STA104 FINAL PROJECT

Class: STA104 - Spring 2021
Professor: Maxime Pouokam
Members: Josemanuel Vega (#ID:916627124)
Kate Johnson (#ID:918200073)
Shih-Chi Chen (#ID:917995392)

1. Introduction

In light of current events and with readily available data made daily over the COVID-19 pandemic affecting us world-wide, we are investigating COVID-19 data that has been made over 3 daily observations and aggregated statistics from those time periods. Our aim is to analyze if it exists a independence between the Parameter (Min, Q1, Med, Mean, Q2, Max) and Time Points (30-Mar-20, 15-Apr-20, 25-Apr-20). To do so, we will use a χ^2 test for independence, and through some analysis of the data beforehand determine to use a parametric or non-parametric method. Once computed, we will use a Tukey cutoff to determine the significant of the value we had computed through pairwise comparisons. Prior assumptions that we have made are that the data itself used to create the aggregated statistics were selected at random, and $e_{ij} \geq 5$ for all i,j.

Interesting question:

(1) Are Parameter and Time Point independent?
(2) Which Parameter and Time Point group has significant difference?

2. Data, and Methods

Data Description:

The data that we are using was collected from the Worldometer website (found here https://www.worldometers.info/coronavirus/) for days March 30, April 15, and April 25. For each of the day, 6 parameters were considered. These parameters are: Min, Q1, Med, Mean, Q2, and Max. Each parameter is summed up from Total cases, Active cases, Total deaths, Critically, and Motality Recovery Ratio.

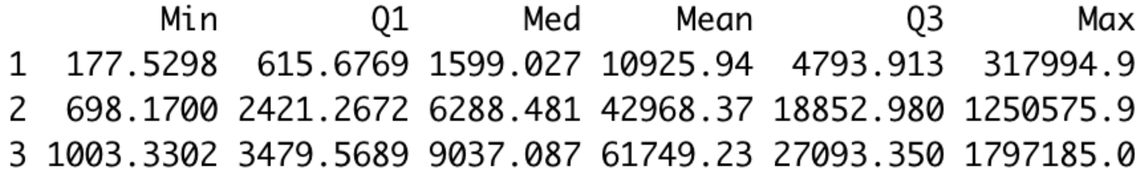
Methods:

(1) Parametric or non-Parametric χ^2 test
(2) Pairwise comparisons by using tukey inspired cutoff.

3. Result

- Contingency table

1 = 30-Mar-20 , 2= 15-Apr-20 , 3=25-Apr-20



```
        Min        Q1      Med     Mean        Q3       Max
1  177.5298  615.6769 1599.027 10925.94  4793.913  317994.9
2  698.1700 2421.2672 6288.481 42968.37 18852.980 1250575.9
3 1003.3302 3479.5689 9037.087 61749.23 27093.350 1797185.0
```

- Check the assumption of parameteric χ^2 test for independent:

(1) Random sample was taken

(2) eij≥ 5 for all i,j

eij Table

From the eij table below, it shows that all eij≥ 5.

```
##          Min        Q1      Med     Mean        Q3       Max
## 1  177.5298  615.6769 1599.027 10925.94  4793.913  317994.9
## 2  698.1700 2421.2672 6288.481 42968.37 18852.980 1250575.9
## 3 1003.3302 3479.5689 9037.087 61749.23 27093.350 1797185.0
```

Since the assumptions are hold, parameteric χ^2 test is used.

- Parameteric χ^2 test:

State the null and alternative hypothesis:

H_0: Time point and Parameter are independent.

H_a: Time point and Parameter are dependent.

Test statistic and p-value:

$\chi^2_{S,OBS} = 38155$

p-value < 2.2e-16

Since p-value is small, we would reject H_0 and conclude that Time point and Parameter are dependent.

- Pairwise comparisons

Since Time point and Parameter are dependent, pairwise comparisons are needed.

(1) Time point 1 (30-Mar-20) and 2 (15-Apr-20).

The Z values for comparing the row values for each column (the Parameter of the subject given the Time point (30-Mar-20, 15-Apr-20)) are show below:

Z_{Min}	Z_{Q1}	Z_{Med}	Z_{Mean}	Z_{Q3}	Z_{Max}
24.16154	38.19976	42.41920	100.25153	88.51343	−148.76474

From these Z values, the most difference is in Max value for Time point 1 (30-Mar-20) vs. Time point 2 (15-Apr-20) because the largest absolute value of Z values obtained.

Tukey inspired cutoff=148.7 7 (Use α=0.01)

There is one significant difference, which is Max value for Time point 1 (30-Mar-20) vs. Time point 2 (15-Apr-20) because absolute value of z-scores are above the Tukey cutoff. In addition, since the z-score is negative, the max value of Time point 1 is less than Time point 2.

(2) Time point 1 (30-Mar-20) and 3 (25-Apr-20)

The Z values for comparing the row values for each column (the Parameter of the subject given the Time point (30-Mar-20, 25-Apr-20)) are show below:

Z_{Min}	Z_{Q1}	Z_{Med}	Z_{Mean}	Z_{Q3}	Z_{Max}
35.51623	52.73514	64.14271	128.25023	106.24858	−189.75930

From these Z values, the most difference is in Max value for Time point 1 (30-Mar-20) vs. Time point 3 (25-Apr-20) because the largest absolute value of Z values obtained.

Tukey inspired cutoff=189.7 (Use α=0.01)

There is one significant difference, which is Max value for Time point 1 (30-Mar-20) vs. Time point 3 (25-Apr-20) because absolute value of z-scores are above the Tukey cutoff. In addition, since the z-score is negative, the max value of Time point 1 is less than Time point 3.

(3) Time point 2 (15-Apr-20) and 3 (25-Apr-20)

The Z values for comparing the row values for each column (the Parameter of the subject given the Time point (15-Apr-20, 25-Apr-20)) are show below:

Z_{Min}	Z_{Q1}	Z_{Med}	Z_{Mean}	Z_{Q3}	Z_{Max}
9.884449	12.800213	22.629120	27.492511	13.587435	−38.854668

From these Z values, the most difference is in Max value for Time point 2 (15-Apr-20) vs. Time point 3 (25-Apr-20) because the largest absolute value of Z values obtained.

Tukey inspired cutoff=38.8 7 (Use α=0.01)

There is one significant difference, which is Max value for Time point 2 (15-Apr-20) vs. Time point 3 (25-Apr-20) because absolute value of z-scores are above the Tukey cutoff. In addition, since the z-score is negative, the max value of Time point 2 is less than Time point 3.

4. Conclusion and Future work

Since the assumptions are hold, the parametric test for independence between Parameter and Time point is used. And from the large test statistic obtained, it shows that Parameter and Time point are not independent. Afterwards, the comparison between Z values and Tukey inspired cutoff shows that Max value for each Time point pair has significant difference.

For the future work, since we just found that the Max value has significant difference between each Time point pair, we would be interested in studying if the mean of total deaths has significant difference between each Time point pair or not.

#Appendix Code

```
cov19_time <- matrix(c(616.01,1842.428,3869.335,23319.4204,11936.83,294523,69
1.01,2348.05,6626.17,41932.7483,17230.6675,1252976.54,572.01,2326.035,6429.09
,50391.3722,21572.745,1818256.28),ncol=6,byrow=TRUE)
colnames(cov19_time) <- c("Min","Q1","Med","Mean","Q3","Max")
rownames(cov19_time) <- c(1,2,3)
table_time <- as.table(cov19_time)

the.test = chisq.test(table_time,correct = FALSE)
eij = the.test$expected
eij
chi.sq.obs = as.numeric(the.test$statistic)
#Group 1 and 2
cov19_time12 <- matrix(c(616.01,1842.428,3869.335,23319.4204,11936.83,294523,
691.01,2348.05,6626.17,41932.7483,17230.6675,1252976.54),ncol=6,byrow=TRUE)
colnames(cov19_time12) <- c("Min","Q1","Med","Mean","Q3","Max")
rownames(cov19_time12) <- c(1,2)
table_time12 <- as.table(cov19_time12)
n = sum(table_time12)
ni. = rowSums(table_time12)
n.j = colSums(table_time12)
all.pjG1 = table_time12[1,]/ni.[1] #all conditional probabilites for row 1
all.pjG2= table_time12[2,]/ni.[2] #all conditional probabilites for row 2
all.pbar = n.j/n #all probabilities regardless of group
all.Zij = c(all.pjG1 - all.pjG2)/sqrt(all.pbar*(1-all.pbar)*(1/ni.[1] + 1/ni.
[2])) #The z-test-statistics
all.Zij
cov19_12<-data.frame(timepoint=factor(c(1,1,1,1,1,1,2,2,2,2,2,2)),parameter=f
actor(c("Min","Q1","Med","Mean","Q3","Max")),totalnumb=c(616.01,1842.428,3869
.335,23319.4204,11936.83,294523,691.01,2348.05,6626.17,41932.7483,17230.6675,
1252976.54))

R=3000
r.perms.cutoff = sapply(1:R,function(i){
    perm.data = cov19_12
    perm.data$timepoint = sample(perm.data$timepoint,nrow(perm.data),replace
= FALSE)
    row.sum = rowSums(table_time12)
    col.sum = colSums(table_time12)
    all.pji = table_time12[1,]/row.sum[1]
    all.pji.= table_time12[2,]/row.sum[2]
    all.pbar = col.sum/sum(row.sum)
    all.Zij = c(all.pji - all.pji.)/sqrt(all.pbar*(1-all.pbar)*(1/row.sum[1]
+ 1/row.sum[2]))
    Q.r = max(abs(all.Zij))
    return(Q.r)
})
alpha = 0.01
cutoff.q = as.numeric(quantile(r.perms.cutoff,(1-alpha)))
#Group 1 and 3
cov19_time13 <- matrix(c(616.01,1842.428,3869.335,23319.4204,11936.83,294523,
572.01,2326.035,6429.09,50391.3722,21572.745,1818256.28),ncol=6,byrow=TRUE)
colnames(cov19_time13) <- c("Min","Q1","Med","Mean","Q3","Max")
rownames(cov19_time13) <- c(1,3)
table_time13 <- as.table(cov19_time13)
n = sum(table_time13)
ni. = rowSums(table_time13)
n.j = colSums(table_time13)
all.pjG1 = table_time13[1,]/ni.[1] #all conditional probabilites for row 1
all.pjG2= table_time13[2,]/ni.[2] #all conditional probabilites for row 2
all.pbar = n.j/n #all probabilities regardless of group
all.Zij = c(all.pjG1 - all.pjG2)/sqrt(all.pbar*(1-all.pbar)*(1/ni.[1] + 1/ni.
[2])) #The z-test-statistics
all.Zij
cov19_13<-data.frame(timepoint=factor(c(1,1,1,1,1,1,3,3,3,3,3,3)),parameter=f
actor(c("Min","Q1","Med","Mean","Q3","Max")),totalnumb=c(616.01,1842.428,3869
.335,23319.4204,11936.83,294523,572.01,2326.035,6429.09,50391.3722,21572.745,
1818256.28))

R=3000
r.perms.cutoff = sapply(1:R,function(i){
perm.data = cov19_13
perm.data$timepoint = sample(perm.data$timepoint,nrow(perm.data),replace = FA
LSE)
row.sum = rowSums(table_time13)
col.sum = colSums(table_time13)
all.pji = table_time13[1,]/row.sum[1]
all.pji.= table_time13[2,]/row.sum[2]
all.pbar = col.sum/sum(row.sum)
all.Zij = c(all.pji - all.pji.)/sqrt(all.pbar*(1-all.pbar)*(1/row.sum[1] + 1/
row.sum[2]))
Q.r = max(abs(all.Zij))
return(Q.r)
})
alpha = 0.01
cutoff.q = as.numeric(quantile(r.perms.cutoff,(1-alpha)))
#Group 2 and 3
cov19_time23 <- matrix(c(691.01,2348.05,6626.17,41932.7483,17230.6675,1252976
.54,572.01,2326.035,6429.09,50391.3722,21572.745,1818256.28),ncol=6,byrow=TRU
E)
```

```r
colnames(cov19_time23) <- c("Min","Q1","Med","Mean","Q3","Max")
rownames(cov19_time23) <- c(2,3)
table_time23 <- as.table(cov19_time23)
n = sum(table_time23)
ni. = rowSums(table_time23)
n.j = colSums(table_time23)
all.pjG1 = table_time23[1,]/ni.[1] #all conditional probabilites for row 1
all.pjG2= table_time23[2,]/ni.[2] #all conditional probabilites for row 2
all.pbar = n.j/n #all probabilities regardless of group
all.Zij = c(all.pjG1 - all.pjG2)/sqrt(all.pbar*(1-all.pbar)*(1/ni.[1] + 1/ni.[2])) #The z-test-statistics
all.Zij
cov19_23<-data.frame(timepoint=factor(c(2,2,2,2,2,2,3,3,3,3,3,3)),parameter=factor(c("Min","Q1","Med","Mean","Q3","Max")),totalnumb=c(691.01,2348.05,6626.17,41932.7483,17230.6675,1252976.54,572.01,2326.035,6429.09,50391.3722,21572.745,1818256.28))

R=3000
r.perms.cutoff = sapply(1:R,function(i){
perm.data = cov19_23
perm.data$timepoint = sample(perm.data$timepoint,nrow(perm.data),replace = FALSE)
row.sum = rowSums(table_time23)
col.sum = colSums(table_time23)
all.pji = table_time23[1,]/row.sum[1]
all.pji.= table_time23[2,]/row.sum[2]
all.pbar = col.sum/sum(row.sum)
all.Zij = c(all.pji - all.pji.)/sqrt(all.pbar*(1-all.pbar)*(1/row.sum[1] + 1/row.sum[2]))
Q.r = max(abs(all.Zij))
return(Q.r)
})
alpha = 0.01
cutoff.q = as.numeric(quantile(r.perms.cutoff,(1-alpha)))
```