Cornell University

Allocation of COVID-19 Vaccinations to Communities of High Socioeconomic Vulnerability

Rachael Adelson

ENGRD 2700

Professor Pender

December 11, 2020

1. Introduction

Over the past year, the 2019 novel coronavirus (COVID-19) has spread throughout the United States, taking the lives of thousands of Americans. The transmission rates and susceptibility levels of the virus are suggested to be higher in communities of high socioeconomic vulnerability. This claim is strongly backed by the similarities between Influenza and COVID-19, as well as the socioeconomic disparities proven to already exist in the Influenza virus. The behavior of the influenza plays a crucial role to help us understand COVID-19 due to the lack of testing resources during the early months of the pandemic; we can use Influenza to closely model the COVID-19 trends of transmission and susceptibility. First, we will examine the driving factors that create these discrepancies in Influenza and demonstrate how these circumstances pertain to the disparities present in COVID-19. Second, we will create two samples of United States counties—one consisting of counties ranked as the most socioeconomically vulnerable and the other containing counties ranked as the least socioeconomically vulnerable—and compare the novel coronavirus transmission rates and vulnerability levels. To do so, we will find the proportion of daily COVID-19 cases for each county in our two samples on a randomly selected day during the pandemic**.** We will construct a sampling distribution of the sample means for our two data sets of counties and compare the two distributions using a Hypothesis Test. Third, we will illuminate statistical flaws in our findings—stressing the lack of COVID-19 resources to communities of high socioeconomic vulnerability during the early months of the pandemic. The lack of resources to socioeconomically vulnerable communities is important to note as the COVID-19 vaccinations begin to rollout across the United States. Lastly, we modify our Hypothesis Test to control for the lack of COVID-19 testing resources during the early months of the pandemic.

2. Influenza

Influenza (Flu) is a common viral respiratory illness that spreads seasonally in the United States each year. In a study done about the health inequities in influenza transmission, it was found that inequities increase influenza transmission amongst low socioeconomic status (SES) individuals (Bansal and Zipfel, 2020). When comparing influenza transmission of low SES individuals to the transmission behavior of the rest of the population, the study considered how the contact patterns of these two groups differ. When analyzing these contact patterns, the following five differences were the drivers of disparities in influenza burden[1] (Bansal and Zipfel, 2020):

1. Higher assortativity amongst low SES individuals
2. Low vaccine uptake
3. Low Healthcare Utilization
4. High Susceptibility from stressful environment factors
5. Low absenteeism from work or school (low-income jobs do not typically give paid leave)

These five differences in contact patterns should also cause socioeconomic disparities in COVID-19 transmission rates and susceptibility levels. Both Influenza and COVID-19 transmit between people in the same manner: people tend to spread both viruses from one person to another when they are in close proximity primarily through droplets caused by talking, sneezing, or coughing (Similarities and Differences Between Flu and COVID-19). Since both viruses behave similarly when spreading through a population, a circumstance that increases the transmission rate and level of susceptibility of Influenza should also do the same for COVID-19.

[1] https://www.medrxiv.org/content/medrxiv/early/2020/04/01/2020.03.30.20048017.full.pdf

However, these five factors are not the only notable ones when accounting for socioeconomic disparities in COVID-19. Housing conditions, public transportation, and type of employment are three more noteworthy differences present amongst low SES individuals that affect their risk of contracting COVID-19 (Bushara, 2020). Poor housing conditions of low SES individuals create inferior sanitation standards, overcrowding, and decreased ability to physical distance—factors that increase the risk for transmission. Furthermore, many low-income individuals rely on public transportation to travel to and from work, increasing their physical contact with individuals and risk to COVID-19 (Bushara, 2020). Lastly, low SES individuals typically work essential jobs—such as construction, transportation, and food service—that prevent them from working at home during the pandemic, which results in increased contact with others.

Lastly, the likelihood of having a chronic health condition increases within socioeconomically vulnerable communities (Fisher and Bubola, 2020). Low SES individuals are about 10% likelier to have a chronic illness, making COVID-19 ten times as deadly (Fisher and Bubola, 2020). While this may not affect the transmission rate of COVID-19 amongst low SES individuals, it increases the severity of symptoms of COVID-19 for these individuals. With many lacking health insurance—as well as access to necessary medical resources—low SES individuals are not only at a higher risk of testing positive and spreading the virus, but also are more vulnerable once they contract it.

3. Hypothesis Test

Factors of socioeconomic vulnerability have been proven to cause disparities in Influenza susceptibility levels and transmission rate between populations of varying socioeconomic status.

Early studies during the COVID-19 pandemic—as well as the mentioned medical studies comparing the spread of influenza and COVID-19—suggest that the same socioeconomic disparities are present in the transmission and susceptibility of COVID-19.

We will now conduct a Two-Sample Hypothesis Test, a widely used technique to test if on average, there is a higher proportion of cases per population in counties of Low SES individuals in comparison to counties of High SES individuals. For our hypothesis test, our two samples will consist of proportions of COVID-19 daily cases per total population of a county —with sample one consisting of counties of low SES counties and sample two consisting of high SES counties. A date during the COVID-19 pandemic was selected at random to tally the daily cases of the counties in our two samples. This selected day had to satisfy the following criteria:

1. The date selected is part of the period during the pandemic in which there was a consistent increase in daily cases and deaths.
2. The date selected occurred before or after the brief decline in cases over the summary.

This set of criteria ensures that the date randomly selected occurred during the height of the pandemic so that conclusions made for that day could be broadened to make general assumptions (for further study of this claim, data from multiple days could be examined and tested). The date April 18th was randomly selected, and the data consisting of the daily cases and deaths were viewed to make sure the April 18th fit the criteria listed above.

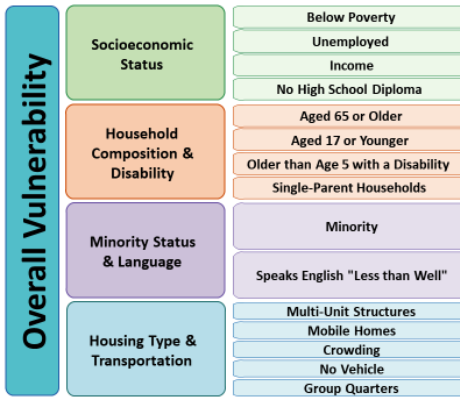

Figure 1: The CDC Social Vulnerability Guidelines broken down into their respective categories. Note that we are only concerned with the first category for our statistical analysis. This figure was taken from the CDC SVI documentation.

To select our two samples of counties, we use the CDC's Social Vulnerability Index, a method of ranking areas across the county based on socioeconomic status, household composition and disability, minority status and language, housing type, and transportation. Figure 1 shows a further breakdown of these respective categories. For the purpose of our study, we are only concerned with the rankings of counties by socioeconomic status, which is calculated considering the following criteria: the percentage of the population below the poverty line, unemployment rate, level of income, and percentage of population with no high school diploma. For our first sample, we will use the approximately 100 of the lowest-ranked counties in the nation by socioeconomic status with our second sample consisting of approximately the 100 highest-ranked counties by socioeconomic status, found by filtering through the "CDC SVI Data and Documentation". [2]

Using the Federal Information Processing Standard Publication codes of the counties in our two samples, we then were able to filter through two separate data sets to find the daily cases on April 18th and the total population corresponding to that county. To create these data sets in CSV

[2] https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html

files, data was used from Census.IRE.ORG and the New York Times GitHub [3] [4]. Once we obtained the cases and population for each county, we calculate a proportion of daily COVID-19 cases per population for every county in our two data sets. Finally, we create a sampling distribution for the sample mean for our two data sets of Low SES counties and High SES counties (Figure 2). Examining the graphs in Figure 2 below, we can make two observations:

1. The mean of the Low SES distribution is less than the mean of the High SES distribution.
2. The Low SES sampling distribution is shifted slightly to the left of the High SES sampling distribution.

These two observations were quite surprising and go against the highly suggested behavior of COVID-19 transmission rates and susceptibility levels; based on the behavior of influenza as well as the early studies of COVID suggesting that disparities exist, we would expect to find that the total cases per county population be greater amongst low SES counties. The inconsistencies likely appeared since we did not control for:

1. The varying levels of COVID-19 testing amongst low SES communities during the early months of the pandemic.
2. The nature of employment amongst low SES individuals.

For the early months of the pandemic (including April 18th), testing in the United States was already scarce. Tests were only available to those who were currently displaying serious symptoms, meaning that individuals could be contagious for up to three days before

[3] http://census.ire.org/data/bulkdata.html

[4] https://github.com/nytimes/covid-19-data/blob/master/live/us-counties.csv

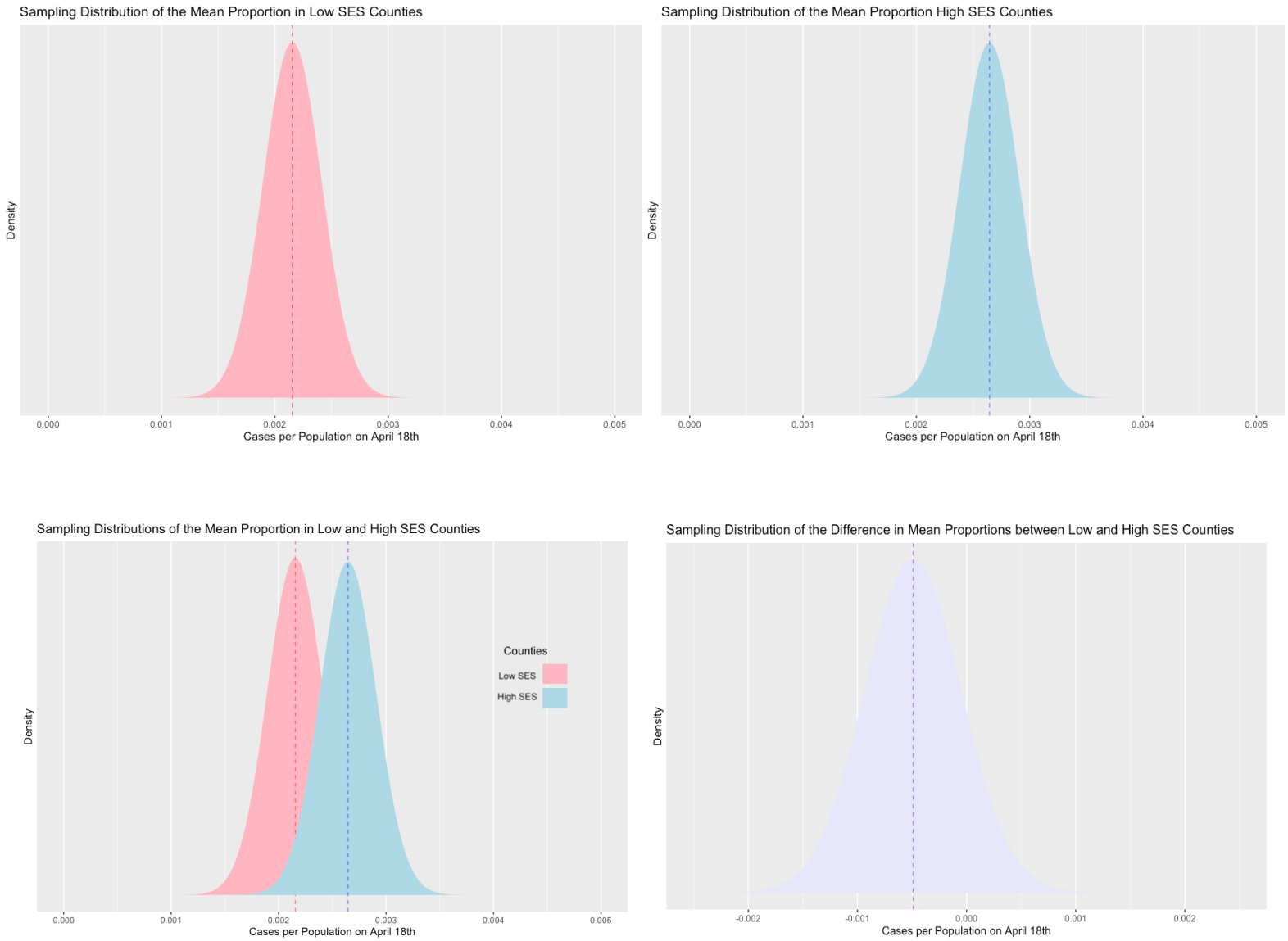

Figure 2: Pictured above is the graphed probability density functions of our two samples. The mean of the sampling distribution of low SES counties is approximately 0.00215, and the mean of the sampling distribution of high SES counties is approximately 0.00265. In the bottom-right corner, we have the sampling distribution of the mean of the difference of our two random variables with a mean value of -0.00049. The results of our hypothesis test were insignificant.

getting a positive test (Dalton, 2020). Even when individuals were symptomatic, they were deterred by time and money related factors. Many low SES individuals do not have health care coverage and could not afford the cost of a COVID-19, which was typically about $100 at major

diagnostic labs (Kliff, 2020). Furthermore, the type of employment for many low SES individuals does not offer paid leave, incentivizing many low SES individuals who were potentially sick to come into work instead of taking the time off to get tested.

In order to control for these factors, we recalculate our sampling distributions using data from exactly four months later on August 18th. By August, testing resources were heavily subsidized and made widely available across the United States in local pharmacies, urgent care clinics, temporary testing sites, and hospitals. The Families First Coronavirus Response Act, which was just days old on April 18th, was now in full swing, giving low SES individuals up two 80 hours of emergency paid sick leave (US Department of Labor).

We then conducted a hypothesis test to discover whether our data was significant evidence to support our claim that on average, the proportion of cases per population is higher in low SES counties than high SES counties. Let H_o be the claim that on average, the proportion of COVID-19 cases per county population is no different in counties of low SES and high SES individuals. Let H_a be the claim that this proportion is greater amongst Low SES counties. We conduct this test at a significance level of $\alpha = 0.01$ and obtained a p-value of $2.267 * 10^{-5}$. In summary, we have strong enough statistical evidence to reject the null hypothesis and conclude that there is significant evidence to support that on average there is a higher proportion of COVID-19 cases in low SES counties, meaning that the transmission rates and levels of susceptibility are higher amongst low SES individuals.

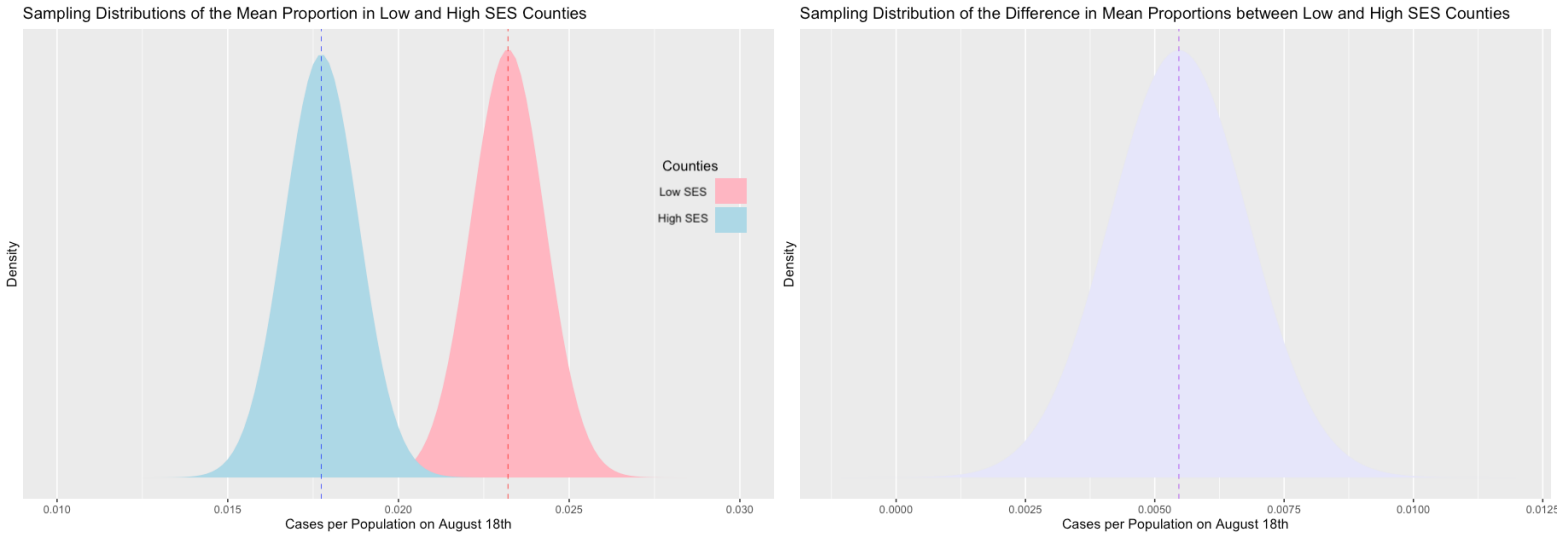

Figure 3: On the left is the graphed probability density functions of our two samples on August 18th. The graph on the right is the sampling distribution of the mean of the difference of our two random variables with a mean value of 0.00546.

4. Conclusion

In summary, our findings do show that there are socioeconomic disparities in COVID-19 transmission rates and vulnerability levels. Initially, our statistical analysis using data from April 18th offered inconclusive results. However, upon further testing, it became clear that this inconsistency was a result of a lack of testing resources and support in Low SES communities. Once these variables were controlled, our data was statistically significant to conclude that the average proportion of cases was higher amongst low SES communities, supporting our original claim that higher transmission rates and susceptibility levels exist in these communities. These findings have several implications. Not only are low SES individuals more vulnerable and likely to contract COVID-19, but they were also given disproportionate support and resources to get tested for the virus during the early months of the pandemic. These findings are not surprising,

yet they are extremely important to address as we begin to allocate and distribute vaccinations across the country.

The distribution of vaccination resources cannot follow the distribution trends of COVID-19 tests this past year. We have proven that Low SES communities are not only vulnerable but are also a highly transmissible demographic. Vaccinating low SES individuals will protect the susceptible while also decreasing the spread of COVID-19 across the county. Time and time again, socioeconomically vulnerable communities are overlooked and forgotten. During the largest pandemic of the century, let us not make this same fatal mistake once again.

Bibliography

"CDC SVI 2018 documentation." *Centers for Disease Control and Prevention*. January 31, 2020

Dalton, Clayton. "Opinion: Early Coronavirus Testing Failures Will Cost Lives." *National Public Radio*, March 14, 2020.

Fisher, Max and Emma Bubola. "As Coronavirus Deepens, Inequality Worsens Its Spread." *The New York Times*, July 3, 2020.

Kliff, Emma. "Most Coronavirus Tests Cost About $100. Why Did One Cost $2,315?" *The New York Times*, June 16, 2020.

"The Coronavirus Does Discriminate: How Social Conditions are shaping the COVID-19 Pandemic. *Harvard Medical School*. May 5, 2020

"What is the difference between Influenza (Flu) and COVID-19?" *Centers for Disease Control and Prevention.*

"Two Sample T-Test." *NCSS Statistical Software*.

R-code used to calculate data and generate graphs:

```
library(dplyr)
SocioeconLow <-SVI2018_US %>% filter(RPL_THEME1>=0.9955)
LowFIPS <- array(SocioeconLow$STCNTY)
LowFIPS<- unique(LowFIPS)
SocioeconHigh <-SVI2018_US %>% filter(RPL_THEME1>=0, RPL_THEME1<=0.0051)
HighFIPS <- array(SocioeconHigh$STCNTY)
HighFIPS<- unique(HighFIPS)

low_proportion <- vector()
for(i in 1: 135){
  county = LowFIPS[i]
  rowpop <- pop %>% filter(pop$GEOID == county)
  rowcases <- counties %>% filter(counties$fips == LowFIPS[i])
  if (length(rowpop$GEOI)==1 && length(rowcases$fips)==1){
    prop = rowcases$cases/rowpop$POP100
    print(prop)
    low_proportion <- c(low_proportion, (prop))}
}

high_proportion <- vector()
for(i in 1: 157){
  county = HighFIPS[i]
  rowpop <- pop %>% filter(pop$GEOID == county)
  rowcases <- counties %>% filter(counties$fips == HighFIPS[i])
  if (length(rowpop$GEOI)==1 && length(rowcases$fips)==1){
    prop = rowcases$cases/rowpop$POP100
    print(prop)
    high_proportion <- c(high_proportion, (prop))}
}
ß
mean_low= mean(low_proportion)
mean_high= mean(high_proportion)
sd_low = sd(low_proportion)
sd_high=sd(high_proportion)
sd_diff = sqrt(((sd_low^2)/132) +((sd_high^2)/129))
zscore = (mean_low - mean_high)/(sd_diff)
pvalue = pnorm(zscore, mean =0 ,sd = 1, lower.tail = FALSE)

print(mean_low)
print(mean_high)
print(zscore)
print(pvalue)
```

```
highplot <- ggplot(data = data.frame(x = c(0.000,0.005)), aes(x)) +
  stat_function(fun = dnorm, n = 129, geom= "area", fill= "light blue",
                args = list(mean = mean(high_proportion), sd = sd_low/sqrt(129))) +
  scale_y_continuous(breaks = NULL) +
  labs(
    title= "Sampling Distribution of the Mean Proportion High SES Counties",
    x = "Cases per Population on April 18th",
    y= "Density") +
  geom_vline(aes(xintercept=mean(high_proportion)),
             color = "blue",linetype= "dashed",size=0.25)
highplot


diffplot <- ggplot(data = data.frame(x = c(-0.0012,0.012)), aes(x)) +
  stat_function(fun = dnorm, n = 129+132, geom= "area", fill= "lavender",
                args = list(mean = mean_low - mean_high, sd = sd_diff)) +
  scale_y_continuous(breaks = NULL) +
  labs(
    title= "Sampling Distribution of the Difference in Mean Proportions between
    Low and High SES Counties",
    x = "Cases per Population on August 18th",
    y= "Density") +
  geom_vline(aes(xintercept=mean(mean_low - mean_high)),
             color = "purple",linetype= "dashed",size=0.25)
diffplot

bothplot <- ggplot(data = data.frame(x = c(0.010,0.030)), aes(x)) +
  stat_function(fun = dnorm, n = 132, geom= "area", fill= "light pink",
                args = list(mean = mean(low_proportion), sd = sd_low/sqrt(132))) +
  scale_y_continuous(breaks = NULL) +
  geom_vline(aes(xintercept=mean(low_proportion)),
             color = "red",linetype= "dashed",size=0.25) +
  stat_function(fun = dnorm, n = 129, geom= "area", fill= "light blue",
                args = list(mean = mean(high_proportion), sd = sd_low/sqrt(129))) +
  geom_vline(aes(xintercept=mean(high_proportion)),
             color = "blue",linetype= "dashed",size=0.25) +
  labs(
    title= "Sampling Distributions of the Mean Proportion in Low and High SES Counties",
    x = "Cases per Population on August 18th",
    y= "Density") +
    scale_colour_manual("Density",
                        breaks= c("Lowest SES Counties","Highest SES Counties"),
                        values = c("light pink", "light blue"))
bothplot
```