A Probabilistic Analysis of Police Violence in the United States using Census Data

https://github.com/farzaank/police

Farzaan Kaiyom - CS109

I. Introduction

Addressing police violence has always been a problem on my mind for a number of reasons. Data shows disparities in police violence, and universal failure to address it. On the other hand, recent comments by certain political figures imply that either disparities don't exist or have been solved. I knew that the data disproved such comments, but I wanted to take a look at to what extent. Further, national racial disparities and my personal proximity to this problem have made it even more important to me. I have always seen homicide rates and disparity statistics floating around, but none are quite comprehensive. As much data as there is, there's not a lot being done with or to it. Thus, I decided to do some probabilistic analysis of police violence data taking into account all variables. I originally wanted to create a calculator of disparities of police homicide rates given many different events (victim's race, armed/unarmed, whether officers were charged, whether they were convicted, etc) to demonstrate where the biggest issues are. This ultimately was something I didn't have time for, as taking into account such a large number of random variables proved more difficult to code than expected. I ended up doing three things in my project: data visualization of police violence in counties in relation to their policies, analysis of police homicide rates in different states of the US, and joint sampling to calculate probabilities given a variety of variables. The joint sampling algorithm gave me an easy route to create the calculator I wanted to create, but I didn't have time to create a GUI for it.

II. Method and Findings

My project is currently a Jupyter Notebook with 3 parts.

The first part of my project was focused on data visualization. The hardest part of this was getting the policy data. I was able to find this in a paper published by Campaign Zero, but it wasn't a CSV, so I had to use Optical Character Recognition software to convert it. The OCR software didn't work perfectly so I had to fix the CSV file myself. After that I was able to use numpy to calculate correlation of various policies to police homicide rates and racial disparities.

The second part of my project involved programmatically calculating police homicide rates in various states based on ethnicity. I used Bayes theorem to derive the probabilities of homicide based on ethnicity (derivation shown on github repo) and assumed a Poisson distribution to calculate a rate of homicide per million. [I plan to add some data visualization to this part, but for now it just shows states' disparities]

The third part of my project was by far the most interesting. Originally, I didn't even want to use the aggregate data because it had so many different variables and didn't spoon feed me the data I wanted. Ultimately I got over it and decided to calculate different Bayesian probabilities based on different conditions in the file (race, armed/unarmed, officer's charges, officer's conviction, etc). I used these probabilities to plot Poissons based on police homicide rates of unarmed people given their race to visualize the huge racial disparity in America. After calculating these probabilities I realized I could use joint sampling and a Bayesian network* to find any probability involving the variables I used. So I did exactly this. After fixing a few large bugs (documented in my code), I had created a joint sampling model that effectively demonstrated trends in US police violence.

*the Bayesian network drawing on my github is missing an arrow between white and black, that I added later

(cont. on page 2)

III. Conclusion

Although I didn't finish my disparity calculator, I now have the tools to create it. Further, I have been able to show trends: namely that force restriction policies decrease police violence and that racial disparities are still very *very* real in America. The future of this project involves completing the calculator and publishing an article in the Daily that I've been working on. I also plan to post my Jupyter Notebook on the Kaggle board for one of the datasets I used so people can use it to find their own discoveries. After all of this I hope that my findings can have some impact so I hope to send my work back to Campaign Zero and Mapping Police Violence, the two organizations that published the datasets I used.

IV. Links to Datasets

https://policeviolencereport.org/
https://www.theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-police-killings-us-database
https://www.joincampaignzero.org/reports [I used Report #3 for departmental policy data]