**RESEARCH ON COVID-19 INFECTIONS**

Jiaqi Cui(Jackie) and Jianyou Lu(Andy)

G11 AP Statistics

*Beijing National Day School, Beijing, China*

***Introduction***

The COVID-19 pandemic has struck the world for many times, causing huge loss and vast hysteria. This study relies on the topic of the COVID pandemic, and will be divided into two separating sections.

The first section examines data in China. According to statistics from the Chinese Center for Disease Control and Prevention, over 1.55 million Chinese got infected during 2022, and over 28518 of them died. Recently, however, WHO(World Health Organization) claims that the infectivity of COVID becomes similar to that of influenza, which is a breaking news reflecting people’s stronger immunity towards COVID. Given the claim, this part of the study aims to evaluate if the infectivity is similar for COVID and influenza in China. The null hypothesis is that the infectivity of COVID is the same as that of influenza in China, while the alternative hypothesis is that the infectivity of COVID is greater than that of influenza. Given that many people are infected by COVID again in 2023, and COVID is still considered as a more infectious virus by the Chinese government, we expect that the infectivity of it is greater than that of influenza. Therefore, the study could have statistically significant evidence to call for extra medical aids and reinforce protection for the COVID pandemic.

The second section provides a prediction for COVID infection. Time series analysis is the most rapidly developed method for predicting infectious diseases in recent years. A time series is a set of data arranged in chronological order. For example, observations of a variable x (t) or a set of measurements will be made at a series of moments t1, t2... tn, ( t is the independent variable and t1<t2<t3... <tn ) is the sequence set of discrete digit combinations, X(t1), X(t2) ... ,X(tn) is the time series. Time series analysis is to predict the future value based on its past value and present value according to the internal relationship between the series. Due to the occurrence, development and outcome of infectious diseases, the migration of time, the interference of external factors, the mutation of the virus itself, and the constant change of internal factors in the human body have formed a set of unique rules. Time series analysis contains the comprehensive effects of various factors, including unknown factors, affecting the incidence of infectious diseases in the time variable. Due to the outstanding advantages of this method, it has been widely used in the prediction and early warning analysis of infectious diseases. Therefore, this study uses the method of time series analysis to model and analyze the epidemic situation in 2022, and predicts the incidence trend of the epidemic, so as to provide a scientific basis for epidemic prevention and control.

***Background Research***

***For Section 1: Effective Reproductive Number***

To analyze the difference in infectivity between COVID and influenza, many researchers use R(Effective Reproductive Number) as a caliber. Explained in the article *Epidemic theory (effective & basic reproduction numbers, epidemic thresholds) & techniques for analysis of infectious disease data (construction & use of epidemic curves, generation numbers, exceptional reporting & identification of significant clusters)*, the Effective Reproductive Number is “the average number of secondary cases per infectious case in a population made up of both susceptible and non-susceptible hosts”. The calculated value of R is often compared with the number 1: if R>1, an infector could transmit the virus to more than one person on average, likely causing a dramatic outbreak; if R=1, the disease is endemic; if R<1, an infector will transmit disease to less than one person on average, which means that the disease will decline or even be eliminated.

***For Section 2: Time series prediction model***

Statistical analysis of time series data has a longer history than forecasting with time series data. Time series analysis originated in 1927, when mathematician Yuel introduced the concept of AR model for the first time. The model was then used to regulate the economic data obtained by observation and prediction. Subsequently, MA model was also proposed on the basis of AR model. Until the prediction work laid down the milestone of time series analysis. Since then, time series analysis has been widely used in the field of engineering, and its theory has gradually developed to a new height. In recent years, with the development of computer technology and information processing technology, the theory and method of time series analysis have been further developed. The steps involved in the modeling process, such as parameter estimation, model recognition and order determination, can be realized simply and conveniently through the computer operation, which is very practical. Early time series analysis methods usually reveal the law of phenomena changing over time through intuitive historical data comparison or drawing observation, namely the so-called descriptive time series analysis. The application of traditional time series analysis in practice is mainly deterministic time series analysis methods, including exponential smoothing method, sliding average method, time series decomposition and so on. However, in real life, the influence of many uncertain factors is getting bigger and bigger. In 1970, Box and Jenkins proposed the time series analysis method based on stochastic theory, which made the time series analysis theory rise to a new height and the accuracy of prediction is also greatly improved. The theory and method of control show great superiority in the processing and analysis of dynamic data, the processing and extraction of complex information and the prediction of future. Time series analysis is the use of this set of data, the application of mathematical statistics to process, in order to predict the future development of things. It includes exponential smoothing method, moving average method (MA), autoregressive model (AR), autoregressive moving average model (ARMA) and the summation autoregressive moving average model, namely ARIMA model, which is widely used at present.

For the study of non-stationary time series, ARIMA model is a very important one. It uses difference to transform non-stationary time series into stationary time series, thus transforming ARIMA process into ARMA process. ARIMA is widely used in many fields at home and abroad. For example, ARIMA model was used in the United States to monitor life expectancy in 2010, and it was also used in Australia to analyze the relationship between the incidence of Ross River virus and climate change. ARIMA model is also widely used in various fields in China. In the medical field, many scholars use this model for epidemic prediction and monitoring. The ARIMA model was also used in this study to analyze and predict the spread of Covid-19.