An Investigation of the COVID-19 Outbreak in Hong Kong and the World

April 2020

ABSTRACT

The coronavirus disease 2019 (COVID-19) is now raging throughout the world. To investigate the spread of the virus in the crowd, data analysis base on the case data provided by Hong Kong (HK) government is implemented in this paper. First, the raw data of HK COVID-19 cases are rearranged based on age, gender, and case classification. Then, age distribution and gender distribution of the HK cases are analysed with statistic charts. Further, the results are extended to the world with hypothesis testing and 95% confidence interval. Based on the results, it can be concluded that males are more likely to be infected by COVID-19 than females, and people at different ages actually vary little in the proportion of being infected. Besides, some suggestions are also proposed to the government. For example, more virus tests are currently necessary.

INDEX TERMS: Age distribution, confidence interval, COVID-19, gender distribution, hypothesis testing

Contents

I. INTRODUCTION

Starting from December 2019, the coronavirus disease 2019 (COVID-19) has swept across all over the world. On 11 March 2020, the World Health Organization (WHO) announced that COVID-19 could be characterized as a pandemic. Up to 21 April 2020, nearly 2.4 million people have been infected and over 160 thousand people died in this catastrophe [1]. The world is facing the greatest challenge since the two world wars. In Hong Kong (HK), the first COVID-19 case was confirmed on 23 January 2020. Since then, more than one thousand positive cases have been reported. After a three-month tough battle, the situation in HK has become much more optimistic. As there is hardly any new confirmed case in HK now, the case data of HK can be used as a typical sample to analyse the COVID-19 situation in the world.

In this paper, all the COVID-19 case data of HK before 20 April 2020 are observed and analysed. The data is from Centre for Health Protection, HK government [2]. The main purpose of this investigation is to find out the relationship between the COVID-19 infection and people age, gender. Then, based on the research results of the data of HK, several suggestions are proposed for HK government and the governments of those countries, which are suffering from the virus eruption.

II. Raw Data Processing

In the data table provided by the HK government, all the COVID-19 cases are listed and numbered in the order of report date, and all the information of every individual case is presented after case number, as shown in TABLE I (Only five cases are presented as an example.). To extract the useful information from this table, the data are rearranged and classified according to gender, age, and case classification. For example, a data table classified by case classification can be given as TABLE II.

TABLE I Part of COVID-19 Case Data from HK Government

No.	Report date	Date of onset	Gender	Age	Name of hospital admitted	Hospitalized/ Discharged/ Deceased	HK/Non-HK resident	Case classification	Confirmed/ probable
1	23/01/ 2020	21/01/ 2020	M	39	Princess Margaret Hospital	Discharged	Non-HK resident	Imported	Confirmed
2	23/01/ 2020	18/01/ 2020	M	56	Princess Margaret Hospital	Discharged	HK resident	Imported	Confirmed
3	24/01/ 2020	20/01/ 2020	F	62	Princess Margaret Hospital	Discharged	Non-HK resident	Imported	Confirmed
4	24/01/ 2020	23/01/ 2020	F	62	Princess Margaret Hospital	Discharged	Non-HK resident	Imported	Confirmed
5	24/01/ 2020	23/01/ 2020	M	63	Princess Margaret Hospital	Discharged	Non-HK resident	Imported	Confirmed

TABLE II Rearranged COVID-19 case data by case classification

Case classification	Case number
Close contact of imported case	23
Close contact of local case	178
Close contact of possibly local case	48
Local case	68
Possibly local case	105
Imported case	604

III. Data Analysis

In this section, the sample data are analysed from different perspectives, such as the proportion of cases in different ages, the proportion of cases of different genders.

3.1 *Age Distribution Analysis*

From the case data table, it is obvious that the age of the infected patients varies in a wide range, from baby to old person. To analyse the age data, a histogram and outlier box plot for all the cases are presented in Fig. 1 (a). It can be observed that the mean age of all the cases is 38 and the standard deviation is 17.8. Also, the outlier box plot shows that the data is somewhat positively skewed, cases of the age from 15 to 40 is the majority of the data. Furthermore, considering that the imported cases dominate in the case table, it is reasonable to process the data of the imported cases and the non-imported cases, respectively. The histograms and outlier box plots of age data for these two groups are presented in Fig. 1 (b) and Fig. 1 (c). The results indicate that most of the imported cases age from 15 to 25, while the age distribution of the non-imported cases is almost normal, with the mean of 43 and standard deviation of 16.3. It can be preliminarily conclude based on Fig. 1 that people in HK age from 30 to 60 are at higher risk of contracting COVID-19. However, it is evident that the ages of the sampled case data also depend on the population age structure. Thus, the population ratio of different age groups needs to be considered together.

According to the HK government statistics [3], the population distribution by age group up to 2019 is presented in TABLE III. Corresponding to the population age distribution, the COVID-19 case numbers in each age group are listed in TABLE IV. The relative % in this table represents the relative case proportion of different age groups, separately considering the local case and imported case. To make the results more intuitive, the relative proportion are presented with bar chart in Fig. 2. It is illustrated that most of the imported cases locate in the age group 15 – 34 and most of the local cases locate in the age group 15 – 64. The results of the imported cases can be attributed to many students returning to HK recently. On the other hand, the results of locate cases are more serious. Generally, people believe that the immunity system of young people is stronger than that of old people. This makes some young people care less about COVID-19, which is extremely dangerous since young people are more likely to be infected based on the data of HK.

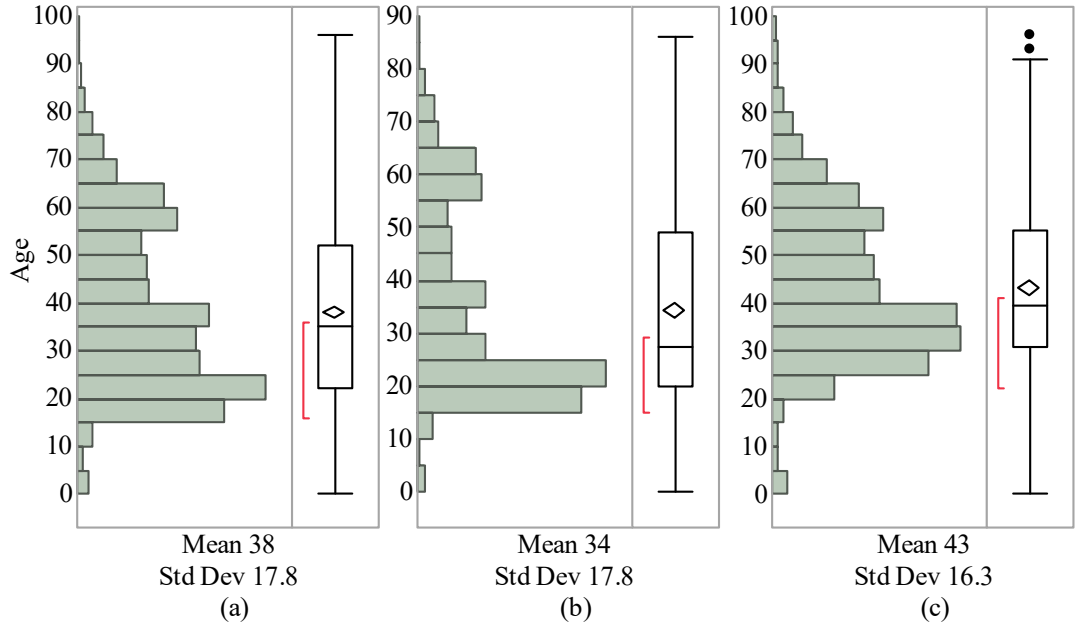

Fig. 1. Histograms and outlier box plots of the age distribution of COVID-19 cases in HK. (a) Data of all the cases. (b) Data of the imported cases. (c) Data of the non-imported cases.

TABLE III HK Population distribution by age group

Age group	Number (,000)	%
Under 15	874.9	11.7
15 - 34	1747.9	23.3
35 - 64	3562.6	47.5
65 and over	1322.0	17.6
total	7507.4	100.0

TABLE IV HK COVID-19 case distribution by age group

Age group	Popula-tion %	Local case	Local case /population %	Relative %	Imported case	Imported case /population %	Relative %
Under 15	11.7	9	0.77	5.5	18	1.54	6.8
15 - 34	23.3	140	6.01	42.6	342	14.68	64.8
35 - 64	47.4	229	4.83	34.2	208	4.39	19.4
65 and over	17.6	44	2.50	17.7	36	2.04	9.0
Total	100.0	422	14.11	100.0	604	22.65	100.0

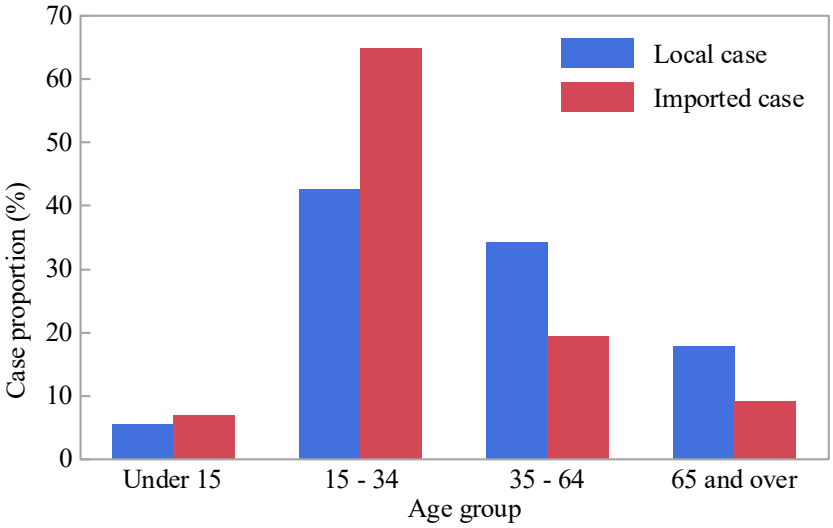

Fig. 2. Relative proportion of COVID-19 cases in different age groups

3.2 *Gender Distribution Analysis*

In addition to age, gender is also an important factor to classify the COVID-19 case data. Based on the data provided by CHP [2], the gender information can be extracted, and the proportion is presented in Fig. 3. Like the age distribution analysis, the gender structure of the whole population needs to be considered. Referring to the government data [3], the population distribution by gender up to 2019 is presented in TABLE V. Then, the case distribution considering population gender proportion is listed in TABLE VI. It is obvious that the infected proportion of COVID-19 of females is much less than that of males in HK.

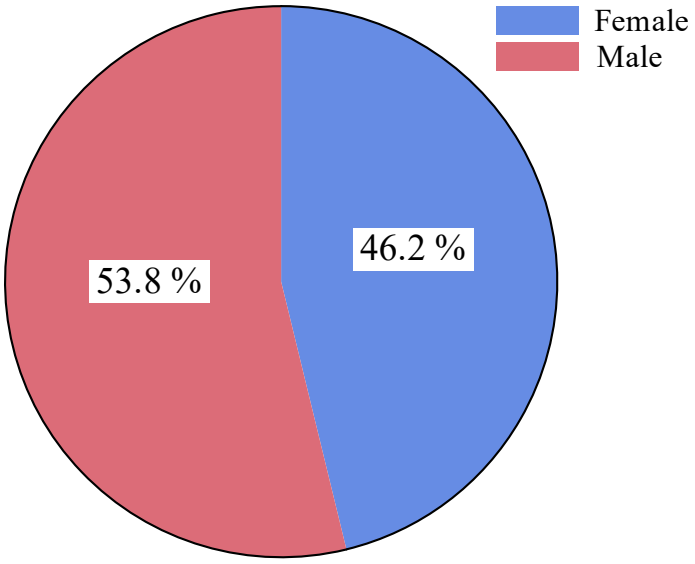

Fig. 3. Gender proportion of COVID-19 cases in HK

TABLE V HK Population distribution by gender

Age group	Number (,000)	%
Female	4084.4	54.4
Male	3423.0	45.6
total	7507.4	100.0

TABLE VI HK COVID-19 case distribution by gender

Gender	Population %	Case number	Case number / population %	Relative %
Female	54.4	474	8.7	41.8
Male	45.6	552	12.1	58.2
total	100.0	1026	20.8	100.0

3.2.1 *Prediction of COVID-19 Infected Proportion of Males and Females in the World*

Based on the data of HK COVID-19 case distribution by gender, it is possible to predict the infected proportions of males and females in the world. To achieve that, two hypothesises are proposed:

$$\text{Null hypothesis, } \mathbf{H_0}: p = 0.5 \tag{1}$$

$$\text{Alternative hypothesis, } \mathbf{H_a}: p > 0.5 \tag{2}$$

where p is the proportion of males in all COVID-19 infected persons in the world. Considering the following assumption:

1. The COVID-19 data in HK can be considered as a random sample of all COVID-19 infected persons in the world.
2. The sample size n = 1026 is large since $np = n(1\text{-}p) = 513 \geqslant 10$.
3. The population size in the world is much larger than the sample size.

Setting the significance level α is 0.05, the test statistic could be

$$z = \frac{0.582 - 0.5}{\sqrt{\frac{0.5 * 0.5}{1026}}} = 5.25 \tag{3}$$

Accordingly, the P-value is smaller than 0.00001. Thus, P-value < α, and the null hypothesis $\mathbf{H_0}$ is rejected. There is convincing evidence that the proportion of males among all COVID-19 infected people in the world is large than 50%.

In addition, the 95% confidence interval of the proportion of males among all the infected people in the world can be predicted as

$$0.582 \pm 1.96\sqrt{\frac{0.582(1-0.582)}{1026}} = (0.552, 0.612). \quad (4)$$

Thus, we are 95% confident that the infected proportion of males among all the infected people in the world is between 55.2% and 61.2%. In other words, when facing COVID-19, males are more dangerous than females.

3.2.2 *Analysis of the difference between different age groups regarding the infected proportions of males and females*

To further analyse the difference of infected proportion between males and females in the world, the HK COVID-19 case data is rearranged according to gender and age group. The results are shown in TABLE VII. Considering the gender difference in HK population, females are 1.19 times than males, the data need to be modified. The data considering the gender difference is presented in the last three columns in TABLE VIII, which is the two-way frequency table.

TABLE VII HK COVID-19 case distribution by age group and gender

Age group	Female	Male	Total
Under 15	10	17	27
15 - 34	230	252	482
35 - 64	194	243	437
65 and over	40	40	80
Total	474	552	1026

TABLE VIII Two-way frequency table of HK COVID-19 case distribution by age group and gender

Age group	Female	Male	Total
Under 15	9 (10.9)	17 (15.1)	26
15 - 34	193 (186.7)	252 (258.3)	445
35 - 64	163 (170.3)	243 (235.6)	406
65 and over	34 (31.0)	40 (43.0)	74
Total	399	552	951

To analyse whether there is difference of infected proportion between males and females in different age groups among the whole population in the world, two hypothesises are proposed:

Null hypothesis, $\mathbf{H_0}$: the infected proportions of males and females are the same for all the four age groups.

Alternative hypothesis, $\mathbf{H_a}$: the infected proportions of males and females are not all the same for all the four age groups.

If there were no difference between these four age groups regarding the infected proportions of males and females, the females infected proportion would be 399/951 and the males infected proportion would be 552/951. Then, the expected case numbers of females and males in the different age groups could be calculated based on those two proportion. For instance, the expected case number of females under 15 years old could be (399/951)*26, the expected case number of males between 15 and 34 years old could be (552/951)*445. Similarly, all the expected case number of females and males could be generated, and the values are listed in the brackets in TABLE VIII.

Considering the following assumption:

1. After compensating the gender difference in the whole population, it is reasonable to assume that the data are from independently chosen random samples.
2. All the expected cell counts are at least 5 as shown in TABLE VIII, it is reasonable to assume that the sample size is large.

Besides, the degree of freedom of this two-way frequency table is 3. Setting the significance level α is 0.05, the test statistic could be

$$\chi^2 = \frac{(9-10.9)^2}{10.9} + \cdots + \frac{(40-43)^2}{43} = 1.98 \quad (5)$$

Accordingly, the P-value is about 0.58. Thus, P-value > α, and the null hypothesis $\mathbf{H_0}$ is failed to be rejected. The evidence does not suggest that the infected proportions of males and females are not all the same for all the four age groups.

These results indicate that there is no significant difference between all these four age groups when it comes to the infected proportions of males and females. Combining the results in section 3.2.1, it can be concluded that in all age groups, no matter young or the aged, males are more easily to be infected by COVID-19 than females.

3.3 *Age and Gender Distribution of COVID-19 Case Data in Different Time Intervals*

Since HK has at least survived from the first wave of outbreak of COVID-19, the case data of HK can be systematically divided into three parts based on time intervals, the early stage (case 1 to case 342), the middle stage (case 343 to case 684), and the later stage (case 685 to case 1026). The rearranged proportion data are listed in TABLE IX and TABLE X. Similarly, the age structure and gender structure of the population has been considered. The proportion data of imported case and local are then presented in Fig. 4 and Fig. 5, respectively. Consistent with the aforementioned results, no matter in imported data or local data, the infected proportion of males is obviously larger than that of females, which means males are more dangerous during this pandemic. As for the age distribution, Fig. 5 illustrates that there is nearly no distinction in infected proportion among people who are over 15 years old, at the early stage. However, when it comes to the middle stage and the later stage, the infected proportion of people age from 15 to 34 increased largely, while the infected proportion of people older than 65 decreased dramatically. If the conclusion in section 3.1 is reused here, it can be derived that people age from 15 to 64 are more likely to be infected by COVID-19. Nevertheless, if this were true, the data of the early stage would be abnormal. Therefore, the data of the age distribution of COVID-19 infected proportion actually indicate that there is no difference between people at young age or at old age. At the early stage, since the medical resources were sufficient, most people with suspected symptoms were tested, and people at different ages have been confirmed COVID-19. Later, the medical resources became scarce and more and more old people did not go to hospital to take a COVID-19 check if they thought they just got a flu or cold, which leads to the decrease of infected proportion among old people. Thus, the conclusion drawn in this section could be that the probability of COVID-19 infection among people older than 15 is the same when only considering age. The proportion difference at the middle stage and later stage indicates that there may be still some infectors in this city, and they are more likely to be the aged.

TABLE IX Imported COVID-19 case data in different time intervals

	Female %	Male %	Under 15 %	15 – 34 %	35 – 64 %	65 and over %
Early stage	40.6	59.4	4.7	54.5	26.4	14.4
Middle stage	43.2	56.8	6.3	72.7	14.1	6.9
Later stage	40.2	59.8	8.6	64.0	19.8	7.6

TABLE X Local COVID-19 case data in different time intervals

	Female %	Male %	Under 15 %	15 – 34 %	35 – 64 %	65 and over %
Early stage	41.2	58.8	5.3	32.5	32.3	29.9
Middle stage	44.9	55.1	4.0	52.4	35.6	8.0
Later stage	40.9	59.1	7.6	49.6	36.1	6.7

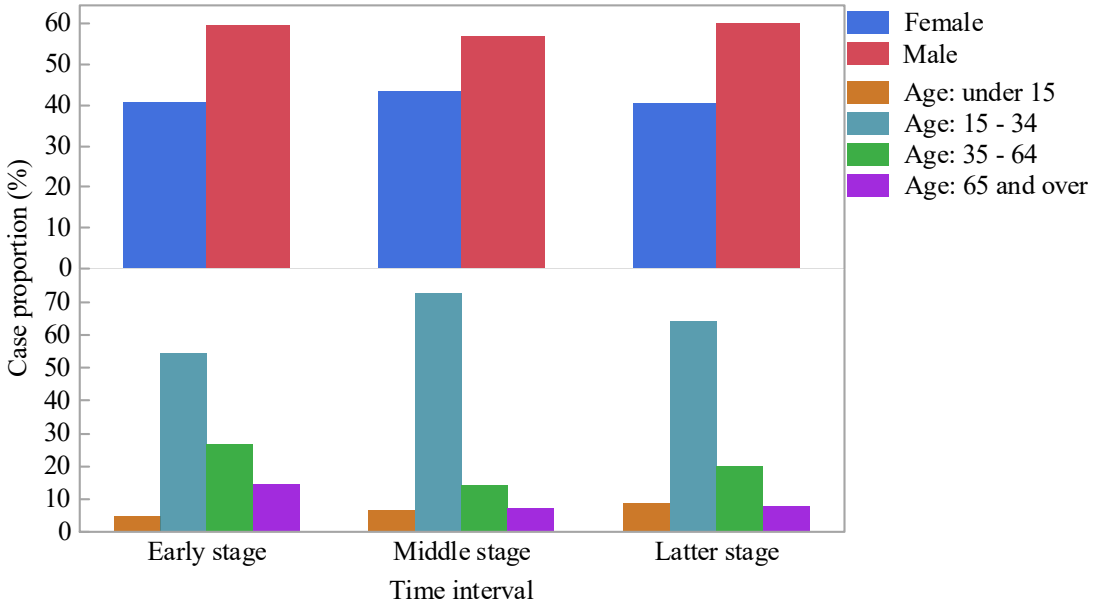

Fig. 4. Imported COVID-19 case data in different time intervals

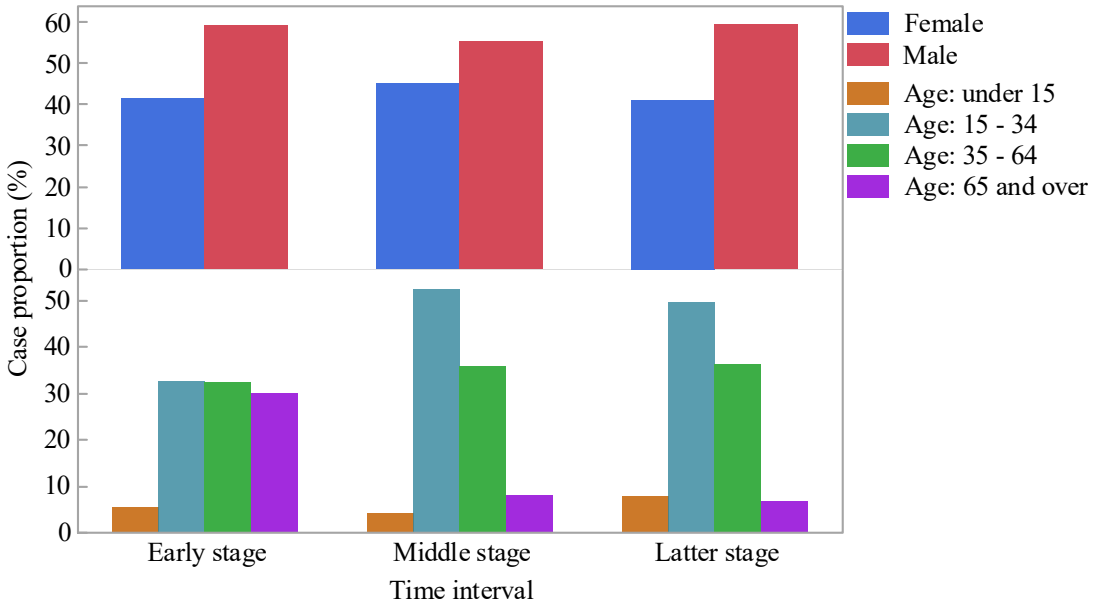

Fig. 5. Local COVID-19 case data in different time intervals

IV. CONCLUSION

The COVID-19 case data in HK are investigated in this paper to find out the relationship between the virus infection and gender or age. Based on the results, several conclusions can be proposed:

1. The infected proportion of males in HK is larger than that of females. The 95% confidence interval for the proportion of males in all COVID-19 infected people around the world is (55.2%, 61.2%), which means males are more dangerous than females when facing COVID-19.
2. There is no significant difference among people at different ages when it comes to the infected proportions of males and females. Males are always more likely to be infected by COVID-19.
3. The COVID-19 situation in Hong Kong may not be as optimistic as we imagined. It is possible that there are still some mild infected persons among the aged group. Thus, larger-scale detection is necessary to find out those infected people, also to find out the asymptomatic carriers

REFERENCE

[1] World Health Organization, *Coronavirus disease 2019 (COVID-19) Situation Report – 92*, 21 April 2020, <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports >.

[2] Centre for Health Protection (HK), *Latest situation of cases of COVID-19 (as of 21 April 2020)*, 20 April 2020. < https://www.chp.gov.hk/files/pdf/local_situation_covid19_en.pdf >.

[3] Census and Statistics Department (HK), *Hong Kong in Figures, 2020 Edition*, April 2020. < https://www.censtatd.gov.hk/hkstat/sub/sp460.jsp?productCode=B1010006 >.