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, grocery stores emptied, 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%. 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 level from 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 the availability of medical science to timely produce a cure, the world needs to rely on ‘toughening it out.’ That does not install confidence across the broad audience 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

Before collecting the various puzzle pieces, the first objective is to define a clear research question. Specifically, what is the actual mortality risk in Seattle, taking into account the population distribution and demographic makeup? Does this value align with the initial estimates from China? And how similar is their current trend to Italy? From evaluating the trajectory relative to China and Italy, this should forecast the impact on this major metropolitan city. An initial hypothesis is that the higher median population age will cause more deaths in China and Italy, than Seattle (US Census, 2013) (Wikipedia, 2020). These countries also have a first-mover disadvantage, versus Seattle has more time to prepare for the outbreak.

According to the CCDC (2020), patients over 60 years of age account for 80% of the fatalities, with a mortality rate of roughly 10-16%. Istituto Superiore di Sanita (ISS 2020), reports that approximately 90% of deaths are over 60 years with the lethality of the disease between 6.4 to 13.2%. Younger constituents are dying at comparable rates between 0 to 0.2%, confirming the highest risk to baby boomers and the silent generation. Using these figure