Data Analysis on Police Shootings

Donglei Cai, Samuel Nelson, Frank Wang, Charlie Zhou

December 14, 2021

1 Introduction

African Americans make up about 14.2% of the U.S. population yet account for about 26.6% of people killed—decedents—by police (source 5). In this study we investigate whether the proportion of decedents who are African Americans differs significantly from the proportion of individuals in the U.S. who are African Americans. Additionally we generate an estimate for the proportion of decedents who are African American using a Bayesian Posterior model.

2 Data Overview

The fatal police shooting data we are using is a census collected by The Washington Post. The data is generated by aggregating the police shooting information from social media postings and police reports and the Washington Post has been collecting this data since 2015 in an attempt to combat under-reporting of fatal police shootings and to help elucidate some questions involving race and fatal police shootings. This data is constantly being managed by the Washington Post on their GitHub and we can access and download the data in csv format from there (see source 1 or this link).

Now let's examine our data further. Since reports of police shootings were vetted using social media posts and smartphone video when body camera or police records were unavailable, selection bias is a possibility as non-affluent areas may be under-represented due to decreased access to technology. This convenience sampling (social media posts are easily searched and videos are easily vetted compared to rumors) may have led to selection bias. While the people making up the dataset are all deceased—and so have no input on the use of the data—family members may have particular beliefs about how the data is used. In terms of the granularity of the data, each row represents one incidence of fatal police shooting and its relevant information. Since our data is very fine-grained and packed with very specific details regarding every single fatal shooting, this will make the interpretation of our finding very robust. One feature that we wish we had is the income of the decedents. This would have given us an indication of their socioeconomic status which is closely tied to systemic racism.

3 Research Questions

In this paper, we will be covering the following two research questions.

Our first research question is whether the probability a decedent is African American is different than would be expected given population proportions. If we indeed found that the rate at which African Americans are shot is different than would be expected given population proportions, then we can recommend targeted police training and try to eliminate racial profiling through educational workshops and drills. We will be using Multiple Hypothesis Testing to answer this question. This method is perfect for determining if the distribution of African Americans who are shot matches the distribution of African Americans in the population.

For our second research question, we are trying to find the rate at which African Americans are shot assuming that police shootings of African Americans are under-reported. Answering this question can help us how bad the racial profiling is for policing. Knowing this fact can help police departments determine how much effort and resources they want to put into educational workshops and drills on avoiding racial profiling. We use a weak beta prior whose ratio reflects a higher rate of African Americans shot than is represented by the data. Additionally we will also use the body camera variable

since the existence of a body camera could be indicative of a police forces attempt at accountability and may affect the likelihood of African Americans being shot.

4 EDA

4.1 Visualization

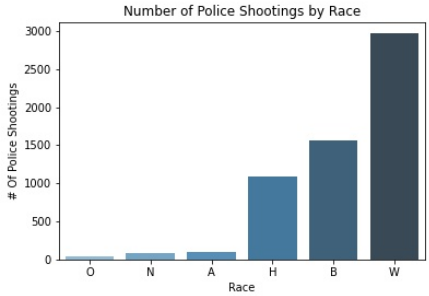

Figure 1: Number of Police Shootings by Race

Trend: The ratio of Black deaths to White deaths is about 1/2. This is much different than their population proportions would suggest. We'll determine if this is significant using multiple hypothesis testing in the next section. Explanation: This visualization is relevant to our research question because we can see that the rate of which African Americans are shot is greater than the population proportion.

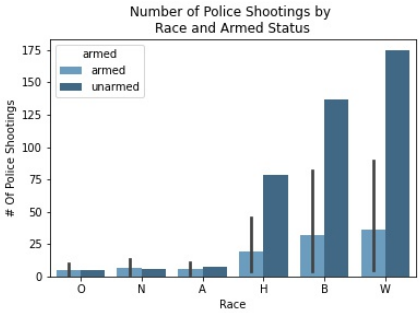

Figure 2: Number of Police Shootings by Race and Armed Status

Trend: The ratios between races are much closer for armed than unarmed. It will be interesting to determine if this is significant. Explanation: This visualization is relevant to our research question because being armed could potentially be thought of as a confounding variable.

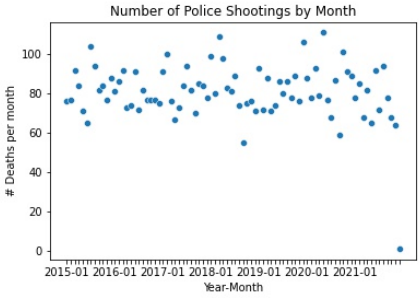

Figure 3: Number of Police Shootings by Month

Trend: There doesn't appear to be a trend in deaths over time. Explanation: This is relevant because it implies there is no effect associated with time that might be changing the rate of police shootings. This is important since we are using a dataset spread out over 6 years to make inferences about death rates now. The only outlier is the month we're in almost certainly because it's not over yet.

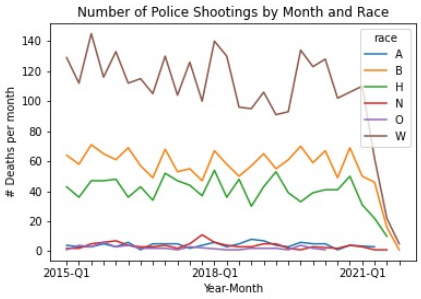

Figure 4: Number of Police Shootings by Month and Race

Trend: We can see that the trend in police shootings over time doesn't change between races. Explanation: This graph is relevant because we are using a dataset collected over 6 years to estimate a parameter (rate at which African Americans are shot) that might have changed over time. Since it seems to stay consistent, we can continue without accounting for time separately.

4.2 Data Cleaning

For Data Cleaning, we first removed all rows with missing armed or race values. Since we are investigating the roles race and armed status play in police shooting, we should deal very carefully with missing values. Interpolation would be a bad idea because we will very likely get wrong data from interpolation. The best thing to do here is to drop all rows with missing values. We have plenty of data left to build a strong model and removing that much data will not negatively impact our model performance and analysis in the end. In addition to dropping all columns wholly unrelated to our research questions, we removed the date column because our EDA above shows that time had no significant effect on relative race proportions.

5 Multiple Hypothesis Testing

5.1 Problem Setup

The ratio of African Americans shot about 0.266 of people who were shot by police. The ratio of African Americans out of the U.S. population is about $\frac{46936733}{331167284} = 0.142$. Both values come from the U.S. Census Bureau; the number of African Americans comes from the table given by source 3 and the total number of people is the amount estimated by the U.S. Census Bureau's Population Clock (source 4) as of Jan 01 2020. Notably our estimation of the proportion of Americans who are African Americans is higher than the Washington Post estimate. This is because they use the 'Black alone' category rather than the 'Alone or in Combination'. We use 'Alone or in Combination' for reasons like the *'hypodescent laws'* which defined people as Black if they "had one drop of African blood." (Source 6) Even if someone is not 100% African American it is still likely they face adversities due to systematic racism and so they may be subject to increased risk of police shootings.

We are assuming each decedent is African American with probability p similar to what Cody Ross (source 2) does on predicting race and police shootings. The authors are looking for the probability of being shot while African American rather than the probability that a decedent was African American. The assumptions are similar. Since each decedent makes up a bernoulli random variable we—under the null hypothesis—model the number of African American decedents as $Bin(N, p_0)$, where N is the number of people shot and p_0 is the proportion of the U.S. that is African American.

The main assumption we make is that each trial (i.e. each incident of a police shooting) is independent of the others. A moments reflection will reveal that this cannot be true since if a white

person is killed they are no longer part of the population and the probability of a decedent being African American goes up under the null hypothesis. We can treat the trials as independent however since the removal of one white person from the population would increase the population proportion of African Americans by 0.00000000043 percentage points (hardly enough to be significant). A related assumption is that each trial has the same probability of 'success'. Shootings aside, things like birth rate can also affect population proportions. The current proportion of African Americans in the U.S. is close to as high as it's ever been at 14.2 %. Since this advantages our null hypothesis, we know that birth rate and other forms of demographic drift will not increase the chance of a type 1 error beyond what we control for. We are interested in finding out if the rate at which African Americans are shot different than would be expected given population proportions.

5.2 Methods and Results

We will be testing five hypotheses instead of one to make our final conclusion robust. The hypotheses we will test are as follows:

Hypothesis Testing Question 1: Is the rate at which African Americans are shot different than would be expected given population proportions?

Hypothesis Testing Question 2: Is the rate at which African Americans are shot different than would be expected given population proportions when the individuals are armed?

Hypothesis Testing Question 3: Is the rate at which African Americans are shot different than would be expected given population proportions when the individuals are not armed?

Hypothesis Testing Question 4: Is the rate at which African Americans are shot different than would be expected given population proportions when the police is wearing a body camera?

Hypothesis Testing Question 5: Is the rate at which African Americans are shot different than would be expected given population proportions when the police is not wearing a body camera?

Testing five hypotheses will allow us to investigate and explore how these five external conditions effect the rate at which African Americans are shot by the police.

The null hypothesis for each question has a binomial distribution where the number of samples are the number relevant to the specific condition N_i—i.e. for question 1 all samples with a recorded race are relevant, for question 2 all samples with a recorded race and that were armed are relevant. N_i is the relevant sample size for the given conditions. The other binomial parameter (the probability) will be the probability of the decedent being African American under the null hypothesis (i.e. the proportion of the U.S. that is African American, 0.142). Since our p-value is the probability under the null distribution of getting the number of observed African American decedents (x) or more, we will calculate $P(S \geq x)$ where $S \sim \text{Binomial}(N_0, 0.142)$. Each p-value will then be calculated as follows.

$$p_i = \sum_{k=x}^{N_i} \binom{N_i}{k} p_0^k (1-p_0)^{N_i - k}$$

All our p values were highly significant. To the point that each was shown as 0.0 when printed in code. That means that for each of our questions the probability of observing the observed number of African American decedents or more was close to zero.

We will use a overall significance level of 0.05 and Holm-Bonferroni procedure to adjust the significance level for each hypothesis test. We'll also use the Benjamini-Hochberg procedure since demonstrating the racial distribution in police shootings under even just one condition is still useful (i.e. while controlling for the false discovery rate if 5% of rejected tests are false discoveries only one is necessary to determine racial bias). The Benjamini-Hochberg procedure allows us to decrease the false discovery rate.

Now we will apply the Holm-Bonferroni Procedure to our five p-values. For ordered (lowest to highest) p-values $P_1, P_2, ..., P_m$ where P_k is the kth p-value, we can adjust the significance level cutoff α for each p-value and control the strong family-wise error rate using the following method. If $P_k < \frac{\alpha}{m+1-k}$, we will reject the null hypothesis. The lowest cutoff is $\frac{0.05}{5} = 0.01$. Our highest p-value is 0. We will reject all null hypotheses and conclude that the probability a decedent is African American does not follow the population proportion regardless of armed status or whether or not there was a body camera.

Secondly, with the Benjamini-Hochberg procedure, our cutoff will be ≈ 0.0. We end up rejecting each null hypothesis because all the p-values will be equal to or less than our cutoff. We will conclude that the probability a decedent is African American does not follow the population proportion regardless of armed status or whether or not there was a body camera.

5.3 Results

We rejected the null hypothesis for each question we tested. We used the Benjamini-Hochberg procedure to control False Discovery Rate—the probability that the null is true given that we reject—and we used Holm-Bonferroni correction to control the more stringent strong family-wise error rate. The Holm-Bonferroni procedure controls the strong familywise error rate like just the Bonferroni adjustement but by ranking the p-values of the tests and adjusting the cutoff linearly, we get higher power (this is a different algorithm from Benjamini-Hochberg even though it sounds similar).

5.4 Discussion

All discoveries were significant. This is likely due to our use of a binomial distribution to represent the selection of decedents that were African American. Given that we controlled for the strong family wise error rate we can make conclusions about individual tests and as an aggregate (i.e. there probability of a decedent being African American does not follow population proportions regardless of conditional circumstances such as if the decedent was armed or if the officer had a body camera). There is one cause for concern in terms of p-hacking. Using a binomial distribution relies on assumptions that we had good reason to believe were not violated (see the beginning of the section for why) but it could be a cause of an overly narrow distribution resulting in artificially high power. Again, we believe that a binomial distribution is a reasonable choice for the null distribution but that is a concern. In the future it might be useful to do non-parametric tests as well.

6 Bayesian Hierarchical Modeling

6.1 Methods

Below, we include a drawing of the graphical model which we used, along with explanations for all the variables and their relationships, and relevant conditional distributions.

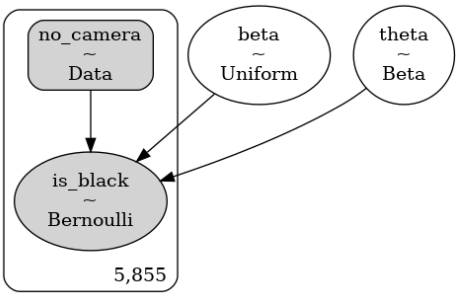

Figure 5: Bayesian Hierarchical Diagram

θ represents the proportion of fatal police shootings where the decedent is African American, β is a parameter which corresponds to how likely the individual will be African American when there is no body camera record, the z_i corresponds to whether each case had a missing body camera record (1 if missing), and the x_i correspond to whether the individual was African American (1 if yes).

θ affects the x_i by definition because it represents the proportion of fatal police shootings where the decedent is African American. β, in conjunction with the z_i, affects the corresponding x_i because they represent the effect of a missing body camera on the chances that the decedent is African American.

The distributions which we use in our model are as follows: $X_i \tilde{}\text{Bernoulli}(\theta \exp(-\beta z_i))$, with priors of $\theta \tilde{}\text{Beta}(2, 6)$ and $\beta \tilde{}\text{Uniform}(0, 1)$. The distribution for X_i is reasonable because it matches with the observation that in the cases when there is the body camera record is missing, the proportion of

African Americans increases noticeably. The prior distribution for β is reasonable because we must have $\beta > 0$ in order to model the observation that among the cases when body camera record is missing, the proportion of African American individuals involved in the shooting is larger. The prior distribution for θ is reasonable because its expected value close to 0.33 implements our prior belief that police shooting of African Americans is higher than the empirical 0.266.

6.2 Results

This method gives a 95% credible interval for the probability a decedent is African American of [0.333, 0.394], which is significantly different than the empirical result: 0.266. Here, a 95% credible interval means that the posterior probability of being in that interval is at least 95%. This leads to the inference that based on our model, the police killings of African Americans are under-reported. In the distribution below you can see that θ rests much higher than the observed estimate.

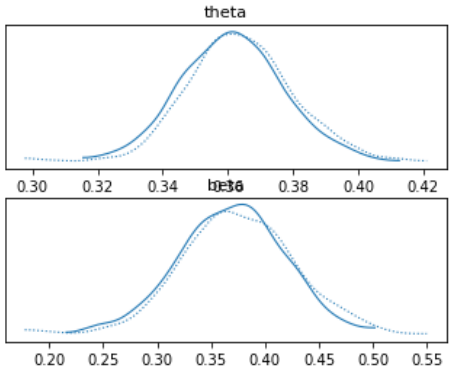

Figure 6: Bayesian Posterior Sample Distribution

6.3 Discussion

One significant limitation of this method is that we were obligated to choose a parameterized conditional distribution for whether or not a decedent is African American, and we also had to choose prior distributions for those parameters. Although we made best efforts to choose reasonable distributions, there is little reason to be confident that our model's distributions represent the true relationships between our variables. Additional data related to the race of the police officers involved may help with strengthening our model with more information.

Another formulation which we initially explored as a sense-check was a "pooled" Bayesian modeling approach, where we did not account for the variables related to body camera records. With that model, the empirical proportion of 0.266 fell within the credible interval which was given. The conclusion based on that model, therefore, was that police shootings of African Americans was not under-reported. However, this model is lacking in detail because it doesn't explicitly account for the existence of body camera records, so the model detailed above is more trustworthy.

7 Conclusion

Overall, our key findings were that the probability a decedent is African American does not follow the population proportion regardless of armed status or whether or not the officer(s) were wearing a body camera, and furthermore, the number of African American decedents is under-reported. Based on these results, we would urge police training to place a stronger emphasis on racial sensitivity training. One limitation of the data that we could not directly account for was more fine-grained information related to the state of the decedent at the time of the shooting, which prevents us from doing deeper analysis on how police may react to similar situations depending on an individual's race. Future studies could be done to investigate this exact question.

8 References

[1] https://github.com/washingtonpost/data-police-shootings/blob/master/fatal-police-shootings-data.csv

[2] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141854

[3] https://www2.census.gov/programs-surveys/decennial/2020/data/redistricting-supplementary-tables/redistricting-supplementary-table-01.pdf

[4] https://www.census.gov/popclock/

[5] https://www.washingtonpost.com/graphics/investigations/police-shootings-database/

[6] https://plato.stanford.edu/entries/race/