Section 3: Week 6: Analyzing Statistics in Data Sets

Nate Bachmeier

TIM-8130: Data Mining

March 15, 2020

North Central University

# Analyzing Statistics in Data Sets

## Why is COVID-19 Topical

The world has become fixated on the Corona Virus Disease 2019 (COVID-19), as it continues to spread in a global pandemic. There is much uncertainty about how the disease will impact the general public leading to city-wide curfews, emptying of grocery stores, and the end of the bull market. One of the drivers of the mass panic comes from the media’s reporting of the mortality rates without sufficient context. For instance, original estimates put the likelihood of death at roughly 1% and have since risen as high as 7% (Wikipedia, 2020). However, these figures treat all entities as homogeneous, despite critical physical distinctions. Consider the differences between the immune systems of teenagers versus senior citizens, and how that changes their ability to cope with illness. Other features, such as pre-existing medical conditions and organ deterioration from cigarette smoking, influence the mortality calculus.

## What is COVID-19

The name novel coronavirus disease 2019 incorporates several artifacts, starting with ‘novel’ meaning new. Corona, meaning crown, refers to the spikes that protrude out of the protein envelope. What makes this instance unique is its ability to infect humans, versus most coronavirus, are lethal to birds and mammals, particularly across the farming industries (Schoeman & Fielding, 2016). Both Severe Respiratory Syndrom (SARs) of 2003 and Middle Eastern Respiratory Syndrom (MERS) of 2012 are recent COVID outbreaks. These diseases target the respiratory system and create symptoms ranging from a cold to pneumonia. Doctors can prescribe medication for these symptoms, but not for the underlying condition. A COVID protein exists as a single-stranded RNA genome that requires host cells to assist with its replication. Some medical professionals believe stopping that process, similar Remdesivir for Human Immunodeficiency Virus (HIV) patients, holds the solution (Cara, 2020). However, even if these existing drugs are successful, it can take years to gain approval from the Food and Drug Administration (FDA). Without a solution for medical science to produce a timely cure, the world needs to rely on ‘toughening it out.’ That does not install confidence across broad audiences as it leads to more questions than answers.

## Modeling the Mortality Rates

A heightened sense of fear exists as people do not understand their real risk of mortality. These challenges stem from media outlets taking a raw average of deaths divided by reported cases. Instead, the communities need preprocessing to cluster and classify members based on physical and demographic features. Kaiser Family Foundation (2016) states that West Virginia (36%), Mississippi (34%), and Kentucky (33%) have the highest rates of pre-existing conditions for nonelderly adults in the country. The number of deaths per thousand should naturally be higher in these states due to the impaired state. Another critical factor is the age of the population, with senior citizens having the highest risk. Perhaps the diversity of the community influences infection rates, such as homogeneous areas in Wyoming, that are unlikely to encounter international tourism. The population density of a city changes how people commute and the volume of interactions that transact every day. Each of these dimensions changes the probabilities of dying, along with the criticality of healthy individuals remaining so. Understanding these risks can increase targeted awareness and, in turn, save lives.

## Research Question

Having a specific research question is a prerequisite to delivering quality results. A core question that has frozen the country is understanding how lethal is this virus? Previous attempts from both mainstream and social media only consider the population as a whole, despite the significance of age. This dependency drives the need for a demographically adjusted average. Once a mechanism for approximating this value exists, a second objective is to discover specific communities within the United States that are at higher risk. For instance, Florida and Arizona are frequent retirement destinations with substantial senior populations. If the data allows for getting more granular, then discovering specific cities before the crisis occurs, can ensure resources are strategically provisioned. Since it is not economically feasible for support to cover the entire nation, these placement efficiencies minimize costs while maximizing aid. That directly translates into lives saved.

## The usefulness of statistics based mining

It can be tempting to begin data exploration with advanced mining algorithms that discover autocorrelations and use optimization theory to predict future values. Instead, starting with simple statistics can be a faster alternative to generate a back of the napkin model. For example, to predict how lethal COVID-19 will become could use an ensemble of classification and regression analysis to devise an estimate that includes all known data points. However, finding and normalizing these disjointed data sources would be tedious and require substantial amounts of effort. This approach also comes with additional risks to the research project because of the initial investment before it could fail fast. A better approach uses high-level correlation information to infer how things might play out. These quick statistical models might not be exact, but it will be in the ballpark. After having these initial insights, the business can more easily justify making more investments to improve accuracy. Descriptive statistics can also be useful for mining as it communicates the shape of the data set and any outliers.

## Statistical Strategies and Analysis

Statistical analysis uses correlations and variance to model the probability that some action will occur. For instance, when a person becomes sick and contagious, they spread the illness to others. A model can describe this interaction with Markov chains and Monte Carlo simulations to estimate the population size that becomes impacted. Researchers need to be careful that their model has a basis in science. Given one hundred random time series, a portion of them will correlate by chance. After confirming the premise is sound, the accuracy of the model needs to consider the likelihood that its prediction would happen. A standard solution is to look at the value distribution, and then see if the z-score (population) or t-score (sample) is within an acceptable multiple of the standard deviation. However, these simple evaluation methods rely on all data points being independent, which is not the case for many real-world scenarios that model time or nature.

## Data Sources

Without a crystal ball to predict the future, the next best alternative is to find correlated systems. China, South Korea, and Italy are at different stages of battling the outbreak and have reports of the macro trends across their countries. While Italy and China have comparable death rates for each age group, South Korea is almost half as high. Bendix (2020) notes that 140,000 Koreans have been tested, and attributes the lower death rate to prevention. An alternative analysis might conclude that more tests increased the denominator by accounting for more mild cases. These data sets feed into four weighted models, one per country and a final that uses the mean rate for each group. The 2018 US Census data contains state and county-level demographics at an age group resolution of 5-year intervals. That information was pivoted and regrouped into ranges of below 40s, 40s, 50s, 60s, 70s, and above 80 on a per county basis.

## Analysis and Results

A transformation combines the census age groups with the correlated mortality models to predict the death rates. For example, California has 80M people with 5M in their 50s, 4M in 60s, 2.5M in 70s, and 1.4M older than 79. Based on the statistical information from Korea, Italy, and China, that suggests 5,022 Californians would die per one percent infection rate. Assuming the worst-case scenario, that everyone in the state becomes infected, between 290 to 760 thousand residents would die. The Golden State would have the most deaths because of its enormous population. A normalization process divides the total death rates by the state population, and then by one hundred to approximate a single percentage change. This linear approximation does not account for the exponential curve that would exist as spreading the disease accelerates the population infection. If COVID-19 contaminated everyone in the United States, then between 2.6 to 6.7 million total people would die. These population losses would be uneven across the country, with Florida losing 1.72% of its population versus Utah only 1%.

## Challenges and Limitations

South Korea, China, and Italy operate differently than America in terms of their political and medical priorities. These distinctions might limit the applicability of their statistics directly to the US Census data. Another concern comes with different accounting methods between the countries, which is partially mitigated by only reviewing the lethal percentages. Despite the availability of other global statistics, such as the number of cases, those figures are less reliable due to economic restrictions that prevent mild cases from being reported. This model does not strongly consider people under forty, as their death rate is below 0.2%. However, the inclusion of more data sources, such as pre-existing conditions, could surface those people. While the results generally align with the hypothesis, a direct evaluation strategy does not exist and, through governmental intervention, unfold differently. Regardless of how the spread progresses, it is unlikely to be uniform and align perfectly with these predictions.

# References

Bendix, A. (2020, March 5). *South Korea has tested 140,000 people for the coronavirus. That could explain why its death rate is just 0.6% — far lower than in China or the US.* Retrieved from Business Insider: https://www.businessinsider.com/south-korea-coronavirus-testing-death-rate-2020-3

Cara, E. (2020, February 18). *Why Are HIV Drugs Being Used to Treat the New Coronavirus?* Retrieved from Gizmodo: https://gizmodo.com/why-are-hiv-drugs-being-used-to-treat-the-new-coronavir-1841770027

CCDC. (2020, February 20). *The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19)*. Retrieved from China CDC: http://www.ourphn.org.au/wp-content/uploads/20200225-Article-COVID-19.pdf

ISS. (2020, March 9). *Confirmed COVID-19 cases in Italy by age*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/2020\_coronavirus\_pandemic\_in\_Italy

KFF. (2016, December 12). *An Estimated 52 Million Adults Have Pre-Existing Conditions That Would Make Them Uninsurable Pre-Obamacare*. Retrieved from Kaiser Family Foundation: https://www.kff.org/health-reform/press-release/an-estimated-52-million-adults-have-pre-existing-conditions-that-would-make-them-uninsurable-pre-obamacare/

Schoeman, D., & Fielding, C. (2016). Coronavirus envelope protein: current knowledge. *Virology Jounal Volume 16 Issue 69*, 1-22.

US Census. (2013). *Population by Age and Sex for China.* International Data Base.

Wikipedia. (2020). *Demographics of Italy*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Demographics\_of\_Italy