Assessing racial bias in stop-and-frisk decisions

Andrew Ligeralde LIGERALDE@BERKELEY.EDU

Biophysics Graduate Group, University of California, Berkeley

Abstract

In this work, we perform statistical analyses on the New York City Police Department (NYPD) Stop, Question, and Frisk (SQF) repository to address the following question: how is racial bias perpetuated in the application of the stop-and-frisk policy? We find, as in previous work, that Black individuals are frisked and searched more frequently than White individuals for criminal possession of a weapon given that no weapon is detected. We also demonstrate that in contexts like this, where Black individuals are overrepresented in the data, metrics of statistical fairness may be misleading and cause inequalities to appear less egregious.

1. Introduction

Stop, Question, and Frisk, sometimes abbreviated as stop-and-frisk, is a policy that allows police officers to stop, question, and search individual citizens and their belongings given reasonable suspicion of a crime (Goel et al. 2017, Badger 2020). It has been reported that exposure to stop-and-frisk has negative long-term psychological effects at the individual level, independent of whether or not it leads to arrest (Badger, 2020).

The motivations behind this work are to understand potentially harmful racial biases in the application of stop-and-frisk, evaluate the effectiveness of interventions designed to mitigate these biases, and explore how these biases may hamper attempts to apply statistical fairness criteria to detect and mitigate inequities across groups. In this set of analyses, we analyze how stop-and-frisk is differentially applied across populations of Black and White individuals. We consider the extent to which the NYPD decision to frisk and search meets statistical non-discrimination criteria with respect to Black and White suspects.

We conduct these analyses within the population of individuals who are stopped, frisked, and searched for criminal possession of a weapon (CPW) in 2020 using data from the New York Police Department's (NYPD) Stop, Question, and Frisk (SQF) repository. The SQF repository is a yearly log of individual stops made across the 77 police precincts of New York, and includes officer-reported details such as time, location, and duration of each stop, whether the suspect was frisked and/or searched, and characteristics of the suspects in question such as demographic information and prior records. As Goel et al. (2016) discuss in detail, the CPW subset of the SQF dataset is useful for assessing fairness as it enables a clear definition of false positive rate (FPR), which corresponds to how often suspects are frisked and searched for CPW when they do not turn out to have a weapon, as well as true positive rate (TPR), which corresponds to how often suspects are frisked and searched for CPW when they do in fact have a weapon.

The corresponding non-discrimination criterion we are interested in is equality of error rates across groups, which typically includes false positive and false negative rates. This is referred to generally as "separation" (Barocas et al., 2019, Hardt and Recht 2021) as well as “equalized odds” (Hardt et al. 2016). Because we are analyzing this data primarily for its potential to reduce unequal harms across groups, as opposed to the motivation of reducing crime rates, we will focus on equality of FPR (as opposed to FNR), as our non-discrimination criterion, which we will refer to as “equalized harms”.

Our main finding for this analysis is that innocent Black individuals are subject to frisk and search more often than innocent White individuals, and a fairness-adjusted (equalized-harms) classifier trained on selected features of stopped individuals can achieve both lower FPRs for a given TPR within both groups as well as more equal decision thresholds between groups. However, we also find that the metric of FPR can make the disparity between Black and White suspects appear less egregious when comparing across groups within individual precincts. We conclude this is likely because FPR disguises the fact that Black individuals are stopped far more than White individuals. The comparatively large sample sizes for Black individuals distort the degree to which stop-and-frisk practices satisfy non-discrimination criteria as a result. A common intuition is that increased data collection regarding discriminatory institutional practices can help us understand harmful patterns, which in turn can guide policy changes that lead to fairer practices for members of protected classes. However, our analysis demonstrates that when the data generation process is itself biased, commonly used metrics to evaluate and enforce fairness can mask these biases and may potentially be misleading without proper consideration of the context in which the data was generated and collected.

2. Results

2.1 Considerations of fairness in stop and frisk decisions

Population level and precinct level detection rates for NYPD frisk and search

Throughout this section, we will use Y to denote a random variable that encodes whether a given stopped suspect actually has a weapon ($Y = 1$) or not ($Y = 0$). $\hat{Y}_{NYPD}(X)$ encodes the decision by the police to stop, frisk, and/or search, with the assumption that this is based on features X of the suspect (e.g., age, race, location). $\hat{Y}_{NYPD}(X) = 1$ encodes a stop, frisk, and/or search, and $\hat{Y}_{NYPD}(X) = 0$ means no frisk and/or search was conducted after the initial stop. A encodes the race of the suspect. The quantity we are interested in is $P(\hat{Y}_{NYPD} = 1|Y = 0, A)$, the FPR of the implied classifier implemented by NYPD officers when conditioning on whether the suspect is Black or White.

We first consider the disparity between detection rates for Black and White suspects at the population level (Table 1). We find that while TPR (rate of frisk and/or search given that the suspect actually has a weapon) is nearly equal between Black and White suspects, FPR (rate of frisk and/or search given that the suspect does not actually have a weapon) is moderately higher for Black suspects.

We also examine FPR by precinct conditional on race (Figure 1). We note two significant trends. First, many precincts have high FPR for Black suspects but have either an FPR of 0 or no innocent suspects logged for White suspects. Second, in many precincts where FPR for black and white individuals are both high, FPRs may appear equal. However, looking at counts of false positives by precinct shows that this is likely due to low sample sizes

	FPR (NYPD)	TPR (NYPD)
Black suspects	0.91	0.96
White suspects	0.86	0.95

Table 1: Population level detection rates for frisk and search conditional on suspect race.

for White suspects — even after normalizing for population by race within each precinct, the number of false positives for black individuals is notably higher than that for white individuals across precincts.

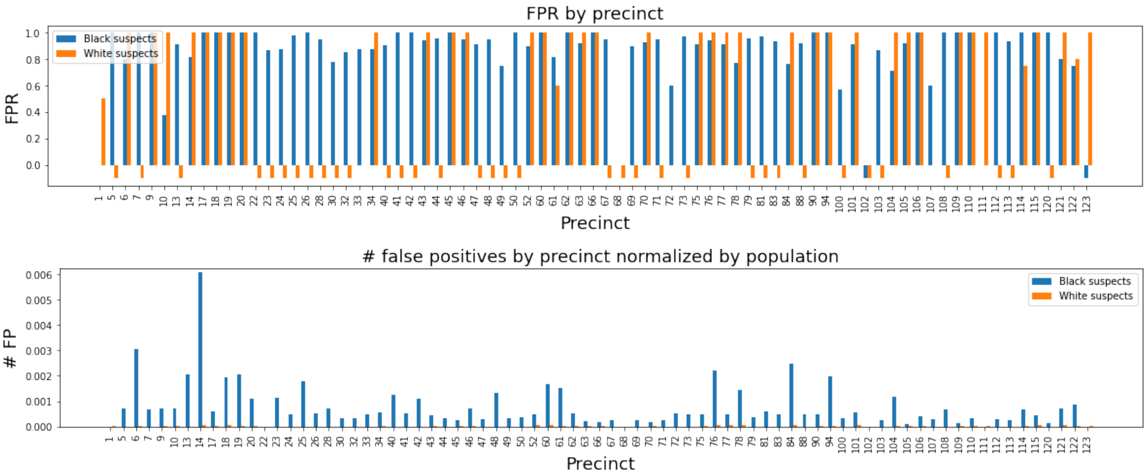

Figure 1: Detection rates by precinct. Top: Negative bars indicate no available data for those conditions. Bottom: Counts are normalized by population of each group within each precinct. The value for Black suspects in Precinct 22 (0.056) is an outlier and was therefore omitted from the plot.

Comparison of NYPD frisk and search to fairness adjusted classifier

Next, we train a logistic regression classifier $\hat{Y}_{model}$ on selected features of the suspects X to predict target variable Y, whether an individual is carrying a weapon. We include those features which correspond to reasonable suspicion (e.g., past offenses, suspicious action), as well as demographic features like age, race, sex, and precinct. For the full set of features included in the data, see the Jupyter Notebook linked in Methods. Our classifier achieves a high TPR for low FPR (Figure 2). Notably, the ROC curves for Black and White suspects are close for the full range of thresholds, suggesting the features selected are not heavily biased by race. By comparison, NYPD frisk/searches incur higher FPRs in comparison to the model for a range of decision thresholds corresponding to the same TPR.

Using our classifier, we derive the implied decision boundaries for NYPD frisks/searches. We assume that officers are trained to consider similar features as are included in the model when evaluating suspicious activity on any given stop. We note that there are a number of

	θ_{NYPD}	θ_{fair}
Black suspects	0.66	0.74
White suspects	0.82	0.82

Table 2: Implied decision thresholds for NYPD frisk/search and thresholds for fairness-adjusted (equal-harms) classifier.

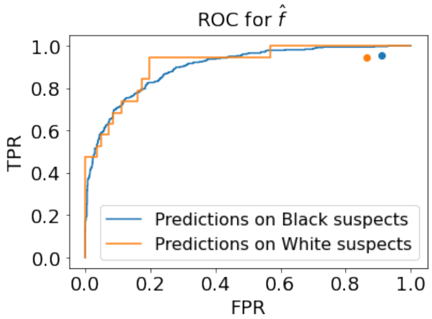

Figure 2: ROC curve for logistic regression model trained on select features. Two plotted points indicate the values for NYPD frisk/search detection rates.

features we do not account for in the model that may play a role in these decisions, such as time of day, historical precedent of crime for a given precinct, and as well as various heuristics and indeterminate factors that are not easily quantifiable or included in the dataset, such as demeanor of the suspect. A more robust logistic classifier, which has a more comprehensive list of features and also accounts for pairwise correlations between features, is described in detail in Goel et al. (2016). We find that based on the previously calculated FPR for each race, the corresponding decision threshold is lower for Black suspects (66% likely to carry a weapon results in frisk/search) than for White suspects (82% likely to carry a weapon results in frisk/search).

Next, we use our classifier to derive a fairness-adjusted classifier. To do this, we turn to the technique described in Hardt et al. (2016), which involves calculating group specific thresholds that equalize FPR across all values of A. We define an "equalized-harms" classifier $\hat{Y}_{fair}$ to be one such that

$$P(\hat{Y}_{fair} = 1|Y = 0, A = black) = P(\hat{Y}_{fair} = 1|Y = 0, A = white). \tag{1}$$

We set the desired FPR for the fairness-adjusted classifier as the FPR of frisk/search for White suspects and find that interestingly, the equalized-harm threshold is higher for White suspects than for Black suspects to achieve the same FPR (Table 2).

One might expect that given the higher numbers of false positives for Black suspects, the fairness-adjusted classifier should have a higher threshold for Black suspects than White suspects. Based on the results from Figure 1, we conclude that this is likely due to there being fewer White suspects. Because FPR increases as sample size decreases while holding the number of false positives constant, the small sample size for White suspects has a greater net effect on FPR than the smaller number of false positives for White suspects. For example, if there is just one frisk/search for a White suspect that occurs in a given year, and the suspect happens to be innocent, the FPR for White suspects will be 1, whereas if there are five frisk/searches for Black suspects, and four of them happen to be innocent, the FPR for Black suspects will be lower at 0.8, despite the high number of false positives for

Black suspects. This effect will lead to a higher threshold for White suspects compared to that for Black suspects when implementing an equalized-harms adjusted classifier if there are far fewer White suspects than Black suspects. We note, however, the threshold for Black suspects in the equalized-harms classifier is still higher than that implied by NYPD frisks/searches.

We also show that the fairness-adjusted classifier with overall FPR = 0.84 (the FPR for frisk/searches for White suspects) does not appear to make a large qualitative difference when examining FPR by precinct conditional on the two groups (Figure 3). The same trends of high FPR across both groups and many precincts having high FPRs only for Black suspects persist. Additionally, the false positive counts by precinct show that in some cases, the number of false positives actually increases for Black suspects under the fairness-adjusted classifier. Altogether, we conclude from this set of analysis that accounting for separation is an inadequate non-discrimination criterion when groups are unequally sized, as may commonly occur in practical settings.

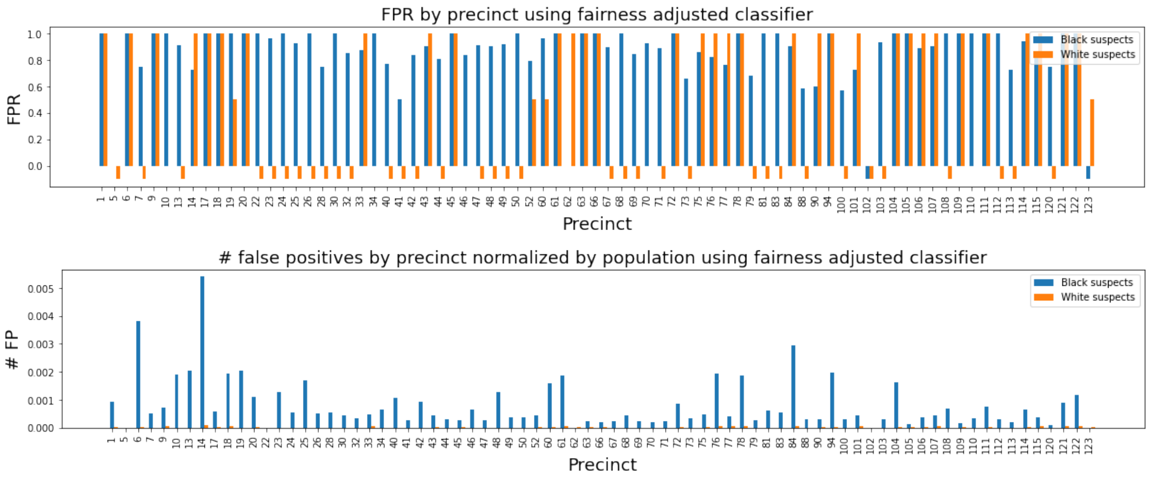

Figure 3: Detection rates by precinct from a fairness-adjusted classifier. Top: Negative bars indicate no available data for those conditions. Bottom: Counts are normalized by population of each group within each precinct. The value for Black suspects in Precinct 22 (0.056) is an outlier and was therefore omitted from the plot.

3. Conclusion

We analyze the detection rates of NYPD stops/frisks for criminal possession of a weapon conditional on race for the year 2020. We use this data to train a logistic regression classifier, from which we derive the implied decision boundaries of NYPD stops/frisks, as well as a fairness-adjusted (equalized-harms) classifier for comparison. We find that on average, innocent Black individuals are frisked and searched more often (higher FPR) than innocent White individuals. We also find that Black suspects are more likely to be searched for a given level of suspicion (lower decision threshold). However, we also find that using solely FPR as a fairness criteria based on these data will distort perception of historical data of police stops, as well as bias statistical decision-making models that are purportedly designed to be non-discriminatory. In examining historical data, we find that while in certain precincts, FPRs are indeed nearly equal across both groups, there are many precincts where only stops of Black suspects incur high FPRs, as well as precincts where FPRs across groups are equal due to the relatively low number of stops for White suspects. We find that the fairness-adjusted classifier has many of the same pitfalls as evidenced in the stop and frisk data, namely, high false positive counts for Black suspects relative to that for White suspects, despite nearly equal FPRs in many precincts. We conclude that this is due to the disproportionately high number of police stops of Black suspects, a trend discussed in detail in Pierson et al. (2020).

Our work corroborates prior findings in Kallus et al. (2018), where the authors derive a similar fairness-adjusted classifier based on NYPD SQF data and demonstrate that when extending the classifier to make decisions on test data at the population level, Black individuals are still disproportionately harmed compared to White individuals despite the considerations of fairness. Similarly, they conclude that this is a result of the inherent bias in stop and frisk repositories, which are generated under a prejudiced historical policy that leads to the gross over-representation of Black individuals in the data. Because these datasets are typically the settings where this type of non-discrimination research is conducted, our work strongly suggests that historical context as it relates to the data generation process may never be ignored in interpretation of quantities that purportedly describe fairness when evaluated on real data. An extension of this analysis on the separation criterion could also to examine independence and sufficiency (Barocas et al., 2019), as well as improve the design of models and classifiers adjusted for these other fairness criteria.

4. Methods

Data and analyses

For the code used to perform all analyses, please see https://github.com/ligeralde/sqf. Stop-and-frisk data used is found at https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page. Population data by precinct is found at https://johnkeefe.net/nyc-police-precinct-and-census-data.

Equalized harms classifier

We derive an equalized harms classifier for CPW based on the group-specific (conditioned on group $A = a$) empirical cumulative distribution function of the score $\hat{R}$ (probability of weapon possession) output by the logistic regression model, conditioned on $Y = 0$, the fact that no weapon is detected:

$$F_a(\theta) = P(\hat{R} \geq \theta | Y = 0, A = a). \tag{2}$$

Here, we treat the ECDF as a function of the threshold θ. The equalized harms classifier is then computed as $\hat{Y}_{fair} = \mathbb{1}(\hat{R} \geq \theta_a)$, where θ_a, the group specific threshold for group a, is given by

$$\theta_a = F_a^{-1}(FPR), \tag{3}$$

and FPR is the desired FPR across all groups.

References

Emily Badger. ”The Lasting Effects of Stop-and-Frisk in Bloomberg’s New York,” The New York Times. (2 March 2020).

Solon Barocas, Moritz Hardt, Arvind Narayanan. Fairness and Machine Learning (2019). fairmlbook.org

Braga, AA, Macdonald, JM, McCabe, J. Body-worn cameras, lawful police stops, and NYPD officer compliance: A cluster randomized controlled trial. Criminology. 2021; 1– 35.

Sharad Goel, Justin M. Rao, Ravi Shroff ”Precinct or prejudice? Understanding racial disparities in New York City’s stop-and-frisk policy,” The Annals of Applied Statistics, Ann. Appl. Stat. 10(1), 365-394, (March 2016)

Sharad Goel, Maya Perelman, Ravi Shroff, David Alan Sklansky; Combatting Police Discrimination in the Age of Big Data. New Criminal Law Review 1 May 2017; 20 (2): 181–232.

Moritz Hardt and Benjamin Recht. Patterns, predictions, and actions: A story about machine learning (2021). mlstory.org

Moritz Hardt, Eric Price, Nathan Srebro. Equality of Opportunity in Supervised Learning (2016).

Nathan Kallus and Angela Zhou. Residual Unfairness in Fair Machine Learning from Prejudiced Data (2018).

Pierson, E., Simoiu, C., Overgoor, J. et al. A large-scale analysis of racial disparities in police stops across the United States. Nat Hum Behav 4, 736–745 (2020).