Project 2: Where did the Data Come From?

Johann Niebuhr, Odilon Rojas, Taylor Blair

Math 141, Week 7

Import Libraries

```
library(tidyverse)
```

Overview

For project one we looked at individual covid precautions compared to the change in Covid cases. In order to create insight, we required three datasets: Covid behavior, Covid cases, and state by state population.

US Covid Behavior

How do people feel about Covid safety percautions

```
behavior_covid <- read_csv("/home/courses/math141f20/Data/covid-19-behaviors/us_covid.csv")


behavior_covid$endtime <- as.Date(behavior_covid$endtime, format = "%d/%m/%Y %H:%M")
behavior_covid <- behavior_covid %>%
  rename(date=endtime)%>%
  select(date, state, i12_health_1, i12_health_2)

sample_n(behavior_covid, 5)
```

```
## # A tibble: 5 x 4
##   date       state        i12_health_1 i12_health_2
##   <date>     <chr>        <chr>        <chr>
## 1 2020-05-08 Florida      Always       Always
## 2 2020-08-06 New Mexico   Always       Frequently
## 3 2020-05-02 New York     Sometimes    Always
## 4 2020-07-24 Pennsylvania Always       Always
## 5 2020-04-03 Texas        Not at all   Frequently
```

Summary

- **Who**
  - Imperial College London Global Institute of Health Innovation
  - YouGov

- **When**
  - Starts: 2020-04-02
  - Ends: 2020-08-24
  - Daily questionares
- **Where**
  - USA USA USA
  - *Other countries were polled, but we are restricting our data to only the US*
- **Why**
  - Imperial College of London research about the ongoing Covid-19 response.
- **How**
  - Surveys given to people online
- **Who is included?**
  - Individuals who recieve YouGov polls (Internet and in the US)
- **What evidence is there that everyone is present?**
  - It's a sample of a population and its deviations are often indicative of current status
  - Weighted to reflect US demographics

Write up

The dataset `behavior_covid` (or `us_covid.csv`) is from The Imperial College London's Global Institute of Health Innovation, in partnership with the British based international polling company YouGov. The Imperial College London is a major research university and according to its website the Global Institute of Health Innovation plays an advisory role for governments and companies. YouGov generally gathers data for market research and opinion polling, they earned their reputation by accurately predicting election outcomes in the UK.

The data was gathered via repeated surveys on the behaviors and satisfaction of people in response to COVID-19. The source of the respondents was presumably the YouGov panelists. YouGov has a panel of millions of people spread globally who are polled for their responses to surveys. People can sign up for the panel and will receive compensation for completing responses. The US panel contains 2 million respondents and from this pool a 1500 respondent sample is selected, "panelists are invited to each survey, based upon their age, gender, race, and education, in proportion to their frequency to the frequency of adult citizens in the most recent American Community Survey," weighted based on "demographics, voter registration status, and 2016 Presidential vote." The American Community Survey is the ongoing data collection project of the census bureau. YouGov reports a 4% margin of error in this process for adults and a 5% margin of error for voter registration.

Since the sample is taken from YouGov's participants, all of those sampled have internet access, at least enough to allow them to be participants. They must also all be people who have an interest in the incentive model used by YouGov. That YouGov weights and samples in order to be representative likely offsets many of the potential over and under representations in the YouGov population. As a data collection company YouGov seems to be fairly well regarded and is at the very least widely trusted by important organizations, the Imperial College London in this case. The College is an institution with a heavy research focus and consistently ranks in the top ten universities in the world.

US Covid-19 Dataset

Covid Cases in US States

```
github_url <- "https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv"
state_covid_stats <- read_csv(github_url)

state_covid_stats <- state_covid_stats%>%
```

```
  subset(select=-(fips))

sample_n(state_covid_stats, 5)
```

```
## # A tibble: 5 x 4
##   date       state                    cases deaths
##   <date>     <chr>                    <dbl>  <dbl>
## 1 2020-10-10 Oregon                   36925    600
## 2 2020-08-25 Northern Mariana Islands    54      2
## 3 2020-05-15 Minnesota                14240    692
## 4 2020-03-28 Washington                4311    193
## 5 2020-04-26 Missouri                  6997    282
```

Summary

- **Who**
  - New York Times
  - State goverments
- **When**
  - Starts: 2020-01-21
  - Up to: 2020-10-21
  - Daily counts of covid cases between these two dates.
- **Where**
  - USA, USA, USA
  - *Other countries were polled, but we are restricting our data to only the US*
- **Why**
  - "*The New York Times is releasing a series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time. We are compiling this time series data from state and local governments and health departments in an attempt to provide a complete record of the ongoing outbreak.*" (Official statement)
  - Create an open source repository for reaserchers
- **How**
  - Reporters and data scientests working for the NYT track press releases along with state databases and aggregate the sources into several CSVs.
- **Who is included?**
  - Hierarchical step by step order to be counted
    - Individuals who were tested
    - Tested positive
    - Results were confirmed to the state government
    - State Goverment counted it in their data
    - NYT kept the data in their counts (*majority of the time*)
  - "Probable cases"
- **What evidence is there that everyone is present?**
  - There isn't, but the data is close/good enough for our purposes and there isn't higher quality open source data avalible.

Write up

An important aspect of our project was to compare state behavior dataset to a historical covid cases dataset from identical time period. The trouble is, state level covid datasets are fragmented and often follow different

guidelines. The time and effort it would take to track down the data from each state would likely exceed the scope of this course. This lead us to the NYT Covid 19 dataset.

The New York Times is a distinguished newspaper, so it can seem rather odd that is the creator of this dataset. To understand why The NYT put effort into creating a massive dataset, it is worth looking at the NYT Mission statement:

We seek the truth and help people understand the world.

As the world gets ever more connected the role of numbers in making sense of noise has risen exponentially. To help readers understand what is happening in the current crisis, the NYT has been producing a steady stream of covid-19 graphics to help visualize growth. Although the NYT is large enough to create a data strictly for internal use, its mission statement encourages sharing knowledge.

Although we filtered the data to the time range of `behavior_covid` (*Which was given and not updated*), the data spans from January 20th (first official case reported, in Snohomish County) to 2020-10-21 (day this was knit: `max(state_covid_stats$date)`).

Data is collected from several sources. Primarly the NYT relies on the data provided by state goverments. This does have several confounding variables.

Confounding Issues with Counting Cases

This subsection is necessary because what counts as a case is an essential part of our project, and Taylor could write pages about this.

As we previously mentioned, Covid-19 datasets are often fragmented and based upon several different standards.

Some states have different thresholds for a case to be reported, states change their records if thresholds are updated (sometimes removing cases), in addition lack of testing can be a confounding variable.

Another issue in recording is if an individual travels from one location where they were exposed to the virus, and die in another. States have different methods of counting deaths and cases so individuals can be double counted or not counted at all. This can be best seen in Nevada which reportedly counted only one in five case by not counting individuals from out of state. The NYT handles these situations by counting only the location where the individual became exposed in a seperate location by overwritting local counts in their CSVs.

State Population

2019 US Population

```
state_pop_stats <-read_csv("~/Math141/Projects/Project 1/nst-est2019-alldata.csv")

state_pop_stats <- state_pop_stats %>%
  select(NAME, POPESTIMATE2019) %>%
  rename(State=NAME)%>%
  rename(Population_2019=POPESTIMATE2019)

sample_n(state_pop_stats, 5)
```

```
## # A tibble: 5 x 2
##   State         Population_2019
##   <chr>                   <dbl>
## 1 New Hampshire         1359711
```

```
## 2 Washington            7614893
## 3 Colorado              5758736
## 4 West Region          78347268
## 5 New Mexico            2096829
```

Summary

- **Who**
  - US government
  - Census Bureau
- **When**
  - The population census data was collected on National Census Day, April 1st 2010.
  - Estimations were made on other dates
- **Where**
  - USA, USA, USA
- **Why**
  - In order to appropriate federal funding in an efficient manner the US gov organises a decennial census of the citizens of the US
  - Required by the US constitution.
- **How**
  - People with clipboards going to every household.
  - Estimations use some funky math with data from other government agencies
  - Official documentation
- **Who is included?**
  - Everyone*! (*In the US*)
  - Individuals that have responded to the census or have been included in suplemmentary statistics.
- **What evidence is there that everyone is present?**
  - It's a census, which means it covers everyone regardless of status.
  - The population estimate can contain some errors, however, It includes approximately 328,293,000 individuals. This means 100,000 individuals would be 0.03%. So it is highly unlikely that the numbers are off by a large enough factor to impact our data.

Write Up

In order to compare states covid cases, we needed to scale case numbers so that our data wouldn't over exaggerate more populous states. Using only a state's total number of cases would result in Wyoming appearing to have handled Covid better than California. Because of this we created a variable for cases per 100,000 individuals.

In order to create cases per 100,000, we wanted an estimate of each state's current population. R does contain a built in data frame of state populations called `state.x77`, however, the population estimates are from 1970. Read more here

---

```
data(state)
cat("Oregon Population in x.77 dataset: ", data.frame(state.x77)[1][37,]*1000)
```

```
## Oregon Population in x.77 dataset:  2284000
```

For context, Oregon's current population is roughly 4.2 million

---

This led us to find the census population estimate 2019. It is worth noting that there is no 2020 population estimate because the 2020 census is ongoing.

We will not focus too heavily on the 2010 census, as that is unlikely to be the greatest confounding variable, and is rather straightforward. The census is that the government counts every individual in a decennial event to reportion federal tools.

The census 2019 population estimate can be broken down into fairly simple formula with 4 variables: $PopBase + Births - Deaths + Migration = PopEst$. As has been previously mentioned, there is no 2020 data, so the estimation is based on the 2010 census and a series of statistics from other government bureaus.

The census bureau does not calculate the birth and death records. Instead it relies on the NCHS, a subdivision of the CDC, which reports US health related statistics. The NCHS collects the annual birth records in addition to death records. These are then added to the population base.

To calculate for migration the census breaks the problem into two categories, domestic and international. Domestic migration is calculated with data from three main sources: tax returns from the IRS, medicare data from CMS, and Social security data from the SSA. By using data from these sources the census can approximate the rate as a percentage at which individuals leave and enter a given state.

International migration contains 5 categories We will not delve into the specifics of each method as they are outside the scope of this project and exact information on methods is not easily accessible.

- *Foreign-Born Immigration*
  - Non US Citizens entering the US
  - Broken into two subcategories: from Mexico, not from Mexico
- *Foreign-Born Emigration*
  - Non US Citizens leaving the US
  - Broken into 7 unequal subcategories based on: gender, country, length stay
- *Migration between the United States and Puerto Rico*
  - "U.S. citizens residing in Puerto Rico" (Puerto Rican citizens are both US and not full US citizens)
  - Based on community surveys and net air traffic
- *Native-Born Migration*
  - US civillian Citizens living in another country
  - Using other countries census and population registers
- *Armed Forces Population*
  - Individuals in military service
  - Data sourced from: Defense Manpower Data Center. No exact methodology given.

There are ways to use the above to calculate change in a given demographic. But because we only use the state level data we will summarise by saying that the Census bureau calculates for each county and the sum of counties in a state is used to make the state population. More information on process can be found here.