Dataism - Week 2

CA Police Scorecard

1. Police Brutality

- In U.S., 1,099 killings by police in 2019.
- About 1,000 every year for the last 10 years.
- Comparison: In Germany, the total number on record since 1952 is 491
- Constant protest. Body cams and diversifying police force have led to little/no change

2. The Project

Campaign Zero Created a metric for scoring California police departments on:

- **Police Violence**
- **Police Accountability**
- **Approach to Policing**

Separate score out of 100 for each category, overall score by taking average.

2. The Project

Goals:

- Help communities, researchers, and policy-makers take informed action
- Make easy to compare police departments
- Track effects of new policies


ENDING POLICE VIOLENCE

Exploration of project

Things to note:

- Presentation
- About the Data
- Github page
- Raw Data
- Getting Involved

3. The Data

Sources:

Deadly force, civilian complaints and arrests, from official databases:

- CA DOJ's Openjustice
- FBI's uniform crime reporting statistics (UCR)
- CA monthly arrests and citation register

Police use of force, use of force complaints, police policy manuals:

- directly from police agencies via public records requests

4. Metrics

Metric: *A numerical quantity used to measure the magnitude of a (qualitative) concept*

Examples:

- GDP has become the standard for measuring economic and social progress
- University Rankings supposedly measure the quality of the education/experience of students
- Credit scores, diversity ratings etc.

4. Metrics

CAUTION!!! Metrics are subject to several pitfalls:

- Oversimplification
- Replacing real goals with optimizing a number
- Capable of being gamed
- What’s included/excluded often depends on human judgement and morals

4. Metrics

Oversimplification: A metric crunches many dimensions of information down into a single dimension.

- Impossible to capture full complexity
- Yet objective mathematical appearance leads to trust in its correctness
- Often don't question what is lost/omitted

NOTE: Campaign Zero does a good job presenting all of the information that is condensed into the score.

On GDP

"The valuable capacity of the human mind to simplify a complex situation in a compact characterization becomes dangerous when not controlled in terms of definitely stated criteria. With quantitative measurements especially, the definiteness of the result suggests, often misleadingly, a precision and simplicity in the outlines of the object measured. Measurements of national income are subject to this type of illusion and resulting abuse, especially since they deal with matters that are the center of conflict of opposing social groups where the effectiveness of an argument is often contingent upon oversimplification. [...]"

-Simon Kuznets

4. Metrics

Optimizing a number:

Replacing a complex goal with optimizing a metric can have negative side effects and eventually obscure/work against the original goal.

Take GDP as a measure of the “strength” of the economy. Doesn’t include volunteer work, household labor, commons, environmental health, quality of infrastructure…

In maximizing GDP growth, these things are neglected or sacrificed

4. Metrics

Metrics can be gamed:

- Once widely adopted, there is incentive to “win.”
- Structure of model can be exploited:
    - Universities pay fees for students to retake SAT
    - Falsify SAT scores, acceptance rates...
    - Change procedure/market heavily in order to reject as many as possible
    - Says little/nothing about quality of education...

Can be hard to escape such models. The winners have no incentive.

4. Metrics

Good to ask “what if?”

- What if GDP factored in public commons, health of environment, household labor…?
- What if a metric is widely adopted? Would there be incentive to game the system? How could it be gamed?
- What if the situation for which the metric was conceived has significantly changed? How much of that change is *because* of the metric?

(CAN/SHOULD ASK SIMILAR QUESTIONS FOR ANY MODEL)

Should we focus on a single metric?

Pros:

- Tells a good story, is easy to understand
- Comparison is easy
- If well-designed, can motivate change/align incentives and help track the effects of new policies.

5. Numerical Proxies

Must choose which quantitative data captures the factors we want to include in the Police Score: (Abstract qualitative => concrete quantitative)

For **Police Violence,** scorecard uses:

- Percentage of non-lethal force/arrest
- Percentage of deadly force/arrest
- Number of of unarmed civilians killed/seriously injured
- Racial bias in arrests and deadly force*
  - *This is a separate metric with it's own numerical proxies

What could be missing? (ex. Explicit reward for de-escalation…)

6. The Police Violence Score

Score from 0-100 (0 = worst, 100 = best):

$$\left(\text{Percentile Less Lethal Force Used per Arrest} + \text{Percentile Deadly Force Used per Arrest} + \text{Percentile Unarmed Civilians Killed or Seriously Injured} + \text{Percentile Racial Bias in Arrests and Deadly Force} \right) / 4$$

- First three are all computed in the same way
- Fourth is more complicated (need method for measuring racial bias)
- **Potential issues**
  - “Per arrest” => a department could improve by arresting more people
  - “Percentile” => grades are relative, good = better than most

6. Police Violence Score

Let's look at the second term: "**Percentile Deadly Violence per arrest**"

Example: Assume just three police departments:

Agency Name	Total Deadly Force Incidents 2016-2018	Total Arrests, 2016-2018	Deadly Force Incidents per Arrest (x10^4)	Percentile Deadly Force Incidents per Arrest
Department A	1	3581	2.8	100
Department B	17	28200	6	66.6
Department C	3	3911	7.7	33.3

"Percentile" = % of all departments that are the same or worse

6. Police Violence Score

Another potential problem:

Suppose City A is small and super safe and only arrests 100 people from 2016-2018. But there was one incident of a fatal shooting by the police. (1%)

Now suppose City B is crazy and people get arrested for all sorts of things, leading to 10,000 arrests and 100 fatal shootings. (1%)

Should City A and City B have the same ‘deadly violence’ score?

6. Police Violence Score

- **PROBLEM: Deadly force, though too common, is statistically very rare**
- General rule of thumb: For very rare events you need very large samples of the population to make statistically significant conclusions
- Practically speaking:
  - For big cities, the estimation of the probability that a police officer will use deadly force is much closer to the “true probability”
  - It would make sense to compare the scores of “similar” departments or to group them together

7. Racial Bias in Use of Deadly Force

Question: How can we measure racial bias in the use of deadly force?

Answer: Compare to the use of deadly force on white people.

- Studies show that nationwide, black men have twice the risk of being killed by a police officer.
- This is as a proportion of the total population:

  2 x (# white victims)/(# white population) = (# black victims)/(# black population)

7. Racial Bias in Use of Deadly Force

In calculating bias score, Campaign Zero asks:

What about the chances of being killed amongst people being placed under arrest?

This happens in the vast majority of deadly force incidents.

Upshot: Racial bias in arrests leads to further skew in the use of deadly force.

7. Racial Bias in Use of Deadly Force

It’s like a game of Russian roulette where black players have

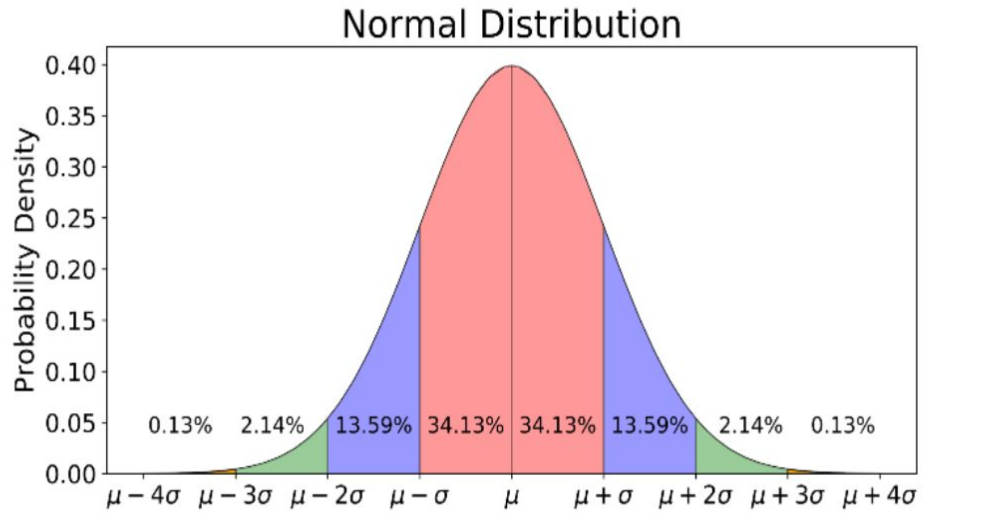
- More loaded chambers (bias in use of deadly force during arrest)
- Have to play more often (bias in arrests)

The bias in deadly violence score looks at both, but we focus on the former.

7. Racial Bias in Use of Deadly Force

Normal Distributions:

- Appears everywhere in nature when a number tends to cluster around an average value
- Example: average height is approximately μ = 170 cm
- Most people are close to average height
- Heights further from average are less and less common

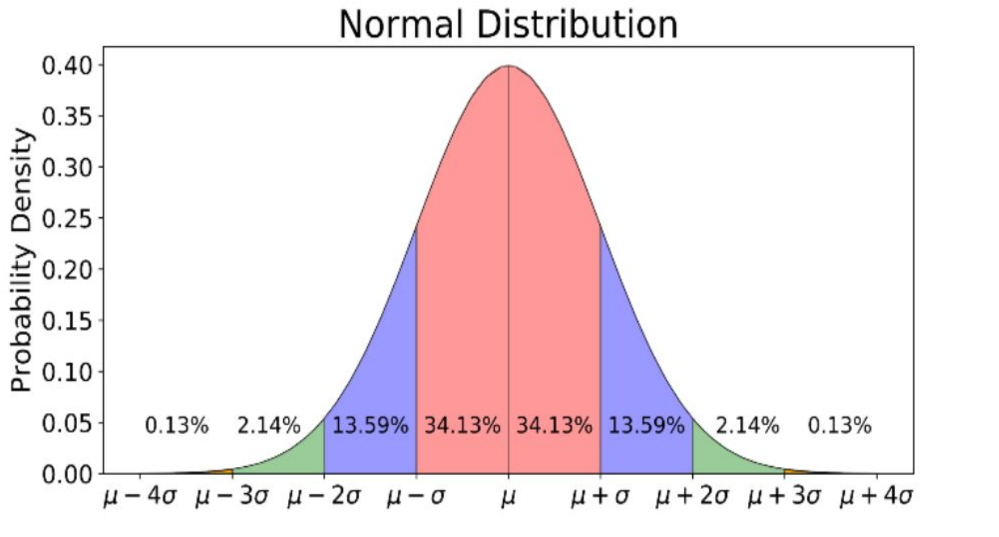


68% of the data is within 1 standard deviation, 95% is within 2 standard deviation, 99.7% is within 3 standard deviations

7. Racial Bias in Use of Deadly Force

Normal Distributions:

- μ = mean/average
- σ = standard deviation
- The standard deviation gives a measure of spread of the data varies around the average
- Most of the time (99.7%) within 3 σ's of average.
- Call values on edges "outliers"



68% of the data is within 1 standard deviation, 95% is within 2 standard deviation, 99.7% is within 3 standard deviations

7. Racial Bias in Use of Deadly Force

Estimating bias in the # of loaded chambers for a specific city:

Idea: Take probability of dying during arrest *if white* as "baseline"

Let $n = \text{\# of black arrests by department,}$

we compare the following:

- actual number black victims of deadly violence
- the expected number of white victims if *n* white people were arrested.

7. Racial Bias in Use of Deadly Force

If $$p = \frac{\text{\# white residents killed by department}}{\text{\# white residents arrested by department}}$$

and *n* white people arrested, then we expect

$$\mu = n \cdot p$$

to be killed by police. (Example: p= 1% chance of being killed if placed under arrest, 1000 people arrested. Expect 0.01 x 1000 = 10 deaths)

7. Racial Bias in Use of Deadly Force

We don't expect *exactly* n*p deaths. Just something "in the neighborhood."

Assumption: The number of deaths of involving white residents is normally distributed with average $\mu = n \cdot p$ and standard deviation $\sigma = \sqrt{np(p-1)}.$

To score racial bias, Campaign Zero measures *the number of standard deviations between the actual number of deaths and the expected number* n˙p.

7. Racial Bias in deadly force

The number of standard deviations between the actual number of deaths and the expected number n˙p.

- There's a formula for that:

$$\frac{\#\ \text{actual black victims} - n \cdot p}{\sigma}$$

7. Racial Bias in deadly force

$$\frac{\#\text{ actual black victims} - n \cdot p}{\sigma}$$

This is called the *z-score*

- To give police departments a “racial bias in deadly force” score, departments are
    - ranked by their z-scores (large = bad, small = good), and
    - graded 0-100 by percentile

7. Racial Bias in Deadly Force

Problems:

- Again not very reliable for small cities.
- Can improve score by
  - increasing n (arresting more black people)
  - Increasing p (either killing more white people or arresting fewer)

$$\frac{\#\text{ actual black victims} - n \cdot p}{\sigma}$$

Possible Discussion Questions

Should police departments know how the score is calculated?

How should the scorecard be used/interpreted?

Should it be modified?

What would happen if the scorecard were widely adopted?

What ethical decisions were made in designing the metric?

What data should police departments make available?

How can we ensure that the data is reliable?

Further resources

- Jupyter notebook for this course (with code for computing racial bias score)
- Separate notebook with assignment (code to compute easier term in police violence score)
- *Donut Economics* - Book on the ways in which GDP and other mainstem economic models can be “weapons of math destruction” + proposed alternatives.

Final Project

Goals:

- Help participants connect and learn from each other
- Allow students to more deeply explore the questions that most interest them
- Produce something concrete that can serve to educate/inspire after the course has ended
- Keep demands minimal
- Would be awesome if leads to longer term collaborative data-activism project
- Presentation/Discussion in last week of class.

Final Project

- My proposal: 4 groups under the broad headings of:
    - Economics
    - Socio-political issues
    - Environmentalism/Sustainability
    - Tech/Data/Futurism
- Would correspond to discussion groups
- Of course all interlinked. More about the kind of questions you're interested in.

Output

- Work together to produce a blog post/jupyter notebook that at least
  - Articulates a problem
  - Proposes a data-driven response (or discusses an existing one)
  - Shares resources for getting involved/going further
- With time, could add:
  - Description of existing data or what data is needed
  - Methods (perhaps w/ prototype code)
  - Whatever else you think is cool

Questions

- Is anyone vehemently opposed?
- How many people are tech-savy?
- Are you OK with groups being chosen for you?
  - Would take requests into account