
| 50–74 | Morons |
| 25–49 | Imbeciles |
| 0–24 | Idiots |

*Source:* Wikipedia.

# Example 2: The Flynn Effect

- What does an IQ test measure?
    - many versions, but typically multiple-choice with verbal, arithmetic and "visuospatial" sections
    - higher IQ = more correct answers

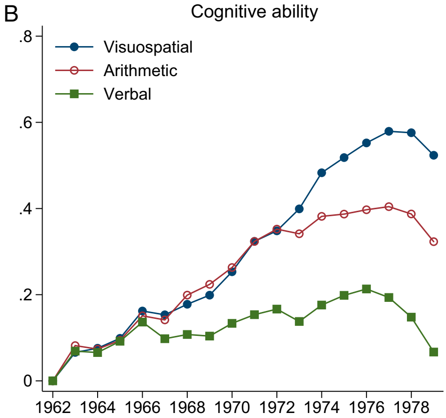


One kind of IQ test item, modeled after items in the Raven's Progressive Matrices test

*Source:* Wikipedia.

# Example 2: The Flynn Effect

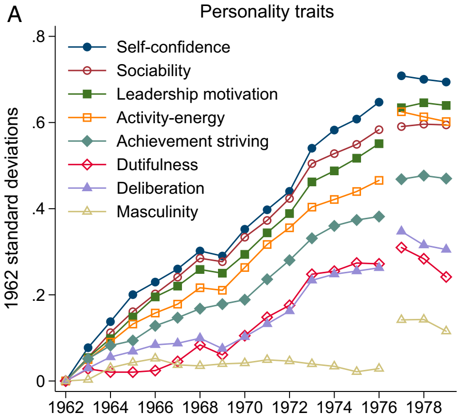
- What does an IQ test measure?
  - many versions, but typically multiple-choice with verbal, arithmetic and "visuospatial" sections
  - higher IQ = more correct answers
- Who takes an IQ test?
  - most of data used in research from conscripts of men in the beginning of mandatory military service
  - discuss: what can go wrong with a figure like this?




Average scores for measures of cognitive ability by birth year for native-born military conscripts in Finland. All scores are depicted in base year SDs, with base year means normalized at zero. *Source:* Jokela, Pekkarinen, Sarvimäki, Terviö, Uusitalo (2017), PNAS.

# Example 2: The Flynn Effect

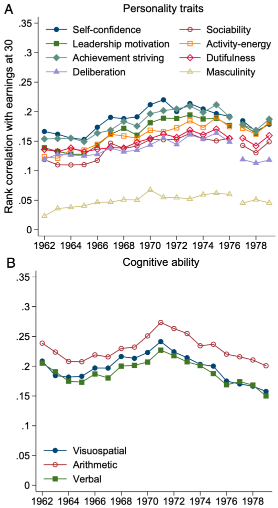
- What does an IQ test measure?
  - many versions, but typically multiple-choice with verbal, arithmetic and "visuospatial" sections
  - higher IQ = more correct answers
- Who takes an IQ test?
  - most of data used in research from conscripts of men in the beginning of mandatory military service
  - discuss: what can go wrong with a figure like this?
- Finnish conscripts also take a personality test
  - 2-h paper-and-pencil test where conscripts are asked to choose whether they agree or disagree with 218 statements concerning their personality
  - questionnaire unchanged between 1982–2001
- Are these large numbers?




Average scores for measures of personality traits by birth year for native-born military conscripts in Finland. All scores are depicted in base year SDs, with base year means normalized at zero. The break reflects a change in test administration. *Source:* Jokela, Pekkarinen, Sarvimäki, Terviö, Uusitalo (2017), PNAS.

# Example 2: The Flynn Effect

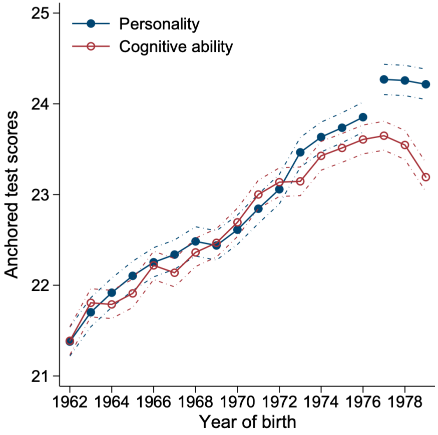
- We quantify the test score changes in two steps
  1. show that the scores predict later income
  2. predict income for everyone using their test scores




The relation of earnings and (A) personality traits and (B) cognitive ability by birth cohort, measured as the within-cohort rank correlation between the test score and annual earnings at age 30. *Source:* Jokela et al. (2017), PNAS.

# Example 2: The Flynn Effect

- We quantify the test score changes in three steps
  1. show that the scores predict later income
  2. predict income for everyone using their test scores
  3. calculate average *predicted* earnings for each cohort
- Interpretation
  - similar "increases" in IQ and personality traits
  - both imply about 12% increase in earnings capacity between the 1962 and 1976 birth cohorts




Average of anchored test scores by birth cohort, with anchoring to average annual earnings at age 30–34 (in 1,000s of 2010 Euros) using the 1962–76 birth cohorts for estimating the prediction model. Dashed lines depict 95% confidence intervals. *Source:* Jokela, Pekkarinen, Sarvimäki, Terviö, Uusitalo (2017), PNAS.

# Summary

- This course is about doing and reading empirical research
  - complements introductory statistics
  - aims to build intuition and learning to ask the right questions
  - these are critical skills for a modern economist
- Today: Asking two types of questions
  - How was this data collected? How reliable is it?
  - Is that a large number? What should we compare it to?
- Next time: Samples and distributions