Final Project: Racial Economic Inequality in the US during the Covid-19 Pandemic

Leuven Wang

April 16, 2021

Abstract

This study investigates the extent to which the COVID-19 pandemic has disproportionately affected black communities in the United States, an important issue that exists at the intersection of racism, economic inequality, and healthcare. The CDC provides over 21 million records of COVID-19 cases in the US, including individual's race and county of residence. Pairing counties with their median household income data from the US Department of Agriculture, we looking at the income of those who have died as well as the racial demographics of those fatalities. Our investigation uses linear regression, maximum likelihood estimators, bootstrap confidence intervals, mean hypothesis tests, goodness of fit tests, and Bayesian credible intervals. We conclude with clashing results that indicate that whilst the COVID-19 pandemic could have impacted different races disproportionately, problems in data collection and suppression make it extremely hard to reach an incontrovertible conclusion. The importance of stringent data collection methods for ongoing pandemics is highlighted.

Introduction

Racial economic inequality in the United States of America hardly needs an introduction. In 2016, the Federal Reserve reported that the median wealth for non-retired African-American households was a tenth of white households (Hanks, A. & Solomon, D. & Weller, C.E. ,2018). In the world's wealthiest country, 1 in 3 black children live in poverty (Lee, H., Esposito, M., Edwards, F., Chun, Y. & Grinstein-Weiss, M. , 2020). The proportion of African American households with a negative net worth is greater than 4 times the proportion of white households that have a negative net worth (Hanks, A. & Solomon, D. & Weller, C.E. , 2018). The problem is chronic and persistent. Not only does it prevent minority communities from achieving the economic power and opportunities that they are entitled to as Americans, but it also feeds into the self-perpetuating cycle of poverty, thereby leading to significant social problems in crime, education, and healthcare, among others areas. The drastic disparity in racial economic status has only been accentuated during the COVID-19 pandemic. The Centers for Disease Control (CDC) concedes that minority groups are at increased risk of contracting COVID-19 and dying from it (CDC, 2021). We can attribute this in part to the lack of healthcare resources accessible by these communities. This can be due to both their personal financial situations and inadequate civic funds in their environment.

This investigation will seek to investigate how the COVID-19 pandemic has impacted black communities disproportionately using data of individual cases from the CDC that records race. We will attempt to determine what proportion of COVID-19 deaths are from the black community and much this differs from their proportion of population. We will also analyze the distribution of income levels of all those who died, based upon their county of residence and see where the average income of a black victim of COVID-19 would fit. We hypothesize that black people will account disproportionately more for COVID-19 deaths and that the average income of a black person who died from COVID-19 would be substantially lower than the average income of all fatal victims of COVID-19.

Data

There are two main sources of data we are using to approach this investigation. The first one comes from the CDC and is a record of over 21 million cases of COVID-19 in the US (retrieved April 16 2021). The dataset offers a place to record each individual's race, gender, age, county of residence, and their status of mortality (whether they are dead or alive). The dataset is updated monthly and at the time of retrieval, contained cases from January 2020 to March 2021. Each record to the dataset is completed by a medical professional who fills out a CDC form for a case and sends it to the CDC. This means that many professionals working for many organizations in many jurisdictions across the US have contributed to the dataset. This means that the dataset only records COVID-19 patients that have sought medical attention and been clinically diagnosed. It does not record the millions of COVID-19 infections that have not reported their condition to any medical authorities or sought help. This is important as we can expect the majority of those infections which have not been reported are uncritical or asymptomatic cases. Given that current estimates project that the US has had 30 million cases so far, the sheer volume of uncritical cases not reported may make mortality rates for this virus seem worse than they really are.

The CDC claims the dataset is undergoing continuous rigorous quality assurance procedures. This involves removing records that contain illogical dates etc... However, some of the data has undergone deliberate suppression to prevent the identification of individual cases. This includes removing the race, location, and date of individual cases where their combination is extremely infrequent. This impacts rural records more given their small population and may subsequently affect our investigation by precluding us from understanding how minority communities fare in those areas. If we make the very general assumption that rural areas are generally less medically equipped, economically bountiful, and have greater social stigmas, then we can expect the results of our investigation to fail to account for the disproportionate suffering of minority groups in those areas. To clean this data, we removed all cases where we did not definitively if the patient had died from the disease or is still alive.

The second piece of data is from the US Department of Agriculture's Economic Research Service. It is a comprehensive list of counties in the US alongside their median household income for 2019. This data is calculated by the US Census Bureau's Small Area Income and Poverty Estimate (SAIPE) program based on estimates from samples collected during the American Community Survey completed each year (US Census Bureau, 2020). It is complete and needed no further cleaning,

From these two datasets, we derived two datasets of our own. The first dataset calculates the mortality rate for over 500 counties across the US. We only used counties that had recorded cases that match our criteria. This means cases where we definitely know the mortality status of the patient and their county of residence. Using each county's FIPS code, a government standard unique identifier, we lined up each county with its corresponding median household income.

Different reporting procedures between jurisdictions and the additional data suppression done by the CDC means that we have achieved unrealistic mortality rates - in some cases, they indicate a 100% fatality rate for those infected. This may be because local authorities report fatal cases exclusively or with less stringency. To mitigate these unrealistic figures, we are limiting our dataset to only include counties with a mortality rate of less than 5%. This is well above present estimates by experts on COVID-19 mortality rates in the US (Ioannidis, 2020) so it should not remove the figures that are more realistic and accurate.

The second dataset records the race and income of all the cases where the victim is known to have died from COVID-19. We cleaned the data so that it only includes cases where we know their race and county of residence. Again using each county's FIPS code, we matched each patient's county of residence with that county's median household income and from thereon we treated that as an indication of their personal level of income. This dataset contains almost 200,000 records.

To reiterate the final central variables of our investigation:

Race: The race of the individual in the record. It can be American Indian/Alaska Native, Asian, Black, Multiple/Other, Native Hawaiian/Other Pacific Islander, or White. We are not recording any cases where we do not know the race of the person in question. For simplicity, we will not count multiple/others races as being black.

Mortality Rate: The number of deaths from COVID-19 in each county divided by the total number of COVID-19 cases in the same county where we know whether the individual is alive or not. Only including counties where there are cases where we know whether the individual is alive or not.

2019 Median Household Income (MHI): The median household annual income of each county in the year 2019, measured in US dollars. Also used as a general estimation for each fatality's income level, based upon what county they lived in.

Summary	**Measurement**
Mean 2019 MHI of COVID-19 Fatalities (USD)	70835
Standard Deviation of 2019 MHI of COVID-19 Fatalities (USD)	16103
Mean Mortality Rate across Counties	0.0158825
Standard Deviation of Mortality Rate across Counties	0.0127028

Fig. 1: Table of Numerical Summaries of Sample Data

Figure 1 is a table showing some numerical summaries of the sample data we have obtained. It lists the center and spread of the MHI of COVID-19 fatalities. This gives us a general understanding of COVID-19 fatalities and what income level the average victim was in. Similarly, the table also shows the mean and standard deviation of mortality rates across counties. This reveals a general idea of how lethal this virus is once contracted and how much we can expect mortality to differ between counties of different wealth.

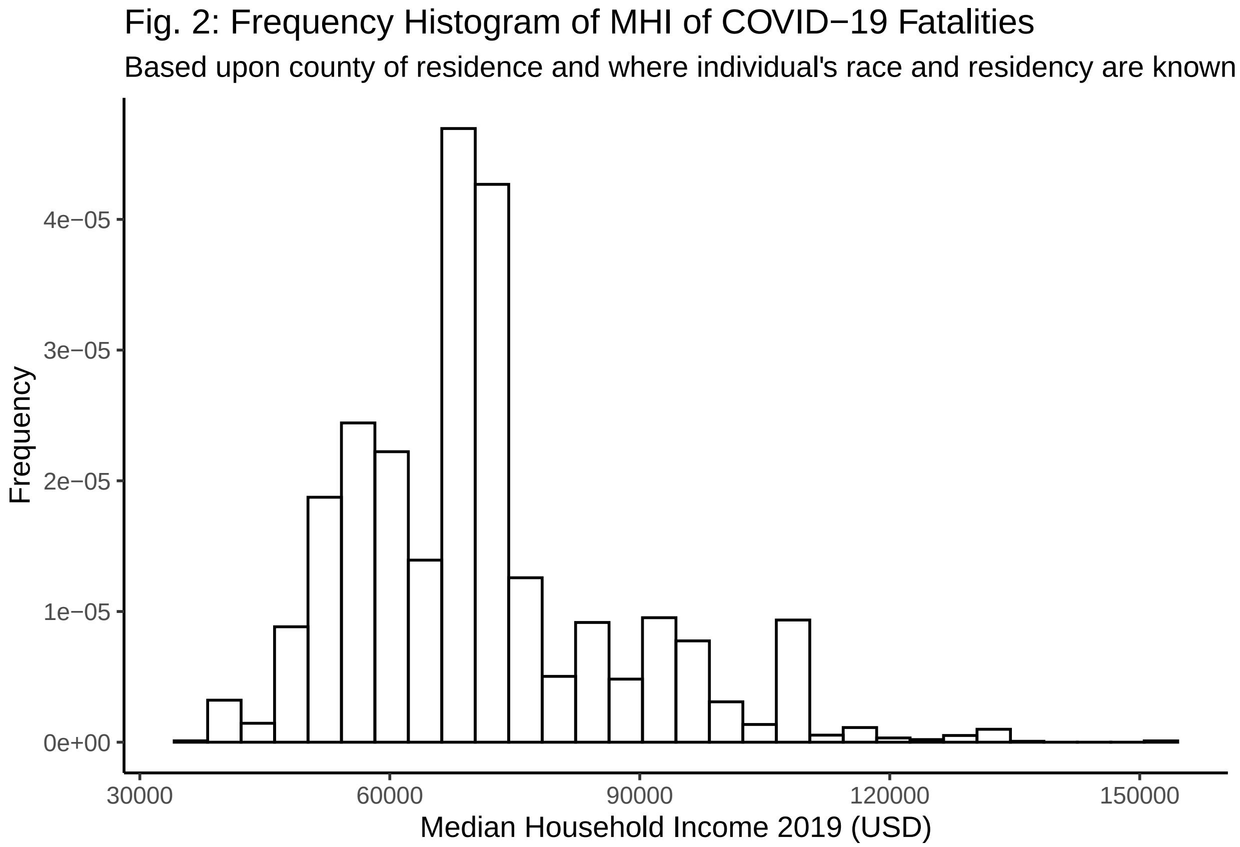


Figure 2 is a density histogram showing the MHI of known COVID-19 fatalities. As can be seen, COVID-19 affects both the rich and poor although death is more frequent at the lower end of the spectrum. This matches our general assumption that the poor would have less access to medical resources or a healthy environment and thus succumb to the disease more frequently.

All analysis for this report was programmed using `R version 4.0.4`.

Methods

We will be using six different methods to study the relationship between racial economic inequality and the effects of the COVID-19 pandemic. We will first conduct a linear regression of mortality rates in each county and its median income. We will also study maximum likelihood functions that model the distribution of incomes from all deaths and set a Bayesian credible interval for the mean income of all COVID-19 deaths. We will conduct a mean hypothesis test to speculate if the mean income from all black COVID-19 fatalities and all general fatalities are equal. We will use bootstrapping to approximate the proportion of COVID-19 deaths that are from black individuals and perform a goodness of fit test to investigate if COVID-19 deaths across all races are proportionate to the makeup of the population.

Linear Regression

A linear regression aims to formulate the relationship between an independent and a dependent variable in a linear equation of the form:

$$Y_i = \alpha + \beta x_i + U_i$$

where Y_i is the dependent variable and x_i is the independent variable. Our linear regression will look at the relationship between COVID-19 mortality rates in US counties and their corresponding 2019 MHI. Our dependent variable is the mortality rate in each county whilst our independent variable is their MHI. Here, α, is the linear model's point of intercept with the y-axis - where the mortality rate of a county would be if it had an MHI of 0 USD. Of course, this is not a realistic scenario but it nonetheless provides a hypothetical starting point for our model. β is the gradient of the line modeling the relationship and its orientation and magnitude will determine whether the variables are positively or negatively correlated and how strong their correlation is. U_i is an independent random variable with a normal distribution of mean 0. It is meant to account for the data's random fluctuations and why they appear on either side of the line of best fit and not always exactly on it.

Maximum Likelihood Estimator

We will attempt to find the maximum likelihood estimator for the mean of the distribution of incomes of individual COVID-19 fatalities. This will hopefully allow us to better understand the extent to which COVID-19 affects different economic classes. Our data is a record of the incomes of a large number of individuals who have died from COVID-19. Our methodology is to fit a distribution to the data and find the optimal parameters that best represent the data. This will be done by first calculating a likelihood function from an assumption on the general type of distribution that can model our data. A likelihood function tells us the odds to which sample data supports particular values for a parameter. We will then use calculus to find the maximum of this likelihood function. The accuracy of our final result ultimately depends on whether the chosen distribution accurately represents the distribution of income and whether the sample data accurately represents the population.

We will assume that our data is a sample of random independent Poisson variables with rate λ. We can make this assumption because income is something that can be counted. Whilst smaller denominations may cause some studies to treat money as a continuous variable, this is merely a question of individual perspective. Since we are limiting our denomination to cents, and the amount of money one makes in a year can indeed be treated as "events" (we base salaries and wages on periods of time), so it is reasonable to consider income as a discrete Poisson variable. We are using the MLE to estimate λ. After derivation, we have found that the MLE for λ is $\bar{x}$, the sample mean.

All derivations regarding the MLE can be found in section 1 of the Appendix.

Hypothesis Test

The mean hypothesis test aims to calculate the probability that the mean of a distribution is a particular value. For this investigation, we will look at the distribution of income of black fatalities of COVID-19 and ask if its mean, μ, is the same as the mean income of the general sample of COVID-19 fatalities. To do this, we will take the mean of our sample data of incomes of general fatal COVID-19 cases, k, and create a null hypothesis that is $H_0 : \mu = k$ and an alternate hypothesis, $H_A : \mu \neq k$. We will assume that the random variable, which is the standardized version of μ, follows a standard normal distribution $N(0, 1)$. This means creating a test statistic:

$$\frac{(\bar{X} - \mu)}{s/\sqrt{n}} \overset{app}{\sim} N(0, 1)$$

where $\bar{X}$ is the sample mean (of the incomes of all black fatalities), s is the standard deviation of the sample, and n is the number of black fatalities in our sample. We can do this because our sample data contains over 20,000 records of the income of black COVID-19 fatalities, making n a very large number. Using our data and assuming the value of our null hypothesis, we will calculate the value of the test statistic and the subsequent probability of achieving it, or a value more extreme, in the $N(0, 1)$ distribution. This is our p-value. The smaller it is, the more considerable evidence there is against the null hypothesis. We will use the pre-specified value of $\alpha = 0.05$ to decide if we should reject or keep the null hypothesis. If the p-value is smaller than α we will reject the null hypothesis, suggesting that the mean income of black fatalities of COVID-19 is different from the mean income of general COVID-19 fatalities. If the p-value is larger than α, we will keep the null hypothesis open for future consideration. This method assumes that the original sample is representative of the entire population.

Bayesian Credible Interval

We want to create a Bayesian Credible Interval containing the mean income from all COVID-19 fatalities. This involves building a posterior distribution of mean incomes which relies centrally on Bayes' Theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Here, $P(\theta|X)$ represents the posterior distribution - a distribution of the mean after accounting for the sample data. $P(X|\theta)$ is the likelihood function. $P(X)$ is the probability of the data given all potential values for the parameter of interest, θ. Finally, $P(\theta)$ represents the prior distribution - how we think θ might be distributed, given no sample data. The parameter we are interested in investigating here is the mean income of all COVID-19 fatalities, μ.

In this portion of our investigation, we are considering the distribution of incomes of COVID-19 fatalities. As we have explained before in the section on maximum likelihood estimation, we consider this data to be distributed as a random discrete variable following a rough Poisson distribution. From this, we will derive a likelihood.

When choosing a prior distribution for μ, we like to factor in that the disease is more likely to be lethal for poorer individuals. Thus, the mean is more likely to be skewed towards the left on the income scale, making a Gamma distribution fitting for its distribution. Let us suppose that μ is most likely to be around 68000 USD and that its prior distribution is $Gamma(6, \frac{1}{12000})$.

Using the prior, likelihood, and sample data, we will build a posterior distribution for μ. Then we will take its 2.5th and 97.5th percentile as the boundaries for its 95% credible interval. We chose a 95% credible interval because it will most likely include the true parameter value, whilst discarding the most extreme values. By the Bayesian framework, we can thus say that there is a 95% probability that μ will be within the interval.

This method depends on many things including how representative the sample data is of the population, the appropriateness of our prior distribution and how we view the distribution of incomes.

All derivations regarding the posterior distribution can be found in Section 2 of the Appendix.

Confidence Interval

We will be attempting to find a 95% confidence interval (CI) for the location of the true proportion of black COVID-19 deaths. A CI is a continuous range of values that we can ascertain, up to a certain degree of confidence, contains the true parameter we are searching for. We will set this confidence interval through bootstrapping - a method by which we resample observations, with replacement, from the original data sample. The original data sample records the race of a large number of individuals known to have died from COVID-19. After building our bootstrap sample, we will calculate the proportion of black deaths. We will repeat this for 1000 bootstrap samples and order them all in a distribution. Then we will pick the 2.5th and 97.5th percentile of that distribution to find the values that mark the boundaries of our 95% CI. This is an empirical bootstrap since we are resampling observations directly from the original data. We have chosen a 95% CI because it offers a strong chance of containing the true parameter whilst at the same time discards the most extreme results. This means that if we were to repeat this experiment numerously, around 95% of our CIs will contain the true parameter values. We are 95% confident that the true proportion of COVID-19 deaths attributable to black people will be within our CI. This method assumes that the original sample is representative of the entire population.

Goodness of Fit Test

A goodness of fit test is similar to a mean hypothesis test. However, instead of considering a hypothesis based on one parameter, it studies how well a model distribution fits a set of observations. The null hypothesis thus becomes “the data fits the distribution” and the alternative hypothesis turns into “the data does not fit the distribution”. For this goodness of fit test, we will consider the null hypothesis that COVID-19 deaths are proportionate to the makeup of the general population.

We will be using a Chi-squared test to conduct this goodness of fit test. Firstly, we will calculate from the sample data the proportion each race accounts for in COVID-19 death. Then we will calculate both the likelihood of these sample proportions, $L(p_0)$, and the likelihood of the percentages of the population that make up each race $L(\hat{p})$. This is because if deaths are proportionate, then the two should be similar. We can then construct a random variable such that it has a chi-square distribution like

$$-2log(\frac{L(p_0)}{L(\hat{p})}) \sim \chi^2_{\alpha-1}$$

where α is the number of racial categories there are. Calculating the probability of achieving the calculated statistic or a value more extreme, under this distribution, we obtain a p-value from which we can make a judgment as to whether to reject the null hypothesis just as we did for the mean hypothesis test.

To reinforce the results of our finding, we will also be conducting a simulation of this test. This means we will generate a large number of datasets of individual races the same size as our original sample, but this time, the racial proportions will be set according to the general population estimates, as our null hypothesis states. For each of these datasets, we will calculate the proportion that each race occupies. Then we will record how many of these datasets generate proportions that are as extreme as the proportions we have observed in our sample data. Extremity is measured by the difference between the sample data proportions and the population estimate proportions. It includes both ends of the spectrum - a race can be disproportionately under or over-represented in COVID-19 deaths. This method assumes that the original sample is representative of the entire population.

Results

The results of our investigation are many and varied. Some of them meet our hypotheses whilst others suggest conclusions in the complete opposite direction. We have presented our results here such that the first four deal with the relationship between racial income levels of fatalities whilst the remaining two results at the end focuses on the racial proportions of fatalities.

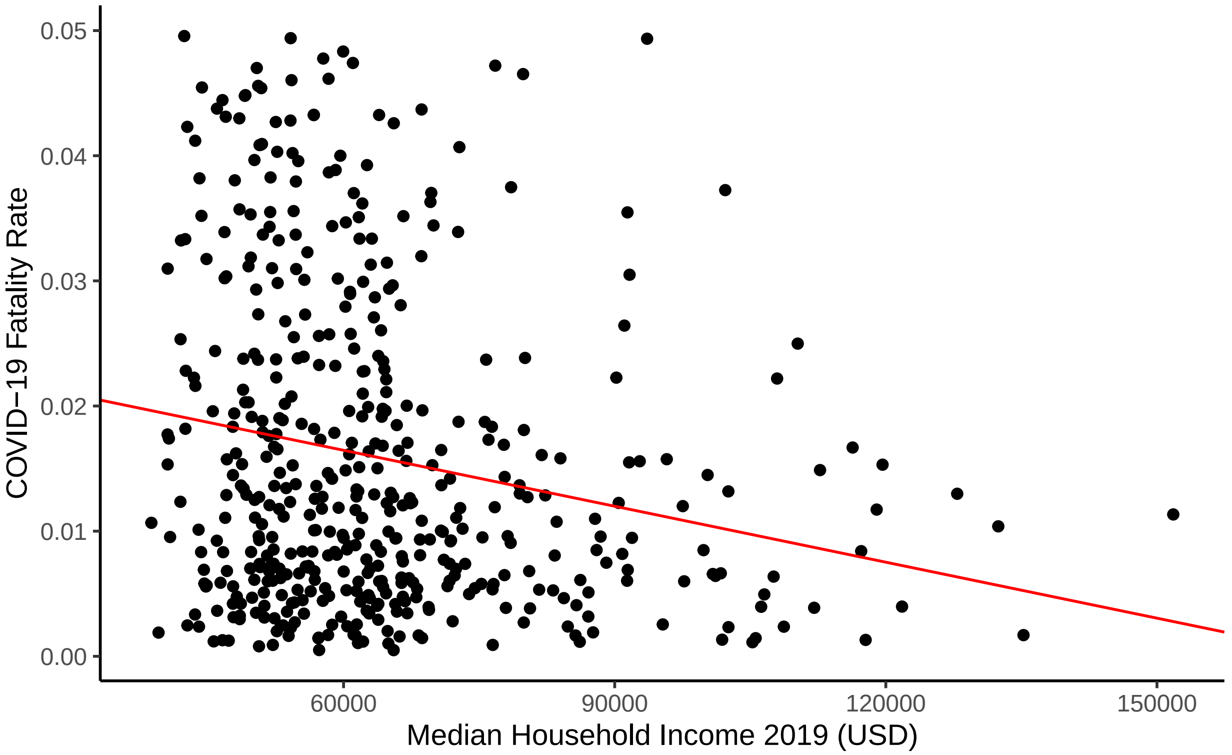
Linear regression.

$\hat{\alpha}$	0.025408
$\hat{\beta}$	$-1.4906698 \times 10^{-7}$

Fig. 3: Table of Numerical Estimates of Linear Regression Model

Figure 3 shows the values of α and β that our model has estimated would best describe the relationship between a county's MHI and its mortality rate. As you can see, the model seems to indicate the presence of a negative correlation - that as MHI increases, fatality decreases. This fits in with our general assumption that wealthier individuals are more likely to survive COVID-19 if infected given their greater access to healthcare and better living environments.

Fig. 4: Scatterplot of MHI and COVID−19 Mortality Rates across US Count
With Linear Regression Model



The above scatterplot graphs each recorded county's MHI and its COVID-19 fatality rate. It also shows the linear regression model derived by our estimation. While fatality rates do vary in between counties with similar income levels, there is a substantial case to be made that as the median income rises, the chances of surviving COVID-19 grow. However, this scatterplot also reveals to us a great discrepancy in the data - that there is much more information on lower-income counties than there are higher-income counties. This may be because simply of how wealth is grouped in the US but nonetheless the fact remains that we have comparatively few high-income data points to analyze.

Maximum Likelihood Estimator

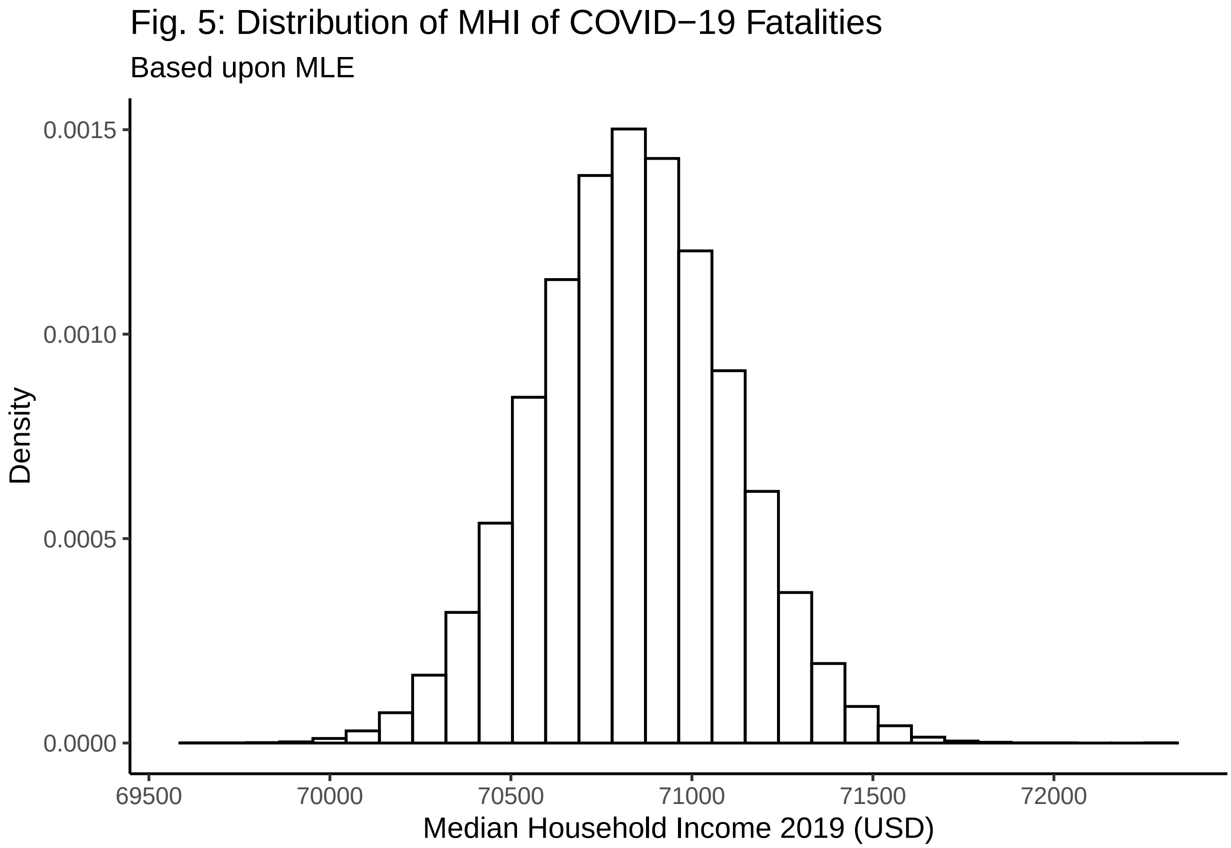


Figure 5 shows the distribution of incomes of COVID-19 fatalities based on our MLE for a Poisson distribution. The distribution peaks in a similar place to the distribution in figure 2. However, we can observe that this distribution is far more symmetrical, thus not reflecting our original conception that poor people are more likely to die from COVID-19 than wealthier individuals. However, it should also be noted that this distribution does not have a spread anywhere near the size of that of the data sample. Thus, it accounts for neither the poorest end of the spectrum nor the wealthiest end. Therefore, this distribution may be most helpful in giving us an understanding of the effects of COVID-19 on the nation's median salary earners.

Hypothesis Test

Statistic	**Value**
Sample Mean Income of all COVID-19 Fatalities	7.0835082×10^4
Sample Mean Income of Black COVID-19 Fatalities	6.6420643×10^4
Standard Deviation of Income of Black COVID-19 Fatalities	1.3575155×10^4
Test Statistic	-48.1142257
p-value	0

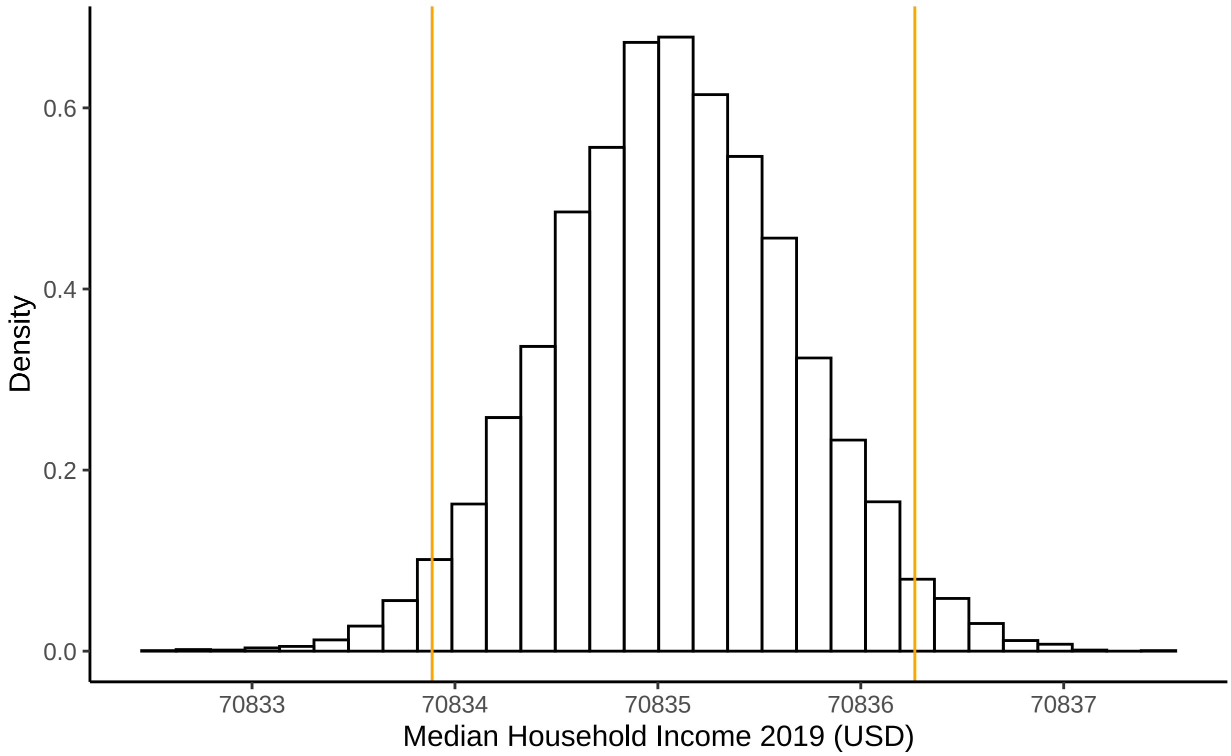
Fig. 6: Table of Statistics on the Hypothesis Test for the Mean Income of Black COVID-19 Fatalities

Figure 6 is a table showing the most important statistics in our mean hypothesis test. Remember that this specific test was to see if the mean income of black COVID-19 fatalities is close to the mean income of all COVID-19 fatalities in our dataset. The p-value is the probability of achieving our test statistic (which is based on our data) or a more extreme value if the two mean incomes are the same. Since it is 0, there is overwhelming evidence against this null hypothesis that the two means are the same. In this case, we had promised to reject the null hypothesis if the p-value is smaller than $\alpha = 0.05$ and our result here is unambiguous. We have no choice but to reject this null hypothesis, thereby concluding that the mean income of black COVID-19 fatalities is not the same, nor close, to the mean income of all general COVID-19 fatalities. If we are to look at figure 6, we would see that in our observed data sample, the mean income of black COVID-19 fatalities is substantially lower than that of the general COVID-19 victim.

Bayesian Credible Interval

Fig. 7: Posterior Distribution of Mean of MHI of COVID−19 Fatalities

Based upon Poisson Income Distribution and Gamma Prior



Percentile	Value
2.5th Percentile	7.0833887×10^4
97.5th Percentile	7.0836266×10^4

Fig. 8: Table of Percentiles for the Posterior Distribution of Mean Incomes of COVID-19 Fatalities

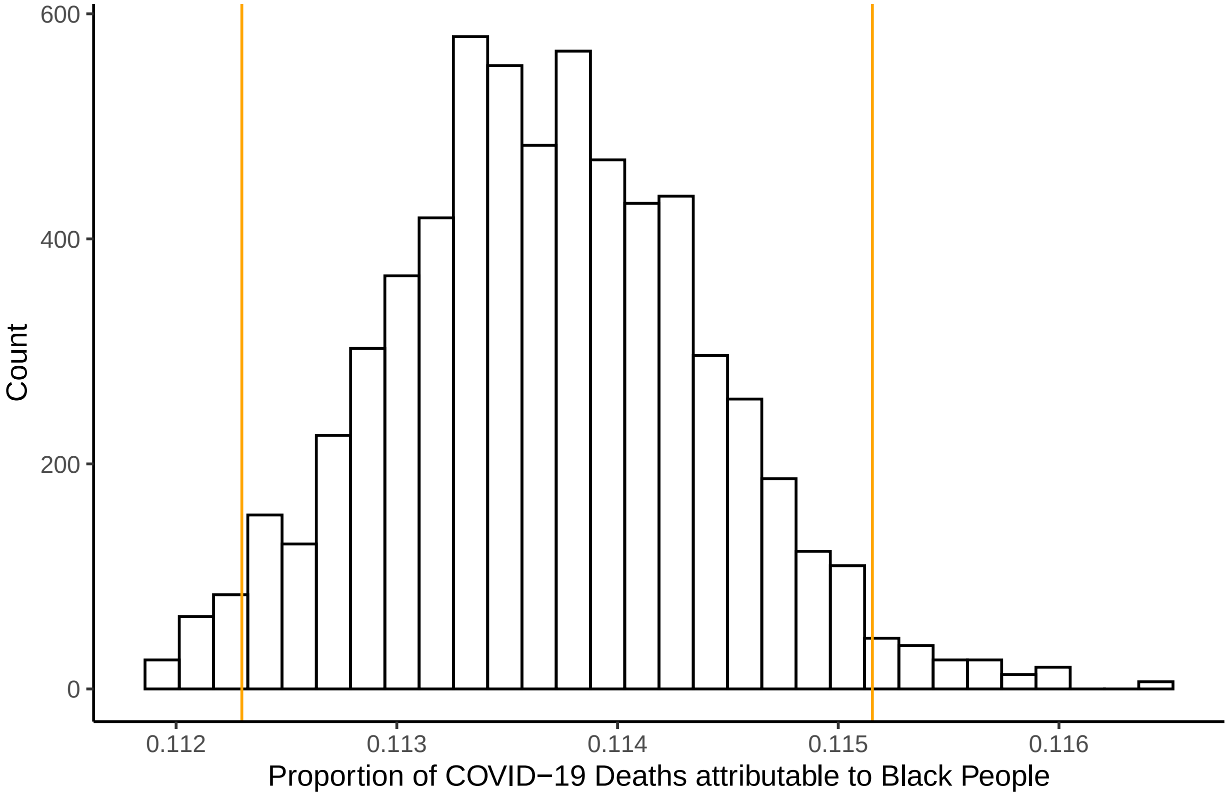
Figure 7 and 8 display the results of our 95% Bayesian credible interval of the mean income of COVID-19 fatalities. This means that there is a 95% probability that the true mean income of all COVID-19 fatalities exists within the boundaries of 7.0833887×10^4 and 7.0836266×10^4 US dollars. These two boundaries are illustrated by the orange lines in figure 12 which also graphically represents the posterior distribution we derived.

Given that the 2019 US Census estimate places the median income of all US households at $68,703 (US Census, 2019) and this range completely exceeds that number, this is quite surprising as we had expected COVID-19 to hurt the poor more. Perhaps this is explainable by the postulation that the wealthiest victims of COVID-19 had such great wealth that in calculating the total mean MHI, we inadvertently skewed the distribution of mean incomes more towards them.

We had mentioned that this method depended on how representative the sample data is of the population, the appropriateness of our prior distribution and how we view the distribution of incomes. The appropriateness of all of these assumptions can be debated. For instance, whilst the sheer scale of the data suggested that it encompasses a substantially diverse representation of the population, we have to bear in mind that it was collected across many different jurisdictions and subject to artificial suppression. Poorer communities could have more disorganized health services that were less stringent in reporting the victim's county of residence, thus causing us to lose out on data.

Confidence Interval

Fig. 9: Bootstrap Sampling Distribution of Proportion of Black COVID−19 Fa



Percentile	**Value**
2.5th Percentile	0.1122977
97.5th Percentile	0.1151546

Fig. 10: Table of Percentiles for the Distribution of Bootstrap Proportion of Black COVID-19 Fatalities

The above graph and table show key elements of the bootstrap distribution of proportions of black COVID-19 fatalities. We can say with 95% confidence, that the true proportion of COVID-19 deaths attributable to black people lies within the interval from 0.1122977 to 0.1151546. This is graphically represented in the distribution graph in figure 9. Our results tell us that black people make up around 11% of COVID-19 related deaths. Current estimates by the US Census Bureau is that black individuals make up 13.4% of the US population. This is unexpected as it shows that black people are actually dying less from COVID-19 than if the proportion of deaths was split following their proportion in the population. This contradicts our hypothesis that black people account disproportionately more for COVID-19 deaths.

Goodness of Fit Test

Race	Sample No. of COVID-19 Fatalities	Sample Proportion of Total COVID-19 Fatalities	Estimated Proportion of Population (US Census Bureau)
American Indian/Alaska Native	274	0.00142	0.013
Asian	7776	0.04038	0.059
Black	21892	0.11368	0.134
White	156880	0.81467	0.763
Native Hawaiian/Other Pacific Islander	50	0.00026	0.002
Multiple/Other	5696	0.02958	0.028
Total	192568	1	1

Fig. 11: Table showing Sample Data COVID-19 Fatality Totals and Proportions and US Census Population Estimates by Race

The table above shows the number of COVID-19 fatalities, distinguished by race, as well as their proportion. It also shows the proportion each race accounts for in the general population, as estimated by the US Census Bureau. This is the main data we will be using to conduct our investigation.

Mathematically deriving a test statistic and calculating its subsequent probability under a chi-square distribution gives us a p-value of 0. This suggests that there is no probability that COVID-19 fatalities are proportionate to the makeup of the population. The same conclusion can also be reached with a simulation, as explained below.

Race	No. of Proportions as Extreme as Sample Data
American Indian/Alaska Native	0
Asian	0
Black	0
White	0
Native Hawaiian/Other Pacific Islander	0
Multiple/Other	0

Fig. 12: Table showing the number of General-Population-Proportionate Simulated Samples exceeding the Extremity of the Sample Data on Racial Proportions in COVID-19 Deaths

In our simulation, we generated 1000 samples of the same size as our data sample, each containing individual races according to the proportions estimated by the US Census Bureau. We then calculated the proportion that each race accounted for in each sample and recorded the number of times that proportion was as or more extreme than the original data sample. The table above lays out the results of that record, grouped around the different races. As you can see, not a single race achieved a single proportion that reached the level of extremity of the original data sample. If we were to take this as a calculation of probability and divide each result by the number of samples, 1000, the simulated p-value would be 0. Thus, there is a very low probability that COVID-19 deaths are proportionate to the makeup of the population and we have every reason to reject this null hypothesis.

All analysis for this report was programmed using `R version 4.0.4`.

Conclusions

We began this investigation with the hypotheses that black people will account disproportionately more for COVID-19 deaths and that the average income of a black person who died from COVID-19 would be substantially lower than the average income of all fatal victims of COVID-19. Using data from the CDC and the US Census Bureau, we assigned fatal victims to their residential county's median household income to estimate their wealth level.

We attempted to model the relationship between county income levels and their mortality rate using linear regression. Then, assuming the distribution of income of all COVID-19 fatalities followed a Poisson distribution, we derived the maximum likelihood estimator of its mean λ. We also used this Poisson assumption and a prior Gamma distribution to build a 95% credible interval containing the true average MHI of all COVID-19 victims. We also conducted a mean hypothesis test to compare the mean average income of black COVID-19 fatalities and all COVID-19 fatalities to see if they were similar. Ignoring economic factors, we also attempted to measure the extent to which COVID-19 disproportionately killed black people. This was done by setting up a bootstrap simulation to find the true proportion of COVID-19 deaths attributable to black people and through a goodness of fit test that analyzed the entire population by race to see how plausible it was that COVID-19 affected all races proportionately.

The results of our investigation are varied. Some meet our hypotheses and others contradict them. The linear regression generally met our expectation that as income levels rise, mortality rates decrease. This matches our expectation that wealthier individuals would have access to better resources and thus, have a greater chance of surviving COVID-19. However, we must account for the fact that there were relatively few counties of great wealth so there is less data on that front than we would have liked. We should also point out that the mortality rate in counties at the lower end of the spectrum varied greatly. How much external environmental factors (such as air quality) contributed to this is unknown.

The Bayesian credible interval and the MLE both attempted to locate the mean MHI of all COVID-19 fatalities. They both agreed that the mean income seemed is most likely to be around $71,000. Given that the median US household income nationwide among the general population was estimated to be $68,703 (US Census, 2019) in 2019, this is a big surprise as it suggests that the average COVID-19 victim was richer than the average American. This contradicts our hypothesis that COVID-19 was more likely to kill poorer people. However, our results were built upon assumed models such as the Poisson distribution and the prior Gamma distribution. Both of these could probably be changed or modified and improved. Also, the distribution of wealth in the US is skewed so greatly to extremes that a comparatively few exceedingly wealthy fatalities probably could have balanced out a greater number of poorer victims.

The mean hypothesis test found that there was 0 probability that the mean income of black COVID-19 deaths could have been equal to the mean income of all COVID-19 deaths. If we look at the sample data alone, we will find that the sample mean income of black COVID-19 fatalities is far below that of the sample mean income of all COVID-19 fatalities. This conclusion matches our hypothesis that the average black victim of COVID-19 is poorer than the average victim of COVID-19 and once again highlights the problem of racial economic inequality which motivated this study. The fact that the p-value in this investigation is 0, indicating no chance of equality at all, stresses just how wide the disparity is.

Our bootstrap confidence interval attempted to locate the true proportion of COVID-19 deaths attributable to black individuals. It determined that this proportion is a little more than 11%. This is a very surprising conclusion to our investigation as the US Census Bureau estimates that black people account for 13% of the general population in the US. This means that black people are dying disproportionately less from COVID-19. This directly contradicts our hypothesis as well as the CDC's official analysis (CDC, 2021). We suspect that this may be due to several reasons. Firstly, we only extracted records that included the individual's race and county of residence. If we make the general assumption that black people live in poorer counties with fewer resources, these counties may have been less stringent in their reporting procedures and not filled in those fields (particularly the field on the county of residence) as often as wealthier counties did. Secondly, the CDC engaged in artificial data suppression methods which targeted attributes such as race.

This second reason also makes it very difficult to verify the veracity of the conclusion of our goodness of fit test which tried to determine the extent to which COVID-19 disproportionately affected all races. Our conclusion is that this is a disease that does have a tendency to affect different races disproportionately although it would seem that this disproportion is negatively oriented most noticeably for white people. Minorities such as blacks, Asians, and Native Americans seem to be disproportionately under affected by COVID-19. The two issues we have noted above with our data probably contribute extensively to this conclusion which goes against our hypothesis.

In conclusion, whilst we are fairly certain that the average income of black COVID-19 fatalities is lower than that of the general population, we can't confirm the extent to which COVID-19 affects poor people more than those of average wealth. However, we have reason to believe that the wealthiest are at much less risk from COVID-19. We also can't concretely speak to whether the COVID-19 pandmic has disproportionately killed more black people and how or if it has disproportionately affected any race although there is evidence to suggest the latter.

Weaknesses

There are plenty of potential pitfalls that this investigation could have fallen in. Firstly, let us begin with external factors beyond our control. That begins with the collection of the data. As we have discussed before, the nationwide directive to collect COVID-19 data is handled by local health authorities. Thus, thousands of nurses, doctors, and other officials across hundreds of organizations and jurisdictions fill out data forms with differing procedures and levels of stringency. It is reasonable to suspect that poorer communities and those more overwhelmed often miss out on important data, thus lessening the weight of the data representing those areas. This could have contributed to the higher mean incomes of COVID-19 fatalities we achieved. Next, this data only includes those cases which have been reported to authorities. COVID-19 is known to be a disease that exudes differing levels of severity with some infected individuals being asymptomatic carriers. As such, those cases reported to the authorities are much more likely to be severe ones - thus raising the recorded proportion of deaths. We also have to consider the data suppression techniques undertaken by the CDC. These measures were taken to protect the identity of individual cases that may be revealed through attributes of race, gender, age, and location. As such, cases suppressed are overwhelmingly in rural areas with smaller overall populations. Again these are often the poorest communities and racial minorities are the ones most likely to have their attributes suppressed. This may account for the disproportionately low measures of fatality we have seen in our minority groups as noted above. Then there is the matter of how we have measured income in this investigation. Throughout this study, we have used the individual's county of residence's MHI as an indication of their income. However, incomes can vary greatly within counties. This is due to both the differing sizes of counties as well as their settings - urban cities account for some of the greatest economic disparities in the world. As such, what we have treated as individual income in this study is only a very, very rough generalization.

Secondly, let us examine the shortcomings that we have made on our part in this investigation. We can begin with our choice of the Poisson distribution to model the distribution of income among COVID-19 fatalities. Whilst an argument can be and has been made that income is passable as a discrete variable, it is also possible to treat it as a continuous one. This may allow for greater flexibility in future investigations. A better understanding of income distribution as a whole would have been very helpful in determining the mean income of COVID-19 fatalities. Our next mistake was that in investigating the race of all COVID-19 fatalities, we excluded cases where we did not know their county of residence. This was needless and done out of convenience as we already had a dataset with the needed attributes. However, it could have removed many records, especially those from poorer communities (for the reasons mentioned above), and skewed the results of our investigation by making our sample data less representative of the population.

Next Steps

For future investigations into the effects of the COVID-19 pandemic from a racial economic point of view, we should strive to remove our mistakes in this study. This means maintaining a more diverse sample that truly represents the population. Furthermore, work can be done to investigate median incomes in the general population versus just the black community, excluding all matters on COVID-19. Beyond examining just death, we can also understand the racial economic inequalities of this pandemic by studying the levels of accessibility communities have towards testing as well as the overall infection rates in different communities.

Discussion

If there is one thing this crisis has proven, it is the importance and need for comprehensive data collection and compilation. Not only will the COVID-19 pandemic be analyzed by researchers for decades or possibly centuries after its end, but a strong system of data collection and analysis will be needed to form smart, flexible, and effective responses to the next major health crisis. Whilst this investigation has provided clashing conclusions on racial economic inequalities during the pandemic, this merely accentuates the need for further efforts and greater rigor in understanding our world.

Bibliography

Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

Centers for Disease Control. (2020, May). *Human Infection with 2019 Novel Coronavirus Case Report Form.* https://www.cdc.gov/coronavirus/2019-ncov/downloads/pui-form.pdf

Centers for Disease Control. (2021, February 12). *Health Equity Considerations and Racial and Ethnic Minority Groups.* https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html#anchor_1595551060069.

Centers for Disease Control. (2021, April). *COVID-19 Case Surveillance Public Use Data with Geography.* https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

Centers for Disease Control. (2021, April). *COVID-19 Case Surveillance Public Use Data with Geography.* https://data.cdc.gov/api/views/n8mc-b4w4/rows.csv. (Last Accessed: April 16, 2021)

Centres for Disease Control. (2021, April 11). *Demographic Trends of COVID-19 cases and deaths in the US reported to CDC.* https://covid.cdc.gov/covid-data-tracker/#demographics.

Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

Hanks, A., Solomon, D. & Weller, C.E. (2018, February 21) *Systematic Inequality: How America's Structural Racism Helped Create the Black-White Wealth Gap.*https://www.americanprogress.org/issues/race/reports/2018/02/21/447051/systematic-inequality/

Ioannidis, J.P.A. (2020, October 14). *Infection fatality rate of COVID-19 inferred from seroprevalence data.* https://www.who.int/bulletin/volumes/99/1/20-265892/en/. World Health Organization

Lee, H., Esposito, M., Edwards, F., Chun, Y. & Grinstein-Weiss, M. (2020, July 27) *The demographics of racial inequality in the United States* Brookings. https://www.brookings.edu/blog/up-front/2020/07/27/the-demographics-of-racial-inequality-in-the-united-states/.

United States Census Bureau. (2019). *2019 American Community Survey 1-Year Estimates: INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS) Table S1901.*

United States Census Bureau. (2019) *Vintage 2019 Population Estimates.*
https://www.census.gov/quickfacts/fact/table/US/PST045219#qf-headnote-a.

United States Census Bureau. (2019, September 15). *Income and Poverty in the United States: 2019.*
https://www.census.gov/library/publications/2020/demo/p60-270.html

United States Census Bureau. (2020, November 28). *2010 - 2019 County-Level Estimation Details.*
https://www.census.gov/programs-surveys/saipe/technical-documentation/methodology/counties-states/county-level.html

U.S Department of Agriculture. (2021, January 5). *Unemployment and median household income for the U.S., States, and counties, 2000-19.*
https://www.ers.usda.gov/webdocs/DataFiles/48747/Unemployment.csv

United States Department of Agriculture. (2021, February). *Documentation.*
https://www.ers.usda.gov/data-products/county-level-data-sets/documentation/

All analysis for this report was programmed using `R version 4.0.4`.

Packages used: tidyverse, openintro, knitr, and kableExtra.

Appendix

Section 1: Derivation of MLE

Assumption: Data is a sample of random independent Poisson variables. Therefore,

$$X_1, X_2, ..., X_n \overset{iid}{\sim} Pois(\lambda)$$

and

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Deriving the likelihood function:

$$L(\lambda) = L(\lambda | X_1, ..., X_n) = f_{X_1,...,X_n}(x_1, ..., x_n)$$

$$= f_{X_1}(x_1) f_{X_2}(x_2) ... f_{X_n}(x_n)$$

$$= \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} ... \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$$

$$= \frac{\lambda^{x_1 + ... + x_n} e^{-\lambda n}}{x_1! x_2! ... x_n!}$$

$$= \frac{\lambda^{n\bar{x}} e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)}$$

Deriving the MLE:

$$l(\lambda) = ln(L(\lambda)) = ln(\frac{\lambda^{n\bar{x}} e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)})$$

$$= ln(\lambda^{n\bar{x}} e^{-n\lambda}) - ln(\prod_{i=1}^{n}(x_i!))$$

$$= ln(\lambda^{n\bar{x}}) + ln(e^{-n\lambda}) - ln(\prod_{i=1}^{n}(x_i!))$$

$$= n\bar{x}ln(\lambda) - n\lambda - ln(\prod_{i=1}^{n}(x_i!))$$

Differentiating to find its maximum:

$$\frac{\partial l}{\partial \lambda} = \frac{n\bar{x}}{\lambda} - n$$

$$0 = \frac{n\bar{x}}{\lambda} - n$$

$$\Rightarrow n = \frac{n\bar{x}}{\lambda}$$

$$\Rightarrow \lambda = \bar{x}$$

Therefore, we have an estimator for λ that $\hat{\lambda}_{MLE} = \bar{x}$. Verifying that this value is indeed the maximum:

$$\frac{\partial^2 l}{\partial \lambda} = \frac{\partial}{\partial \lambda}(\frac{n\bar{x}}{\lambda} - n) = -\frac{n\bar{x}}{\lambda^2}$$

$$\frac{\partial^2 l}{\partial \lambda}(\bar{x}) = -\frac{n\bar{x}}{\bar{x}^2} = -\frac{n}{\bar{x}}$$

Since n is the number of variables and $\bar{x}$ is the sample mean of the data which has to be positive as we are measuring the median household income, we can confirm that this value has to be negative. As such, it is concave down, thus confirming that our estimator is the maximum. So,

$$\hat{\lambda}_{MLE} = \bar{x}$$

Section 2: Derivation of Posterior Distribution

Assumptions:

Prior Distribution: $\lambda \sim Gamma(6, \frac{1}{12000})$. Therefore,

$$f(\lambda) = \frac{(\frac{1}{12000})^6}{\Gamma(6)}\lambda^{6-1}e^{-\frac{\lambda}{12000}} = \frac{(\frac{1}{12000})^6}{\Gamma(6)}\lambda^5 e^{-\frac{\lambda}{12000}}$$

Data is a sample of random independent Poisson variables. Therefore,

$$X_1, X_2, ..., X_n \overset{iid}{\sim} Pois(\lambda)$$

and

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Derivation of Posterior Distribution:

$$P(\lambda|data) = \frac{P(data|\lambda)P(\lambda)}{P(data)} = \frac{P(X_1...X_n|\lambda)P(\lambda)}{P(X_1,..,X_n)}$$

$$= \frac{P(X_1|\lambda)P(X_2|\lambda)...P(X_n|\lambda)P(\lambda)}{P(X_1,...,X_n)}$$

$$= \frac{\frac{e^{-\lambda}\lambda^{X_1}}{X_1!}\frac{e^{-\lambda}\lambda^{X_2}}{X_2!}...\frac{e^{-\lambda}\lambda^{X_n}}{X_n!}\frac{(\frac{1}{12000})^6}{\Gamma(6)}\lambda^5 e^{-\frac{\lambda}{12000}}}{P(X_1,..,X_n)}$$

$$= \frac{\frac{e^{-\lambda n}(\lambda^{X_1}\lambda^{X_2}...\lambda^{X_n})\frac{1}{(12000)^6}\lambda^5 e^{-\frac{\lambda}{12000}}}{(X_1!X_2!...X_n!)\Gamma(6)}}{P(X_1,...,X_n)}$$

$$\text{Let } C = \frac{\frac{(\frac{1}{12000})^6}{(X_1!X_2!...X_n!)\Gamma(6)}}{P(X_1, ..., X_n)}$$

$$\text{Therefore } \frac{\frac{e^{-\lambda n}(\lambda^{X_1}\lambda^{X_2}...\lambda^{X_n})\frac{1}{(12000)^6}\lambda^5 e^{-\frac{\lambda}{12000}}}{(X_1!X_2!...X_n!)\Gamma(6)}}{P(X_1, ..., X_n)} = Ce^{-\lambda n}(\lambda^{X_1}\lambda^{X_2}...\lambda^{X_n})\lambda^5 e^{-\frac{\lambda}{12000}}$$

$$= Ce^{-\lambda(n+\frac{1}{120000})}\lambda^{X_1+X_2+...+X_n+5} = Ce^{-\lambda(n+\frac{1}{120000})}\lambda^{n\bar{X}+5}$$

Notice that this follows the form of a $Gamma(\alpha, \beta)$ distribution.

$$\text{Let } \frac{1}{\beta} = n + \frac{1}{12000} \Rightarrow \frac{1}{\beta} = \frac{12000n+1}{12000} \Rightarrow \beta = \frac{12000}{12000n+1}$$

$$\text{and let } \alpha - 1 = n\bar{X} + 5 \Rightarrow \alpha = n\bar{X} + 6$$

Therefore, the posterior distribution for λ is

$$\lambda \sim Gamma(n\bar{X} + 6, \frac{12000}{12000n+1})$$