STA 104 Take-Home Project

Zheng Wang (916184463)
tzwwang@ucdavis.edu
Junxi Chen(917528808)
jxichen@ucdavis.edu

Amy T. Kim Instructor
STA104: Nonparametric Statistics
University of California, Davis
March 2nd, 2021

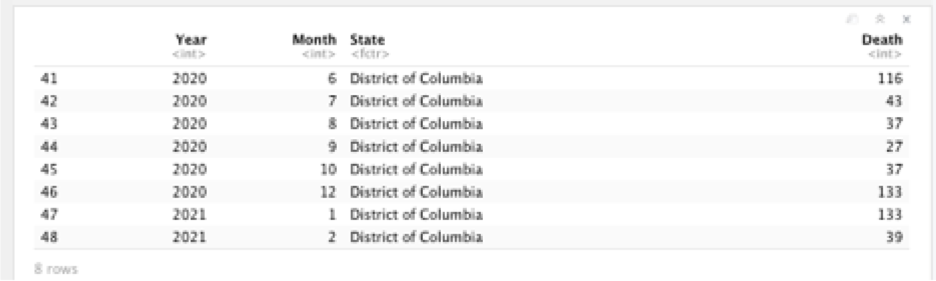
Question 1

1.Introduction:

For this first question, we have the data about different states for different death rates because each state has different regulation and infrastructure. And we choose the state of New York, Columbia, Mexico, and New Jersey. Then, we are trying to test the difference and independence in death rate for various groups, and to know if there is a difference in death rate between each state because of different policies. Hence, we use the simultaneous inference with cutoff, Kruskal-Wallis Test, and permutation test to our goal.

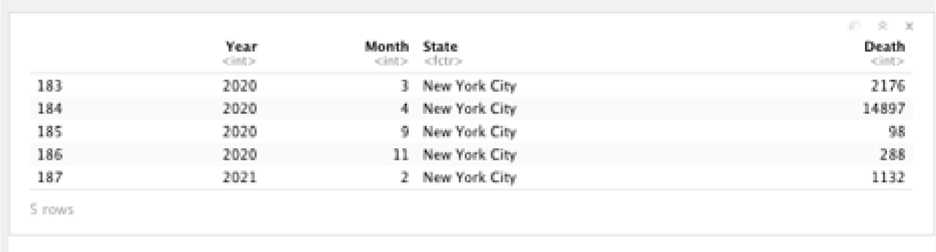
2.summary of data

We took the subset of those four states



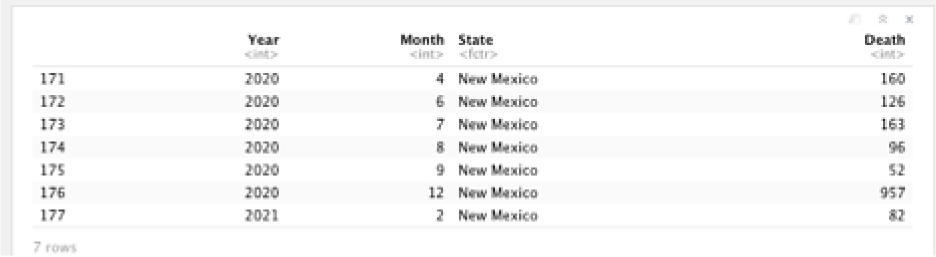
	Year	Month	State	Death
	<int>	<int>	<fctr>	<int>
41	2020	6	District of Columbia	116
42	2020	7	District of Columbia	43
43	2020	8	District of Columbia	37
44	2020	9	District of Columbia	27
45	2020	10	District of Columbia	37
46	2020	12	District of Columbia	133
47	2021	1	District of Columbia	133
48	2021	2	District of Columbia	39

8 rows



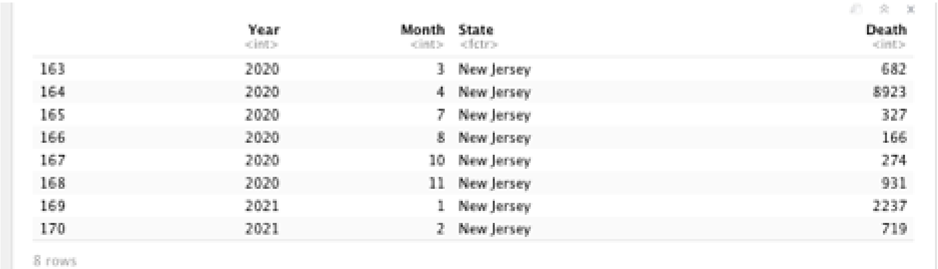
	Year	Month	State	Death
	<int>	<int>	<fctr>	<int>
183	2020	3	New York City	2176
184	2020	4	New York City	14897
185	2020	9	New York City	98
186	2020	11	New York City	288
187	2021	2	New York City	1132

5 rows



	Year	Month	State	Death
	<int>	<int>	<fctr>	<int>
171	2020	4	New Mexico	160
172	2020	6	New Mexico	126
173	2020	7	New Mexico	163
174	2020	8	New Mexico	96
175	2020	9	New Mexico	52
176	2020	12	New Mexico	957
177	2021	2	New Mexico	82

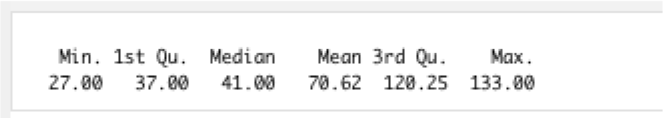
7 rows



	Year <int>	Month <int>	State <fctr>	Death <int>
163	2020	3	New Jersey	682
164	2020	4	New Jersey	8923
165	2020	7	New Jersey	327
166	2020	8	New Jersey	166
167	2020	10	New Jersey	274
168	2020	11	New Jersey	931
169	2021	1	New Jersey	2237
170	2021	2	New Jersey	719

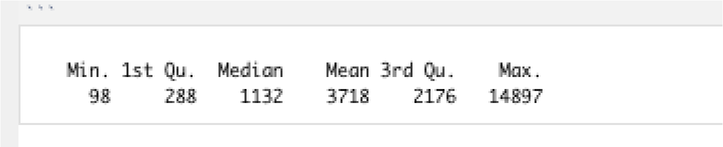
8 rows

The mean for state of Columbia death rate is 70.625, standard deviation is 47.46107
The mean for state of New York is 3718.2, and the standard deviation is 6302.866
The mean for the state of New Mexico is 233.7143 and the standard deviation is 321.5021
The mean for the state of New Jersey is 1782.375, and the standard deviation 2958.465.
And the histogram and boxplot of those states of death rate
Summary of death in Columbia:



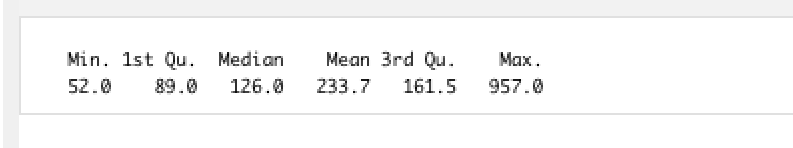
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  27.00   37.00   41.00   70.62  120.25  133.00 
```

Summary of death in new york



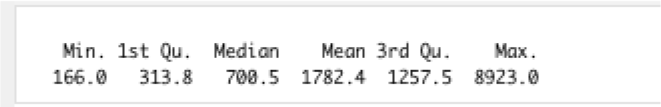
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     98     288    1132    3718    2176   14897 
```

Summary of death in mexico:

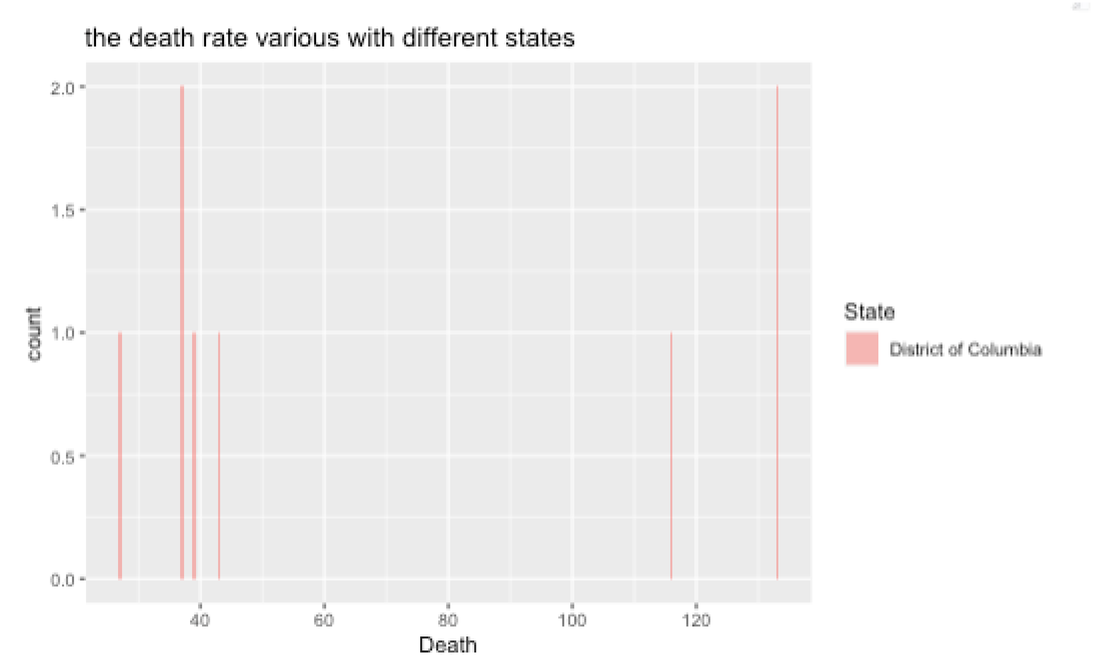


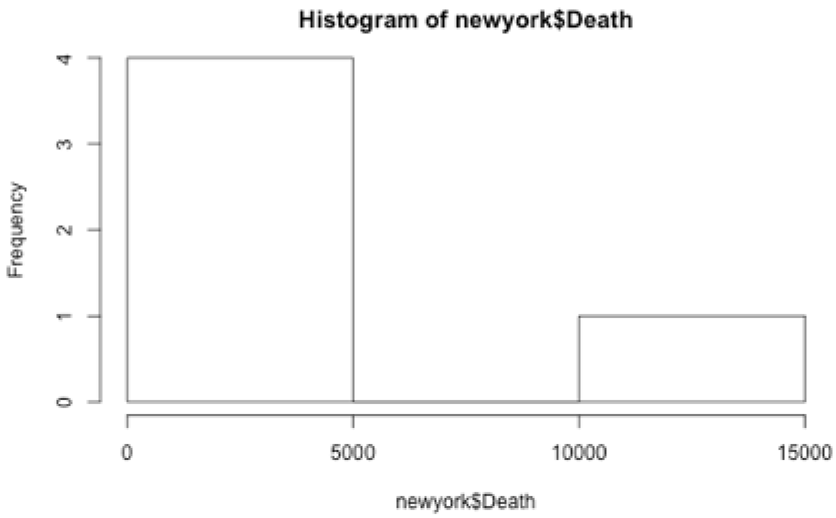
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   52.0    89.0   126.0   233.7   161.5   957.0 
```

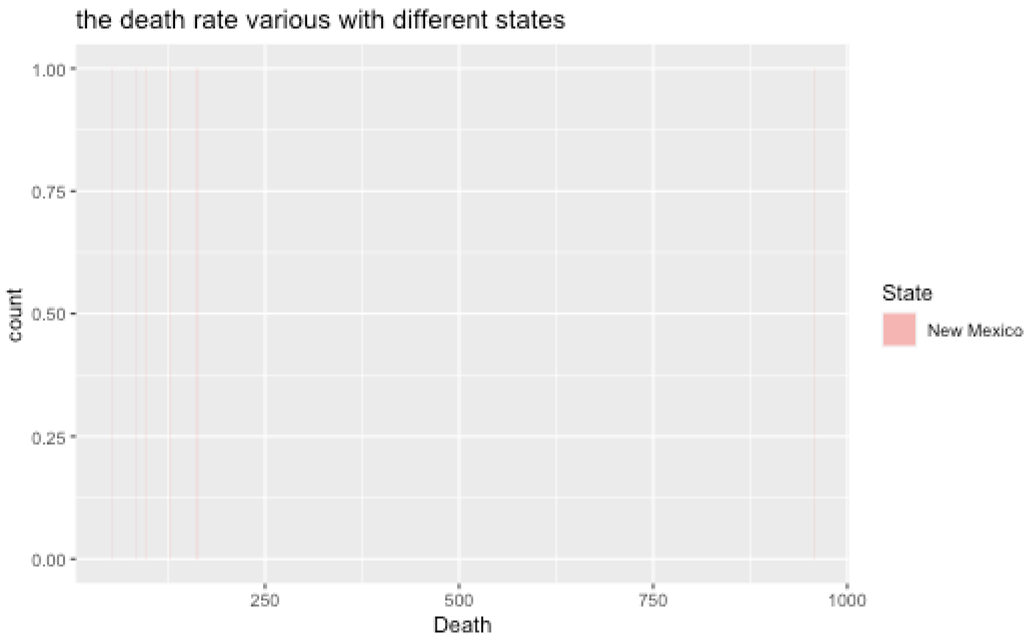
Summary of death in new jersey

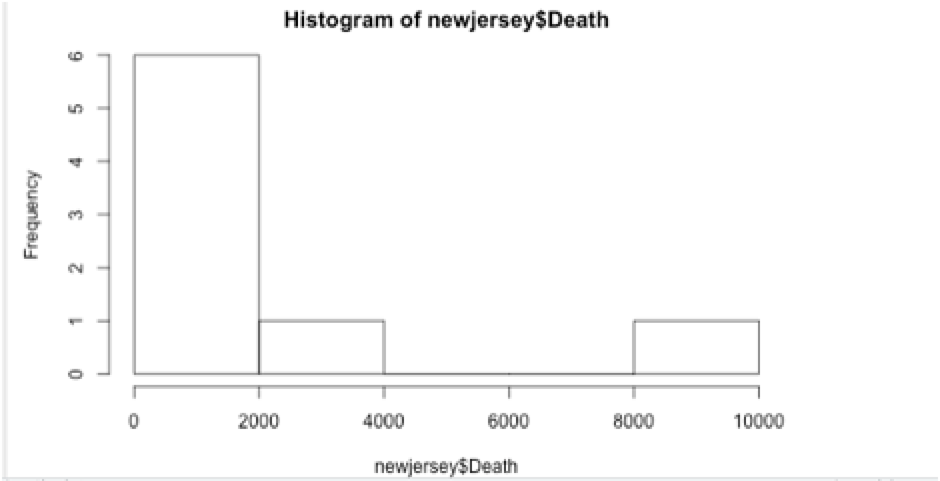


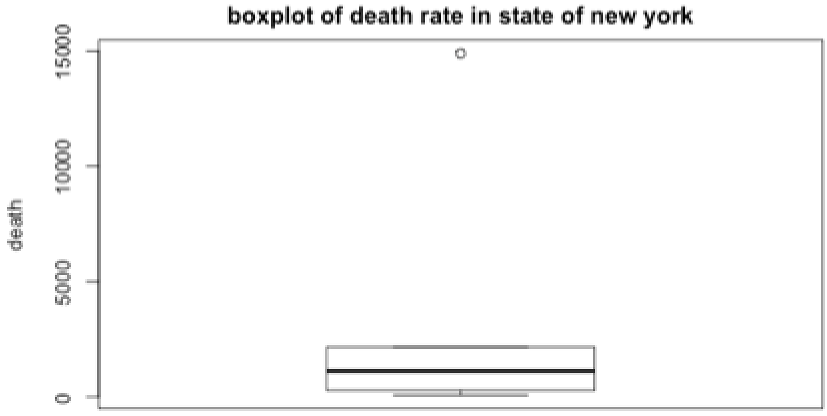
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  166.0   313.8   700.5  1782.4  1257.5  8923.0 
```


the death rate various with different states
2.0
1.5
1.0
0.5
0.0
count
40
60
80
100
120
Death
State
District of Columbia

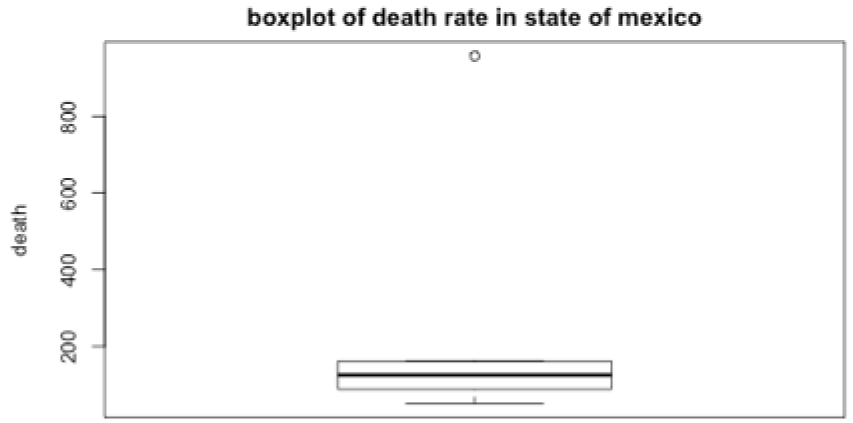

Histogram of newyork$Death
4
3
2
1
0
Frequency
0
5000
10000
15000
newyork$Death

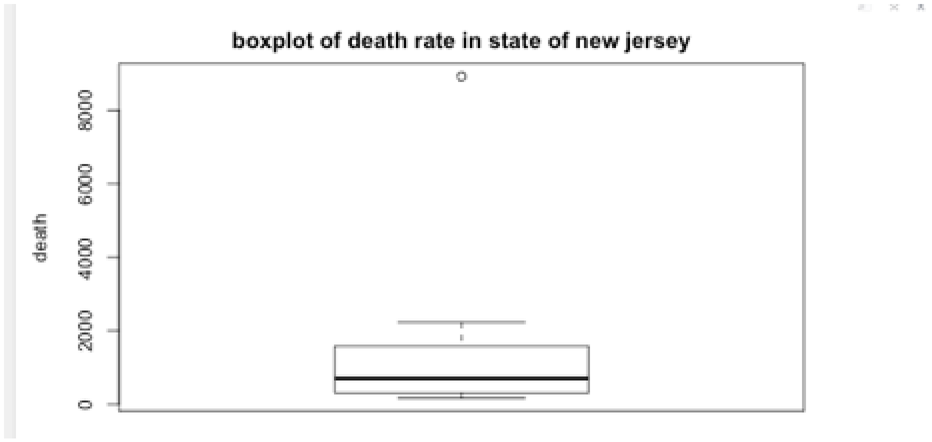

the death rate various with different states
count
1.00
0.75
0.50
0.25
0.00
250
500
750
1000
Death
State
New Mexico

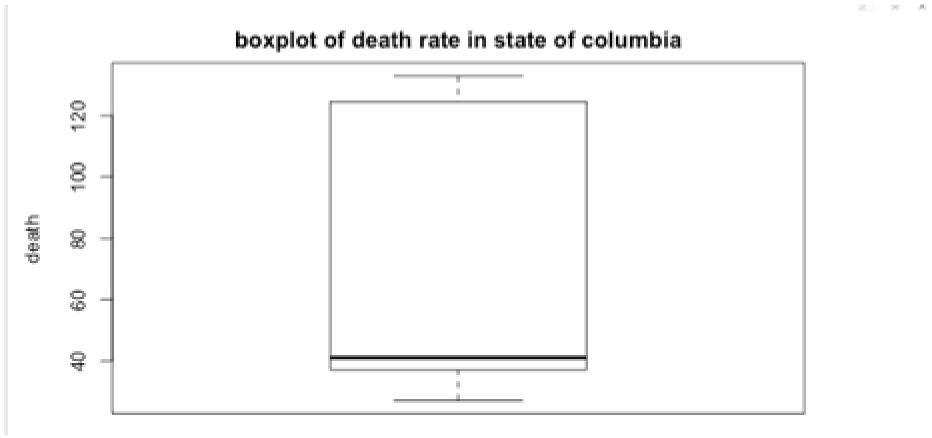

Histogram of newjersey$Death
Frequency
0
1
2
3
4
5
6
0
2000
4000
6000
8000
10000
newjersey$Death


boxplot of death rate in state of new york
death
0
5000
10000
15000

3.Analysis:



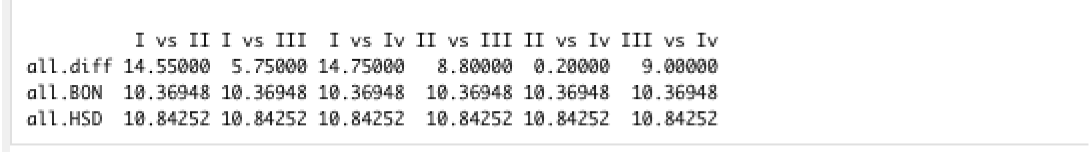




We can see the distribution of those data are different from each other. The distribution of death numbers in Columbia has more variance compared to other groups, and because those data have different ranges. And as we can see, all those distributions have large variances, the data is mainly located in the begin and end.

3.Analysis:

Null hypothesis: we assume that the death rate is dependent with the state variables.
Alternative hypothesis: the death rate is independent of the different states.
For another one, null hypothesis is there are no difference for those four groups an alternative hypothesis is that at least one of them is different.
and we calculate by the permutation test, for that specific permutation, our test-statistic was 0.4623811, and we assume the R is 2000, we calculate the p value is 0.10. Furthermore, if we use the Kruskal-Wallis test to calculate this, assuming the R = 2000, we calculate the p value is close to 0.
And we also use the simultaneous inference to calculate the difference existing for different groups.



```
          I vs II I vs III  I vs Iv II vs III II vs Iv III vs Iv
all.diff 14.55000  5.75000 14.75000   8.80000  0.20000   9.00000
all.BON  10.36948 10.36948 10.36948  10.36948 10.36948  10.36948
all.HSD  10.84252 10.84252 10.84252  10.84252 10.84252  10.84252
```

Hence, we can see from the data, if we assume alpha is 0.05, group one(Columbia) has a significantly different average rank with second group(New York) and fourth groups(New Jersey) because we can see from this that the value is larger than the cutoffs. And group one has the largest difference with Group four. And p value is less than 0.05.

4.Interpretation:

If we assume the alpha is 0.1, 90% confidence, the p value that we calculate from the KW test is close to zero, and permutation p value is also less than the alpha Hence, this is less than the alpha so that we reject the null hypothesis and conclude that the death rate is independent with the different states. And the differences exist in those groups. There are at least one pair of groups that are different from others. And we also can conclude that group one (Columbia) has a significant difference with group two(New York) and group four(New Jersey).

Conclusion:

If the alpha is assumed by 0.1, 90% confidence level, we can conclude that the states are independent with the death amounts. And the Columbia death has a significant difference with New York and New Jersey. But because the p value that we calculate for this is not that close to each other. Hence, for different alpha maybe can create a different result. And also because of the data that we choose, the month, and the number of years, this may cause the conclusion not to be accurate. In this way, if we can control those variables, this conclusion can be precise.

Question 2

1 Introduction

The following paper addressed the question that regarding Deaths involving COVID-19, whether the two factors, Age and Sex, are independent variables or not for the mid-age patients. The result of the paper can impact the detailed information about COVID-19, and can give the doctors more data to find whether a kind of patient needs to be taken to hospital earlier than the others. From the other results we know that the aged patients have higher death rates, we still want to know that whether the sex may have some effect on the death rate of different relations. I take data of 30-39 years, 40-49 years, and 50-64 years out and form a new dataset. I removed the first 3 month data, since at the first 3 month, doctors are not familiar with this new versus and the equipment like ventilator or ECMO machine are not enough, which may cause the death rate higher than later data. I use the Permutation Test for Independence and Multiple Comparisons in Contingency Tables to find the dependence. Since the dataset's form is not like the usual for, so I form a new dataset with two variable, sex and age, and I repeat them by the death number of each group, so that we can use the Permutation Test to test whether the two factors are independent.

2 Summary of Data

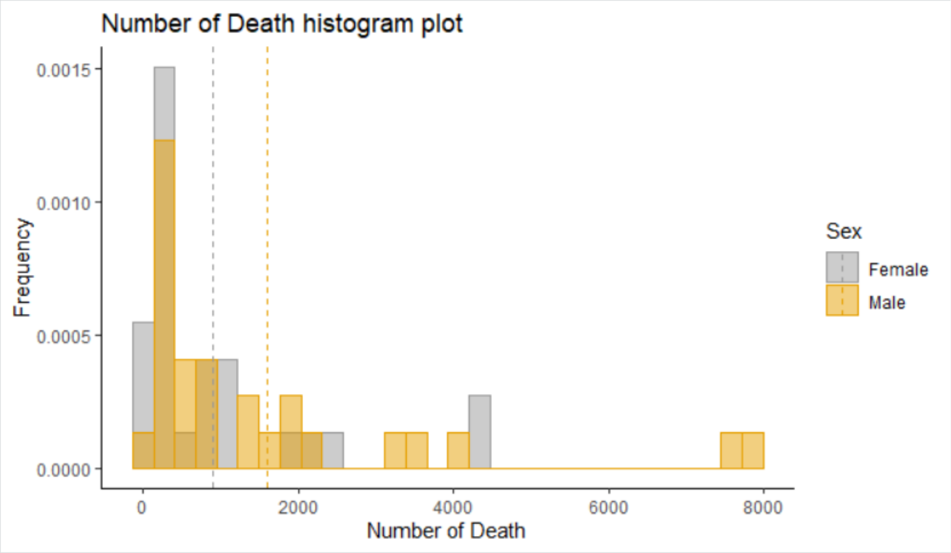


Figure 1: Histogram of death number of female and male.

We can see that the distribution is right long tailed and the mean of females is a little less than the mean of males.

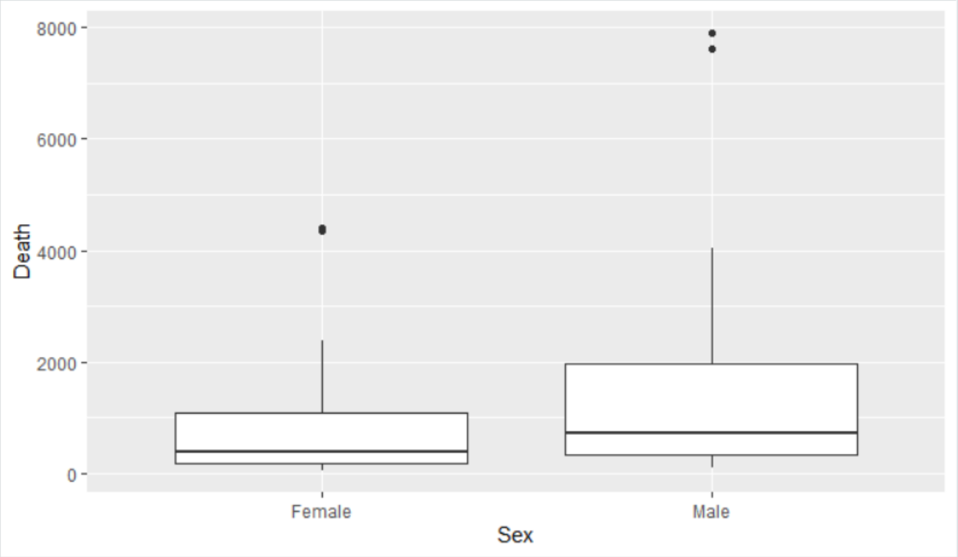
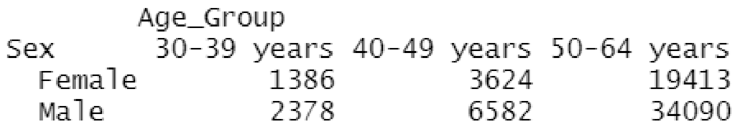


Figure 2: Boxplot of death number of Female and Male.

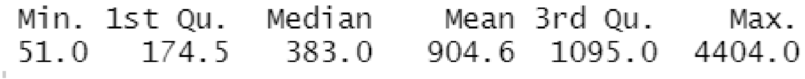
The median of males is also greater than females.



```
           Age_Group
Sex         30-39 years 40-49 years 50-64 years
  Female           1386        3624       19413
  Male             2378        6582       34090
```

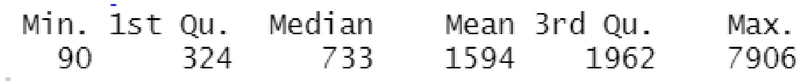
Figure 3: Contingency Table of the total death of different sex and age_group

We can see that in the age groups 18-39 years and 40-64 years, the death number of males is almost twice the death number of females. The death number of males is also greater than females in the 0-17 group



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
51.0   174.5   383.0   904.6  1095.0  4404.0 
```

Figure 4: summary of Death Number of Female



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  90     324     733    1594    1962    7906 
```

Figure 5: summary of Death Number of Male

3 Analysis

Hypothesis:

H_0:Sex and Age are independent regarding to the death number of COVID-19

H_A:Sex and Age are dependent regarding to the death number of COVID-19

Test Statistics & P-value:

$\chi^2_{S,obs} = 2.91$ permutation p-value = 0.24125

Result:

We set α = 0.05, since the permutation p-value is greater than α, we fail to reject H_0.

4 Interpretation

Since we fail to reject$H_{0,}$we can conclude that regarding deaths involving COVID-19, the two factors, Age and Sex, are independent variables for the mid-age patients. So, if in reality, we would observe our data or more extreme

24.125% of time. Thus we can say that the sex and age will not affect the death rate of each other.

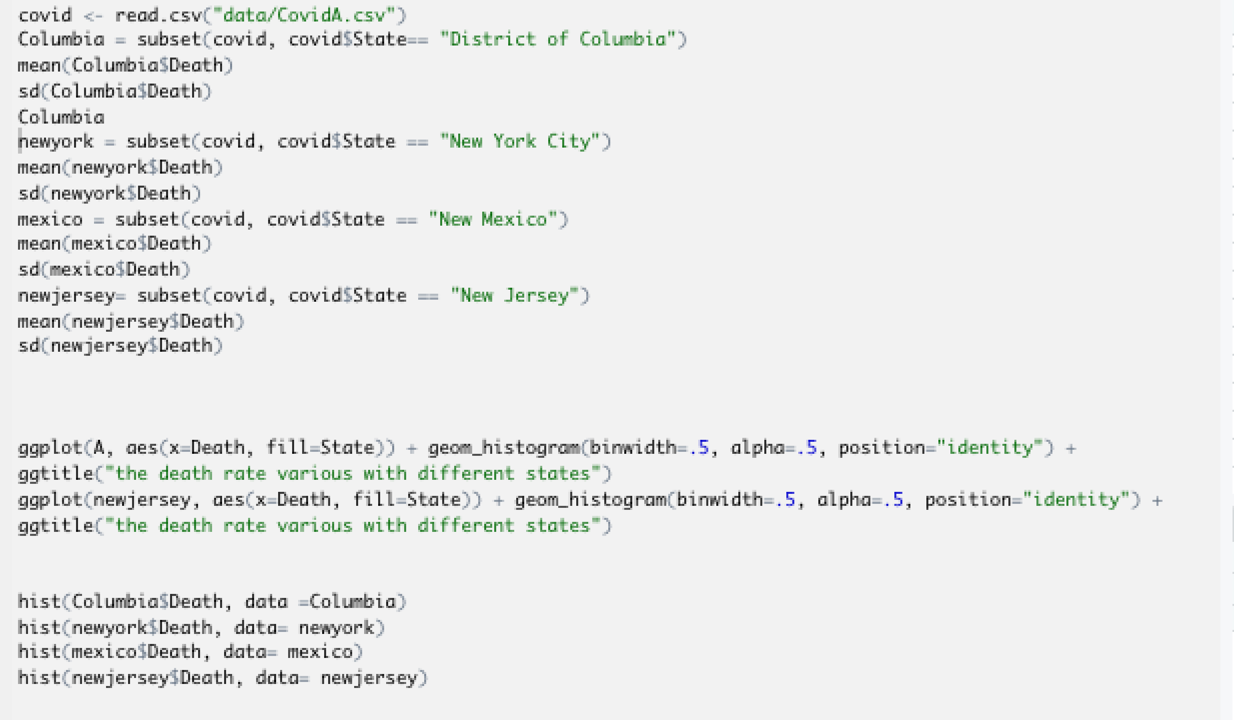
5 Conclusion

In real, when hospital having patients in mid-age, they should treat both Female and Male in the same way, since the sex will not affect the death rate of the same age. Since we already use the permutation test for independence, if the government does the same test, they may have the same result. However, we just test for the mid-age patients and their death rate is not as large as the elders, the government can do the same test on the elder people, their death rate is much higher than the mid-age and youth, they may find some different results. If they can find the age and sex are not independence for the elder patients, they can use the Pairwise comparisons to find the cutoff the dependence and form a more efficient way to treat the elder patients.

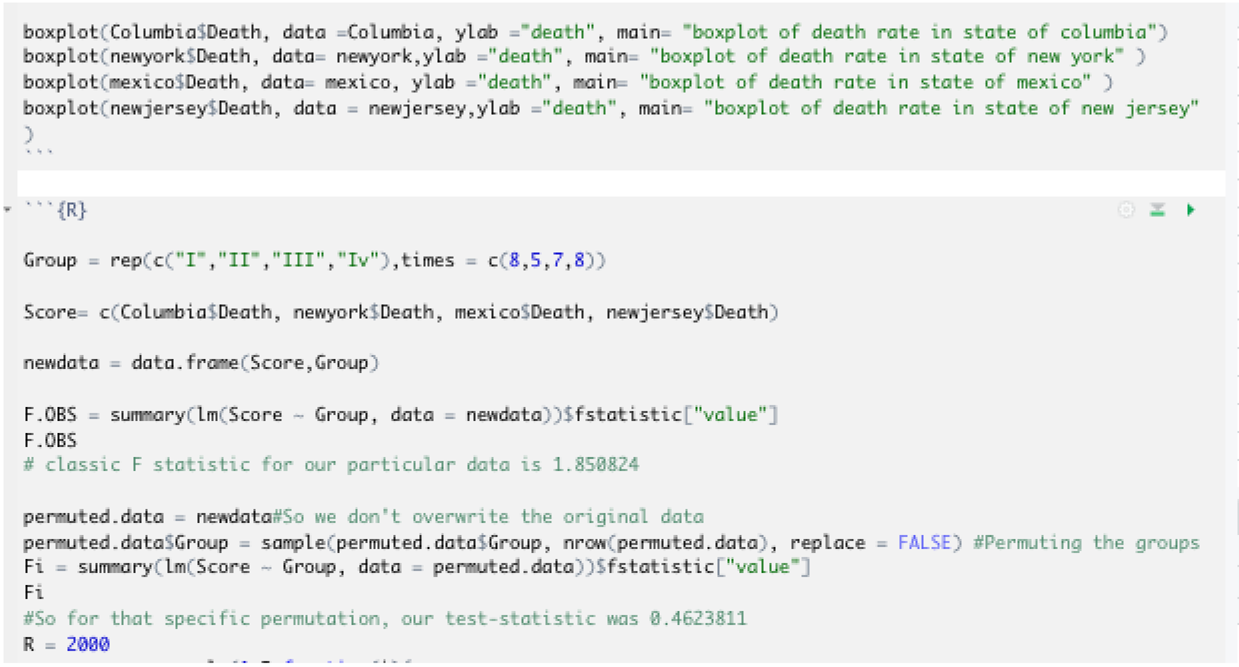
Appendix II

Code of Question 1:

Appendix of data1

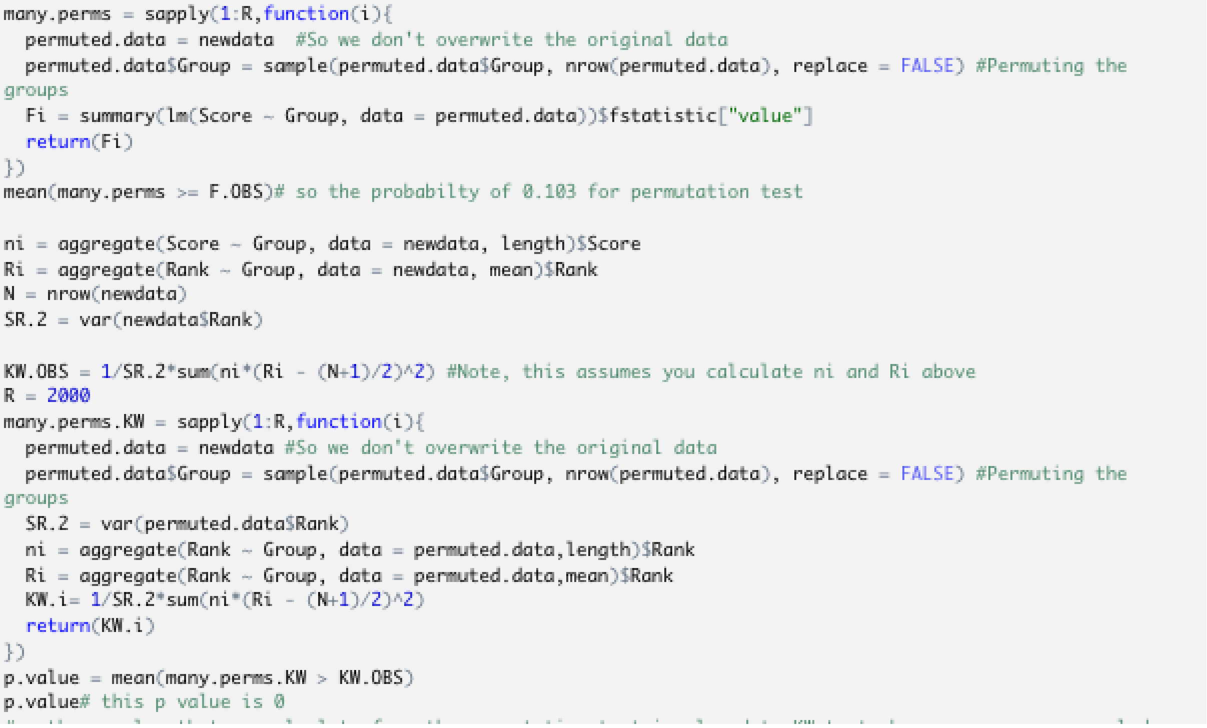


```
covid <- read.csv("data/CovidA.csv")
Columbia = subset(covid, covid$State== "District of Columbia")
mean(Columbia$Death)
sd(Columbia$Death)
Columbia
newyork = subset(covid, covid$State == "New York City")
mean(newyork$Death)
sd(newyork$Death)
mexico = subset(covid, covid$State == "New Mexico")
mean(mexico$Death)
sd(mexico$Death)
newjersey= subset(covid, covid$State == "New Jersey")
mean(newjersey$Death)
sd(newjersey$Death)



ggplot(A, aes(x=Death, fill=State)) + geom_histogram(binwidth=.5, alpha=.5, position="identity") +
ggtitle("the death rate various with different states")
ggplot(newjersey, aes(x=Death, fill=State)) + geom_histogram(binwidth=.5, alpha=.5, position="identity") +
ggtitle("the death rate various with different states")


hist(Columbia$Death, data =Columbia)
hist(newyork$Death, data= newyork)
hist(mexico$Death, data= mexico)
hist(newjersey$Death, data= newjersey)
```

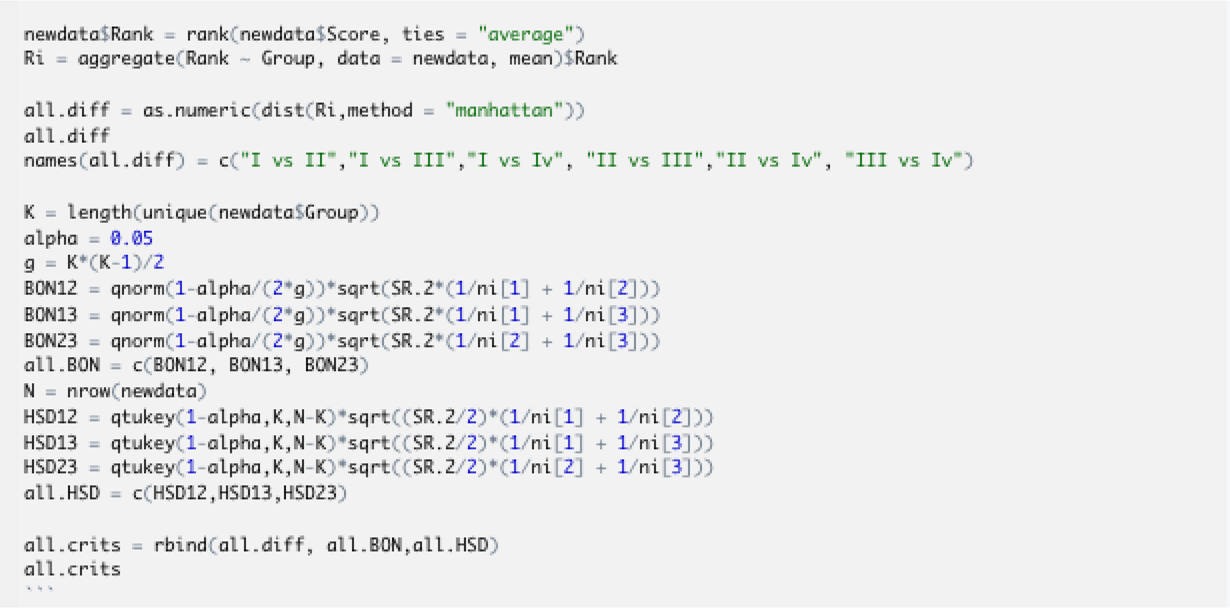


```
boxplot(Columbia$Death, data =Columbia, ylab ="death", main= "boxplot of death rate in state of columbia")
boxplot(newyork$Death, data= newyork,ylab ="death", main= "boxplot of death rate in state of new york" )
boxplot(mexico$Death, data= mexico, ylab ="death", main= "boxplot of death rate in state of mexico" )
boxplot(newjersey$Death, data = newjersey,ylab ="death", main= "boxplot of death rate in state of new jersey"
)
```

```{R}

Group = rep(c("I","II","III","Iv"),times = c(8,5,7,8))

Score= c(Columbia$Death, newyork$Death, mexico$Death, newjersey$Death)

newdata = data.frame(Score,Group)

F.OBS = summary(lm(Score ~ Group, data = newdata))$fstatistic["value"]
F.OBS
# classic F statistic for our particular data is 1.850824

permuted.data = newdata#So we don't overwrite the original data
permuted.data$Group = sample(permuted.data$Group, nrow(permuted.data), replace = FALSE) #Permuting the groups
Fi = summary(lm(Score ~ Group, data = permuted.data))$fstatistic["value"]
Fi
#So for that specific permutation, our test-statistic was 0.4623811
R = 2000
```



```
many.perms = sapply(1:R,function(i){
  permuted.data = newdata  #So we don't overwrite the original data
  permuted.data$Group = sample(permuted.data$Group, nrow(permuted.data), replace = FALSE) #Permuting the
groups
  Fi = summary(lm(Score ~ Group, data = permuted.data))$fstatistic["value"]
  return(Fi)
})
mean(many.perms >= F.OBS)# so the probabilty of 0.103 for permutation test

ni = aggregate(Score ~ Group, data = newdata, length)$Score
Ri = aggregate(Rank ~ Group, data = newdata, mean)$Rank
N = nrow(newdata)
SR.2 = var(newdata$Rank)

KW.OBS = 1/SR.2*sum(ni*(Ri - (N+1)/2)^2) #Note, this assumes you calculate ni and Ri above
R = 2000
many.perms.KW = sapply(1:R,function(i){
  permuted.data = newdata #So we don't overwrite the original data
  permuted.data$Group = sample(permuted.data$Group, nrow(permuted.data), replace = FALSE) #Permuting the
groups
  SR.2 = var(permuted.data$Rank)
  ni = aggregate(Rank ~ Group, data = permuted.data,length)$Rank
  Ri = aggregate(Rank ~ Group, data = permuted.data,mean)$Rank
  KW.i= 1/SR.2*sum(ni*(Ri - (N+1)/2)^2)
  return(KW.i)
})
p.value = mean(many.perms.KW > KW.OBS)
p.value# this p value is 0
```



```
newdata$Rank = rank(newdata$Score, ties = "average")
Ri = aggregate(Rank ~ Group, data = newdata, mean)$Rank

all.diff = as.numeric(dist(Ri,method = "manhattan"))
all.diff
names(all.diff) = c("I vs II","I vs III","I vs Iv", "II vs III","II vs Iv", "III vs Iv")

K = length(unique(newdata$Group))
alpha = 0.05
g = K*(K-1)/2
BON12 = qnorm(1-alpha/(2*g))*sqrt(SR.2*(1/ni[1] + 1/ni[2]))
BON13 = qnorm(1-alpha/(2*g))*sqrt(SR.2*(1/ni[1] + 1/ni[3]))
BON23 = qnorm(1-alpha/(2*g))*sqrt(SR.2*(1/ni[2] + 1/ni[3]))
all.BON = c(BON12, BON13, BON23)
N = nrow(newdata)
HSD12 = qtukey(1-alpha,K,N-K)*sqrt((SR.2/2)*(1/ni[1] + 1/ni[2]))
HSD13 = qtukey(1-alpha,K,N-K)*sqrt((SR.2/2)*(1/ni[1] + 1/ni[3]))
HSD23 = qtukey(1-alpha,K,N-K)*sqrt((SR.2/2)*(1/ni[2] + 1/ni[3]))
all.HSD = c(HSD12,HSD13,HSD23)

all.crits = rbind(all.diff, all.BON,all.HSD)
all.crits
```
```

## Code of Question 2:

library(readr)
library(dplyr)
CovidB <- read_csv("CovidB.csv")
```

```
DBused = filter(CovidB, Age_Group == "0-17 years"
                | Age_Group == "18-29 years"
                | Age_Group == "30-39 years"
                | Age_Group == "40-49 years"
                | Age_Group == "50-64 years"
                | Age_Group == "65-74 years"
                | Age_Group == "75-84 years"
                | Age_Group == "85 years and over" )
DBused = filter(DBused, Month != 3, Month != 4,Month != 5)
DBused = filter(DBused, Age_Group != "65-74 years"&
                       Age_Group != "0-17 years"&
                       Age_Group != "75-84 years"&
                       Age_Group != "85 years and over"&
                       Age_Group != "18-29 years")
Tused = xtabs(Death ~ Sex + Age_Group, data = DBused)
Sex = rep(c("Female", "Male"), times = c(1386 + 3624 + 19413, 2378+6582+34090))
Age_Female = rep(c("30-39 years","40-49 years", "50-64 years"),
          c(1386, 3624, 19413))
Age_Male = rep(c("30-39 years","40-49 years", "50-64 years"),
               c(2378, 6582, 34090))
Age = append(Age_Female, Age_Male)
DB_Form = data.frame(Sex, Age)
z_table = table(DB_Form)
library(ggplot2)
library(plyr)
mu <- ddply(DBused, "Sex", summarise, grp.mean=mean(Death))
ggplot(DBused, aes(x=Death, color=Sex, fill=Sex)) +
  geom_histogram(aes(y=..density..), position="identity", alpha=0.5)+
  geom_vline(data=mu, aes(xintercept=grp.mean, color=Sex),
            linetype="dashed")+
  scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
  scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
  labs(title="Number of Death histogram plot",
      x="Number of Death", y = "Frequency")+
  theme_classic()
ggplot(DBused, aes(x = Sex, y = Death)) + geom_boxplot()
summary(DBused$Death[which(DBused$Sex == "Female")])
summary(DBused$Death[which(DBused$Sex == "Male")])

the.test = chisq.test(z_table)
```

```
eij = the.test$expected
chi.sq.obs = as.numeric(the.test$statistic)
R = 4000
r.perms = sapply(1:R,function(i){
  perm.data = DB_Form
  perm.data$Age = sample(perm.data$Age,nrow(perm.data),replace = FALSE)
  chi.sq.i = chisq.test(table(perm.data),correct = FALSE)$stat
  return(chi.sq.i)
})
perm.pval = mean(r.perms >= chi.sq.obs)
```