Chapter 1

Racism in Policing

1.1 Introduction

In recent years, data on policing has become more accessible than ever before. Through endeavors like The Stanford Open Policing Project [3], it is now possible for a wide array of stakeholders to scrutinize the police practices of in their communities.

In this chapter, you will use statistical programming tools like R and Python notebooks to analyze large amounts of police data. These are datasets that would either be extremely slow or impossible to open in Excel or Google Docs.

1.2 Objectives

By the end of this module students will be able to:

1. Differentiate between quantitative and qualitative variables.
2. Determine the mean and median of quantitative data.
3. Create visual representations of data.
4. Understand proportions.
5. Apply these concepts to crime data to search for racial bias in policing.
6. Use Bayes theorem to calculate conditional probabilities in real-world contexts.
7. Use Google Colab/Python and large datasets to aid in these calculations.

1.3 Understanding The Issue

California provides an informative example of the political process surrounding police data, and how state level data has led to local impact.

In 2016 the Racial Identity and Profiling Act (RIPA) was passed, which enabled the widespread collection of racial identity in police interactions [1]. Before RIPA, the only cities in California

that were required to collect data on the race of citizens involved in police stops were cities whose police forces were under consent decrees- meaning that the federal government had previously found widespread misconduct and mandated that the police force collect that data.

RIPA formed an advisory board comprised of community stakeholders and academic researchers. Each year the board analyzes stop data and releases reports on disparities in policing, which are then used by communities around the state to advocate for changes in both funding and accountability measures for police. As a result of the data, cities like Berkeley, Oakland, San Diego, and Los Angeles have started taking steps towards significant changes to police practices.

Note: Although *race* is regarded as a social construct by scientists from a wide variety of backgrounds [2], we acknowledge the very real racialized oppression that many groups experience in the United States. The use of race as a variable in this data is borne out of the necessity of shedding light on this oppression. In some cases police report the race of the person they stop, and in some cases it is also self-reported, which can lead to some uncertainty.

1.4 Cui Bono: Who Benefits?

When police practices are not scrutinized, the public loses. Minoritized communities have unjustly born the burden of racialized policing, from the such as was the case in Ferguson, Missouri. When court cases are brought against police for misconduct, taxpayers ultimately pay out settlements to victims.

Two classes of organizations benefit from a lack of accountability in policing:

1. Police and police unions: when these groups are not held accountable they save money in additional training and settlement payouts.

2. A host of private companies like the GEO Group, Core Civic, or LaSalle Corrections that make money off of both the prison industry and exploiting prison labor.

1.5 Big Problem: Racially-biased Policing

What is racially-biased policing? How do we know if police are treating specific groups differently? How can we detect racially-biased policing from crime data?

1.6 Math Topic I - Understanding data

We have two large data sets, adapted from the Stanford Open Policing Project [3], giving information on vehicular stops in Hartford, CT and Philadelphia, PA between April 1, 2014 and September 29, 2019. Below are the first two rows from the Philadelphia data set.

date	time	age	race	sex	searchfrisk	contraband	arrest
2014-04-01	0:00:00	20	white	male	FALSE	NA	FALSE
2014-04-01	0:04:00	33	black	male	FALSE	NA	FALSE

We call each row an **observational unit** or **observation**. In this case, each observation is one vehicular stop. Each column is called a **variable**; here we have 8 variables. First, the date and time of the stop are provided. The next three variables give the age, race, and sex of the driver of the vehicle.

The `searchfrisk` variable is `TRUE` if the vehicle was searched or the driver was frisked and `FALSE` otherwise. If either the vehicle was searched or the driver was frisked, `contraband` is `TRUE` if an item was found which was illegal for the driver to possess and `FALSE` otherwise. If no search or frisk was conducted, contraband cannot be found, so `NA` (for "not available") is written. Finally, the driver is either arrested (`TRUE`) or not arrested (`FALSE`).

The `age` variable is a **numerical** variable since the response is a number and it makes sense to perform arithmetic operations on these numbers, such as addition, subtraction, or computing averages. The last six variables are not numerical. Since the responses to these variables fall into different categories, these are called **categorical** variables. We won't be analyzing the `date` and `time` variables, but whether they are considered numerical or categorical depends on the situation.

1.6.1 Summarizing Data

Numerical data

Suppose we are given the following values of a numerical variable:

$$2, 10, 3, 6, 8, 3, 4, 5, 2, 3, 5, 6, 1, 7, 9, 8, 10, 8.$$

These values could represent the cost of 18 different taxi rides or 18 students' scores on a quiz or the number of petals on 18 different flowers.

We often want to find the "center" or "middle" of the values. To compute the **mean**, we simply add up the values and divide the total by the number of values. In this case, we get

$$\frac{2+10+3+6+8+3+4+5+2+3+5+6+1+7+9+8+10+8}{18} = \frac{100}{18} \approx 5.56.$$

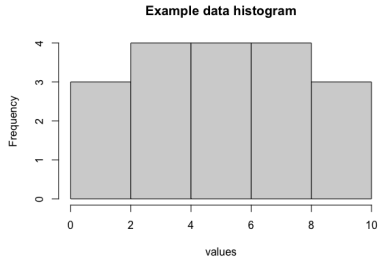
We could say that, on average, each taxi ride cost \$5.56, or that the average score on the quiz was 5.56. But sometimes the mean can hide useful information. Suppose we have a class of 10 students and one student has 10 cookies, while the other 9 students have 0 cookies. The mean number of cookies each student has is

$$\frac{10+0+0+0+0+0+0+0+0+0}{10} = 1.$$

So, on average, each student has 1 cookie. But the "average student" definitely does not have one cookie–in fact, almost all of the students have 0 cookies! The **median** is the middle value once the data is sorted (if there is an even number of values, the median is the mean of the two middle values). For our initial example, we first sort the data:

$$1, 2, 2, 3, 3, 3, 4, 5, 5, 6, 6, 7, 8, 8, 8, 9, 10, 10.$$

Since there is an even number of values, we take the mean of the two values in the middle, which are 5 and 6. Hence the median is $\frac{5+6}{2} = 5.5$. Notice that half of the values are less than the median and half are greater. But what if we want to know more about the data we have than just the mean and median? We can create a **histogram**. The one below is for our example data.



On the x-axis, we've grouped our values into intervals–the leftmost interval goes from 0 to 2, inclusive. The height of the bar in this interval (called the **frequency)** is 3, since there are three values $(1, 2, 2)$ in our data in this interval. The next interval goes from 2 to 4. The interval includes 4 but it does not include 2. The frequency here is 4, because our data included three 3s and one 4. Likewise, the frequencies in the last three intervals are 4, 4, and 3, respectively. These intervals include their upper endpoint but do not include their lower endpoint.

Let us apply these ideas to the `age` variable from the Hartford data. Since there 9630 stops recorded in this data set, computing the mean and median by hand would take a long time! Instead, we'll use `R`, a programming language heavily used in statistics. The functions `mean` and `median` give the mean and median of the variable `age`, while `hist` creates the histogram. Just hit "Evaluate `R`" below each block of code to see the results!

```
<sage language="R">
<input>
age <- hartford$age
mean(age)
</input>
</sage>

<sage language="R">
<input>
median(age)
</input>
</sage>

<sage language="R">
<input>
hist(age)
</input>
</sage>
```

Categorical data

If our data is categorical, none of the previous methods work–we can't take averages or sort the data from least to greatest. Nor can we create intervals and use these to make histograms.

However, there are still a couple things we can do! We can summarize data in a **table** or make a **bar chart**.

Consider the `sex` variable from the Hartford data set. We first make a table in `R`

```
<sage language="R">
<input>
sex <- hartford$sex
table(sex)
</input>
</sage>
```

From this table, we can see exactly how many of the stopped drivers were male and female. Then, we can use these numbers to determine the **proportion** of stopped drivers who were female by dividing the number of females (which we just found in the table) by the total number of stops (which we know is 9630). We can get a visual representation of this same data by making a bar chart.

```
<sage language="R">
<input>
barplot(table(sex))
</input>
</sage>
```

1.6.2 Proportions

Are there more traffic stops in Hartford or Philadelphia?

This question seems easy to answer. We can determine the number of stops in each city between April 1, 2014 and September 29, 2016 using the `nrow` function.

```
<sage language="R">
<input>
    nrow(hartford)
nrow(philadelphia)
</input>
</sage>
```

There were a lot more stops in Philadelphia than in Hartford! In fact, there were

$$\frac{678445}{9630} \approx 70$$

times as many traffic stops in Philadelphia than in Hartford. Does this mean people in Philadelphia are worse drivers or commit more crime? Not necessarily–there are many more people in Philadelphia, so it makes sense that there are more traffic stops. Using data from the U.S. Census Bureau, we see that Philadelphia had approximately 1,555,000 people in 2015 and Hartford had 125,000 [4, 5]. Therefore, there were

$$\frac{1567000}{124000} \approx 13$$

times as many people in Philadelphia than in Hartford.

Well, that's interesting. Philadelphia had about 13 times the population, but 70 times the number of traffic stops as Hartford. These data show us that, there were many more traffic stops

in Philadelphia than in Hartford in this time frame, *even when we control for population.* Unfortunately, the data cannot tell us *why* this is the case. Further research, with the help of experts in policing, history, crime, and a slew of other subjects may help us.

Are Black drivers stopped more often than drivers of other races?

Now let's look at the racial makeup of the stopped drivers in each city. The data here isn't perfect; there are only six possible options for race: `asian/pacific islander`, `black`, `hispanic`,`white`, `other`, and `unknown`. Many, many people do not fit neatly into one of these categories, but we do what we can with the data we have. We use the `table` function to get the numbers we want.

```
<sage language="R">
<input>
    table(hartford$race)
    table(philadelphia$race)
</input>
</sage>
```

Using these tables and the total number of stops in each city, we see that

$$\frac{3589}{9630} \approx 37.3\% \text{ and } \frac{435548}{678445} \approx 64.2\%$$

of the drivers stopped were Black in Hartford and Philadelphia, respectively. We now compare these percentages to 2015 population data from the US Census Bureau. For Hartford, 37.3% of the drivers stopped were black and 38% of the population was black–these numbers seem to indicate that black drivers were not stopped more frequently than drivers of other races [4]. In Philadelphia, however, 64.2% of the drivers stopped were black while only 42.4% of the population was Black [5]. It seems as though Black drivers were disproportionately stopped in Philadelphia during this time. Let's use the data on searches, contraband, and arrests to give us more insight.

Are stopped drivers searched more often if they are black?

Once a driver is stopped, officers may or may not search them or their vehicle. We can use the `table` function we've used before in a new way to help us analyze if a driver's race is related to whether or not they are searched.

```
<sage language="R">
<input>
    table(hartford$race,hartford$searchfrisk)
</input>
</sage>
```

In Hartford, we already saw that there were 3589 stops of Black drivers. In our new table, we see that 925 of these drivers were searched in some way (look at the `TRUE` row and `black` column). The remaining 2664 were not searched. For white drivers, 835 were searched out of a total of 3486 who were stopped. Therefore,

$$\frac{925}{3589} \approx 25.8\%$$

of stopped Black drivers were searched while

$$\frac{835}{3486} \approx 24.0\%$$

of stopped white drivers were searched. Stopped Black drivers seem to be searched at slightly higher rate than stopped white drivers. But perhaps the Black drivers have contraband at a slightly higher rate as well?

We use the `table` function again, but this time include whether or not contraband was found.

```
<sage language="R">
<input>
    table(hartford$race,hartford$contraband,hartford$searchfrisk)
</input>
</sage>
```

The resulting table is a little more complicated–in fact, you probably see two tables! In the first table, all the entries are 0. The reason for these zeros is that we first consider all drivers who were not searched or frisked. There is no way to find contraband in this case, so the entry in the `contraband` column is `NA`–neither `TRUE` nor `FALSE`–so all of the entries are 0.

The second table includes all of the drivers who were searched or frisked. Notice that contraband was found for 8 of the 925 searched or frisked Black drivers, or 0.9%, and for 10 of the 835 searched or frisked white drivers, or 1.2%. We see that the proportion of white drivers who possess contraband is slightly higher than that for Black drivers. It is therefore possible that officers are searching Black drivers on less evidence than they are for white drivers.

In summary, in Hartford between April 2014 and September 2016, it seems as though black drivers were not stopped disproportionately. Stopped black drivers were searched or frisked at a slightly higher rate than stopped white drivers, and contraband was found on a slightly higher percentage of searched or frisked white drivers than searched or frisked black drivers. While we cannot make any definitive conclusions based on our analysis, it seems as though there may was a small anti-black bias in vehicular stops and searches, but there did not seem to be large-scale, widespread racial disparities. You should edit the blocks of code above and see what conclusions you can make for the Philadelphia data.

1.7 Math Topic II - Bayes Theorem

If events A and B can occur simultaneously (i.e., they are not independent), you can calculate their joint probability. An example two such events are being dealt a 6 card in a game of cards, and being dealt a hearts card. (These are dependent because it is possible to be dealt a six of hearts.)

The probability of the two dependent events A and B happening, $P(A, B)$, is the probability of $A, P(A)$, times the probability of B given that A has occurred, $P(B|A)$. This is a result you may remember from a high school Algebra 2 course.

$$P(A, B) = P(A)P(B|A)$$

Putting this distribution in terms of event B, he probability of A and B is also equal to the probability of B times the probability of A given B.

$$P(A, B) = P(B)P(A|B)$$

Setting the two equal:

$$P(B)P(A|B) = P(A)P(B|A)$$

Solving for $P(A|B)$,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

This theorem is ubiquitous in statistical inference and machine learning/artificial intelligence.

1.7.1 Using Bayes Theorem to calculate false positives

One consideration that policy makers and public health experts have to take into account when selecting tests for a disease is the false positive and false negative rate. These can be thought of as the probability of testing positive when a person does not have the disease (false positive) and the probability of testing negative when a person actually has the disease.

Let's say we are trying to find the probability of a false negative. This is arguably the scarier scenario from a public health scenario, as we saw with asymptomatic carriers of the COVID-19 virus. If someone thinks they are negative but are actually positive, they might spread the disease to many others!

Let's say the Umbrella corporation is attempting to test a new rapid COVID test. They start by randomly selecting 100,000 people and administering two tests: their new rapid test and a 100 pct. reliable (but slow) Super PCR test. (In reality PCR tests are not 100% effective)

Note: "Negative" means *tested* negative, not actually negative, "positive" means *tested* positive, not actually positive.

$$P(infected|negative) = \frac{P(negative|infected)P(infected)}{P(negative)}$$

In the broader population, the prior belief of the Center For Disease Control is that the positive rate for the disease is 1 percent. Therefore, $P(infected) = 0.001$. The manufacturers of the test have taken a random sampling of people and administered the test. Of the 100,000 people, 98,500 people tested negative. Therefore $P(negative) = 98,500/100,000 = 0.985$.

Because there are many more non-infected people, and a low negativity rate for the test, the company decides to take the 1,500 people who tested negative using the Super PCR test. All of these people are actually infected, because the other test is 100 pct. accurate. Of these 1,500 people who are actually infected, 1,000 tested positive and 500 tested negative using the rapid test. Therefore, $P(negative|infected) = 500/1,500 = 1/3$.

We now have everything we need to make the false negative calculation.

$$P(infected|negative) = \frac{P(negative|infected)P(infected)}{P(negative)} = \frac{(1/3)(0.001)}{(0.985)}$$

$P(infected|negative) = 0.00032833$. This might not seem like a lot of people, but consider the implications for a city of 1,000,000 people; that's $0.00032833 \times 1,000,000 = 328$ people walking around the city, spreading disease!

1.7.2 Using Google Colab and Python to analyze a large police stop dataset

For the following exercises, follow this link.

Make a copy of the notebook by going to File → Save a copy in Drive, then navigate to your Google Drive and follow along.

1.8 Solving for Change

We kind of included the "solving for change" part in our "math topic" sections. Not sure if/how we should edit to follow this outline properly.

1.9 Reading Questions

1. What are some limitations of this data in determining if police departments are racially biased? What additional data would help?
2. What actions can we take to accurately assess if a police department is racially biased? If a particular police department is shown to have a pattern of racial bias, what are some ways we can address the issue?

1.10 Exercises

For the following three exercises, edit the blocks of code in the **Understanding data** section to show results for stopped drivers in Philadelphia rather than in Hartford.

1. What was the mean age of stopped drivers in Philadelphia during this time? What about the median? Create a histogram of the `age` variable from the `philadelphia` data.
2. Create a table and bar chart of the `sex` variable from the `philadelphia` data.
3. In the **Are drivers searched more often if they are black?** section we analyzed, for the stopped Hartford drivers, how often white and black drivers were searched after they were stopped, and how often contraband was found among those who were searched. Apply the same techniques to the Philadelphia drivers. Were stopped white or black drivers searched at a higher rate? Was contraband found at a higher rate among white or black searched drivers?

For the next two exercises use Bayes theorem and the Colab notebook/ Nashville data.

1. Recent data has estimated the worldwide percentage of Spam emails as 28.5% [6]. A new software company states that their product can detect 98% of emails as spam. Sometimes (2%) of the time, the filter incorrectly labels non-spam emails as spam (false positive). With these percentages in mind, what is the true probability that an email, if labeled spam, is actually a non-spam email? (Hint: There are many ways you can approach this, but it may make sense to use A to model an event that an email is labeled spam, and B to represent that the email actually is spam.)
2. Are white motorists more likely to have a warning issued than Hispanic motorists? Use the Colab notebook to answer this question.

1.11 References

Bibliography

[1] Racial and Identity Profiling Act. Retrieved August 30, 2021, from https://post.ca.gov/Racial-and-Identity-Profiling-Act

[2] Gannon, M.. Race Is a Social Construct, Scientists Argue. Scientific American. Retrieved August 30, 2021, from https://www.scientificamerican.com/article/race-is-a-social-construct-scientists-argue/

[3] 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.

[4] U.S. Census Bureau, American Community Survey Demographic and Housing Estimates (2015), https://data.census.gov/cedsci/table?q=&g=1600000US0937000&y=2015&tid=ACSDP1Y2015.DP05.

[5] U.S. Census Bureau, American Community Survey Demographic and Housing Estimates (2015), https://data.census.gov/cedsci/table?g=1600000US4260000&y=2015&tid=ACSDP1Y2015.DP05.

[6] Spam e-mail traffic share 2019. (n.d.). Statista. Retrieved August 29, 2021, from https://www.statista.com/statistics/420400/spam-email-traffic-share-annual/