Estimating the Recovery Periods of Symptomatic COVID-19 Patients.

Junho Kim-Lee

May 12, 2020

Abstract

Coronavirus disease 2019, or more commonly known as COVID-19, is an infectious respiratory disease first identified in December of 2019 in Wuhan, China. Since then, the disease has turned into a pandemic. Due to the fact that the virus is novel, not much is known about how it spreads and how humans respond to this particular strain of coronavirus. The purpose of this paper is to analyze the recovery rates of confirmed COVID-19 patients. Specifically, I will be conducting an analysis of the relationship between the rate of confirmed cases and the rate of recovered cases to deliver with a high level of statistical certainty what is the average duration that a human being will be infected with COVID-19. By being able to predict the recovery rates of confirmed patients by rigorously proving a relationship between the rate of confirmed and recovered cases, we will be able to better anticipate how the virus will progress as well as support increased hospital loads. In this paper I will show that the average duration of infection is approximately 2 to 3 weeks.

1 Introduction

Coronavirus disease 2019, or more commonly known as COVID-19, is an infectious respiratory disease first identified in December of 2019 in Wuhan, China. Since then, the disease has turned into a pandemic. Due to the fact that the virus is novel, not much is known about how it spreads and how humans respond to this particular strain of coronavirus.

The recovery rate of confirmed patients is the single biggest factor of how rapidly hospitals can decrease their loads. Possessing a thorough understanding of how humans recover from COVID-19 remains a key part of how well we can predict how the virus will affect us over time.

Data was pulled from the Humanitarian Data Exchange Novel Coronavirus (COVID-19) Cases Data set, between the dates January 22nd, 2020 to May 9th, 2020. In their initial form, the datasets showed total confirmed cases, recoveries, and deaths over time. In other words, they were extremely representative of a cumulative distribution function. I used the `diff` function in R to convert these datasets to a probability distribution function by making them show daily new cases instead of total cases.

For the purpose of this study, I chose to analyze the recovery rates of 4 regions: the Hubei Province of China, South Korea, Italy, and Spain. This decision was based on two

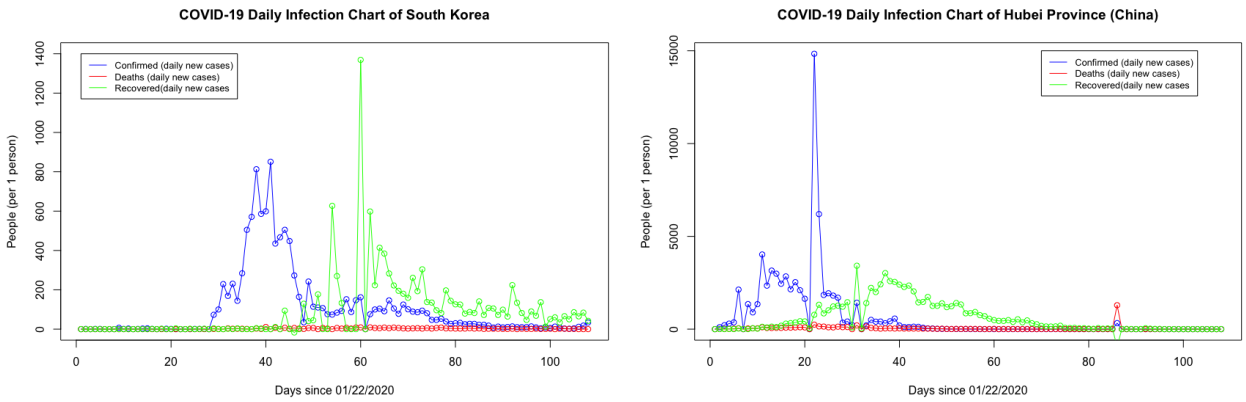

Figure 1: *Daily new confirmed cases, deaths, and recoveries in South Korea and Hubei.*

main factors. Firstly, the prospective region had to have an abundance of both confirmed and recovered cases, in order to have a more robust dataset. Secondly, the prospective region had to have passed their respective peaks, both for confirmed and recovered cases. This was so that the peaks of both curves could be compared, something that would not be possible if the peaks had yet to pass.

The Hubei Province was chosen because it contains Wuhan, the origin of COVID-19. Because they were infected so early, they passed the peak early as well, making them a good candidate for analysis. South Korea was chosen because they went through a well-contained period of mass infection and then mass recovery early on, leading to two well-defined peaks for confirmed cases and recovered cases. Finally, Spain and Italy were chosen because those regions had the highest rate of infection while simultaneously being a majority past the peak.

2 Method

From an optical analysis, I deduced that the plots of the daily new confirmed and recovered cases took on the form of a bell curve.

I started my analysis by first confirming that these plot's curves were indeed bell curves by utilizing Q-Q plots and showing that the R^2 value was sufficiently high.

By proving that the curves were either normal or gamma distributions, I could use properties about them including the formula for their expectation to calculate the on what days the peaks occurred, since the expectation of a normal or gamma distribution is the peak of the bell curve. I made the assumption that, on average, confirmed patients would recover at a similar rate. Using this assumption, I was able to deduce that the peak in recovered cases would correspond with the same patients who got sick when the number of confirmed patients peaked. Therefore, the difference between these two peaks would reveal the average duration between becoming a confirmed case and recovering from the disease. Note that this number does not include the period of incubation, where a patient may be carrying the disease without knowing it. Rather, this number focuses on the duration of time that the patient is symptomatic, and therefore a burden to hospitals.

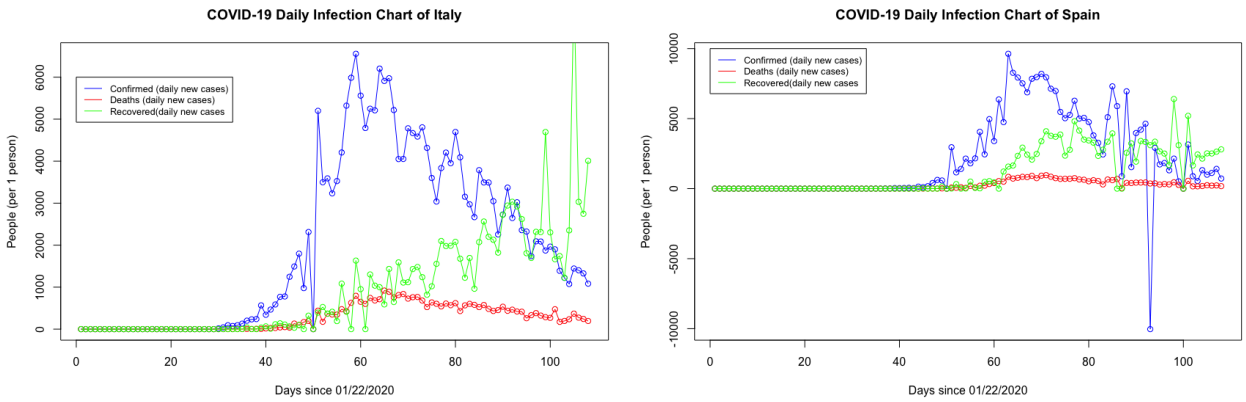

Figure 2: *Daily new confirmed cases, deaths, and recoveries in Italy and Spain.*

I calculated the average sickness duration using this method for all 4 regions with a confidence interval, using the formula for finding a confidence interval for the mean of two distributions. I calibrated my confidence interval to 95% for each country, to produce 4 confidence intervals that could then be examined separately or cohesively.

3 Using Q-Q plots to confirm the bell-curve

Using R, the following datasets from Humanitarian Data Exchange were imported:

`time_series_covid19_confirmed_global.csv`
`time_series_covid19_deaths_global.csv`
`time_series_covid19_recovered_global.csv`

I converted these datasets to represent daily new cases instead of total cases, and the result was a confirmed cases and recovered cases curve that looked bell-shaped. I further confirmed my suspicions by plotting the data for both confirmed and recovered cases as a histogram, and then overlaying a bell curve with the data's mean and standard deviation on top.

The results were promising, so Q-Q plots were made to further confirm that these curves could be handled as bell-curves. I made Q-Q plots for both a normal distribution and a gamma distribution to see which fit better. In all cases, the plot matched up well with $y = x$. Furthermore, I added a linear regression to the plot, which resulted in lines extremely similar to $y = x$, both for normal and gamma distributions. In order to choose which distribution fit better, I looked at the R^2 value to see which plot was more linear. In most cases, the gamma distribution fit better. However, since both gamma and normal distributions have their peak at the mean, it did not matter which one it was. Rather, it was more important to show that the curves were faithful to at least bell-shaped distribution.

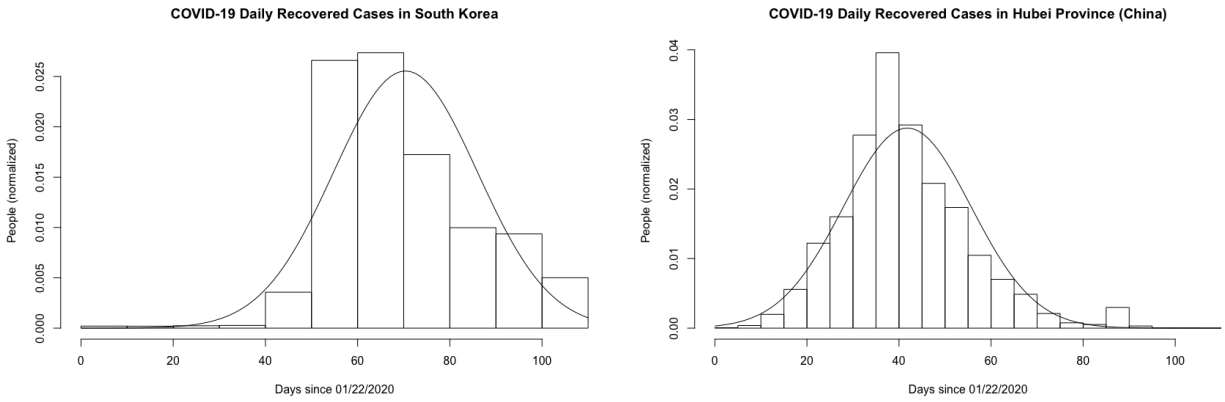

Figure 3: *Select histograms: South Korea and Hubei "recovered" curves*

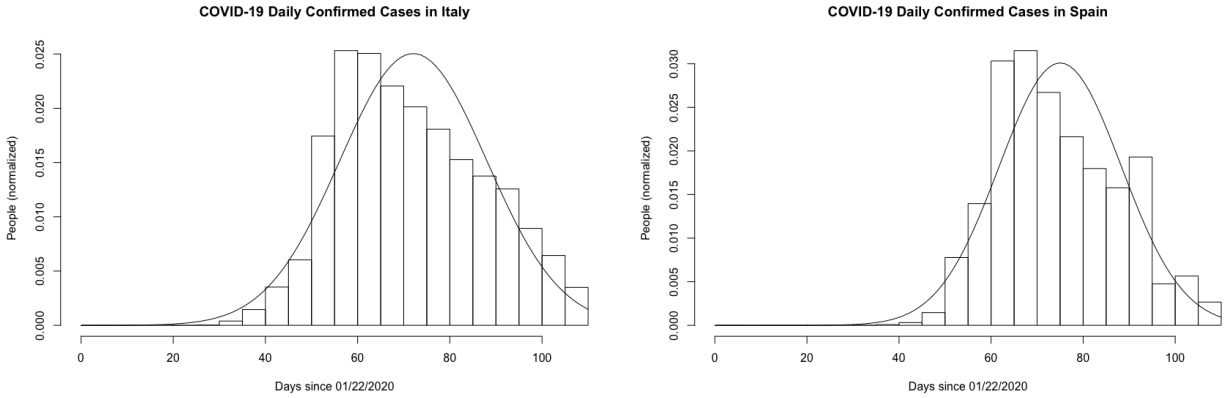

Figure 4: *Select histograms: Italy and spain "confirmed cases" curves*

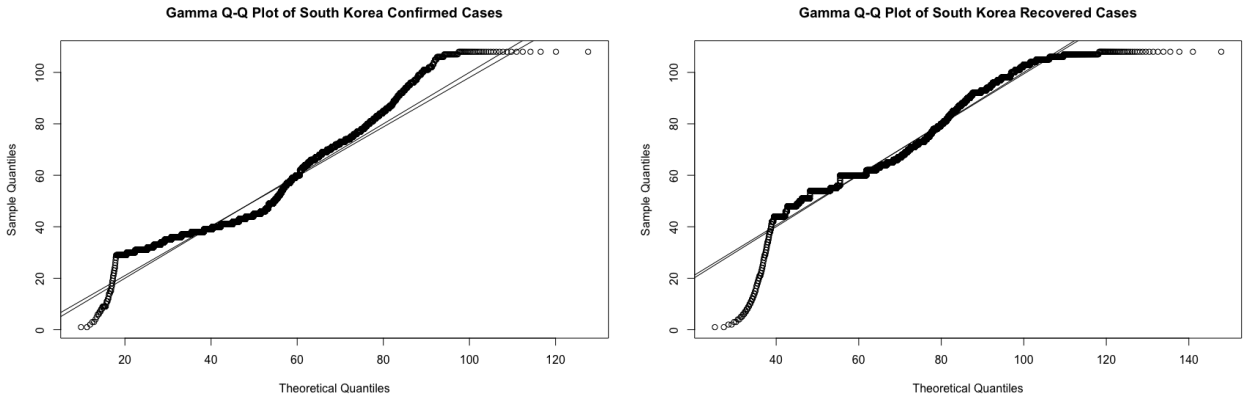

Figure 5: *Gamma Q-Q plots for Hubei*

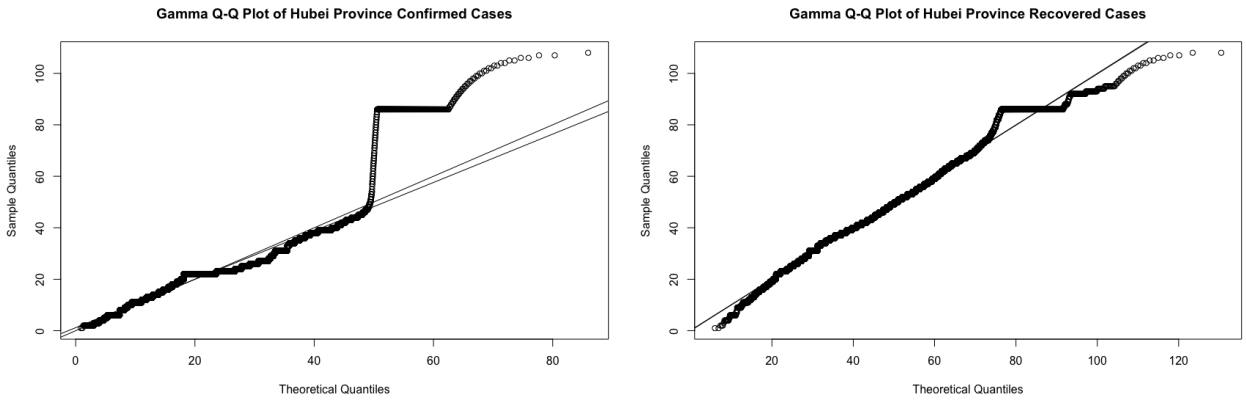

Figure 6: *Gamma Q-Q plots for Spain*

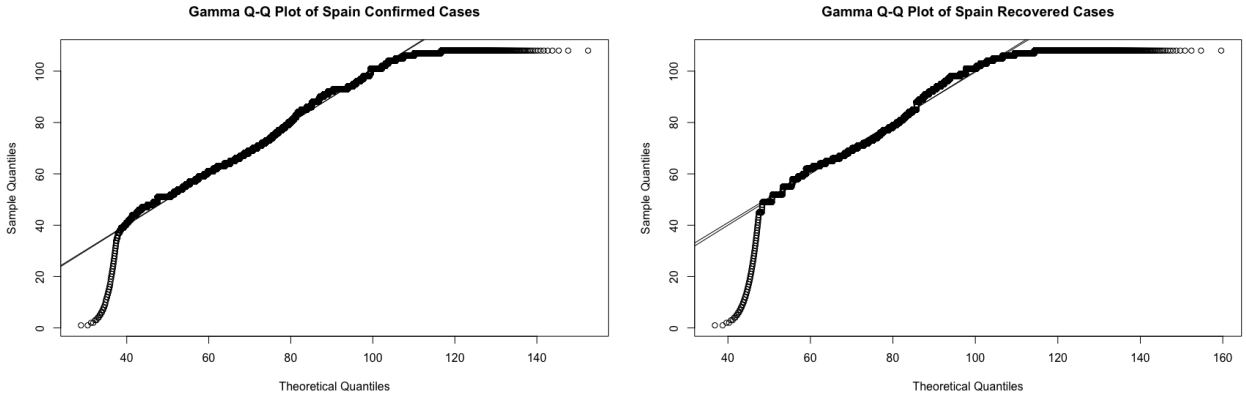

Figure 7: *Gamma Q-Q plots for Italy*

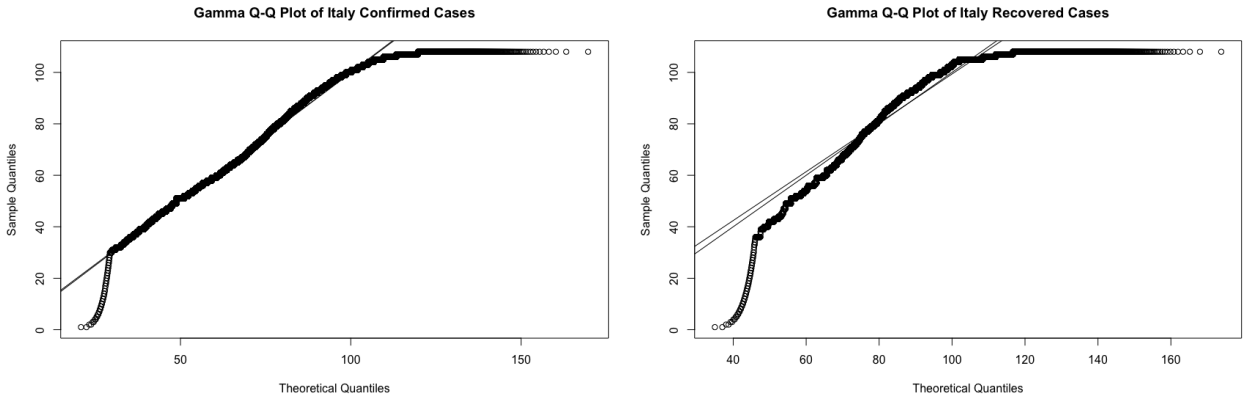

Figure 8: *Gamma Q-Q plots for South Korea*

	South Korea		China (Hubei Province)		Spain		Italy	
	Confirmed	Recovered	Confirmed	Recovered	Confirmed	Recovered	Confirmed	Recovered
Mean	47.33361	70.38842	19.83982	41.80145	75.028	84.45128	72.0687	87.48468
Std. Dev.	14.7077	15.61889	9.121146	13.86793	13.25558	13.57465	15.9318	15.51706
Normal R^2	0.8516	0.9497	0.8243	0.9679	0.9799	0.9658	0.981	0.9407
Normal intercept	3.65194	1.790285	1.827	0.6752755	0.7565464	1.456261	0.6872069	2.6334136
Normal x	0.92285	0.974556	0.90791	0.9838456	0.9899165	0.982756	0.9904646	0.9698986
Gamma R^2	0.9257	0.9594	0.8839	0.9921	0.9852	0.9559	0.9825	0.9006
Gamma intercept	1.789645	1.440976	1.18682	0.1640639	0.5581128	1.8823504	0.6339519	4.4614921
Gamma x	0.962194	0.97953	0.940181	0.9960759	0.9925613	0.9777109	0.9912036	0.9490027

Figure 9: *Data from Q-Q plots*

	S. Korea Lo	S. Korea Hi	Hubei Lo	Hubei Hi	Spain Lo	Spain Hi	Italy Lo	Italy Hi
Days	22.6258677	23.4837523	21.8326432	22.0906168	9.3225029	9.5240571	15.2815783	15.5503817

Figure 10: *Results of the study*

4 Building a Confidence Interval

With bell curves, finding the average sickness duration per region was merely a matter of taking the difference between the peak of both the confirmed cases and recovered cases curves. This would effectively reveal the duration in which a patient is symptomatic, starting with a positive test to being cleared as recovered.

Since the sickness duration of each region is just the difference between the mean of two normal distributions, I could use the formula to solve for a confidence interval of 95% without the true variance.

$$\begin{aligned}
\textbf{95\% confidence interval} &= [(\bar{X}-\bar{Y}) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}, (\bar{X}-\bar{Y}) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}] \\
&= [(\bar{X}-\bar{Y}) - z_{.975}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}, (\bar{X}-\bar{Y}) + z_{.975}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}] \quad (1) \\
&= [(\bar{X}-\bar{Y}) - 1.96\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}, (\bar{X}-\bar{Y}) + 1.96\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}]
\end{aligned}$$

filling in S_1, S_2, n_1, n_2 from the data table.

5 Results and Error Analysis

My results for the 4 countries are as follows in figure 10.

Due to the large sample size, the 95% confidence interval is extremely tight. The results seem to indicate that the average duration that a patient is symptomatically sick can range from anywhere from 9 to 23 days. The average of this interval is approximately 17 days. There are many factors at play that could explain these differences. However, I will mention the biggest potential causes for error both with the data and with my method.

Firstly, while Spain and Italy were mostly past the peak, they were not completely past the peak. Due to this, the mean of the recovered cases peak is slightly skewed in both cases. For example, the mean for recovered cases for Spain is approximately 84, even though by visual

analysis it is clear that the peak should be well past day 84. Additionally, Spain included a weird outlier point in the negatives. This came with the dataset online, but it clearly cannot be possible. This large negative number could have also skewed the data analysis.

As for what I could have done better, the study would have benefited from a larger sample size encompassing more than 4 regions. While each region had an ample amount of cases, having more cases to compare with would have been better. Furthermore, the testing for Q-Q plots could have been more rigorous. Instead of looking at the linear regressions and accepting that the linear regressions are nearly similar to $y = x$, a more rigorous and mathematical approach could have been applied. However, I could also counter my own point by saying that because the purpose of Q-Q plots was only to make sure that these curves represented a bell curve enough, this level of ambiguity may be acceptable. Additionally, my assumption that each region is composed of n independent and identically distributed random variables may need to be followed-up on, as this is a major crux of how I conduct my confidence interval testing. Changes to my initial assumptions will most likely greatly affect the results.

Lastly, it is a fact that testing is not 100% anywhere, so these results only represent a subsection of each region's population. More thorough testing would lead to more accurate results. Also, it is important to make the distinction that the results of this paper can only be used in relation to hospital load. This is because the number of confirmed and recovered cases will only mainly represent those actively being treated for COVID-19, while many are infected at home and unreported. Also, because this model does not factor in incubation, it is only good to track how long a patient is symptomatic and therefore a burden to the healthcare system.

6 Conclusion

In conclusion, this study shows that the average duration of time that a patient is symptomatically sick is close to 2 to 3 weeks. This information can be used to further prepare hospitals to handle high volumes during this pandemic, while also giving hope to patients that the virus is, majority of the time, only temporary.

7 Code Appendix

All the calculations were done in R. Attached is all the various snippets of code that I used. The full script is available at: `https://github.com/JunhoKimLee/ENGRD2700`

CSV loading:

```
# load data
setwd("~/Documents/ENGRD2700/project/")
confirmed = read.tcsv(file = "time_series_covid19_confirmed_global.csv")
deaths = read.tcsv(file = "time_series_covid19_deaths_global.csv")
recovered = read.tcsv(file = "time_series_covid19_recovered_global.csv")

# initialize [days_since_start], which represents the number of days since
```

```
#01/22/2020
dates = as.POSIXlt(confirmed$Country.Region, format="%m/%d/%y")
days_since_start = dates$yday-21
```

Plots pdf of confirmed cases, deaths, and recovered:

```
# [plot_daily] takes 3 vectors and plots the # of new cases each day.
plot_daily = function(c,d,r) {

  c_confirmed = diff(c)
  c_deaths = diff(d)
  c_recovered = diff(r)
  days_since_start2 = days_since_start[c(2:109)]

  plot(days_since_start2, c_confirmed, type="o", col="blue",
  main = "COVID-19 Daily Infection Chart of Italy",
  xlab = "Days since 01/22/2020",
   ylab = "People (per 1 person)") #this title needs to change for every run
  points(days_since_start2, c_deaths, col="red")
  lines(days_since_start2, c_deaths, col="red")
  points(days_since_start2, c_recovered, col="green")
  lines(days_since_start2, c_recovered, col="green")
  legend(0, 6000, legend=c("Confirmed (daily new cases)",
   "Deaths (daily new cases)", "Recovered(daily new cases"),
         col=c("blue", "red", "green"), lty=1, cex=0.8)
         #this positioning needs to be adjusted for every run
}
```

Converts vector to coronavirus data to histogram format:

```
# [to_histogram] takes a vector and converts it so that it may be read as a
# histogram.
# Note: this function is so horrendously inefficient but it's the best that
# works for now.
to_histogram = function(vec) {
  my_hist = c()
  for (i in (1:length(vec))) {
    cat("This is loop number:",i)
    print("")
    this_val = vec[i]
    for (j in (1:this_val)) {
      my_hist = append(my_hist, i)
    }
  }
  return(my_hist)
}
```

Main function that does qq-plot analysis

```
# [qq_analysis] takes a vector, converts it into a histogram, and creates a
# normal and gamma qq-plot with regression analysis.
qq_analysis = function(vec) {
  # convert our new-case data to histogram
  korea_data = to_histogram(diff(vec))

  #find the mean and std. dev.
  mu = mean(korea_data)
  std = sd(korea_data)
  print("Mean and std:")
  print(mu)
  print(std)

  # plot the histogram with normal distribution overlay
  hist(korea_data, main = "COVID-19 Daily Recovered Cases in Spain",
  #change this every run
       xlab = "Days since 01/22/2020", ylab = "People (normalized)",
       freq = FALSE)
  x = seq(0,109,1)
  curve(dnorm(x, mean = mu, sd = std), add=TRUE)

  # qq normal plot
  title = "Q-Q Plot of Spain Recovered Cases"
  #change this too
  n = length(korea_data)
  quantiles = (1:n-.5)/n
  x = qnorm(quantiles,mu,std)
  plot(x,sort(korea_data),main=paste("Normal",title, sep = " "),
  xlab = "Theoretical Quantiles", ylab= "Sample Quantiles")
  abline(0,1)
  # regression
  reg = lm(sort(korea_data) ~ x)
  abline(reg)
  print(summary(reg))

  # qq gamma plot
  my_beta = std^2/mu
  my_alpha = mu/my_beta
  x2 = qgamma(quantiles,shape=my_alpha,scale=my_beta)
  plot(x2,sort(korea_data),main=paste("Gamma",title, sep = " "),
  xlab = "Theoretical Quantiles", ylab= "Sample Quantiles")
  abline(0,1)
  reg2 = lm(sort(korea_data) ~ x2)
```

```
  abline(reg2)
  print(summary(reg2))
}
```

References

[1] "Novel Coronavirus (COVID-19) Cases Data." Humanitarian Data Exchange, 2020, data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases.

[2] "Naming the Coronavirus Disease (COVID-19) and the Virus That Causes It." World Health Organization, World Health Organization, www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.