Jesse Geman
Data 2020
May 3rd, 2018

Data 2020: Final Project

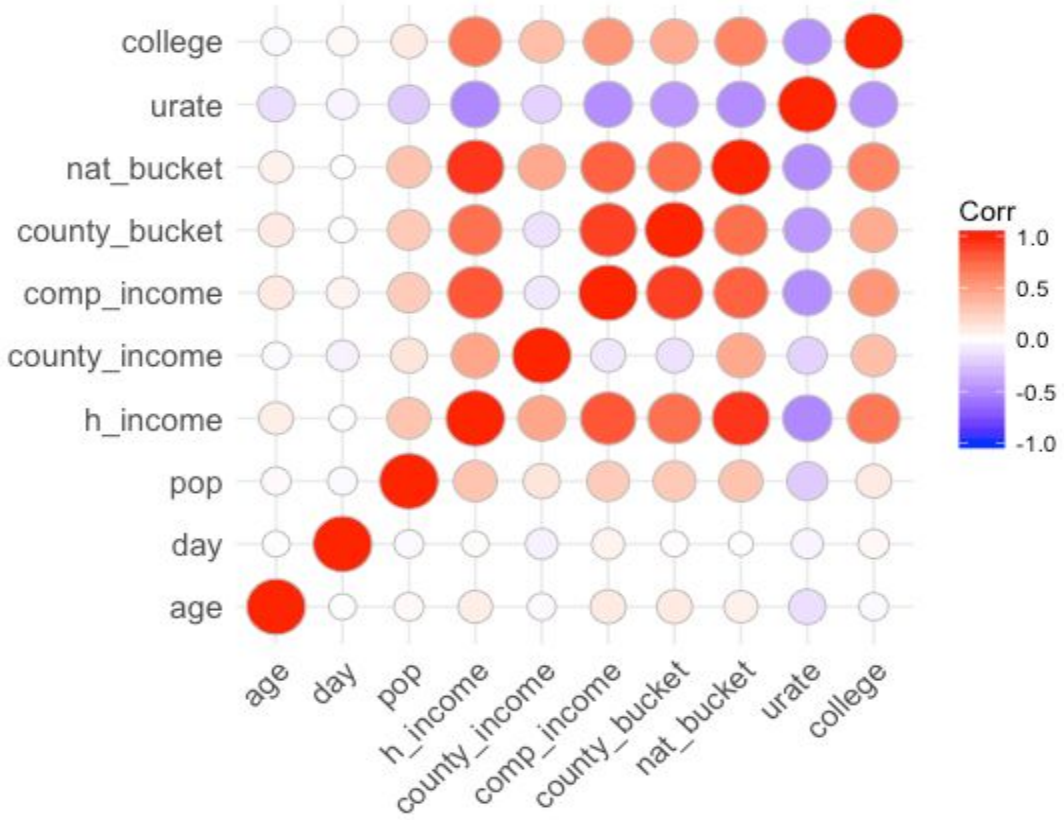
Exploratory Analysis

The goal of this project is to see to how much of a role race plays in police killings. Since we do not have the racial profile of the police officers who are involved in these killings we must limit our conclusions to how the race of those killed by police plays an explanatory role.

In an effort to become better acquainted with the data I created a number of graphs and plots. The motivation being that visualizing the data might explain some of the correlations, trends or oddities that might otherwise go unnoticed if I began by just building models. Additionally, I designed a hypothesis test to see whether the demographics of the counties mirrored demographics of those killed by police in the counties. And while I did make some good progress, there is more exploratory analysis to be done!

Visualization

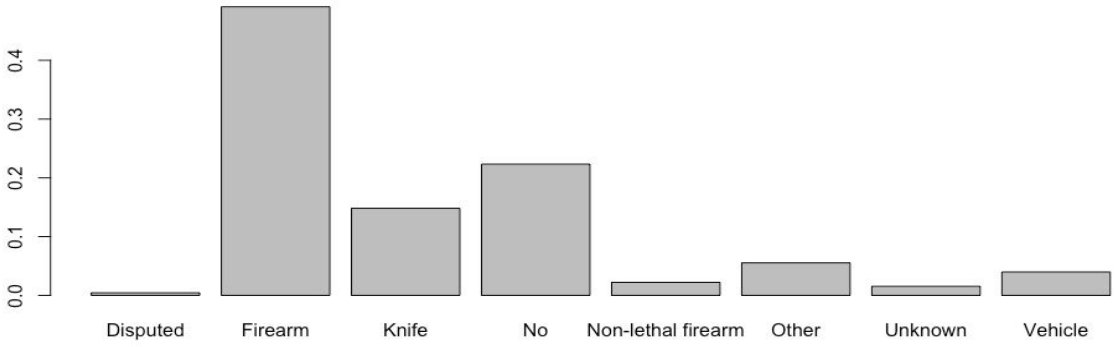
I started my analysis by building a covariance matrix using the police killings dataframe (non-numeric values are not included).



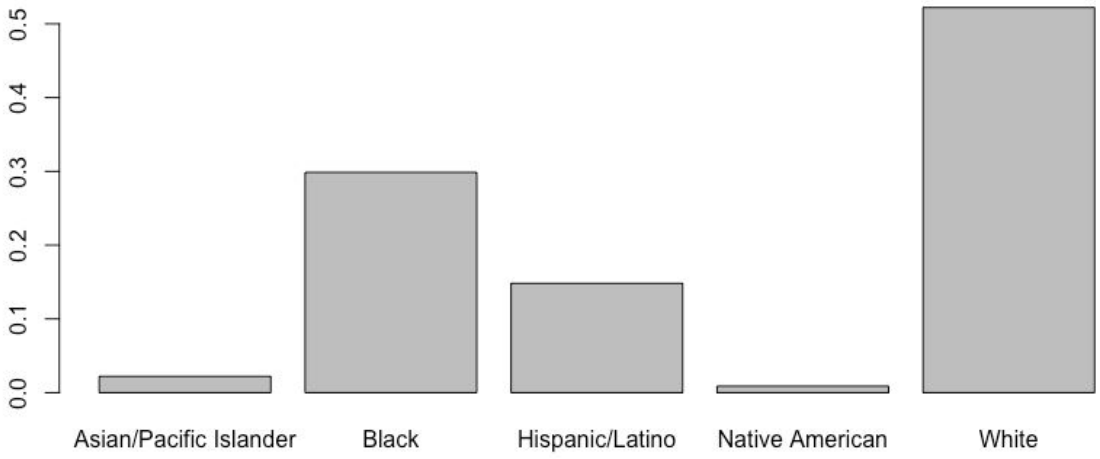
Jesse Geman
Data 2020
May 3rd, 2018

Nothing is particularly surprising here, we see correlation where we would expect. For example, the level of education is positively correlated with median household income and negatively correlated with the rate of unemployment.

I then made a number of histograms to try and see what types of trends might exist between the victims.

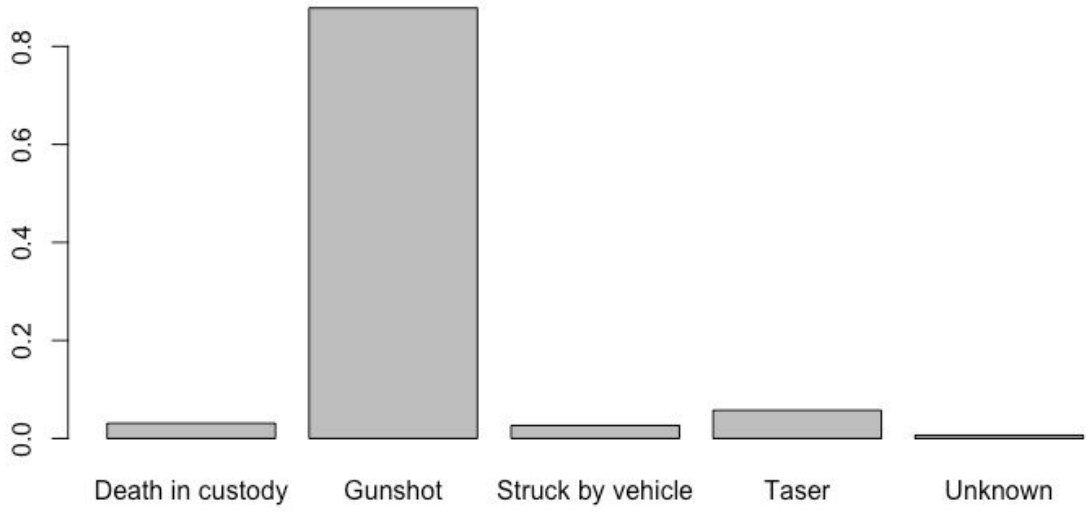


How/whether the deceased was armed.

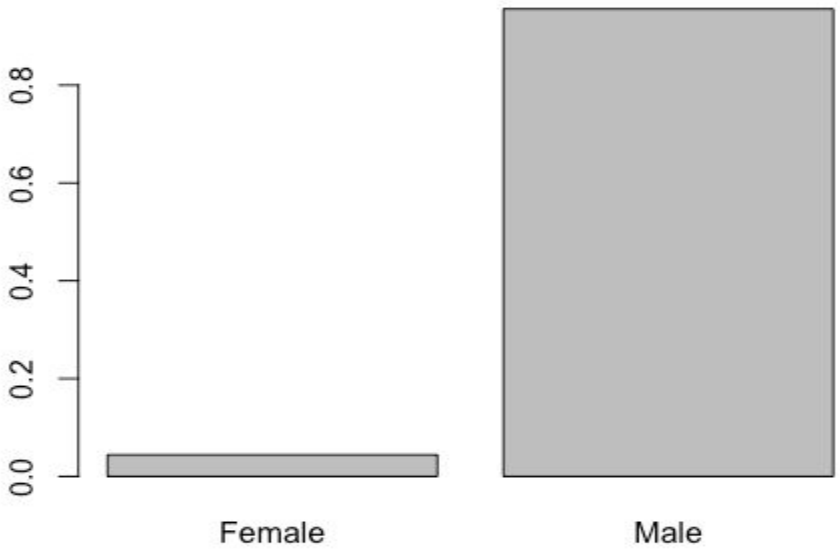


Victims by race (For comparison: 72% of Americans are White, 17% are Hispanic/Latino 12.6% are Black, 5% are Asian/Pacific Islander, 0.9% are Native. In this case Hispanics/Latinos might also be categorized as White or Black). These numbers speak for themselves.

Jesse Geman
Data 2020
May 3rd, 2018

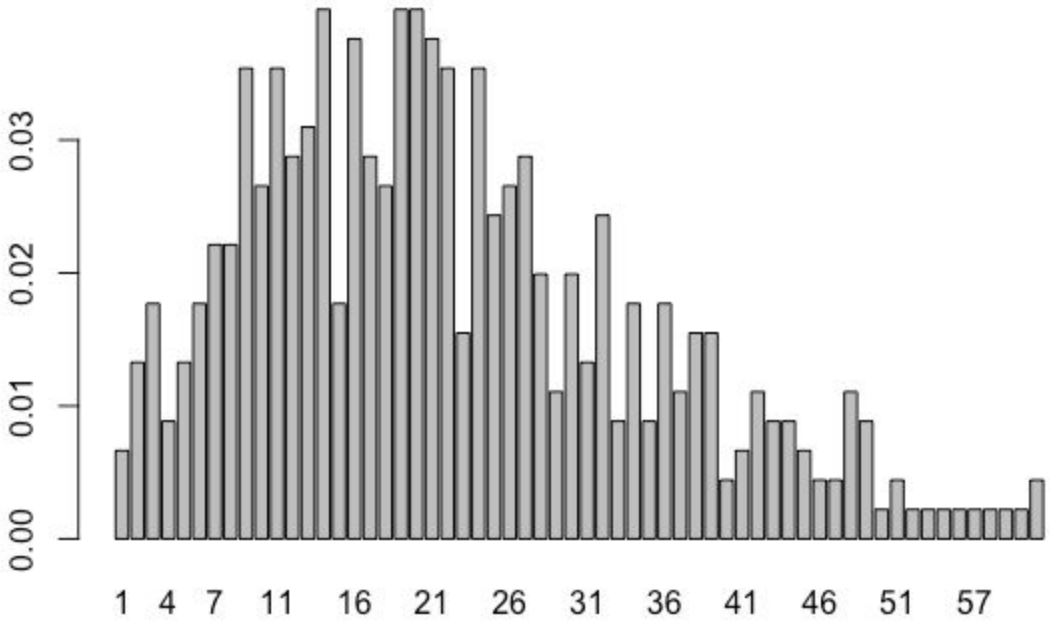


Cause Of Death



Victims by Gender

Jesse Geman
Data 2020
May 3rd, 2018



Victims by Age

Hypothesis Test

While I was able to learn more about the data through the various graphs and plots above, I wanted to get a sense of how “surprising” these killings were given the racial makeup of the counties. I decided to create a simple, Monte Carlo hypothesis test with the following the null hypothesis: the distribution on the racial makeup of those killed by police in a given county is the same as the distribution on the racial makeup of the county’s population.

To begin, for each county I used the race of the victims (either Asian/Pacific, black, hispanic, white, or native) to create an empirical distribution where each probability corresponds to the percentage of people of that race we expect to die when a police killing occurs in that county. After removing entries where the race of the victim was unknown and the rows that contained NA values, there were only six counties with more than one recorded killing in the dataset. Thus, most of the empirical, county distributions had one race that had probability 1 of being killed by the police and four races that had probability 0 of being killed by the police. To reiterate, I ended up with 415 vectors (there were 415 unique counties with recorded police killings) of length five where most of the elements were 0 and either one element was 1, or two

elements were 0.5 (the case where there were two recorded police killings and the victims were of different races).

I then measured the L1 distance between the 415 empirical distribution vectors obtained above and the county demographic distribution vectors (where each element gives the fraction of that county's population that belongs to a specific race). The demographic distribution vector values were taken from the police killings csv.

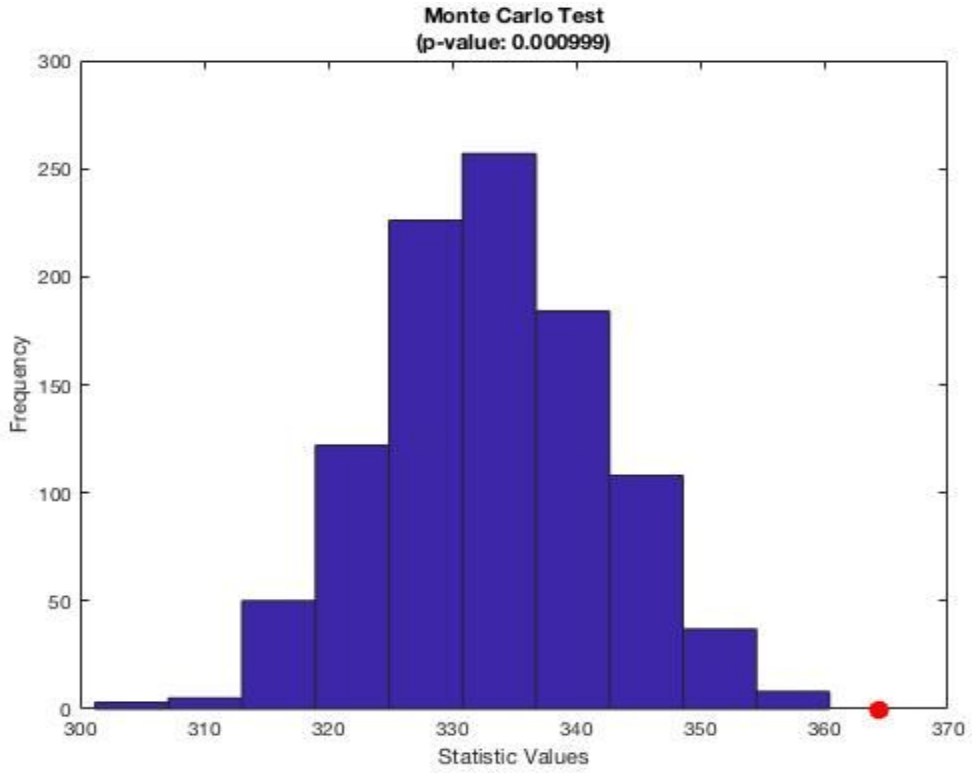
Finally, I summed the 415 L1 distances I obtained for each county. This sum, the sum of the L1 distances between the distribution of the race of the victims and the demographic distribution for the county, is the test statistic.

After calculating the observed statistic value, I generated 1000 sample statistics under the assumption of the null hypothesis. That is, for each county I used the demographic distribution to sample the race for a hypothetical victim. In counties where there were multiple police killings I took multiple samples. I then calculated the L1 distance between the generated empirical sample distribution (obtained the same way I obtained the empirical distribution for the observed killings) and the demographic distribution for that county. Finally, I summed the 415 L1 distances to get a new generated sample statistic value. It is worth reiterating that these statistics were all generated using the model given by the null hypothesis: That for each county the distribution on the racial makeup of those killed by police would be the same as the distribution on the racial makeup of the county. As I mentioned above, I generated 1000 sample statistics.

Below is a histogram of the 1000 statistic values. The red dot corresponds to the statistic we got from the actual data. Obviously, it is significant and we can conclude the distribution on the races of those killed by police differs quite a bit from the distribution on the racial makeup of the county. The p-value for this test was 0.000999. [1]

[1] The code for my hypothesis test can be found in the following pages. It was done in Matlab.

Jesse Geman
Data 2020
May 3rd, 2018



Going Forward

While it is evident that certain racial groups are disproportionately killed by police, we cannot make any causal inference just yet. There could be confounding variables. With this in mind, it would be interesting and useful to see how other factors might relate to police killings. For example income levels. It would also be useful to dig deeper, for example what if we just looked at killings where the victim was unarmed. It's unfortunate that we don't know more about the deceased, what levels of education had they obtained, or what income bracket were they in? Still, we have enough data where we should be able to glean more insights.

```
% C: an nx5 table of probabilities. Each row is five non-negative numbers
% that add up to 1, representing the proportions of the five ethnicities in
% one census tract. There are as 415 rows(we removed rows containing NA
% values and rows where the race of the victim was unknown).

% E: same as C, except that the proportions come from the population of
% victims. Rows in E correspond to the rows in C, i.e. they come
% from the same census tract.

% Define a statistic 'S', which is the L1 distance between E and C

% Build a null distribution for S by creating "surrogate" versions of E.
% For example, let SE be an nx5 table with rows that correspond to the rows
% of C and E, except that the entries are random and come from random
% samples from the distributions represented in C. Each row of SE is
% determined from 'ESums' selections from the C distribution, where ESums
% is the number of victims recorded in the corresponding tract.

% Get the data (Tract data and victim data)

C=csvread('RacesOfCounties.csv',1);
E=csvread('RacesOfVictims.csv',1);

NumSamples=1000;
[NumTracts,NTypes]=size(C);

% Clean C, meaning force total in each row to be 1

CSums=sum(C,2);
for col=1:NTypes
    C(:,col)=C(:,col)./CSums;
end

% Get number of attacks in each row, and then normalize rows of E

ESums=sum(E,2);
for col=1:NTypes
    E(:,col)=E(:,col)./ESums;
end

S=sum(sum(abs(E-C))); % The observed value of the statistic

% Make surrogate NumSamples surrogate E's and compute, for each, a
% surrogate S ('SS')

SS=zeros(NumSamples,1);

for samp=1:NumSamples

    ES=zeros(NumTracts,NTypes);  % ES will hold the current surrogate E

    for row=1:NumTracts
        r = mnrnd(ESums(row),C(row,:)); % Multinomial selection of
                    % number of victims, but from the distribution in C

        % load the victim numbers into the surrogate, ES
        for col=1:NTypes
            ES(row,col)=r(col);
```

```
        end
    end

    % Normalize the rows of ES (make them probability distributions
    ESSums=sum(ES,2);
    for col=1:NTypes
        ES(:,col)=ES(:,col)./ESSums;
    end

    % Compute the L1 distance between C and the surrogate ES
    SS(samp)=sum(sum(abs(ES-C)));

end

% Display Results

figure(1)
close(1)
figure(1)
hist(SS);
hold on
scatter(S,0,100,'filled','r')
hold off
pvalue=(sum(SS>=S)+1)/(NumSamples+1);
disp(['p-value: ',num2str(pvalue)])
```