# Assessing Balance of Eviction Ethnicity in Massachusetts

Joseph Palin

A Thesis in the Field of Information Technology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2019

# Abstract

This project analyzes Massachusetts housing court records, testing if there is ethnic bias in who is evicted. The project starts with semi-manually scraping court records from the Massachusetts court website. Because the website disallows fully automatic scraping of records, tools were developed to facilitate and expedite the downloading of data without violating the terms and conditions for the page.

With source web pages downloaded, the records were processed to extract eviction features: plaintiffs, defendants, location of eviction, and which housing court. Largely through format shifting, nearly a gigabyte of records was parsed and pared down to about 20MB stored in a usable data frame. Incomplete or unparsable records were removed. Once the raw data was processed, addresses were geocoded and turned into census tracts to add probabilities of ethnicity for each record, based on location. Additionally, last names were matched against census database records to extract the corresponding probabilities of ethnicity based on surname. These pieces of data were then combined to create a refined estimate of ethnicity for each data point.

The estimated ethnicities were compared with the underlying demographics to determine if there was a meaningful difference in evictions between how many people of each ethnicity were evicted, and what should be expected based on the underlying population. Depending on how the model was tested, there were two different results. Either Caucasians are over-evicted relative to their population, or Hispanics and African Americans are over-evicted relative to their populations.

# Acknowledgements

To my brother, without whom I didn't have a project, to my advisor, without whom I wouldn't have a thesis, and to my parents and grandparents, who have given me the freedom to pursue my path in life, thank you.

# Contents

# List of Figures

# List of Tables

# Chapter 1: Background

A few years ago I was avidly following algorithms in the news, what they were, but also how they were applied. Were they being applied in an ethical manner? The infamous trolley problem was recirculating because self-driving cars would need to make choices in the event of an impending crash. Should a self-driving car protect its passengers life at the expense of pedestrians on the road? Is there a balance of people and ages which tips this question in a particular direction? Also freshly in the news, an algorithm used for criminal sentencing raised the question of if rescinding access to race data was enough to make a suitably flexible model race blind (O'Neil, 2016b). Race was not explicitly hard-coded into the model, but could be reconstructed by a suitably flexible model which had access to enough features and a training set with judgements biased by human observation of race. In teaching, should teachers being judged by algorithms be able to assess the validity of the algorithm they were being judged by. In both cases, was there enough merit in not disclosing an algorithm's intellectual property to justify keeping the mechanisms hidden from the people whose lives were being affected. In this manner, I came across an article by MathBabe about BISG (O'Neil, 2016a), Bayesian Improved Surname Geocoding, which led to a longer article on the Los Angeles Times (Koren, 2016). Both articles describe BISG, an algorithm for guessing someone's ethnicity from information that is legally obtainable about that individual.

BISG dates back to the late 2000s. Marc Elliot, working for the Rand corpo-

ration, needed a better means of imputing missing race data for the sake of healthcare analysis. Previous standards for estimating race used an individual's place of residence or their last name, combined with the appropriate census statistics on residence or last name to estimate race. In 2009, Elliot published findings that showed a Naive Bayes combination of location and surname information significantly improves the estimate of ethnicity by "41% and 108%" over "single-source surname and address methods" (Koren, 2016). Because Elliot and his peers had access to an extensive healthcare network, they were able to demonstrate this with a health network database containing hundreds of thousands of self reported ethnicity statistics.

After BISG was published by the healthcare industry, it was picked up by the Consumer Finance Protection Bureau (CFPB) for doing analysis of the auto-lending industry. Within the auto-lending industry, it is illegal to make loans based on someone's ethnicity, and to that extent, it is illegal to ask for and obtain the ethnicity. The CFPB wanted to know if the auto-lending industry was discriminating based on race, despite that this was prohibited, and then needed to review auto-loans to see if minority loans were indeed given less favorable loan terms. Ironically, because race is not a valid assessment quality for loans, race data was not retained by the loan industry, and the CFPB had to devise a method to reverse engineer an estimate of ethnicity from other factors, last name and location of residence (CFPB, 2014). With this data, although highly politically contentious, the CFPB went on to use the BISG technique to show probable discrimination against minorities in the auto loan industry, and subsequently levied multi-million dollar fines based on that evidence.

Parallel to loans and healthcare, evictions and their effect on society have become a prevalent topic. Matthew Desmond published "Evicted" in 2016, which argues that evictions are highly destructive, and disproportionately punishing to women with children (Desmond, 2017). Not only can evictions cause homelessness, but they can

also induce poverty, creating a feedback loop which keeps people from lifting themselves up in society. Exploring the implications of eviction being its own feedback loop, the Harvard Access to Justice Lab, a legal clinic partially involved in using technology to expand access to the legal system, subsequently asked if minorities are disparately impacted in the housing market? First we need to unpack, what is disparate impact? Disparity comes in at least two legal forms: disparate treatment, and disparate impact. The former, disparate treatment, is taking an action which discriminates against a protected class; it is an active format of discrimination. Disparate impact, however, is more subtle. It is possible to take actions which, on the surface, do not portend to harm any section of the population, but still have unintended discriminatory effects. Disparate impact is this unintentional cause of harm to a protected class through an otherwise unbiased rule or requirement (Wikipedia, 2019a). This can subsequently be deemed illegal if there is not substantiated reason for the rules causing the discrimination.

An example of disparate impact could be given in the form of a new automotive factory in the 1960s which only hires applicants who have earned a PhD. Because 77 percent of PhD graduates in the 1960s were white men, this policy would disproportionately favor employment of white men even though the policy does not intentionally target any specific ethnicity. At this point, there is a question of if the imposed requirement is necessary to adequately fill the factory positions. If the the requirement is imposed because the jobs for the factory demand engineering knowledge at the level of a PhD, then this could be a legal form of disparate impact. If however, the job was for line workers who only needed to be physically fit and educated at a high school level, this would be unnecessarily selecting for a predominantly white male population of workers at the expense of other genders and ethnicities. This latter case would be an example of illegal disparate impact since an unnecessary

qualification would be causing the hiring bias.

While the example pertains to the disparate impact in the workplace, does disparate impact legally cover housing issues? If yes, it would be an extension of the Fair Housing Act, and this issue remained unsettled until this decade. As recently as 2012, there were cases making their way through the court system which raised the issue of disparate impact applying in housing matters, but the cases were all settled out of court. But, in 2015, a case did make it to the Supreme Court, Texas Department of Housing and Community Affairs v. Inclusive Communities Project, and by a 5-4 decision, the verdict was that the Fair Housing act did create cause for disparate impact to be applied to the housing sector (Wikipedia, 2019b).

Bringing BISG and disparate impact together, attorneys local to Boston could not ascertain for certain what was happening in housing court, but observing the defendants in court, it was believed that defendants were minorities more often than they should have been. The situation thus presented itself like in the auto-lending case. Racial data is not kept in the official court records, so any analysis based on the race of evicted individuals has to be reconstructed from other information. Despite that attorneys had a hunch, and despite that there was belief that local landlords used the public evictions records as justification for not renting to individuals, none of this had been proven.

Given the new standard for disparate impact, the immediate task is assessing if there is statistical proof of disparate impact. If so, the first condition for determining illegal disparate impact would be satisfied, and potential actions could be taken to ameliorate any harm being done.

# Chapter 2: Ethics

The first part of this analysis was getting approval for the data. Working with the Access to Justice Lab in Harvard Law School, we thought this part would be straight forward, but it turned out that we needed to get approval from an Institutional Review Board (IRB) committee. Because the subjects of the research are humans, it was deemed necessary to do a review. All human subjects research requires an IRB ethics review to safeguard against advancing research at a cost of abuse or harm to research subjects.

We applied for an exemption based on the grounds that the information from the courts was publicly accessible, the information from the census demographics was publicly available, and the surname data from the census was also publicly available. Nothing that we were using was data that was otherwise hidden or considered sensitive. As well, the methodology we wanted to use, the Bayesian improved surname geocoding (BISG), had been used before and the technique itself was publicly published. The difference between the previous work and the present work was that we would be using the BISG algorithm on housing records and not on medical or automotive data. That said, research has and can go afoul, our exemption request was initially declined, and we were required to apply for approval from the IRB committee through Harvard Law school.

Applying for approval first consisted of completing the IRB training. I was the only person who would gain access to a pre-downloaded archive of court data who

didn't have prior IRB training and approval, so that was the first immediate hurdle, IRB training through Citi Program (CITI Program, 2017). Training consisted of reading and understanding various modules on ethics as they relate to the study at hand. The first module, required for anyone seeking approval for working with human subjects, gives examples of famous violations of ethics in the past, and uses these to motivate the three principles we come back to repeatedly: Respect for persons, Beneficence, and Justice. In understanding each of these, we are given the definitions. Respect for Persons:

> "Respect for persons incorporates at least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection. The principle of respect for persons thus divides into two separate moral requirements: the requirement to acknowledge the autonomy and the requirement to protect those with diminished autonomy." (Department of Health, Education, and Welfare, 1979)

Elaborating on autonomy:

> "Autonomy means that people must be empowered to make decisions concerning their own actions and well-being. According to the principle of respect for persons, researchers must acknowledge the "considered opinions and choices" of research subjects. In other words, individuals must be given the choice whether to participate in research, and they must be provided sufficient information and possess the mental competence to make that choice." (Department of Health, Education, and Welfare, 1979)

Certainly it is the case that this research would be studying human subjects, and we would not be sending informed consent forms to everyone who had been

evicted in the state of Massachusetts. Nor would we be taking special consideration for those who had diminished autonomy, children or adults of limited autonomy. Considering this, we push on to the other two components that had to be considered in human subject research, beneficence and justice. Elaborating on Beneficence:

> "Persons are treated in an ethical manner not only by respecting their decisions and protecting them from harm, but also by making efforts to secure their well-being. Such treatment falls under the principle of beneficence. The term 'beneficence' is often understood to cover acts of kindness or charity that go beyond strict obligation. In this document, beneficence is understood in a stronger sense, as an obligation. Two general rules have been formulated as complementary expression of beneficent actions in this sense: (1) do not harm and (2) maximize possible benefits and minimize possible harms." (Department of Health, Education, and Welfare, 1979)

Here, we seemed to be on good footing. The potential use case of this research is to restrict access to eviction records if they are harming a protected class of citizens. The supposition going into this research is that landlords are using housing eviction records to not rent to minorities. Because it is hard to prove this and can not really be halted, the goal is to assess the harm being done preceding this claim with the initial eviction. If it can be shown that the original evictions are done in a manner which is biased against minority populations, legal action regarding disparate impact could allow for the sealing of eviction record (so they would no longer be publicly accessible, akin to divorce records which are deemed to not be of public interest, or juvenile records which are sealed to safeguard a protected class of citizens). Point being, the goal was to show harm already exists, and remove the harm from happening, both directly and indirectly by measuring potential bias, and barring future difficulty in finding housing. If the research reaches no conclusive results, no harm is done.

If it does reach conclusive results, either nothing or beneficence. The last major component in the IRB training, Justice:

> "Who ought to receive the benefits of research and bear its burdens? This is a question of justice, in the sense of "fairness in distribution" or "what is deserved." An injustice occurs when some benefit to which a person is entitled is denied without good reason or when some burden is imposed unduly." (Department of Health, Education, and Welfare, 1979)

Justice in this context is seen as equitable distribution of both benefit and burden. This is explained in the training through the context of medical trials and the like, where people may be sick and might be in the control or test populations of new treatments. If one group is untreated, or another group gets a poisonous medicine, there is a clear burden to be carried by a subset of the group. As might happen with medical research, if the test populations are not constructed judiciously, the benefits may only apply to Caucasian males, or children, or some other selective group, therefore inequitably spreading benefit.

All the above considerations apply, but most germane to the study in question is the amount of research time consumed. The needed materials are publicly available through census data, or Massachusetts housing court records. Acquiring the latter takes more time, but it is still free. One advisor raised the question of harm that could be done to eviction subjects. Would it be possible that this technique for unmasking ethnicity be made public and amount to a loss of privacy for the "test subjects". While there is this point of potential loss of privacy for these subjects, there is nothing to stop a motivated individual from using the techniques to do this already. Zip code level demographics are already trivial to acquire. Drilling down to neighborhoods with census tract data can be done on census.gov. Finding surname ethnic breakdowns is a bit more tedious than neighborhood level data, but still straightforward. Given that

the tools are readily available, there is no real loss of previously held privacy here. A curious note is that the the MA court records are not generally searchable in bulk. The court system is leery of research being done on incomplete or erroneous data. This does present an issue, but it only presents an issue of harm if the research results in eviction records being obscured from those who would use it. Arguing that this is a comparatively small population, and a small burden at that, particularly when credit reports could achieve similar assessments on small scale, the incurred burden appears to be minimal.

Further breakdown of the ethics review dials in to two points. The above points are all important, and presented to anyone who might be doing research on human subjects. What then exactly is the definition of "research" and "human subjects" as it applies in this academic setting? Research, as defined by the common law relevant to the IRB:

> "Systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge." (Office for Human Research Protections, 2018)

Contrasting with a description of things which are not research:

> "Scholarly and journalistic activities (for example, oral history, journalism, biography, literary criticism, legal research, and historical scholarship). Public health surveillance activities. Collection and analysis of information, biospecimens, or records by or for a criminal justice agency for activities authorized by law or court order or criminal investigative purposes. Authorized intelligence, homeland security, defense, or national security mission operational activities" (Office for Human Research Protections, 2018)

It appears that this project squarely fits under the heading of research. While there are legal records analyzed, plausibly for the sake of criminal justice, at the core, the goal is to understand the workings of the system through the use of statistical techniques. It is a systematic investigation of the Massachusetts housing eviction records, and it is intended to expand the understanding of how the housing system works relative to the underlying population.

Having qualified as research, does this research qualify as being on human subjects, as certainly the records are about things which have happened to humans:

> "Human subject" means a living individual about whom a researcher (whether professional or student) conducting research: Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens. (CITI Program, 2017)

This is essentially where the need for the IRB becomes moot. The research being done here isn't about human subjects. On the first point, there are no biospecimens. While there is information, it is not gained through interaction or intervention with the relevant individuals. On the second point, none of the information which is used is private information. It could be argued that we are generating or identifying private information (not biospecimens), but to that point, we only create probabilistic models of what a person's ethnicity is likely to be. There is no statement of what a person's ethnicity is, and through repeated sampling, any individual's ethnicity is likely to be estimated to be in multiple different categories, if not eventually every different possible category.

Ultimately, this position is one that was agreed upon by the Institutional Review Board committee. Upon the initial request with the IRB, they asserted that

this might be human subjects work and required everyone exposed to the data to have done an IRB training. But, after having reviewed the proposal and seeing that everyone had done the IRB training, the eventual ruling was that the work was not "human subjects" work, and did not actually need IRB approval (IRB17-1931). The research was free to progress.

# Chapter 3: Data Gathering

The principal pieces of data gathering focused on the census records for demographics, the census records for surname ethnicity frequencies, and ultimately the eviction records.

The surname ethnicity records were the easiest to deal with. Searching the Internet for census surname data will bring up hits on surname data from the last census. Doing this again now, the data for the one hundred-sixty thousand most common last names from the 2010 census came up in a .csv file. When I initially did this search, I had less luck and found the file as a Stata file. I do not use Stata and didn't have a working knowledge of how to deal with this file type; this was at first an irritation, but quickly overcome. Searching for how to open Stata files in python provided a stackexchange link and relevant code Cox (2013). Admittedly, now that I do not need this feature, and no longer use it, I also know that the Python Pandas library has a feature for reading Stata data, very simple and straightforward now when it was once more challenging.

The second part of data retrieval was eviction records. The law lab that I was working with had extensive access to eviction records they had recorded for another project, but I didn't have access to those records initially. I decided I could scrape some data myself from the Massachusetts court website. If you go to the website, it is an antiquated website, with nothing in the way of JavaScript loading aspects of the page after the fact. It was a prime target for traditional data scraping using

something like the beautifulsoup or requests library. Simply create a program to accept the login, find the relevant boxes to be filled and searched, and then parse the results for links to data pages to be stored.

All of this was straight forward enough, but the MA court website has a strict policy of not allowing data scraping. Their official stance is on the data served:

> The case information contained within this web site is generated from computerized records maintained by the Massachusetts Trial Court and is deemed to be public information. While every effort is made to assure the data is accurate and current, it must be accepted and used by the recipient with the understanding that no warranties, expressed or implied, concerning the accuracy, reliability or suitability of this data have been made. The Trial Court, and the developers of this web site assume no liability whatsoever associated with the use or misuse of the data contained herein. The case information from this web site is not the official record of the Trial Court. (Massachusetts Trial Court, 2019)

Access to the Massachusetts eAccess site by a site data harvester or any similar software intended to discover and extract data from a website through automated, repetitive querying for the purpose of collecting such data is expressly prohibited.

There are multiple relevant issues here. At the end, there is the assertion that it is against the will of the court to write software to scrape content from the page. Ethically, this ruled out two preferred scenarios. One, write an extension for chrome which would either fully automate the search and downloading of data, or at least write an extension which would automatically open, save, and close a window once a list of links had been searched for and displayed on a page. Two, writing a fully automated script to login and systematically download every case file by case number

or type of case and date was also disallowed.

In the end, for the sake of partial automation and following the letter of the intended use, I would search for the desired criteria, open the pages 10 to 15 at a time, open them, and save them. Running in the background, but not interacting with the site, I had a script running which would move the files to my desired local repository. Once the files were saved to the local repository, they were opened up via beautifulsoup. Sifting through the page's DOM (hierarchical HTML structure), the immediately relevant information was extracted for further processing. Originally this was the defendant name and address, but later included plaintiffs as well.

For this first round of data gathering, the need for data munging proved to be minimal to nonexistent. The Eastern housing court, especially within the last two years, keeps very clean data records. All addresses are complete. Street number, name, and suffix are stored in separate fields, as were city and zip code. Names were similarly consistently structured, "Last, First" with middle names or initials following the first name, but not being comma delimited.

All of the above consistencies made it very easy to scrape several hundred finalized eviction records from 2018. I could download the data, and Python would transfer and turn the data into a dataframe and then a .csv file for me. The hardest part of this, other than simply writing the code, was making sure I didn't exceed the window opening rate the website had, or lingering on a page for too long. Because there have been attempts to robotically scrape the website in the past, opening tabs was rate limited, and opening too many tabs/records too quickly would earn you a reminder that the website was not allowed to be robotically harvested. As well, if I did a search with many results, it would eventually time out if I spent too long spot checking results on other pages, again with the reminder not to robotically scrape the page. The final annoyance to circumvent here is that records searches would

only provide the first 100 or so results. Even if your search netted you 250 records (usually searching over a multi-day period), only the first piece of that search would be accessible. Consequently, I was working with 25 records per page, following the pattern of load several results, save several results, and repeat until having iterated through the first 100. Then, once arriving at the 100 record cutoff, run a new search from whenever the old records stopped displaying. It was a bit slow, but still fast enough to net a few hundred records.

Ultimately, this set the stage for later work, but the several hundred data points I downloaded were obviated by the much larger data dump I received from the law school after my IRB training and after the "human subjects" issue was resolved. The data set grew by more than two orders of magnitude, and I eventually had nearly two hundred thousand data points to start the assessment with. One minor change was required with the new data set. Managing the data was still html scraping at the heart of it, but details changed with switching from html only to web page complete files. A quick change to the DOM parsing and beautifulsoup code rapidly turned my new 800MB dump of files into a usable table of data.

The third piece of data gathering, and very related to the records processing, was geocoding. In one sense, turning someone's census tract into an estimation of their ethnicity was straightforward. Look at which census tract they live in. Retrieve the ethnic breakdown for that neighborhood. Assign those probabilities to a person. The difficulty with this comes from not knowing what census tract someone lives in. Using maps to find census tracts is completely infeasible for more than one or two people. It turns out you can go to census.gov and they have a tool for identifying census tracts, but this is also too unwieldy and slow. The preferred solution to this problem is a Python implementation of the census geocoder (Python Community, 2019). Taking the addresses parsed from the eviction records, those strings can be

stitched back together and identified by the censusgeodcode library.

The censusgeocode library proved to be an annoying bottleneck in processing. I started work on an a slow/archaic laptop. I thought the laptop was old and that was what was rate limiting, so I tried a significantly newer and faster desktop. To an extent, the old hardware had been part of the problem. Unfortunately the processing rates continued to be slow on the new device. This wasn't a very big issue when I was dealing with the small amount of data I had downloaded by hand. Geocoding took a couple hours. But when I had one hundred-sixty thousand data points, that ran considerably slower. Worse yet, if an address was a valid address, it took on average more than a second to find the census tract, but if an address couldn't be found, it took a few more additional seconds for the software to determine that there was no good return value. This meant that the first run through geocoding literally took days, and I lost about half of the data set to unfindable addresses.

After this first run through the data, and doing some cursory analysis, I decided it was time to do a better job cleaning the data and geocoding. This data cleaning ended up being one of the more tedious aspects of the project. The Mass Court system has admitted that they don't want the court database to be scraped and analyzed for the sake of errors and anomalies in the data, and these errors are definitely present. Of the first thousand or so eviction records, there are only 10 or so which can be salvaged. Of the rest, often there is a number and a street listed, or a number and a city, but frequently not the street number, name, and suffix, city and zip code. The city can be done without if the zip code exists, but without either, the data can not be geocoded. Likewise, without the street number, name, or suffix, the data can not be accurately geocoded and must be ignored/removed.

Additional to omissions in address data are actual pieces of information which the geocoder does not know how to handle. Addresses with letters attached to the

numbers will not process. Sequences of numbers (as sometimes eviction processings were for adjacent homes) would crash the system with the connecting hyphens. Basement or lower level denotations don't register as numbers. Municipalities within larger cities may not register. Simple fixes exist for most of these problems. Splitting the string which is the address will always give some form of a numeral in the first position. If there are multiple addresses concatenated together, it was always by hyphen or slash. Splitting an object containing a hyphen or slash and taking the preceding piece will give something that can be turned into a valid street number. Checking for just the numeric part, and tossing out any letter associated with an apartment or condo will give a number that the census geocoder can handle. While this may introduce a slight variation on the truth of where an eviction took place, it won't change the census tract.

The hardest part of an address to clean up is the street name given its high variability. Because there is an easy heuristic to process the street number, we started there. Then working from the most general piece of the address, the zip code, state, and city are the next easiest things to process. If the zip code exists, we keep it. If the zip code does not exist we look to a city designation. Checking the city is a bit trickier as we're looking for something other than a string of 5 numerals. This is still a somewhat straightforward problem to create a heuristic for. Searching the internet for valid city names in Massachusetts, Wikipedia had a convenient list of cities (Wikipedia, 2019). Without even needing to save the entire page, I could cut and paste the table, with associated values, and then edit out everything that wasn't a city with some light cleaning. Checking the list of cities in the data against the list of municipalities from Wikipedia, I could quickly ascertain that almost every city listed was a correctly entered functional city. The exceptions included one eviction which appeared to regard a city in Connecticut and a few evictions in locations which were

subordinate to larger municipalities (e.g. - East Arlington as opposed to Arlington). In the case of the city in Connecticut, that data was removed from the data set. In the case of subordinate municipalities, the East, West, or otherwise unneeded designation was removed. Other cases were found where data points had no city or zip code, and the street had been improperly marked as the city; with no way to tell which were in the state a particular street was, all of these data points had to be discarded.

After a quick processing of the state designation (all data in the set were listed as MA), the last thing to process was the street addresses. This ended up being a very manual process. Picking apart a street name came with a multitude of variations. Checking the two or three pieces of address I had left, and knowing this had to be street name or suffix, I started with the trailing piece, the suffix.

With high probability, for data points where I had more than two pieces of information to deal with, the last piece of information proved to be extraneous. Often times the person logging the court records had made a designation like "basement" or "4a" or "side alley". All of these things would be useful for finding a residence in the real world, but not useful for geocoding. I constructed a set of all the possible trailing pieces of information, and eliminated anything that was clearly not a street suffix. There were thousands of strings which needed to be evaluated as street suffixes, but when ordered alphabetically, it was very easy to quickly sort through the list picking out the tens of options which were clearly suffixes, or plausible false positives for street suffixes. Since the extraneous bits weren't useful, the non-suffixes were removed while still being able to keep the remaining aspects of that data point.

Having cleaned all the "easy" parts of the addresses, the last part of data cleaning addresses was dealing with the principle street names. Creating a set of these, and scrolling through them, the only obvious issue that remained was parenthetical notes which had been transcribed into the streets. Checking these manually, the

notes from the remaining data indicated duplicate addresses. When sometimes a building lived at the intersection of two streets, two addresses could be given. Because these locations were literally describing the same buildings, parenthetical notes were removed.

At the point where an attempt to clean all aspects of an address had been made, what remained was fed through the census geocoder. While not all constructs would be deemed valid addresses, it wasn't (at least initially) worth the effort to refine the data cleaning any further. If it took forty-four hours to have the census geocoder run through the data, any reruns through that data would waste otherwise valuable time. Alternatively, spending a minute or more to decipher and manually look up an address becomes an egregious use of time when you have more than one hundred thousand valid data points, and even if it's a mere one thousand data points that need looking into, at a one per minute rate, that's multiple work days spent solely looking up addresses, many of which will not be resolvable.

# Chapter 4: Code Implementation

How to implement code became a big issue. On the one hand, an underlying notion of Python is that you should not be spending your time implementing things that already exist in the language. If you need to do a list sorting, you can implement that yourself, but you're unlikely to have a better implementation unless you know a lot about the data you are working with, how it is structured, sorting in general, and how you might be able to find a better specific case algorithm to do the job. Because of this, I tended to err in the direction of using off the shelf tools. If I needed to lookup an address, I let the geocoder do what it was good at. If I needed to search a table, I assumed that pandas had the best query functionality for finding things in its tables. If I needed to resample a cohort of data, iterate through the rows multiple times. While this was sufficient for original analysis, it presented problems.

The geocoder, when processing the entire data set, ran on a time scale of days. At first I thought this was my old laptop. I timed address look-ups, and at one second, I had to set up a batch of computations and let it run through the night and over the weekend. As noted before, realizing this needed to be faster, I first tried running the computations on a faster machine hoping for an order of magnitude speedup. Sadly, the faster machine didn't translate into a faster lookup. Because I didn't need to clean the data often, this only went through a couple iterations. While doing the look-ups sequentially, I realized that many addresses occurred more than once in the data set. Not only were there frequently multiple tenants in a household, but often times

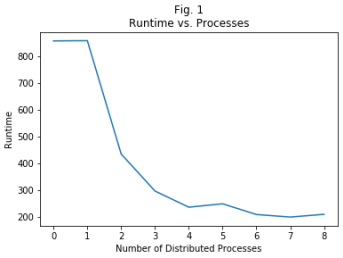
addresses had evictions more than once. Checking every search against a dictionary of prior searches and storing every search as it was made meant no search ever needed to be run more than once. Not only did duplicate addresses not need to be searched, but faulty addresses didn't need to be searched (note that false addresses usually take longer than correct addresses, presumably because there is more to search to verify that something doesn't exist, especially given that there is some ambiguity on how a term can be searched). This only resulted in a small speed up, about 30%, but 30% of multiple days is a noteworthy savings.

I did not realize this until later, but the real time savings on this would have been batch look-ups. While the census geocoder allows for individual look-ups, it also has a feature that allows for batch look-ups of size up to ten thousand addresses at a time. I tested the run time of this once I discovered it, and when combined with the time savings from eliminating the numerous repeat address, it is almost two orders of magnitude speed up, reducing multiple days of processing down to about two hours. With parallel processing, this could easily be reduced to under an hour to achieve a full one-hundred fold improvement on run time. This leads to the next topic.

Bootstrapping is the main technique I used in analyzing data. At its heart, bootstrapping as a statistical tool requires one to resample their data repeatedly to get an empirical representation of the data. When the technique was first invented in the late 1970s by Brad Efron, it wasn't wholeheartedly embraced because computing power was limited, and analytical mathematical approaches were more wieldy if the required assumptions were met (Wikipedia, 2019). Now almost two decades into the next millennium, repeated sampling is a much easier task for a personal computer to undertake.

For most of the tests that I ran, I repeated a simulation one thousand times to get a good picture of the situation and adequately prepare for hypothesis test-

ing. Although this amount of repetition is considered a comparatively low amount of resampling when high accuracy is needed, almost none of the confidence intervals I constructed ended up needing high accuracy (Chihara & Hesterberg, 2011). Because the data frames were all roughly the same size, several early simulations shared a common single iteration run time around 1-2 seconds. Running one thousand simulations meant the standard wait time for a test was in the ballpark of 15 to 20 minutes.




The obvious way to approach fixing this wait time is to use more computing power. At its most rudimentary, I could queue up multiple Python notebooks and give them each a quarter of the work. In a more sophisticated way, that's what I did with Ipyparallel. The software allows for multiple cores on a computer to be used for the same task, in parallel, while collecting the results in a central location for processing (IPython Development Team, 2018). Instead of tasking one core on my machine with running one thousand simulations, I can task each of four cores with running two hundred fifty simulations. Because the cores can run simultaneously, and each as is as fast as the other, the work takes roughly one quarter the time. Hypothetically, if I queue up $n$ cores on my computer, I can reduce the work time by a factor of $n$. In practice, the number of cores on a computer is limited, and in addition to running Python, the computer has others tasks to do. If Python calculations take away from vital tasks elsewhere on the computer, it can be expected that periodically the computer will interrupt Python to do work that needs to be done. Testing to see what that threshold was, I ran calculations on different numbers of cores to see where the point of steeply diminishing marginal returns was. It ended up being between four and eight as the former saw a factor of four improvement, while the latter saw a factor of improvement only marginally better than that, and I opted to use four cores

as my go-to cluster size for speed up.

The last and least glamorous major speed up comes from simple coding efficiency. As noted before, I generally assume that code implemented in packages is fairly efficient, and is not something that I work too hard at out-performing. Particularly with a library like Pandas which is widely used for data analysis, I assume that most operations are fast, and the bulk of time consumed is spent on the actual volume of work. The catch here is that function calls take time, and the act of calling the function generally takes more time. In the simulations I ran, I often needed to iterate over every row of a data frame, and perform a calculation for each row. The naive approach to doing this one thousand times is to run through every row of the data frame, do one calculation in each row, and then repeat the whole process one thousand times. This is very easy to think through and read through. But, the numpy library was doing the bulk of the computations for me within each row iteration, and yielded a means of improving performance by reducing function calls.

Before optimization, I would iterate through M rows of a data frame, do one unit of work creating a 1x6 array per row. At the end of every pass through the data frame, I would collapse the combined Mx6 dimensional array into a 1x6 dimensional array, and then repeat this process 1000 times until I effectively had a 1000x6 array, 1000 data points over 6 ethnicities. As a radical improvement on this methodology, in any simulation that required every row of a data frame to be used, I could do 1000 units of work in each row with one function call and create a 1000x6 array. After one pass through the data frame, I had an Mx1000x6 tensor which I could collapse into a 1000x6 array equivalent to that from the previous process. This eliminated almost a thousand method calls to the data frame, and more than a million calls to the numpy random multinomial function. Despite the added work with each call of the mutinomial function, this reduced run times from a hardware parallelized four

minutes to an efficient code expedited ten seconds.

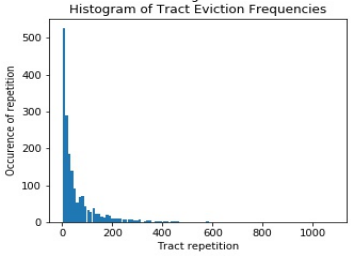
# Chapter 5: Data Analysis

What is the goal? We are trying to find out if there is a difference between the underlying populations of the locations, and the ethnicities of the people being evicted. Do any of the ethnicities get evicted at rates greater or less than their relative populations, and if so, which ethnicities, by how much, and is it statistically significant?

On our first pass through the data, as a default baseline, we assume that the underlying distribution of ethnicities of evictions should model after the inherent distribution of the census tract.

How are evictions spread out? Are the percentages persistent across districts, or are there some districts which have lots of evictions while others have just a few? It turns out that most districts have very few evictions, while others have a multitude. The first 100 tracts have no more than 2 evictions, while the tracts with the most evictions have somewhere between 200 and a little more than 1000 evictions over our time span.

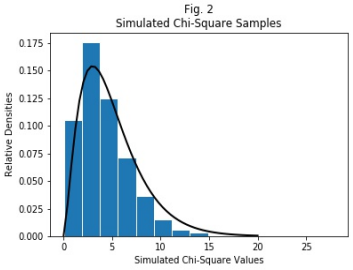
Fig. 1
Histogram of Tract Eviction Frequencies



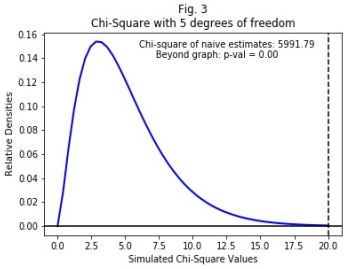

Returning to baseline analysis, the first test is to run through the data and ensure that our reasoning is sound and code is functioning as expected. Iterating through our data frame, we sum all of the baseline predictions for probabilities within each tract. Parallel to this, we create a multinomial sample for each tract specification.

Taking the square of the difference between the observed and expected samples, and then dividing by the expected counts, we expect to see a chi-square distribution with 5 degrees of freedom. Because we know how many samples we are taking, and because we have six categories, the last category is never a free variable.

Notice we get strong agreement between the simulated values we sampled and the actual chi-square distribution. Despite that there are thousands of underlying components to the simulation, they are functioning as a cohesive singular multinomial, and in turn realistically generating the appropriate empirical chi-square distribution.




Using this underlying distribution, we get a preliminary estimate on whether the observed distribution we are using to simulate with matches that of the chi-square. Summing the observed probabilities over the entire data set and creating the chi-square value, we get a chi-square value of 5990. Turning this into a p-value relative to the chi-square with 5 degrees of freedom, we get a p-value of 0, representing that 0 percent of the time we would expect to see observations as extreme or more extreme than we observed simply by random chance. Given that this is a very extreme observation, initially, we are inclined to think that there may be a difference between the underlying demographics and the evicted population.

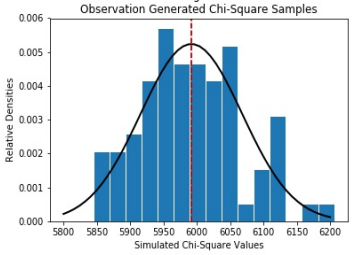



Rather than just accepting that the sum of the probabilities of individuals gives us a good estimate of the population and using the classical chi-square test, we also have the sampling distribution available to us. This next test is iterating over

our data set a hundred times, sampling from the naive estimate of ethnicity from each data point, summing the results, finding the corresponding chi-square value, and then repeating. These results should match with what we saw previously.

The overlaid distribution is a normal curve with the same mean and standard deviation as our bootstrap results. What we see is that the mean of the samples is very nearly the original chi-square value we found without sampling. We also get a range of values which show what might have happened with the same underlying ethnicities, given random chance. This gives a way to create error estimates on the original 5990 chi-square value. The 2.5% and 97.5% marks on the bootstrap samples give us a 95% confidence interval of [5870, 6170] on the chi-square estimate. This translates to a range of possible p-values, with the largest p-value being observed at the low end chi-square value of 5870, yielding a p-value orders of magnitude below any cut-off threshold, and functionally identically 0. Our bootstrap agrees with the classical test that our eviction data is probably fundamentally different from the underlying demographics.

Fig. 4
Observation Generated Chi-Square Samples




One possible issue with this analysis is that we're doing our analysis across all of the census tracts at the same time. Does the analysis demonstrably change if we were to do something like calculate chi-squares values within each census tract and sum for a total value. Because we would not be strictly calculating chi-square values, but rather sums of chi-square values if we iterate over every tract, we'll need to compare with something different. Parallel to the computation of chi-square sums with the observed probabilities, we'll also compute chi-square sums with the expected probabilities. Taking the difference, if the two sets of probabilities generate the same outcome, differences across resamples should be both positive and negative, and gen-

erally symmetric across the y-axis.

Instead of seeing something which is symmetric around the y-axis, what we see is something that is centered around about 15,400. Looking up and down two standard deviations, we see that all of our statistics fall far from the origin, and there is no expectation that this statistic is ever negative. Our 95% confidence interval spans from about 14,750 to a little under 16,500, and by no reasonable threshold is our expected difference near 0.

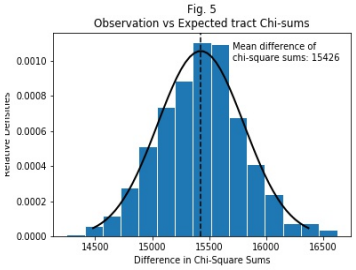
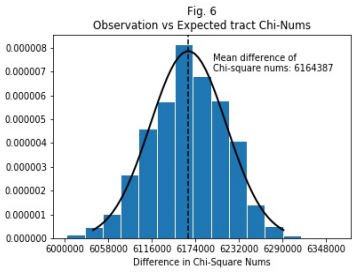


Fig. 5

Of note, summing the chi-square values over the individual tracts causes our test statistics to be significantly larger than they were when we created just one chi-square. Considering the difference between the two techniques, the former would be larger on account of observed counts being significantly different from expected counts (large numerators being squared). The latter, the sums of chi-squares of small districts, will be large because we eventually divide by expected counts, and expected counts can be very small if you have just one or two data points in a census tract. The problem is such an issue that there are times when there are no expected counts in calculating a chi-square value, and the calculation simply can't be done. To deal with this issue, I added a small .01 probability to the expected counts to make sure there was never division by zero. Unfortunately, the resultant division still magnifies any calculated numerator by a factor of 100.

Hoping to avoid the issue of numerical instability and dividing by near-zero numbers, I repeated the process, but created a new test statistic which was the sum of the squares of the difference between the observed and expected counts, all summed over the available tracts. The result is radically increased test statistics. Despite not dividing the chi-square numerators, there are enough large terms not being diminished

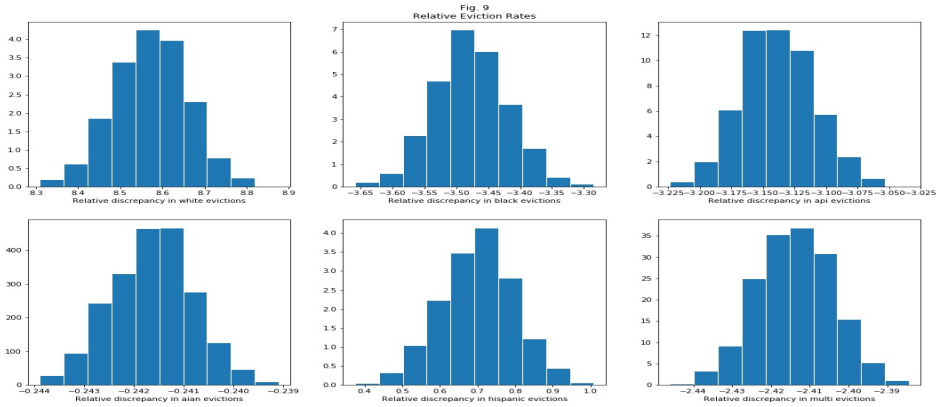
such that the test statistics are still exceedingly large. Regardless, performing the comparison of distributions in this manner, and eliminating the issue of numerical instability, we get the same conclusion as all the previous tests. The likelihood of the eviction data following the same distribution as the population is negligible.



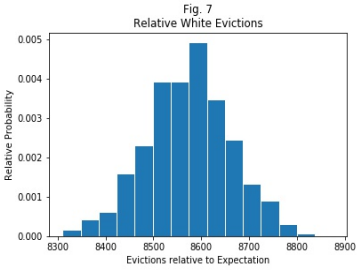

Here we have a mean around 6.16 million with a standard deviation of about 52 thousand. The associated 95% confidence interval runs from about 6.05 million to 6.25 million. We believe the two distributions are different.

Having tried resampling this data in several different ways, the answer is consistently that the function we are using for modeling the evictions is distinct from the function modeling the underlying demographics. This is something we wanted to confirm, but we're also interested in what this means in terms of the different ethnicities. It's completely plausible that both models stem from similar underlying predictions, just that one model might be higher variance than the other, yielding larger chi-square values. To solve this mystery, we'll use resampling and bootstrapping to estimate the balance of evictions relative to the expectation for each ethnicity.

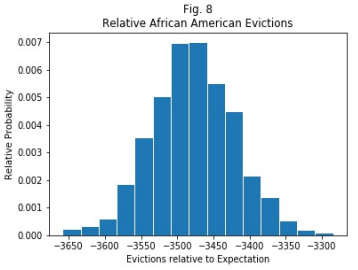
First method, to reduce computation time, we sample the straight probabilities, resampling across the entire data set. What should we expect to see? Given the underlying population, we should expect to see, out of every 111,000 evictions, 63,679 Caucasian evictions, 16,959 African American evictions, 1,880 Asian or Pacific islander evictions, 42 American Indian or Alaskan native evictions, 27,822 Hispanic evictions, and 685 multi -racial evictions. What we actually see is something far different.




Starting with Caucasian evictions, the model predicts that Caucasians will be evicted on average about 8570 times more often than they should given their presence in the population. The data suggests that, on average, this is an over-eviction rate of 13.45%, and the 95% confidence interval suggests that we should expect the eviction rate to be between 13.15% and 13.75% over what we expect for the demographics.




Looking at African American evictions, we see something different occurring. Relative to the underlying demographics, the model suggests that African Americans are under-evicted by 3480 for every 111,000 evictions. The data suggests that, on average, this is an under-eviction rate of 20.51%, and the 95% confidence interval suggests that we should expect the eviction rate to be between 19.8% and 21.1% under what we expect for the demographics.



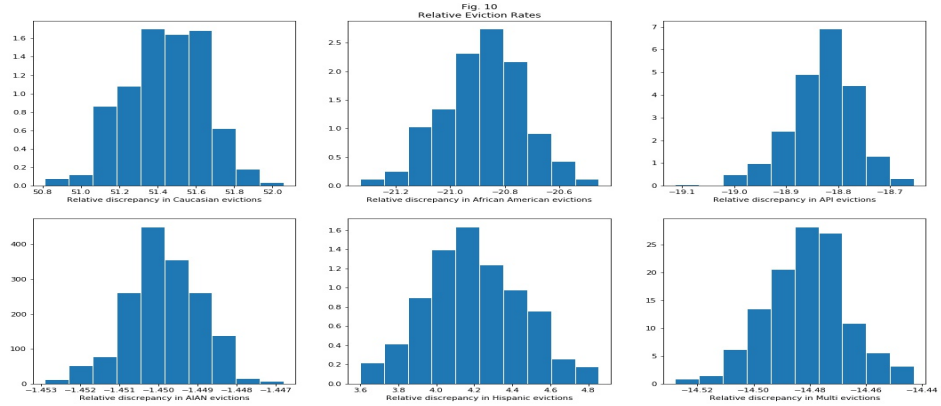

Looking more quickly at the remaining 4 classes: We see that Caucasian evictions

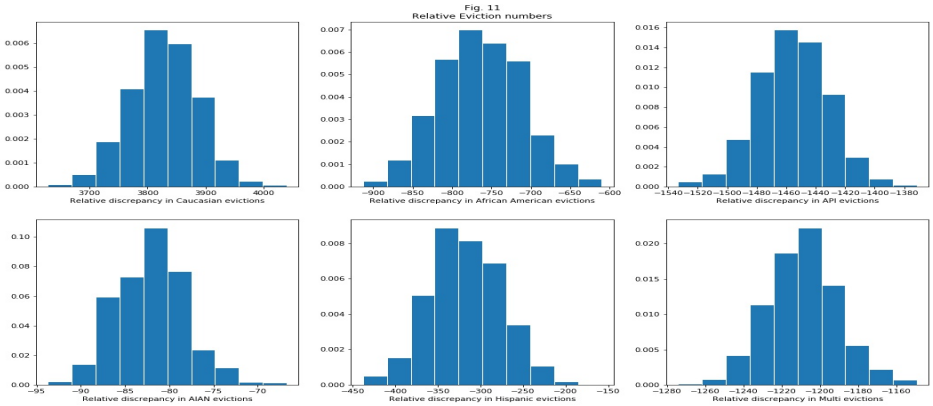
rates are well over their expected value. African American, Asian Pacific Islander, American Indian and Alaskan Native, and Multi race evictions are all firmly below zero and registering as under-evicted. Hispanic evictions are above expectation, but are not registering as different from zero by an egregious margin like all the other ethnicities. The mean for Hispanic evictions was about 700, with a 99% confidence interval ranging from about 460 to 935. Despite that this corresponds to an over-eviction rate between just 1.6% and 3.3% over expectation, this is still a statistically significant over-eviction.

Given that the expectation, from public defenders, was that most of these predictions would be the negative of what the model suggested, I thought it worthwhile to try running the same simulations, but this time enforcing that each tract be used in the resampling. Previously, the resampling allowed for resampling which might not represent all the tracts, especially where there had only been a few evictions, and possibly therein bias toward Caucasian evictions in a state which is predominantly Caucasian. Running the simulation while respecting which tracts the data come from yields the same results we observed above. Caucasians and Hispanics are evicted at a rate greater than than their underlying demographics, while the rest are under-evicted.

The next thing to consider is that maybe different ethnicities are biased in different parts of the state. We proceed to break down the the state into different pieces, Western, Central, and Eastern, roughly in line with the regions covered by the courts. We also further subdivide the Eastern region to look at just metro Boston.

Starting with the Eastern court: In the Eastern district, we see that Whites are over-evicted, and every other ethnicity is under-evicted. All of these results yield 99% confidence intervals well outside a range including zero. The estimated range on Hispanics gives the closest results to not being decisively under or over with a 99%

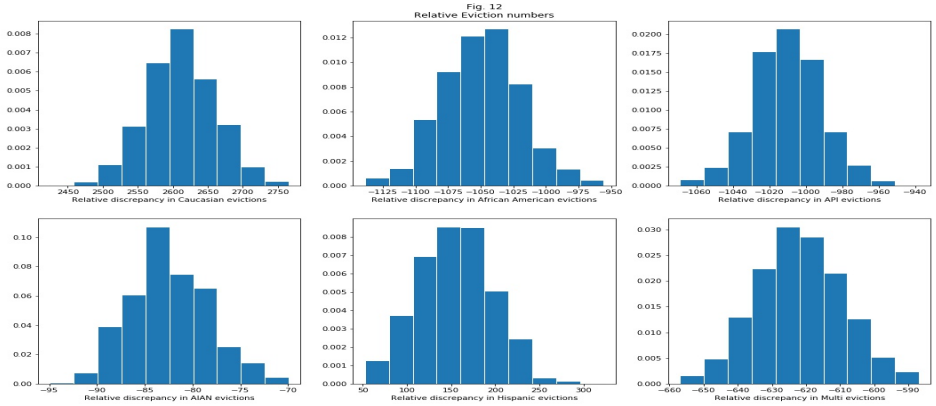


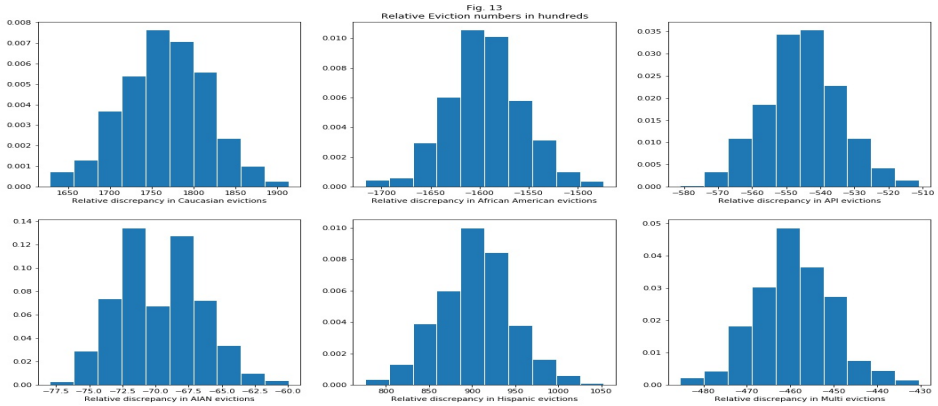




confidence interval ranging from 1.5% to .7% under the expected rates of evictions. Otherwise, all eviction rates are expected to be more polarized.

Next up is the Central court: here, Caucasian and Hispanic eviction counts suggest they are over-evicted. In this case, the Hispanic evictions 99% confidence interval range from .2% to .9% over-evicted. It is a statistically significant rate of over-eviction, although much less than the Caucasian rate of over-eviction closer to 4%. All other estimations are more than 4% below the expectation for the demographics.

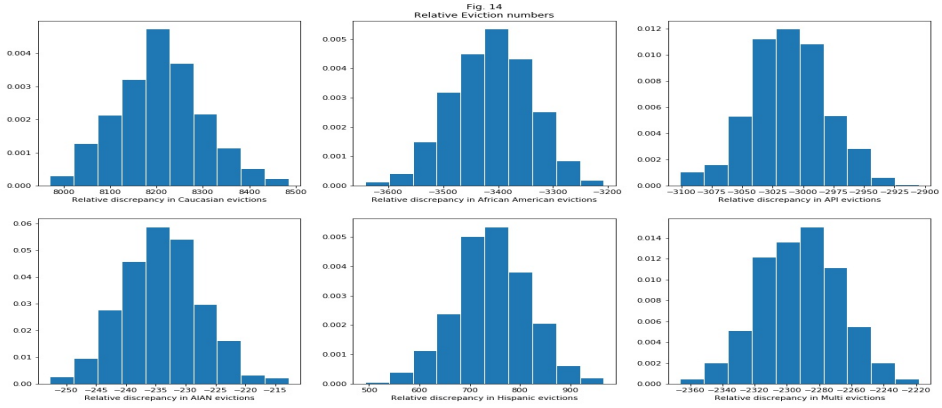
One more court, the Western court: here again we see Caucasian and Hispanic



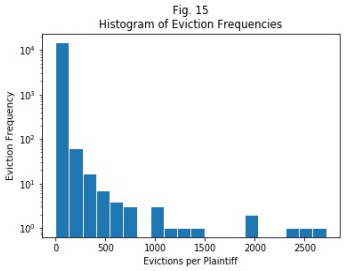




eviction rates higher than the local demographics. All probabilities testing whether these distributions could be aligned with the underlying demographics are again essentially zero as the 99% confidence lie completely on one side or the other of zero, not approaching zero from either side. Noteworthy in the Western courts is that the Hispanic over-eviction rate is estimated at 2.8% to 3.6%, higher than Caucasian estimated rate of 2.5% to 2.9%.

As a last analysis akin to the stratification by court, because I was told by a lawyer in Boston that evictions were overwhelmingly minorities, I ran the test on just a subset

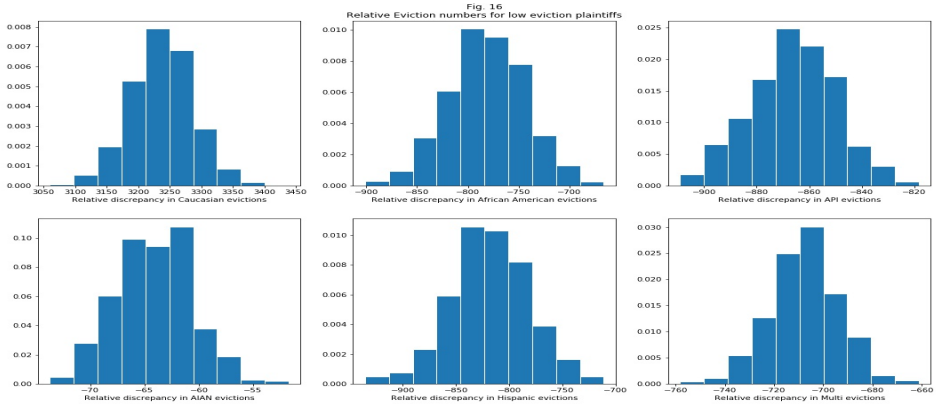


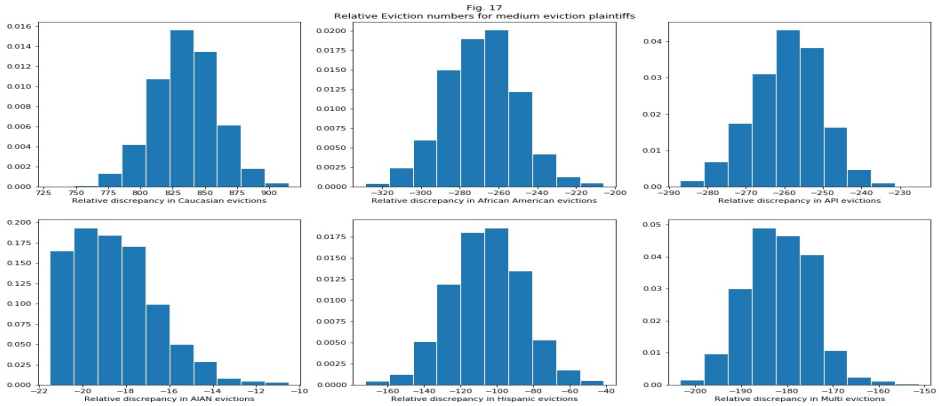

of the eastern court district which represents Metro Boston. As to be expected, the results followed very closely to the Eastern court results, with all confidence intervals being far from 0, and Caucasians and Hispanics were again showing over-eviction.




Having stratified by court system/geography, there is the question of whether there is bias in evictions based on plaintiffs. For this next analysis, I binned the plaintiffs by type. Some plaintiffs showed up in court for a mere 1 eviction. This was by far the norm. On the high end, there were plaintiffs who evicted more than 1000 residents. To see what the empirical distribution is, we draw a histogram of the plaintiff eviction rate counts, with a log scale on the y-axis. Two evictions is the 50th percentile of plaintiffs, and four evictions is the 75th percentile.

The next tests on this page and the next, show eviction rates for plaintiffs below the 90th percentile, between the 90th and 95th, between the 95th and 99th, and above the 99th percentile. The first two plaintiff classes show the enduring pattern that Caucasians should be over-evicted. However, the last two plaintiff classes show




that Caucasians and Hispanics, are over-evicted.

So far, the model has not been able to lend credence to the sentiment expressed by local trial lawyers. Wondering what effects I might not be accounting for, I realized that the underlying baseline being used for comparison is the ethnic breakdown of the local population. But, the local population and the local renting population are noticeably different. The average Caucasian rate of home ownership, as opposed to renting, is listed as 72.5%. In contrast, African American, Asian and Pacific Islanders, American and Alaskan natives, Hispanics, and Multi-race citizens have

Fig. 18
Relative Eviction numbers for medium-high eviction plaintiffs

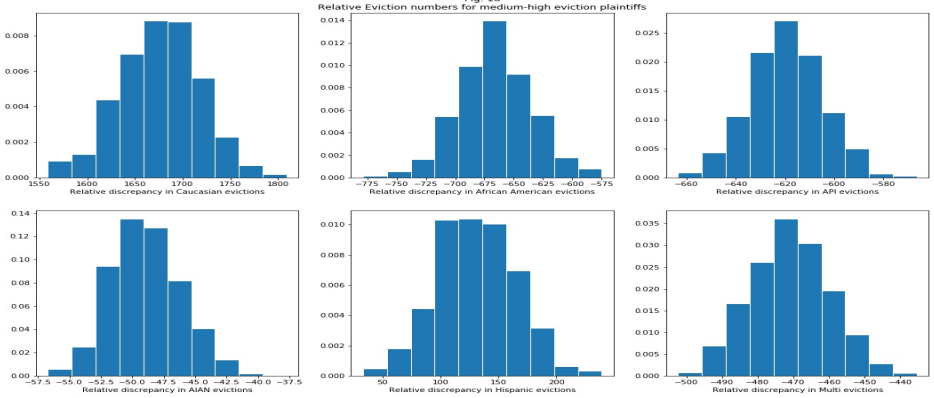
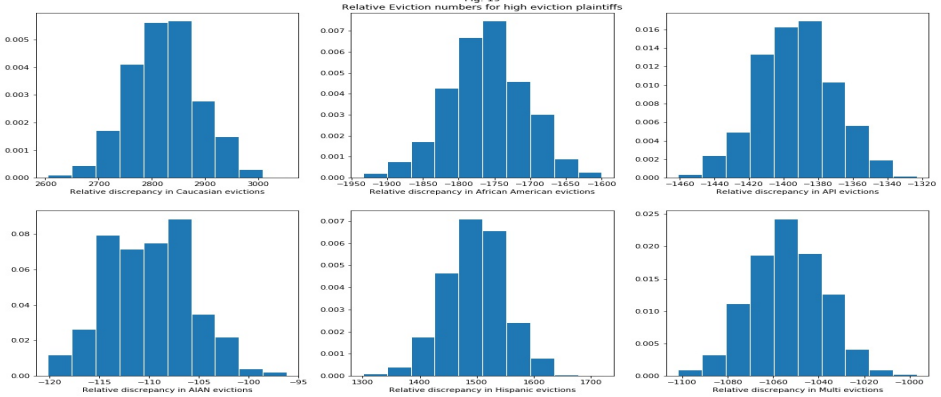


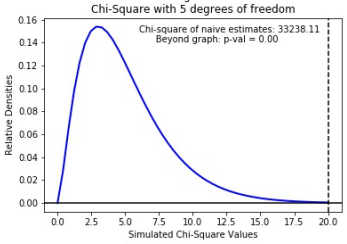

Fig. 19
Relative Eviction numbers for high eviction plaintiffs



average home ownership rates of 41.9%, 57%, 57%, 46%, and 57% (Carmel Ford, 2017). Having established that the different variations on our resampling all give demonstrably the same results, and because generating a sample from each data point is more interpretable, we'll restrict our reanalysis to use only the more routine tools.

Adjusting for the new rental population, we check that our observed data still disagrees with the underlying rental population. Because our chi-square statistics still match well with the corresponding distribution with 5 degrees of freedom, that serves as the basis for our p-value calculation. The p-value is close enough to zero that Python rounds the calculation to zero, and we believe that the the rental population and the eviction population are fundamentally different. The next question is, are revised rental ethnic proportions different from the eviction proportions, or again, are we witnessing a function that is markedly different for a reason such as its variance? We rerun the same simulations and highlight the trends.

Fig. 20
Chi-Square with 5 degrees of freedom




As in previous analysis, we see that Caucasians are modeled as being over-evicted relative to the rental population. All other ethnicities are under-evicted relative to their rental proportions. The results here were the same across all three major court regions and the specially sampled metro Boston region. Testing not by region but by plaintiff type, the results were again similar to what they had been in the original analysis. Either as above, Caucasians were over-evicted, or in the high eviction plaintiff stratification, Hispanics as well as Caucasians were both over-evicted. Hispanics are the only ethnicity with an alternating eviction frequency. Note that they are under-evicted with 99% confidence level 7.5% to 8.8% when looking at geographical regions, but are over-evicted by 1.1% to 1.6% if the data is divided by plaintiff class

and we look at high eviction plaintiffs.

At this point, a question was if the model could predict what was being anecdotally observed in the court. Because the goal is to determine difference from the underlying population, any formula which combines the location ethnicity probabilities with the surname ethnicity probabilities is inherently biasing surname data to be more like the location data. Wondering about this, I reran the tests using the surname probabilities without using a Naive Bayes update with the location data to see what would turn up. If this provided the same results as previously, I would think the model had already predicted everything that it might be capable of predicting. Somewhat expectedly, we once again see that the eviction population is registering as distinct from the underlying rental population. The chi-square is even greater, and the p-value is even smaller. The two distributions are even more distinct than in previous analysis. Given that the surname ethnicities are more dissimilar to the baseline geographic probabilities than any mixture of surname and geographic probabilities to the same baseline, this result was to be expected.

Measuring the relative frequency of evictions for each ethnicity, we see new behavior. In the Eastern Housing district, Caucasian, American Indian and Alaskan native, and multi racial citizens all experience statistically significant over-eviction. In the Central Housing Court, African American, American native, and multi-racial citizens experience statistically significant over-eviction. In the Western Housing Court, it is Caucasian, Asian and Pacific Islander, American Indian and Alaskan native, and multi-racial people who experience statistically significant over-eviction.

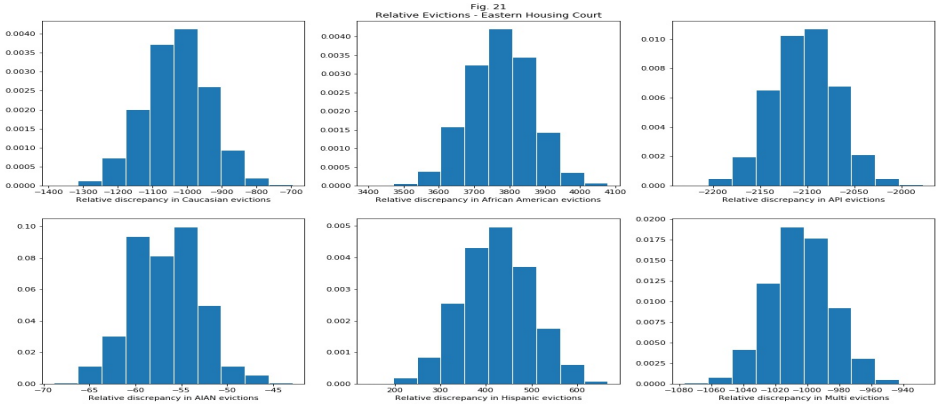
Having seen a change in the analysis, the question goes to what is the appropriate way of combining the underlying demographics of a region and what is gleaned from an individual's last name. In the surname ethnicity data, the census publishes the top 220 thousand last names, the associated probabilities, and then the number of

people with that particular last name. Looking at degenerate cases, if a last name was exceedingly rare, if only one person had that last name, when combining the last name and geographic information about an individual, it would make complete sense to use only the last name data probability estimates. In contrast, if there was a last name which everyone had, the last name information would become uninformative, and the location data would remain as the best and sole indicator of individual ethnicity.

As a final addendum to the above logic, surname predictions are known to be particularly strong indicators of ethnicity in Hispanic and Asian Pacific Islander populations. Rerunning the analysis, clustering these two populations together and comparing them with the remaining ethnicities, we see that with the surname ethnicities compared to the location ethnicities, even for the cases when surname should be by far a better indicator, we do not observe that minorities are disparately impacted.

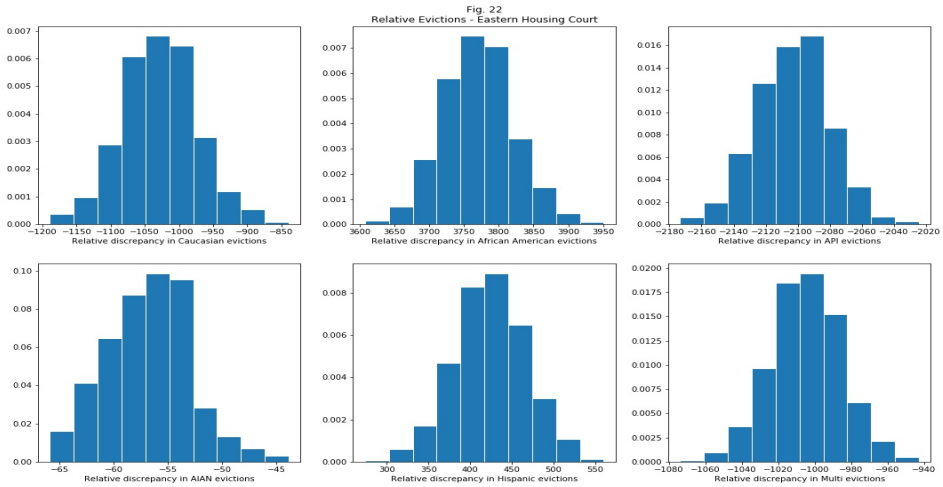
All of the sampling and bootstrap analysis above has been using expected counts for a census tract as the baseline to compare to. However, a plausibly more appropriate baseline to compare to factors in the relative population of an area. Sum the absolute counts of the different ethnicities in each of the census tracts for a region (such as the Eastern housing), and resample using the proportion given the absolute counts as our baseline. The assumption here is that within a zip code, town, or court district, eviction counts should vary randomly from census tract to census tract, but seeing more evictions in one tract than another is likely influenced by chance and population more than any systematic process of intentionally evicting people within a particular census tract. Therefore, we should rerun the preceding analyses and use the ratios of the absolute populations over a geographic region. We will skip the chi-square analysis, it again shows distinct difference between expectation and observation, and go straight to the bootstrap, repeating for court regions.

There are two ways to cut this analysis. One, assume that within a region, location




of eviction is random, and the data points in one tract may be representative of other data points in nearby tracts. Running this analysis, we see something very different from what we had seen up until this point. Looking at 99% confidence intervals, it looks like Caucasian, API, AIAN, and Multi-race individuals are under-evicted. Balancing that out, African Americans and Hispanics are over-evicted with confidence intervals spreading from 37.5% to 42.3% for African Americans, and 3.2% to 8.9% for Hispanics.

The second way to cut this, is to assume that within any tract, the evictions in that tract are particular to that specific location. In this case, we resample from empirical distribution, and eviction frequencies are maintained within a tract. In this case, even with the stricter sampling method, we see almost identical results. Most ethnicities are under-evicted, with the two exceptions of African Americans, who are evicted at a rate within a 99 percent confidence interval of 38.4% to 41.2%, and Hispanics who are evicted at a rate within a 99 percent confidence interval of 4.4% to 7.6%.




Looking at both of the other major court districts, a similar outcome is observed; some combination of African Americans and Hispanics are over-evicted. In the Central court district, Hispanics are predicted to be over-evicted between 69% and 73.5% of the time. In the Western court district, African Americans are predicted to be over-evicted between 8.1% and 14.7%, with Hispanics being over-evicted a very high 110% to 113% more often than expected given local rental populations.

Redoing the analysis of plaintiff categories, we see similar trends to the analysis by court region. Testing an empirical sampling distribution, or running a bootstrap, across three of the four plaintiff categories, we see consistent over-eviction of Caucasian, African American, and Hispanic populations. In the other category, the medium eviction rate, we see Caucasian and Hispanic over-eviction. Worth noting, in this analysis, the Hispanic eviction rates are statistically significant, but not at the exceedingly low thresholds we've been observing. The average Hispanic over-eviction rate is about 3.5%, but the 99% confidence intervals range from slightly above or below, depending on the run, 0% to about 6.6%.

As a final analysis, we test the eviction rates across the entire state and all court districts. Once again, the empirical sampling within tracts and the bootstrapping across tracts yield similar results. African Americans and Hispanics are both over-evicted. African Americans are evicted at a rate about 24% more often than would be expected, and Hispanics are evicted at a rate about 52% more often than would be expected. In contrast, Caucasians were under-evicted by 12%, Asian and Pacific Islanders by 68%, American Indians and Alaskan natives by 80%, and Multi-racial individuals by 75%.

# Chapter 6: Statistical Methods

The three principle methods used in this paper are BISG ethnicity calculations, chi-square tests, and boot-strap and resampling tests.

BISG: Bayesian Improved Surname Geocoding. BISG is an algorithm which takes an estimate of person's ethnicity based on their last name, an estimate of the person's ethnicity based on their place of residence, and uses the two pieces of information together to come up with a Naive Bayes improved third estimate of the person's ethnicity based on the assumption of independence of the two pieces of known information (Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2013). Using the Naive Bayes method, you assume the pairwise probabilities are independent, multiply the probabilities, and then normalize the results so they sum to one.

In more detail, to get a first approximation of the ethnicity of someone you have never met, you can use their address. Using their address, look up the ethnic breakdown of the neighborhood they live in. Census data is freely available down to the level of census tracts; census tracts range in size from about 1200 to 8000 people, so it's a fairly local approximation which might encompass as little as a few blocks or a large apartment complex.

While location data can give a very good first approximation of someone's ethnicity, particularly if the census tract heavily features a particular ethnicity (or just minorities in general if you are not distinguishing further than white or non-white), depending on the census tract, location data might just as likely do a poor job dis-

tinguishing between different ethnicities. A census tract that is half white gives you no better than a 50% chance of correctly guessing if a resident in the census tract is a minority. As an alternative to using location data, last names serve as plausible means of approximating someone's ethnicity. Usefully, last name data is even more accessible than location data, and the census keeps a table of the 150 thousand most prevalent last names and their associated ethnic probabilities. Given this second approximation of ethnicity based on last name, we multiply the corresponding ethnicity probabilities together, normalize the numbers so they again sum to 1, and use this new estimate as an improvement on the original estimations.

As an example, we can try to ascertain the ethnicity of someone with the last name of Williams living at 1000 Massachusetts Ave in Cambridge, MA. Using the address, 65% of people in this census tract identified as white, 4.5% as African American, 0% American Indian or Alaskan Native, 22% Asian or Pacific Islander, 5.3% Hispanic, and 1.1% identified as being of more than one race. Note that these probabilities do not sum to one. Not all recorded data had race reported, and non-reported data functions as the discrepancy between observed race and 100% tabulation. (We could normalize these figure so they sum to one and are valid probabilities, but we will wait to make this correction during the normalizing phase at the end of the computations).

Next we check the census statistics on people with a last name of "Williams". The given surname constitutes .56% of the U.S. population. Of that fraction of the population, 48.5% of people are white, 46.7% are black, .37% are Asian or Pacific Islander, .78% are Alaskan natives or American Indians, 2% identify as belonging to two or more races, and 1.6% identify as Hispanic.

Looking at either piece of information by itself, we are unlikely to be able to resolve the ethnicity of our random person. The location data suggests that the individual is white, but does not convincingly rule out Asian or Pacific Islander. In contrast, the

Table 1: Naive Bayes Example

| Ethnicity | Cauc. | Afr. Amer. | API | AIAN | Hispanic | Multi | |
|---|---|---|---|---|---|---|---|
| Location | .6553 | .0454 | .0000 | .2285 | .0530 | .0113 | |
| Surname | .4852 | .4672 | .0078 | .0037 | .0160 | .0201 | |
| Joint | .3179 | .0212 | .0000 | .0008 | .0008 | .0002 | .3411 |
| Posterior | .9321 | .0622 | .0000 | .0024 | .0024 | .0006 | |

surname data suggests that the individual might be white or African American, but does sufficiently delineate the two options.

The next step now is to combine these two pieces of information. Multiplying the corresponding probabilities of ethnicity based on location and surname (in the first two rows of the table above), we calculate what is known as the joint probability. These computations need to sum to 1, so we will scale them by their combined sum called the marginal probability (the .3411 at the end of the row is the sum of all the values preceding in the row for joint probabilities). Dividing each ethnicity entry in the row of joint probabilities by .3411, we arrive at the posterior estimate of ethnicity probabilities. In the above example, we initially had trouble discerning if our unknown individual was white or African American, or white or Asian Pacific Islander, but after combining the two pieces of information we have strong indication that the mystery individual is in fact white. Given that our data only gives us a probabilistic profile of a person, this is often not a reliable way of predicting a single individual's ethnicity, but it is a reliable way to predict the ethnicity of many individuals, assuming individual errors are smoothed out in aggregate, while improving on historical techniques using just last name or home address information.

The second prominent test featured in the analysis is the chi-square test. The chi-square test provides a means of testing how likely a data set was generated from a specific from a potential source, taking into account random fluctuations. In tabulating evictions for a census tract, the underlying expectation might be that for every

Table 2: Chi-Square Example

| Ethnicity | Cauc. | Afr. Amer. | API | AIAN | Hispanic | Multi | |
|---|---|---|---|---|---|---|---|
| Expected | 50 | 20 | 7 | 8 | 10 | 5 | |
| Observed | 55 | 23 | 7 | 5 | 10 | 0 | |
| (Exp. - Obs.)$^2$ | 25 | 9 | 0 | 9 | 0 | 25 | |
| Diff.$^2$ / Exp. | .5 | .45 | 0 | 1.125 | 0 | 5 | 7.075 |

100 people who are evicted, we expect 60 people to be white, 20 to be black, 5 API, 3 AIAN, 10 Hispanic, and 2 multi-racial individuals. Instead, we might observe 55 white evictions, 23 black evictions, 7 API, 5 AIAN, 10 Hispanic, and no multi-racial evictions. Certainly our observations were different from our expectations, but we need a means of quantifying how different, for which we use the chi-square statistic.

A chi-square statistic is calculated as the square of the difference of the observed and the expected counts, divided by the corresponding expected counts, and then summed.

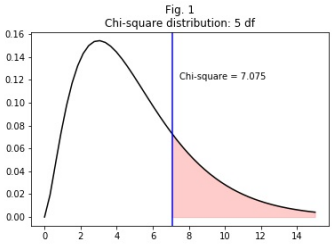
The first row is the data we expected. In this case, our data sums to 100, but it could sum to larger or smaller numbers. For reasons of numerical stability, none of the expected counts can be 0, and it is recommended in the literature that the expected not be less than 5, but otherwise, expected counts are unrestricted.

The second row of the table is what we actually observe. The observed counts should sum to the same value as our expected counts (100 in our case), but it need not be the case that all of the observed data be non-zero or even as large as 5.

Some of our observations are greater than the expectation, while others are less than the expectation. In the third row, taking the square of these differences, we are partially measuring the magnitudes of the differences between observation and expectation, and not allowing for some differences being positive and some being negative to zero out our measurement of deviation.

The fourth row of the table scales the third row by the value of the expectation.

When we have larger initial expectations, it is reasonable to expect larger fluctuations, but the squared differences are in turn diminished more by the greater initial expectation, putting everything on a similar scale.




With our statistic in hand, 7.075 in this case, how do we make sense of what it means? This statistic is a chi-square statistic, and is intended to be compared against a chi-square distribution. To do so, we first need to calculate the number of degrees of freedom in creating our distribution. As we saw above, our expected counts were taken out of 100. Similarly our observations were also taken out of 100. In this case, knowing 5 of the ethnic categories was enough to deduce what the 6th category was. Because of this, we had 5 degrees of freedom (had the number of observations not been fixed at 100, we would have had an additional degree of freedom as we would need to know all six counts to calculate our statistic).

Looking at the chi-square distribution with 5 degrees of freedom, we see the distribution itself in black. If our observations had exactly matched the expected counts, all of the differences would have been zero, and the chi-square statistic would also be zero. As the observations deviate more from the true expectation, the squared differences get larger, and so does the test statistic. What the distribution above shows is that the most likely event to occur is a chi-square statistic of about 3, some deviation from expectation, but not a lot. The further to the right of the graph, the less likely those statistics are to be observed. The measure of an observation's likelihood is the sum of all the probability densities representing outcomes more extreme than what we observed, the shaded red area. Summing that area of probabilities, in this case, we get a probability of about .215 of seeing an observation as extreme or more extreme than the one we saw. If 21% of situations deviate more from our expectation than

the one we saw, with a 10% threshold or less, you would not reject the hypothesis that our observation was a random observation from our expected distribution; our observation is a result that is sufficiently likely to occur by chance.

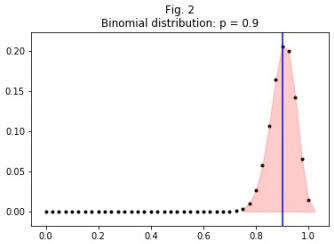
Bootstrap analysis. The fundamental idea behind the bootstrap is that we will use our observed data as a substitute for the actual unknown distribution, and then use a sampling distribution of a desired parameter to create an estimate of the behavior of that parameter. This aspect is known as the "plug-in principle." In trying to create a sampling distribution for parameters, we will plug in our observed data for the original distribution(Efron & Tibshirani, 1993). After assuming that our data mimics properties of the underlying distribution, we begin to create new samples, of equal size to the original data set, with replacement, from our observations. Just as repeated sampling from the underlying distribution would show variation in test statistics, we anticipate similar variation in sampling from the observed data. From the variation, we calculate percentiles on the resampled test statistics and use those for our confidence intervals. The 2.5 percentile and 97.5 percentile of the resampled statistics serve as the bounds for the 95% confidence interval for a statistic (and similarly for confidence intervals with greater or lesser degrees of certainty). Because we are explicitly looking for the variation in calculating our test statistics, this brings us to the second principle of the bootstrap, Monte Carlo simulation. In order for the bootstrap to work, we must use sampling with replacement(Efron & Tibshirani, 1993). Just as sampled values can repeat from the underlying distribution, they must also be able to repeat in the resampled distribution. If we didn't use sampling with replacement, there wouldn't be any variance in the test statistic, and we wouldn't be able to create confidence intervals or otherwise gain insight from our data.

Restrictions of the bootstrap. As noted by Chihara and Hesterberg, "1000 bootstrap samples are enough for rough approximations, but more are needed for greater

accuracy." (Chihara & Hesterberg, 2011) They go on to note that while 200 to 1000 bootstrap samples were recommended by Efron and Tibshirani when they used the bootstrap in 1993, in a more modern era of computing, on the order of 10,000 to 15,000 bootstrap samples are desired if you want the 95% confidence intervals with 2.5% percent tails to be accurate to within .25% (Chihara & Hesterberg, 2011). Fortunately, in our case, for any intervals we construct, the observations in question are so far beyond even the 1% or .1% confidence levels that we can safely get by with 1000 samples.

What does all this extra work get us; why not use classical methods? In the classical method, avoiding all the extra computation, we could assume that the underlying evictions numbers for each category are roughly normal over a period of time, find the sample standard deviation, and then calculate a p-value based different observed samples. There are a few problems with this. The first problem is that our distribution may not be normal, and in fact it is not. As well, the distribution we are concerned with may not be symmetric, which again, it is not. So, using a normal approximation may give a reasonable approximation, but ideally we would use something better.

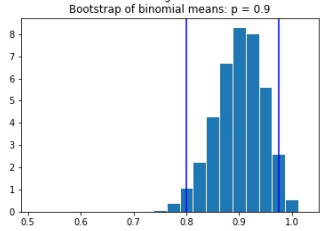
As a better approximation of what our eviction rates look like, we can use a binomial model if we want a 1 vs. other or 1 vs. many categorization. Or we can use a multinomial model to deal with multiple classifications at the same time, although still the estimate of average predictions classically comes with a normal approximation to the associated confidence intervals. To the right, we can see that given data with a mean of .9, the 95% confidence interval for where that mean might have fallen is symmetric around our observation, .9. This is usually reasonable, except it predicts that an average eviction rate in this hypothetical case may be greater



than 100%.

Looking through the literature, there are numerous ways to correct for this problem. Several different tests deal with this issue, but possibly the simplest is to use the bootstrap. In this case, the sample was 40 points, with 90% of them being "successful trials." We can re-sample from those points, take means of the resamples, and then construct the bootstrap confidence interval. Seen in figure 3, you can observe that the sample means are not symmetric around .9, although that is still the mode of the sample means. However, the greatest observed mean is 1, so no predictions are too high. The lowest 2.5% mark is .8, with the 97.5% mark being .975. The range of sample means are not symmetric around the mode, and this is reflected in the 95% bootstrap confidence interval.

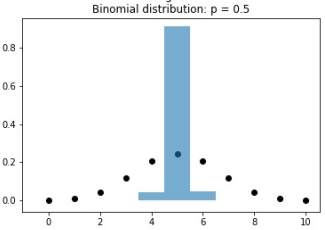
Fig. 3
Bootstrap of binomial means: p = 0.9




As noted previously, this type of result could be obtained with more sophisticated confidence interval construction methods, but solving this problem leads us to a second property. Our data only have estimates of probabilities and not actual counts. What do we do if our data predicts success with .6 probability, and failure with .4 probability? If we look at the majority probability in each case to determine counts, what should look like a binomial distribution with probability .6 of success suddenly looks like a discrete uniform distribution with probability 1.

Alternatively, if we accumulate all the probabilities and construct a single multinomial from the summed probabilities (likely rounded to the nearest integer), this also fails to capture the behavior of the data. As an example, if 5 people are predicted white with probability .99, and black with probability .01, while 5 more are

Fig. 4
Binomial distribution: p = 0.5



predicted white with probability .01, and black with probability of .99, in actuality, we expect 5 white and 5 black data points with almost no variance. If we sum probabilities and use a binomial distribution, we end up with the same expected counts, but a much higher variance of the distribution. Looking at the visual, the conglomerate multinomial, or binomial in this case, has values and probabilities represented by the the black dots. In contrast, resampling from the probabilistic representation of individuals yields a distribution with much tighter variance, and is representative of the actual beliefs.

In the different possible approaches to handling this data, we see that the classical assumptions break down. Even in trying to force the data to behave in a way that fits into the classical analyses, the techniques break in one way or another. Fortunately, the bootstrap does not share these issues, and because it only requires several minutes of computational power, we opted for bootstrap resampling through the predominance of our analysis.

# Chapter 7: Conclusion

Given the BISG assessment of race, it appears that Hispanics and African Americans are significantly over-evicted relative to the Caucasians, Asian and Pacific Islanders, American Indians and Alaskan natives, and Multi-racials. While original analysis suggested that Caucasians were over-evicted relative to the other ethnic categories, this analysis treated all census tracts equally despite that some census tracts are much larger than others. Accommodating for the disparity in census tract sizes, the estimation of over-eviction shifted sharply against African Americans and Hispanics.

Problems that arose: how to accurately assess the rental population? Without rental statistics to transform census tract ethnicity breakdown to the relative census tract rental ethnicity breakdown, the national rates of renting had to be used as a proxy. It is possible that the analysis changes once a more detailed analysis can be done on a tract by tract basis.

The underlying idea of BISG is that it should be used as a Naive Bayes approach assuming that location and surname are independent for a given population. This may be a reasonable approximation for people receiving healthcare, or for people applying for auto loans, but it isn't clear how accurate this is for people being evicted from their housing.

BISG was intended to improve data imputation as part of a larger analysis of data sets with gaps. Overestimation of a race in one place, and underestimation in another

may very well cancel out in the larger analysis because the larger analysis does not depend on local trends.

In contrast, with eviction differentiation, we are directly comparing BISG approximations with the underlying demographic; we are assessing how much the naive Bayes approximation changes the estimation of ethnicity relative to the local demographic. Given this, the model would benefit from being refined further. The census surname database gives race proportions for the common last names, but it also gives the absolute proportions of the last names. There is a compelling reason to think that exceedingly rare last names will play a larger role in estimates of individual ethnicities than the location data. Similarly, exceedingly common surnames should contribute less to an ethnic profile while location data becomes comparatively more important.

Anyone looking to advance this eviction analysis in the future could spend time in the court surveying how independent the two data components actually are. Even though the goal here could be described as avoiding time and money intensive data collection in person in favor of mathematical modeling, a small sampling of data to further develop the relationship between features in the model could present a large time and cost savings in the long run if it convincingly examines the relationship between location, surname, and surname frequency.

As a final note, anyone who seeks to use this analysis for demonstrating disparate impact should note that the model may be able to show a statistically significant distinction between how different ethnicities are effected, but does not provide a mechanism. The root cause as to why different ethnicities are evicted at different rates is unknown.

# References


Carmel Ford (2017). Home ownership by race and ethnicity. http://eyeonhousing.org/2017/12/homeownership-by-race-and-ethnicity/.

CFPB (2014). Using publicly available information to proxy for unidentified race and ethnicity. *CFPB*.

Chihara, L. & Hesterberg, T. (2011). *Mathematical Statistics with Resampling and R*. Wiley.

CITI Program (2017). Social & Behavioral Research Investigators. https://www.citiprogram.org.

Cox, N. (2013). Pandas read stata with large dta files. https://stackoverflow.com/questions/19744527/pandas-read-stata-with-large-dta-files.

Department of Health, Education, and Welfare (1979). The belmont report. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.htmlg.

Desmond, M. (2017). *Evicted: poverty and profit in the American City*. Broadway Books.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

IPython Development Team (2018). Using ipython for parallel computing. https://ipyparallel.readthedocs.io/en/latest/.

Koren, J. R. (2016). Feds use rand formula to spot discrimination. the gop calls it junk science. https://www.latimes.com/business/la-fi-rand-elliott-20160824-snap-story.html.

Massachusetts Trial Court (2019). https://www.masscourts.org/eservices/home.page.2.

Office for Human Research Protections (2018). 45 cfr 46.102 the federal regulations - sbe. https://www.ecfr.gov/cgi-bin/retrieveECFR?gp=&SID=83cd09e1c0f5c6937cd9d7513160fc3f&pitd=20180719&n=pt45.1.46&r=PART&ty=HTML.

O'Neil, C. (2016a). Bisg methodology. https://mathbabe.org/2016/08/29/bisg-methodology/.

O'Neil, C. (2016b). *Weapons of Math Destruction: how big data increases inequality and threatens democracy*. Crown.

Python Community (2019). censusgeocode. https://pypi.org/project/censusgeocode/.

Trevor Hastie, Robert Tibshirani, Jerome Friedman (2013). *The Elements of Statistical Learning*. Springer Series in Statistics.

Wikipedia (2019). Bootstrapping. https://en.wikipedia.org/wiki/Bootstrapping_(statistics).

Wikipedia (2019a). Disparate impact. https://en.wikipedia.org/wiki/Disparate_impact.

Wikipedia (2019b). Fair housing act. https://en.wikipedia.org/wiki/Fair_Housing_Act.

Wikipedia (2019). Municipalities in Massachusetts. https://en.wikipedia.org/wiki/List_of_municipalities_in_Massachusetts.