The Stanford Open Policing Project and Statistical methods for Calculating Racial Bias.

Caleb Berman

Overview

In this report I give a summary of current methods of statistically determining racial bias in police traffic stops outlined in Pierson et al. and their limits. I then provide statistical analysis on what conclusions can be made on racial bias derived from a dataset on traffic stop outcomes in the state of Texas.

Background on the Stanford Open Policing Project

The Stanford Open Policing Project is a current interdisciplinary team of researchers at Stanford University working to aggregate and clean Traffic stop data from across the United states. The goal behind this project is to use this data to understand bias in policing across America. Currently data on policing in America is very hard to work with for many reasons. Firstly, much of it is not available to the public. Secondly policing data across the United States has no standardization by any regulatory body. This means that the data is extremely messy and has many different and varying features. Thirdly, there's evidence that a sizable amount of the data is misreported in those fields such as the race of a subject of a traffic stop. The Stanford Open Policing Project hopes to solve these problems by aggregating all available traffic stop data to one central source. In addition to this they work to clean much of the data to have a few standardized columns that make it easy to work across the many datasets.

Currently the aggregate metadata contains 255 million rows of traffic stop data collected from local and statewide police forces from a total of 42 states. It represents a tremendous opportunity for big data analysis as over 50,000 Americans are pulled over every day on average. By combining cleaned data from across the country researchers involved with the open policing project have two goals: to draw large scale statistical conclusions regarding the state of bias in policing in America that one would be unable to do given a much smaller local dataset as well as provide the open source data to other data scientists and policymakers to draw conclusions from.

Statistical Methods for Quantifying Bias

Considering that racial bias is a complex sometimes-unconscious decision making process by humans, directly measuring it proves to be quite difficult. In an effort to quantify widespread systemic bias the stanford open policing project outlines the following methods to indirectly measure racial bias in the context of traffic stops that statisticians have developed.

Stop Rate Benchmark

The core idea behind this kind of analysis is to compare the observed distribution of police activity (traffic stops, searches, arrests, etc.) by race to the distribution of a given population's race. One of the simplest ways of achieving this is to normalize the number of people in each racial group getting stopped in traffic by their makeup in the geographic locations population. For an unbiased system, you would expect that the rate at which each racial group is stopped would be equal.

This method has its flaws in that it doesn't take into account any background factors on a race's driving behavior that may have led to the stop. For example if any race happens to drive more or faster then one might more traffic stop data from that race.

Veil of Darkness test

The Veil of Darkness test is another method for controlling for the stop rates of different races. The central idea behind this test is that when it is dark outside, police on average will have worse vision and therefore will not be able to tell different races apart before they are stopped. To make this comparison independent of the actual time it is (because different races may be driving more frequently at different times in the day) the fact that the sun sets at different times throughout the year can be exploited and you can compare the amount of any given race stopped at one time while it's light out vs the number of that race stopped while its dark out at one given time. This makes sense as driving patterns are typically tied to what work hours you have at your employment and not when the sun has gone down. For an unbiased system you'd expect these numbers to be equal as there shouldn't be a difference in stoppage rates for when police can tell whether or not it is any particular race behind the wheel.

The outcome of this test is limited in the same way that the stop rate benchmark was limited. It doesn't take into account all driving behavior including what types of cars each race drives which may be linked to race. Furthermore if there is any race-based change in driver behavior over the course of the year then this test isn't going to pick up on it. One proposed improvement the open policing project makes to this test is to consider stops occurring the day directly before and after daylight savings time when the date has been largely left on unchanged while the sunset time has changed drastically.

Outcome test

One way to get around having to incorporate the differing driving and background behavior of each race is to look at the end outcome during a traffic stop. That is if a certain race has been

already stopped, what is the chance that the race will have further investigation by the police (searching their car) and what amount of those searches are successful in finding illegal substances. The ratio of searches successful in finding contraband and the total number of attempted searches is known as the "hit rate." For an unbiased system we'd expect races to have equal hit rates. This is because hit rates represent the threshold of evidence that a given police officer has before they decide to search a person. For example if a race theoretically had a lower hit rate than another, police were less sure evidence-wise that the race had some sort of contraband in their car. Furthermore, this kind of test doesn't rely on population demographic data that would have to be estimated from another source as we only draw our data from the sample space of drivers that have been stopped.

The outcome test has one major flaw in the form of a concept known as infra-marginality. A very oversimplified explanation of infra-marginality is if populations you compare hit rates between have different distributions of risk, then there may be a disparity between hit rates even in the absence of bias.

Threshold test

The threshold test is a more recently developed indicator of bias that takes a bayesian approach. The basic idea is to observe the department of the officer, the race of the driver, if a search was conducted, and if contraband was found. Statisticians then model the probability that a person is in fact carrying contraband, x_i by the discriminant distribution, ~ disc(ϕ_{rd}, δ_{rd}).
Where ϕ_{rd} is the expected value of the distribution and δ_{rd} is the level at which one is able to tell the difference between drivers who are carrying contraband or not. We can then model the decision by a police officer to conduct a search on a stopped persons car as when x_i is greater than some threshold, $t_{r,d}$. If we find that this threshold is different for different races then it is evidence for discrimination.

Findings from the paper

In their paper, A large-scale analysis of racial disparities in police stops across the United States, the researchers in the stanford open policing project found that while hit rates were comparable for black and white drivers by location, conducting a threshold test revealed a much lower threshold needed for black drivers than white drivers, indicating discrimination. Another result worth noting is that they noticed a significant drop in search rates for drivers in states where recreational marijuana was legalized.

Data analysis

Here I conduct a small number of statistical bias tests on one of the datasets available from the Stanford Open Policing Project. The dataset I chose to analyze was state trooper traffic stops in Texas. I specifically chose this dataset because it had a large amount of data to draw from (N = 27,426,840) over a large range of dates (01/01/2006 to 12/31/2017). In addition to those factors, one question that was unexplored in the analyses done by the Stanford open policing project was the distribution of bias among the officers themselves. In other words are police departments systemically racially biased due to a smaller number of very biased individuals or is traffic stop bias more uniformly distributed among the officers? This is one such question that I hoped to answer.

One major problem with this dataset is that there was evidence that minority drivers were incorrectly labeled as white in the unprocessed data. To get around this limitation the Stanford open policing project mapped the race feature using a separate dataset of surnames collected from the U.S. census.

Analysis code considerations

All code to generate tables, plots, and fit models can be found in the appendix of this report. Given that the statewide dataset had close to 30 million rows, taking up greater than 7 GB of disc space a majority of the data analysis was done in sqlite3. In order to do so I modified the csv downloaded from the Stanford open policing project into a database file using the command line. This had benefits in analyzing the data more efficiently in addition to the added benefit of demonstrating some recent SQL skills I'd like to showcase.

Stop rate analysis

To begin the analysis I first wanted to take a look at what biases calculating a stop rate would uncover. To do this I needed race-level demographic data on the population of Texas. I assumed that this data would be the most accurate in years in which the United States census occurred. Thus I limited the traffic stop data to the year 2010 and joined it with demographic data retrieved from the United States 2010 census. The results can be seen in the table below.

Race	Year	num_stopped	num_searched	num_warned	num_cited	population	Stop rate	Search rate	Warning rate	Citation rate
asian/pacific islander	2010	39327	344	28263	14725	9.806769e+07	0.000401	3.507781e-06	0.000288	0.000150
black	2010	251410	8564	199643	91248	2.967176e+08	0.000847	2.886246e-05	0.000673	0.000308
hispanic	2010	780653	19373	610913	305427	9.454731e+08	0.000826	2.049027e-05	0.000646	0.000323
other	2010	3121	65	2406	1111	2.967176e+08	0.000011	2.190635e-07	0.000008	0.000004
unknown	2010	72798	1497	58155	21059	2.640284e+08	0.000276	5.669845e-06	0.000220	0.000080
white	2010	1377987	20595	1111716	415219	1.139094e+09	0.001210	1.808016e-05	0.000976	0.000365

Table 1. Rates of police action as a proportion of the demographic makeup.

We see that while traffic stops of white people occur at a higher rate of 1.43 times the rate of traffic stops for black people, cars of black people are searched at a higher rate of roughly 1.6 times that of white people. This is evidence that an outcome test may have more significant results compared to the benchmarks presented here.

Outcome analysis

To look at the bias in outcome for car searches I calculated the hit rate for both white and minority drivers illustrated in this table below.

Race	Number Searched	Number Contraband Found	Hit Rate
Minority	283946	77230	0.271988
White	259601	88936	0.342587

Table 2. Hit Rates calculated for Minorities and White identified people across all of Texas.

One interesting result I found is that when plotting the hit rate as a function of year you see an increase in hit rate for both Minority and White drivers. One possible interpretation of this result is that a police officer's evidence baseline for searching someone's car has become more stringent in recent times.

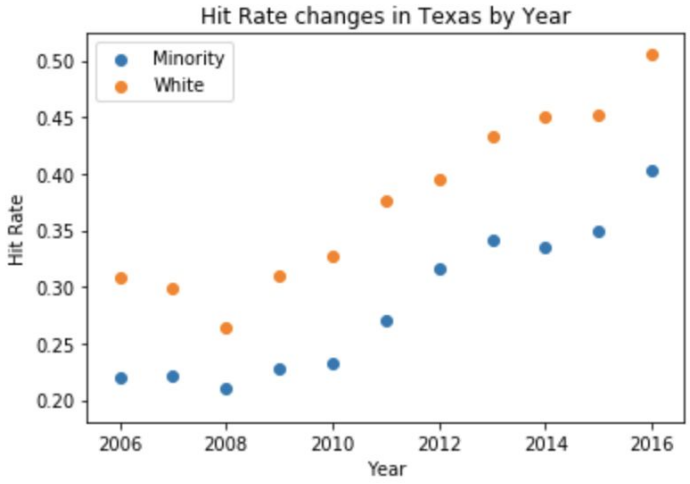

Figure 1. Hit rate as a function of time.

In order to get a better understanding of the prevalence of these hit rate disparities across geographical locations I then calculated a separate hit rate for each county in texas given by the plot below.

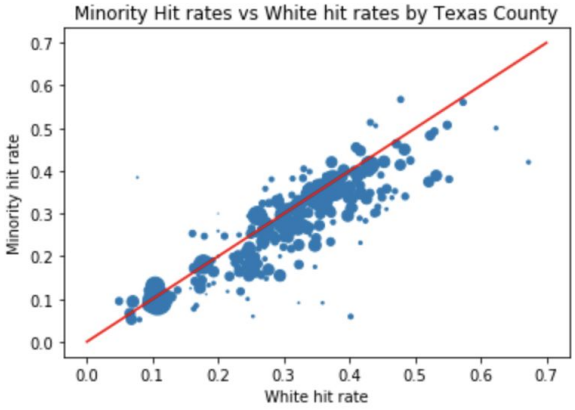

Figure 2. Minority vs White hit rates for each county in Texas. Here the size of each point represents the number of traffic stops in that county and the red line represents an equal proportion of hit rates. Counties were excluded if they had absolute hit rates (1 or 0) as this was deemed incomplete data.

This plot clearly shows that hit rates for white drivers skew higher than the hit rates for minority drivers giving some indication of bias of outcome for police car searches. Specifically 192 counties out of 253 counties in Texas exhibited higher hit rates for people identified as white compared to minorities.

To investigate the discrepancy in bias between individual officers I produced a similar plot for the hit rates for each officer_id within this dataset as shown below.

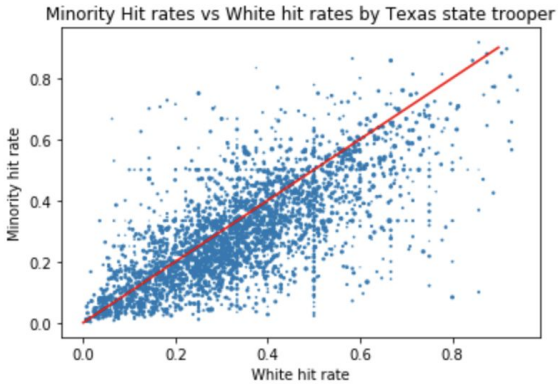

Figure 3. Minority vs White hit rates for each state trooper in Texas. Here the size of each point represents the number of traffic stops made by each officer and the red line represents an equal proportion of hit rates. Officers were excluded if they had absolute hit rates (1 or 0) as this was deemed incomplete data.

This plot is more unwieldy due to the high number of state troopers making car searches. In order to answer the question if this perceived bias was driven by a few very biased officers or by a large amount of officers exhibiting small amounts of bias I attempted to categorize the distribution of potentially racially biased officers by plotting the distribution of each officers difference in hit rate (White hit rate - Minority hit rate).

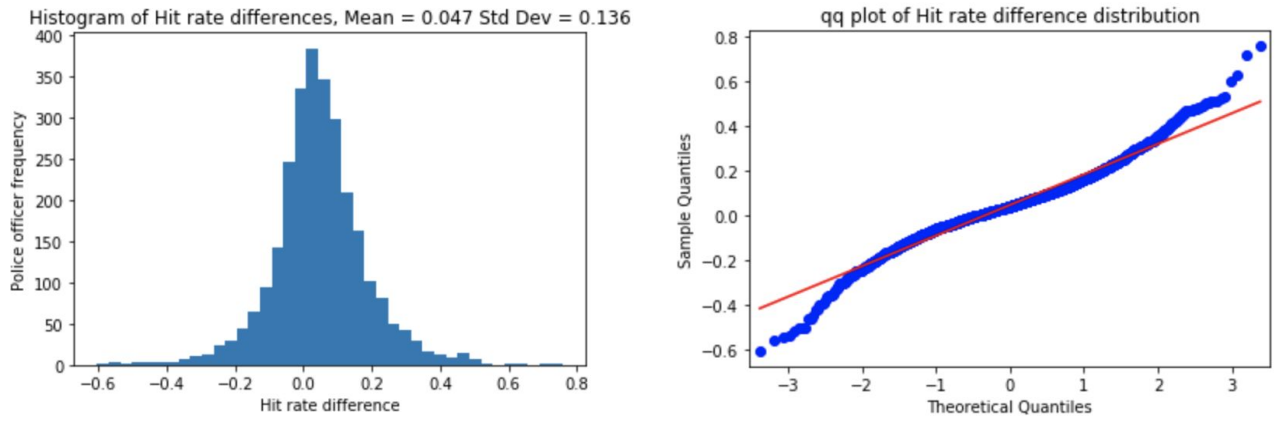

Fig. 4 Distribution of Hit rate difference among officers. The QQ plot was fit using ~N(0.047, 0.136).

These plots demonstrate that this distribution can be decently approximated by a Normal distribution translated rightward very slightly. I think this points to evidence that a large amount of outcome bias in car searches stems from mostly small amounts of racial bias among a majority of police officers.

Lastly in order to confirm the statistical significance of the claims of outcome bias I fit a logistic regression model on the basis of race and gender to predict if contraband was found.

Logit Regression Results

Dep. Variable:	Contraband_found	**No. Observations:**	543547
Model:	Logit	**Df Residuals:**	543544
Method:	MLE	**Df Model:**	2
Date:	Fri, 11 Dec 2020	**Pseudo R-squ.:**	0.004862
Time:	14:24:49	**Log-Likelihood:**	-3.3299e+05
converged:	True	**LL-Null:**	-3.3462e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9297	0.008	-119.654	0.000	-0.945	-0.914
is white	0.3287	0.006	55.448	0.000	0.317	0.340
is male	-0.0644	0.008	-8.393	0.000	-0.079	-0.049

Table 3. Logistic regression results. Model was fit using the Statsmodels Python module.

Here the coefficient of the “Is white” term in the logistic regression implies that based on the data, if a state patrol officer searches the car of an already stopped white person they will have a greater probability of success in finding contraband. This implies that cars driven by white people are under-searched when compared to those driven by minorities. We can see that the coefficient of racial difference is much greater than other populations differences such as gender as the coefficient of gender difference has a magnitude 5 times smaller then the coefficient representing racial differences. Additionally as can be seen in the results table this conclusion is statistically significant with $p < 0.05$.

Conclusions

As evidenced by the analysis above, there exists significant bias in outcome for minorities vs non minorities in car searches. Future analysis in the form of veil of darkness and threshold testing would need to be done to show that this evidence of bias isn’t a result of inframarginality.

Works Cited

E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, and S. Goel. “A large-scale analysis of racial disparities in police stops across the United States”. Nature Human Behaviour, Vol. 4, 2020.

Simoiu, C., Corbett-Davies, S. & Goel, S. The problem of infra-marginality in outcome tests for discrimination. Ann. Appl. Stat. **11**, 1193–1216 (2017).

United States Census Bureau. 2010 Census.U.S. Census Bureau. 2010. Web. 1 January 2013 <http://www.census.gov/2010census/data/>.