104 Final Project

Group members: Weilin Cheng, Jeff Lee, Hengyuan Liu

Instructor: Maxime Pouokam

TA: Rui Hu

03/11/2022

Content

I. Introduction

The emergence of COVID-19 has changed how everyone lives their lives because of how dangerous and easily spread it is. As a result, getting any information we can on this deadly virus can be helpful to the general population. We are using a data set on COVID-19 from a rural city in an African country, Cameroon, from March 30, 2020. The data was captured on the day after the specified date. Our data consists of the variables gender(Male, Female) and age(10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+). We will be using the parametric and permutation tests to check for independence between gender and age. We found our data set to have two different groups: Infected cases and Death cases. We decided to take the infected cases and test if there is any independence/dependence between gender and age groups.

First, we will summarize the data by using box plots, bar plot, and histograms to explore how the data is distributed and to find out whether the data deviate from normality. To find the test statistics, we need to construct a contingency table with two categorical random variables: Gender and Age Periods. Second, we set our null hypothesis, which is if gender is independent of age periods in the infected cases. To test our hypothesis we constructed a contingency table to compute our test statistics Chi-Square and permute our P-value. Finally, we will interpret our results to figure out whether gender is independent of age groups in the infected cases.

II. Materials and Methods

A. Materials

First of all, we will find the mean and standard deviation of the infected for each variable. We start with the variable **Age**:

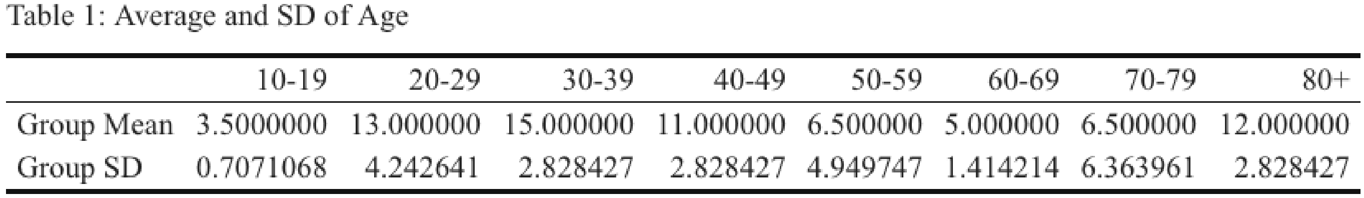


Table 1: Average and SD of Age

	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
Group Mean	3.5000000	13.000000	15.000000	11.000000	6.500000	5.000000	6.500000	12.000000
Group SD	0.7071068	4.242641	2.828427	2.828427	4.949747	1.414214	6.363961	2.828427

As we can see above, we observe that the age group 30-39 has the largest mean, and the age group 50-59 has the largest standard deviation. This indicates that the age group 50-59 has data that are more spread out from the mean of the overall age group data set. Meanwhile, age group 10-19 has the smallest mean and standard deviation which means the data of age group 10-19 are clustered around the mean of the overall age group data set.

Next, we have the variable **Gender**:

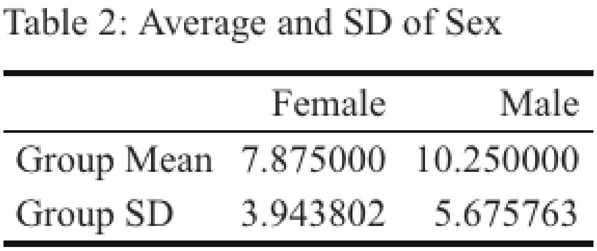


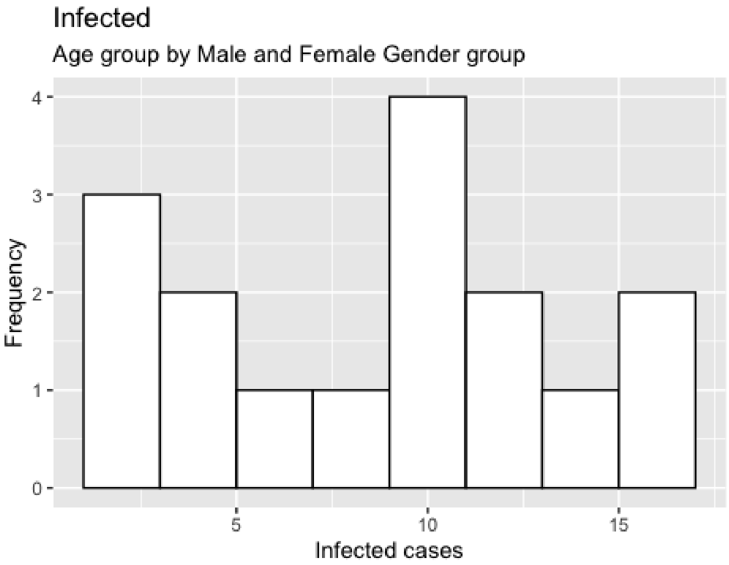
Table 2: Average and SD of Sex

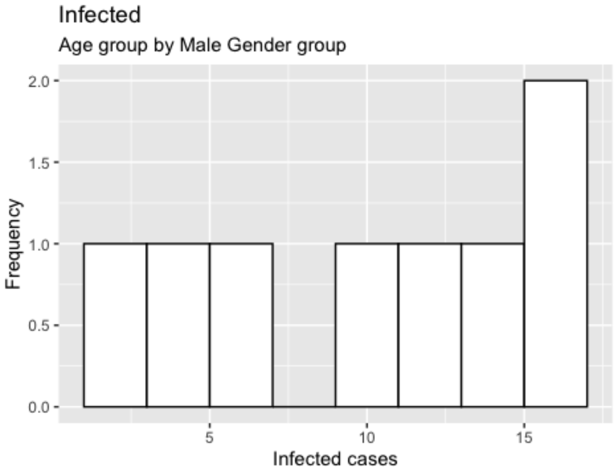
	Female	Male
Group Mean	7.875000	10.250000
Group SD	3.943802	5.675763

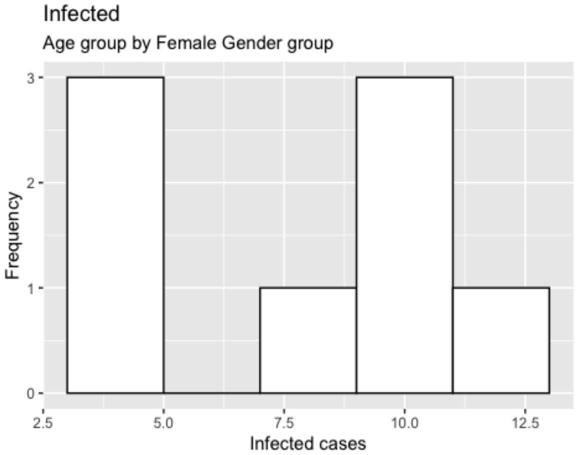
As we can see above, there are far more males infected than females infected. Also, the standard deviation for males is also larger than females, which means the infected case for males is more spread out.

Let's use some plots to visualize the data to do father analysis.

1. Histogram







Looking at our histograms above, we can observe that our histograms of Age group by Male Gender group is left-skewed and gender Female histogram is non-symmetric distribution. Our Age group by Male and Female Gender group histogram seems to have a non-symmetric distribution.

2. The bar plot of infected by age group is shown below:

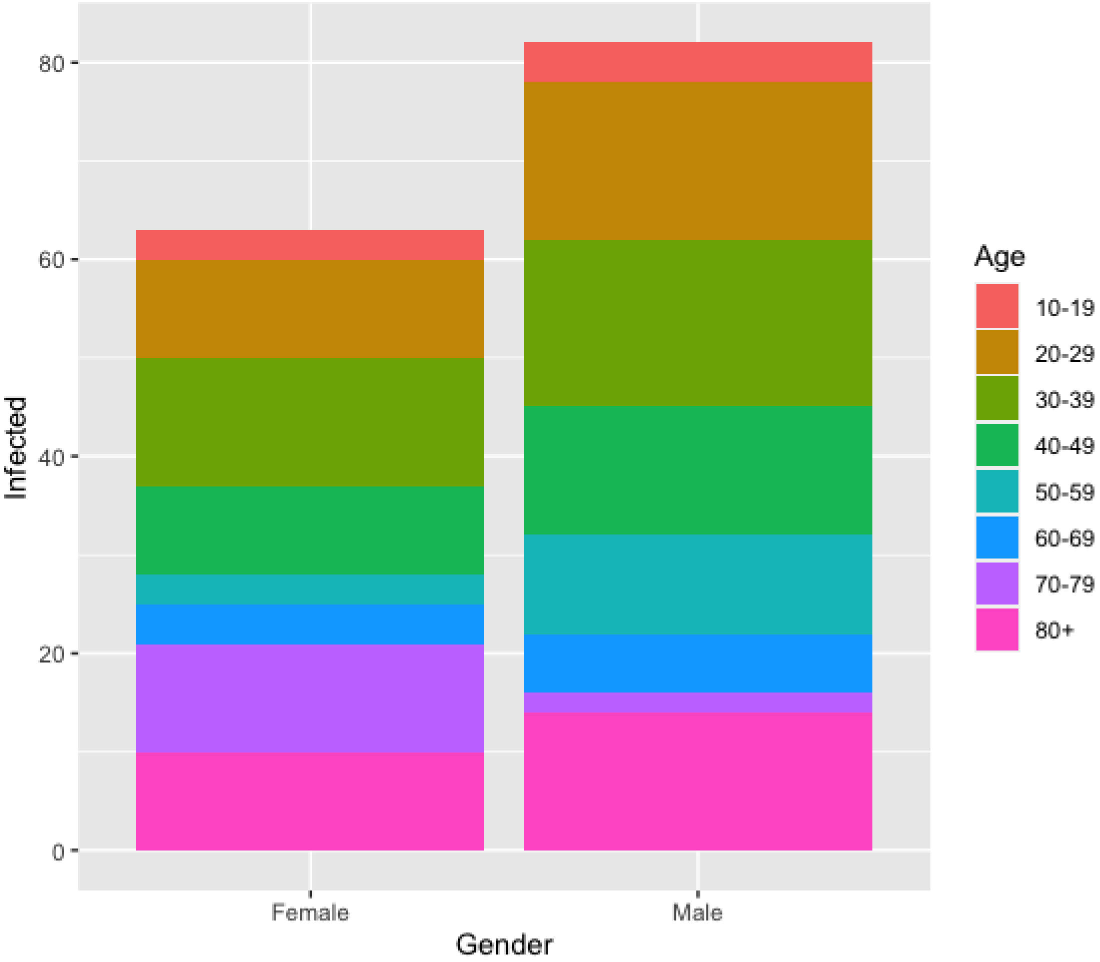


Figure 1

Interpretation of Figure 1:

Based on the histogram above, we noticed the age group of 70-90 years of the infected count are smaller for Males than Females. The age group of 80+ years, 60-69 years, 50-59 years, 40-49 years, 30-39 years, 20-29 years, and 10-19 years of the infected count are larger for Males than Females. The ages 10-19 and 60-69 were also on the smaller side for both Males and Females. Unfortunately, there is not much correlation besides the information provided.

3. The bar plot of infected by gender is shown below:

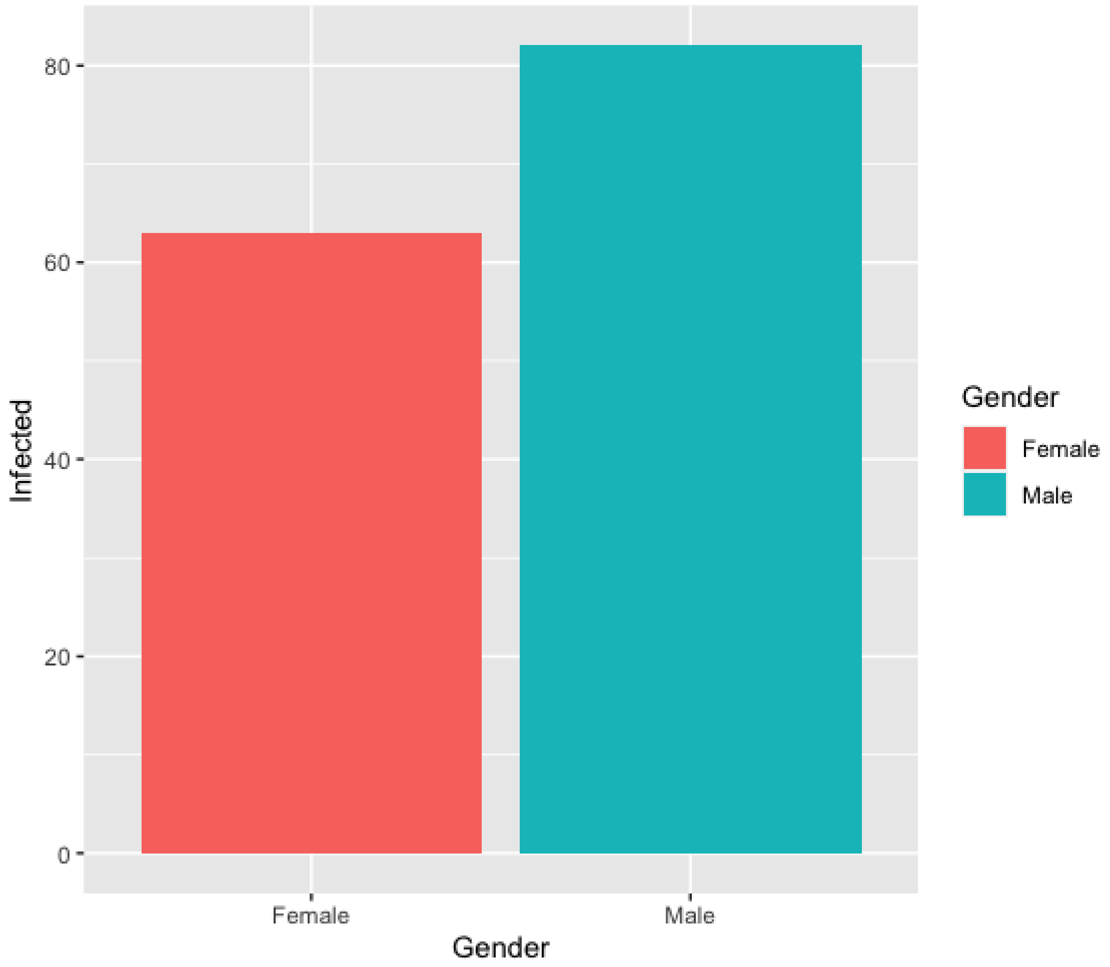


Figure 2

Interpretation of Figure 2:

Based on the histogram above, we can see that the portion of infected cases for males is larger than females from the Covid-19 virus.

1. The box plot of infected by age group is shown below:

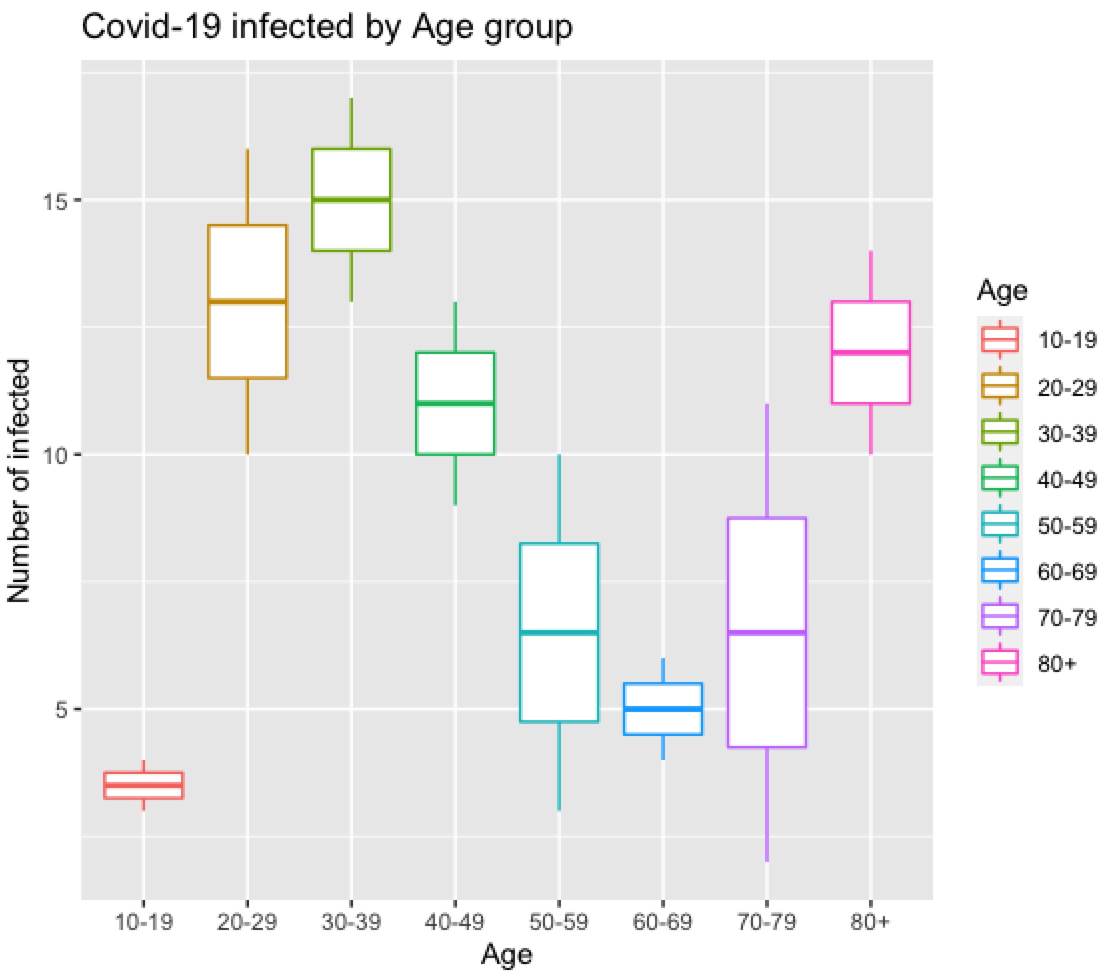


Figure 3

Interpretation of Figure 3:

As we can see in the box plot above, the different age groups do not seem to follow any trend. However, the number of people infected between groups 20-29, 30-39, 40-49, and 80+ are higher than the other age groups in this country. It might indicate that people from the age groups mentioned have increased chances of getting infected by COVID-19.

3. The box plot of infected by gender is shown below:

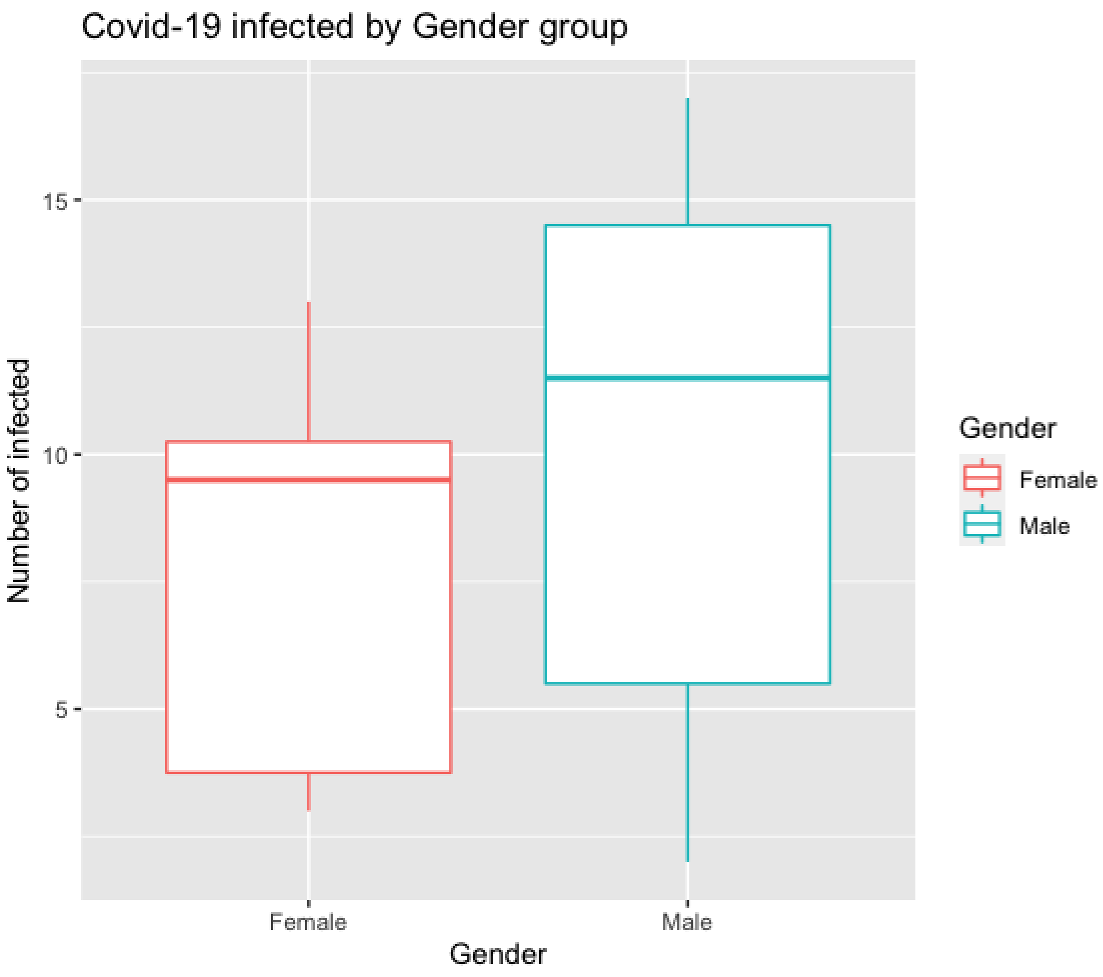


Figure 4

Interpretation of Figure 4:

From the box plot above, we can see that males account for a much larger portion of those who are infected than females from COVID-19.

B. Method

In this section of the report, we need to determine the method we will use for our data and answer the following questions: Are Gender and Age Groups independent in relation to COVID-19 infection?

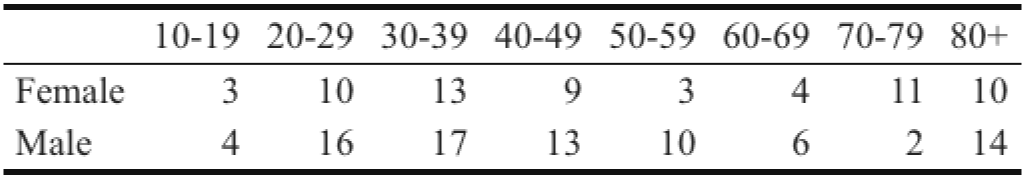
Assumption for Parametric chi-square:

1. Random sample was taken
2. e_{ij} is at least larger than 5, for all i, j
3. n_{ij} should not have vastly different magnitudes

We start the analysis by looking at the contingency table and check if any assumption is violated for Parametric chi-square.

- The 2x8 Contingency Table for n_{ij} (Gender and Ages):

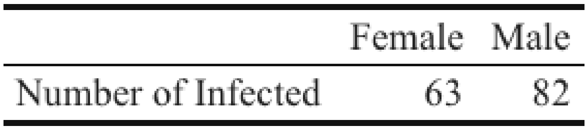
Table 3: Contingency Tables



	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
Female	3	10	13	9	3	4	11	10
Male	4	16	17	13	10	6	2	14

- The marginal sums of $n_{i.}$ for gender:

Table 4: Marginal Sum for Gender

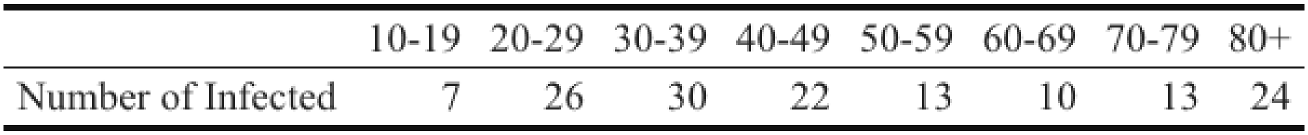


	Female	Male
Number of Infected	63	82

Table 4 is the sum of each row for two genders.

- The 1x8 marginal sum of $n_{.j}$ for Ages Periods:

Table 5: Marginal Sum for Age



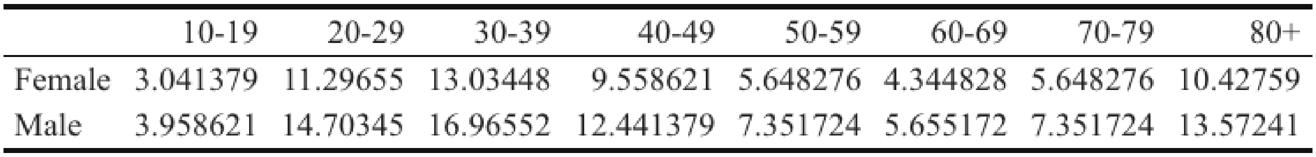
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
Number of Infected	7	26	30	22	13	10	13	24

Table 5 is the sum of each column for eight age periods. For example, for groups 10-19 we have 3 females and 4 males infected by Covid-19; therefore, the total number of infected for this group is 7.

According to Table 4 and 5, the n_{ij} have vastly different magnitudes, the parametric chi-square test is violated.

- The 2x8 Contingency Table for e_{ij}:

Table 6: Expected value of infected counts for each grid



	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
Female	3.041379	11.29655	13.03448	9.558621	5.648276	4.344828	5.648276	10.42759
Male	3.958621	14.70345	16.96552	12.441379	7.351724	5.655172	7.351724	13.57241

For Table 6, we choose to construct the contingency table because there are two categories of random variables (Genders and Age Periods). e_{ij} is the expected count on if the null hypothesis is true, which is equal (row total i)*(column total j)/n. Since not all of the values of e_{ij} are greater than or equal to 5, it implies we may not have a chi-square distribution. The assumptions of a parametric chi-square test may be violated. Therefore, based on the violation for 2 and 3, we have to use a permutation test to approach this question.

III. Result

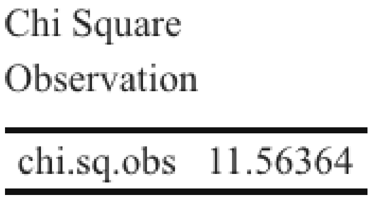
1. First, we construct our hypothesis about our test.

   Null Hypothesis: Gender and age groups are independent regarding the infected cases.

Alternative Hypothesis: Gender and age groups are dependent regarding the infected cases.

2. Second, we find the value of test-statistic

We construct the test by computing the Chi-Squared test-statistics in R:



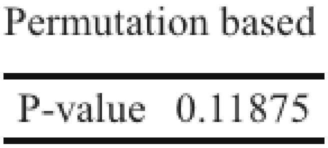
Chi Square
Observation

chi.sq.obs	11.56364

Figure 11

The test statistics is χ^2 = 11.56364

3. Now, we repeated Permutation Test 4000 times to find the permutation-based P-value:



Permutation based

P-value	0.11875

Figure 12

Interpretation of Figure 12:

After performing the permutation test R=4000 times, we get the p-value of 0.11875 which is quite large. Therefore, we fail to reject the null and conclude that there are no differences in some groups. We now need to find the Zij of each group and compare them with the cutoff to see which groups are significantly different.

Chi-Squared Test and Permutation P-value Interpretation:

The test statistic that we computed was $\chi 2$ = 11.56364. Based on the Chi-square value, we repeated the permutation test 4000 times to find the permutation-based p-value=0.11875, which turned out to be larger than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there it's independence between genders and periods of Age in infected cases.

Our above p-value tells us that if the group of Gender(Female and Male) was independent of the Age Periods, we would observe our data (or more extreme/dependent) with a probability of 0.11875 of the time.

IV. Conclusion & Future

In this project, we wanted to determine if the gender of the overall number of infected COVID-19 patients were independent of age groups. By observing our histograms, box plots, and bar plot we could see that there was no general trend in the age groups, while more males were infected than females. We then constructed a contingency table and determined that since not all of the values of e_{ij} are greater than or equal to 5, we have to use a permutation test to approach this question. We then used the chi-squared test to conduct the permutation test, eventually leading to us failing to reject the null hypothesis. As a result, we concluded that gender and age groups are independent regarding the COVID-19 infected cases in Cameroon. Since they are independent of each other, we do not have to further explore their dependencies.

Although we found there is no dependence between age and gender in infected cases, there could be some extraneous factors (i.g. different age distributions in a population, time periods, geography, etc.) that influence our observations. Since our COVID-19 data used here is from a rural city in Cameroon, it has less evidence to support that there is independence between age and gender for the whole population of Africa in the pandemic. Thus, we still need to be aware that regardless of which age and gender you are, there are still many different factors that can affect the infection rate of COVID-19.

Code appendix

```
knitr::opts_chunk$set(echo=FALSE)

library(ggplot2)

library(knitr)

library(kableExtra)


#Summary For Infected Group by Graph

#Read data

Covid = c(4, 16, 17, 13, 10, 6, 2, 14, 3, 10, 13, 9, 3, 4,

      11, 10, 1, 2, 4, 8, 8, 3, 1, 9, 1, 4, 11, 4, 2, 1, 9, 8)

Gender = rep(c("Male","Female"),times = c(8,8))

Gender

Age = rep(c("10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", "80+"))

Age

Infected=c(4, 16, 17, 13, 10, 6, 2, 14, 3, 10, 13, 9, 3, 4, 11, 10)

Gender_M=c(4, 16, 17, 13, 10, 6, 2, 14)

Gender_F=c(3, 10, 13, 9, 3, 4, 11, 10)

total.data.sub= data.frame(Infected,Gender,Age)

total.data.sub

total.gender=data.frame(Gender_F, Gender_M, Age)
total.gender

# 1. boxplot

#infected by Age group

ggplot(total.data.sub, aes(y=Infected, x = Age, group = Age, color = Age))+ geom_boxplot() +
ylab("Number of infected")+

  xlab("Age") + ggtitle("Covid-19 infected by Age group")

#infected by Gender group

ggplot(total.data.sub, aes(y=Infected , x = Gender, group = Gender, color = Gender))+
geom_boxplot() + ylab("Number of infected")+

  xlab("Gender") + ggtitle("Covid-19 infected by Gender group")



#2. Barplot

ggplot(total.data.sub) + geom_col(aes(Gender, Infected, fill = Age))

ggplot(total.data.sub) + geom_col(aes(Gender, Infected, fill = Gender))



#3 Historgram

ggplot(total.data.sub, aes(x =Infected)) + geom_histogram(binwidth = 2, color = "black", fill =
"white") + ylab("Frequency") + xlab("Infected cases")+ggtitle("Infected", subtitle = "Age group by
Male and Female Gender group")

ggplot(total.gender, aes(x =Gender_M)) + geom_histogram(binwidth = 2, color = "black", fill =
"white") + ylab("Frequency") + xlab("Infected cases")+ggtitle("Infected", subtitle = "Age group by
Male Gender group")

ggplot(total.gender, aes(x =Gender_F)) + geom_histogram(binwidth = 2, color = "black", fill =
"white") + ylab("Frequency") + xlab("Infected cases")+ggtitle("Infected", subtitle = "Age group by
Female Gender group")

Gender_M



#Analysis Data for Infected Group

Gender_Infected = rep(c("Male","Female"),times = c(82,63))

Age_Infected = rep(c("10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", "80+",
                    "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", "80+"),
                  times = c(4, 16, 17, 13, 10, 6, 2, 14, 3, 10, 13, 9, 3, 4, 11, 10))
Age_Infected
total.data.infected = data.frame(Gender_Infected,Age_Infected)
total.data.infected


#1. Contingency tables
zee.table = table(total.data.infected)
zee.table
knitr::kable((zee.table))%>% kable_styling(bootstrap_options = "striped",
                                   full_width = F, position = "center")
#To get the row sums by functions rowSums and colSums
ni. = rowSums(zee.table)
ni.
knitr::kable((ni.))%>% kable_styling(bootstrap_options = "striped",
                                        full_width = F, position = "center")
n.j = colSums(zee.table)
n.j
knitr::kable((t(n.j)))%>% kable_styling(bootstrap_options = "striped",
                                        full_width = F, position = "center")
#2. Calculating the test statistic
the.test = chisq.test(zee.table,correct = FALSE)
eij = the.test$expected
eij
knitr::kable((eij))%>% kable_styling(bootstrap_options = "striped",
                                   full_width = F, position = "center")
```

```
chi.sq.obs = as.numeric(the.test$statistic)

chi.sq.obs

knitr::kable((chi.sq.obs))%>% kable_styling(bootstrap_options = "striped",

                                  full_width = F, position = "center")


#3. Permutation P-value

R = 4000

r.perms = sapply(1:R,function(i){

  perm.data = total.data.infected

  perm.data$Age_Infected = sample(perm.data$Age_Infected,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)

perm.pval

knitr::kable((perm.pval))%>% kable_styling(bootstrap_options = "striped",

                                      full_width = F, position = "center")
```