STA135-Final Project

Shih-Chi Chen, YiJin Chi, Youwei Wang

6/5/2021

1. Introduction

The COVID-19 pandemic has caused millions of deaths worldwide since it first started in 2019. Until now, this pandemic problem still exists in many countries including USA. The goal of our project is to analyze if it exists a significant difference between the mean of different features (Total Cases, Total Deaths, Active Cases, Critical, Motality Recovery Ratio) and Time Points (30-Mar-20, 15-Apr-20, 25-Apr-20) by using MANOVA. Afterwards, in-depth analysis on if it exists a significant difference between individual means of each feature and time points is applied by using ANOVA.

2. Data, Models, and Methods

Data Description:

The COVID-19 data used here is publicly and available from Worldometer website https://www.worldometers.info/coro, and those data were captured on the next day to these specified dates: March 30, April 15, and April 25, 2020. COVID-19 total cases less than 500 or countries with missing data were omitted from the analysis in order to keep a good representability of each variable. Number of countries included in the analysis was 56 countries on March 30, 82 countries on April 15, and 91 countries on April 25.

The dataset include the variables: total cases and total deaths that refers to total cases and deaths confirmed with COVID-19; active cases refers to total number of open cases (mild, serious, or critical); critically ill cases refers to number of serious/critically ill cases; mortality recovery ratio refers to the ratio between total deaths to totalrecovered patients.

Methods:

Apply the One-Way MANOVA analysis as the multivariate generalization of univariate analysis of variance to compare the vector mean of different features as a whole between three groups by using Time Point as the dependent variable and see if there is a significant difference. Here, we assume there is no outliers (check for the outliers and decide not to remove them), multivariate normality, homogeneity of variances, and linearity.

Afterwards, we apply the One-Way ANOVA to analyze individual means of each feature and time points.

3. Results

MANOVA Assumptions:

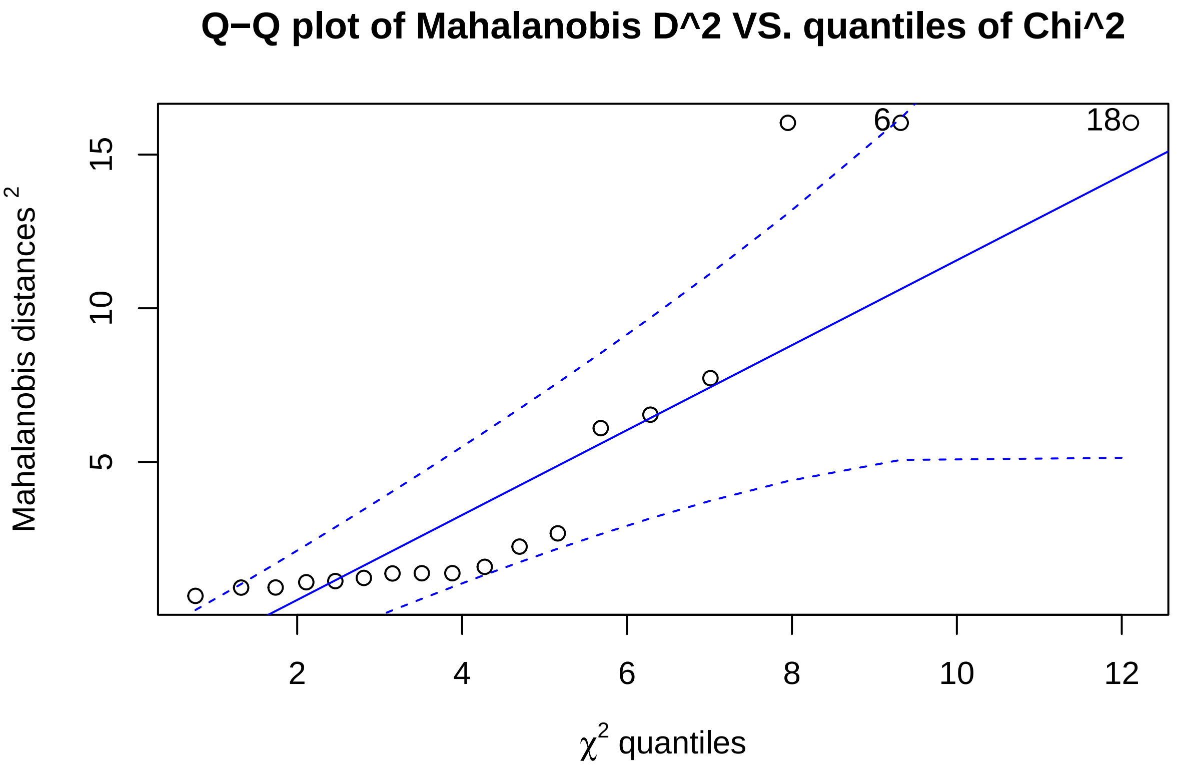
(1) Observations are randomly and independently sampled from the population.

(2) Each dependent variable has an interval measurement.

(3) Dependent variables are multivariate normally distributed within each group of the independent variables (which are categorical).

(4) The population covariance matrices of each group are equal (this is an extension of homogeneity of variances required for univariate ANOVA).

(5) There are no univariate or multivariate outliers.



```
## [1] 18  6
```

We find some outliers on the qqplot under $\alpha = 0.05$, however, we decide not to remove it because of the small sample size. When we have small sample size, we often do not want to lose a data point given the proportion it represents on the sample size.

Also, from the qqplot, it shows that the normality is violated, but we will assume the assumption of MANOVA (normality) holds for further study.

One way MANOVA analysis

MANOVA is the multivariate generalization of univariate analysis of variance.

Objective: Here, we want to compare the mean of different features as a whole between three groups by Time Point and see if there is a significant difference.

Test the following hypotheses: $H_0 : \vec{\mu_1} = \vec{\mu_2} = \vec{\mu_3}$

H_a : Not all the vector means among different groups are the same.

```
##                   Df  Pillai approx F num Df den Df Pr(>F)
## factor(timepoint)  2 0.45492  0.70663     10     24 0.7096
## Residuals         15
```

From the above table, we can find the test for the equality of the mean vectors has a very large p-value (0.7096) > α=0.05, so we would fail to reject the H_0.

One Way ANOVA

We look at the individual means using a Bonferroni adjustment to the level of significance. If alpha= 0.05, then to examine for differences between the individual means using ANOVA, a level of significance of 0.05/5 = 0.01 could be used.

Total_Cases ~ timepoint $H_0 : \mu_{totalcase1} = \mu_{totalcase2} = \mu_{totalcase3}$

H_a : At least one μ is not equal.

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## factor(timepoint)  2 6.028e+10 3.014e+10   0.408  0.672
## Residuals         15 1.108e+12 7.385e+10
```

Fail to reject H_0.

Total_Death ~ timepoint $H_0 : \mu_{totaldeath1} = \mu_{totaldeath2} = \mu_{totaldeath3}$

H_a : At least one μ is not equal.

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## factor(timepoint)  2 1.720e+08  85983170   0.407  0.673
## Residuals         15 3.166e+09 211081165
```

Fail to reject H_0.

Active_Cases ~ timepoint $H_0 : \mu_{activecase1} = \mu_{activecase2} = \mu_{activecase3}$

H_a : At least one μ is not equal.

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## factor(timepoint)  2 3.816e+10 1.908e+10   0.363  0.701
## Residuals         15 7.879e+11 5.253e+10
```

Fail to reject H_0.

Critical ~ timepoint $H_0 : \mu_{critical1} = \mu_{critical2} = \mu_{critical3}$

H_a : At least one μ is not equal.

```
##                    Df    Sum Sq  Mean Sq F value Pr(>F)
## factor(timepoint)  2   9554665  4777333   0.201   0.82
## Residuals         15 356316537 23754436
```

Fail to reject H_0.

MR_ratio ~ timepoint $H_0 : \mu_{MRratio1} = \mu_{MRratio2} = \mu_{MRratio3}$

H_a : At least one μ is not equal.

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## factor(timepoint)  2    496   248.2   0.852  0.446
## Residuals         15   4368   291.2
```

Fail to reject H_0.

Based on 5 One-way ANOVA outputs, there are all end up with fail to reject the H_0, and we conclude that individual means of each feature and time points are not significant difference under Bonferroni adjustment $0.05/5 = 0.01$.

4. Conclusion and Future work

Conclusion

First, we use qq plot, residual plots, etc to detect outliers and check normality. There are some outliers and the normality seems violatated, but for the academic purpose here, we assume the assumptions hold.

Second, we run One-way MANOVA testing to see if vector means of multiple features are the same among different groups by time point. The p-value from the result shows that there is not enough evidence to conclude the vector means among difference groups are the same.

Third, we further run One-way ANOVA testing to see if individual feature means are the same among different groups by time point. None of the p-value from anova testings makes us confident enough to say that any of the feature means is the same among different groups.

Future Work

From our analysis, even though we did not find significant evidence to support the difference in either the vector means among groups or individual feature means among the groups, we should keep in mind that our sample size is very small and may not be critical enough for the finding. In the future, we could gather more relevant data to increase the size of our sample pool and rerun the test to see if the new result would have enough evidence show that the vector means or individual feature means would differ among groups by time point.

Appendix: All code for this report

```r
knitr::opts_chunk$set(echo = F, fig.align = 'center')
#set the data
cov19_all<-data.frame(timepoint=factor(c(1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3)),Total_Cases=c(515,954.8,

#outliers
the.data<-subset(cov19_all, select = -c(1)) #remove time point
md <- mahalanobis(the.data, center = colMeans(the.data), cov = cov(the.data))
alpha <- .05
cutoff <- (qchisq(p = 1 - alpha, df = ncol(the.data)))
#outliers
names_outliers_MH <- which(md > cutoff)
library(RVAideMemoire)
mqqnorm(the.data,main='Q-Q plot of Mahalanobis D^2 VS. quantiles of Chi^2')
# MANOVA
save <-manova(formula = cbind(Total_Cases, Total_Death,
    Active_Cases, Critical, MR_ratio) ~ factor(timepoint), data = cov19_all)

summary(save)
mod.fit0<-aov(formula = Total_Cases ~ factor(timepoint), data = cov19_all)
summary(mod.fit0)
mod.fit1<-aov(formula = Total_Death ~ factor(timepoint), data = cov19_all)
summary(mod.fit1)
mod.fit2<-aov(formula = Active_Cases ~ factor(timepoint), data = cov19_all)
summary(mod.fit2)
mod.fit3<-aov(formula = Critical ~ factor(timepoint), data = cov19_all)
summary(mod.fit3)
mod.fit4<-aov(formula = MR_ratio ~ factor(timepoint), data = cov19_all)
summary(mod.fit4)
```