nCoV-2019

Daniel Vogel

3/6/2020

About the Data Source

Source: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Content

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it's unclear how easily or sustainably this virus is spreading between people - CDC

This dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. Please note that this is a time series data and so the number of cases on any given day is the cumulative number.

The data is available from 22 Jan, 2020.

Column Description

2019_ncov_data.csv

Sno - Serial number Date - Date and time of the observation in MM/DD/YYYY HH:MM:SS Province / State - Province or state of the observation (Could be empty when missing) Country - Country of observation Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardised currently. So please clean them before using it) Confirmed - Number of confirmed cases Deaths - Number of deaths Recovered - Number of recovered cases

Acknowledgements

Johns Hopkins university has made the data available in google sheets format here. Sincere thanks to them.

Thanks to WHO, CDC, NHC and DXY for making the data available in first place.

Analysis of Global nCoV Cases in Order to Predict the End of the Pandemic

This dataset is updated daily with globally reported cases of the nCoV virus. We plot this data in various ways to determine if the nCoV spread is following a predictable pattern. The 2020 Summer Olympics is scheduled to start in Tokyo on July 24. We will develop a simple model to see if the nCoV cases in China will have subsided by then.

Assumptions

Note:The original data had multiple observations for some dates. In these cases, we will only consider the latest observation for that date. This is valid since these are always increasing quantities.

The ncov dataset has 3 columns for observation counts: Confirmed, Recovered, Deaths It is not clear from the data if Recovered cases are counted in Confirmed. It looks like Confirmed cases are ever-increasing so I will assume a "Recovered" patient is also "Confirmed". This will allow us to see a downward trend in cases in the future.

```
## Some info about the dataset which will change with new datasets
LatestObservation<-max(ncov_df$ObservationDate)
FirstObservation <-min(ncov_df$ObservationDate)
DaysObserved <- 1 + max(ncov_df$YEARMD) - min(ncov_df$YEARMD)

print("The following data will change based on the dataset used as the dataset is updated daily")
```

```
## [1] "The following data will change based on the dataset used as the dataset is updated daily"
```

```
# formatted print statements to use if echo=FALSE
sprintf("Lastest Observation Date: %s", LatestObservation)
```

```
## [1] "Lastest Observation Date: 03/02/2020"
```

```
sprintf("First Observation Date: %s", FirstObservation)
```

```
## [1] "First Observation Date: 01/22/2020"
```

```
sprintf("Total Days Observed: %s", DaysObserved)
```

```
## [1] "Total Days Observed: 41"
```

```
# aggregate Country/Region -> inside china , outside china.  Add a bolean column called "inchina" to th
ncov_df$InChina<-(ncov_df$Country=="China"|ncov_df$Country=="Mainland China")

# use dplyr summarise to aggregate
InOut_df<-ncov_df %>% group_by(YEARMD, InChina) %>% summarise(Total=sum(Confirmed))

# Lets make a Tidy frame with the same data for printing and plots
Tidy_df<-pivot_wider(InOut_df, names_from=InChina, values_from=Total)

# rename from TRUE/FALSE to InChina, OutChina
names(Tidy_df)[names(Tidy_df)=="TRUE"]<-"InChina"
names(Tidy_df)[names(Tidy_df)=="FALSE"]<-"OutChina"

# Lets look for Maximum Count to see the scale and calculate the Mean later
InChinaMax<-max(Tidy_df$InChina)
OutChinaMax <-max(Tidy_df$OutChina)

# Formatted print.  Uncomment if you turn off Echo.
sprintf("Total Confirmed Inside China: %s", InChinaMax)
```

```
## [1] "Total Confirmed Inside China: 80026"
```

```
sprintf("Total Confirmed Outside China: %s", OutChinaMax)
```

```
## [1] "Total Confirmed Outside China: 10283"
```

```
summary(InOut_df)
```

```
##      YEARMD             InChina            Total
##  Min.   :2020-01-22   Mode :logical   Min.   :    8.0
##  1st Qu.:2020-02-01   FALSE:41        1st Qu.:  487.8
##  Median :2020-02-11   TRUE :41        Median : 3097.5
##  Mean   :2020-02-11                   Mean   :23038.7
##  3rd Qu.:2020-02-21                   3rd Qu.:43821.8
##  Max.   :2020-03-02                   Max.   :80026.0
```

Aggregate Data Used For Plots

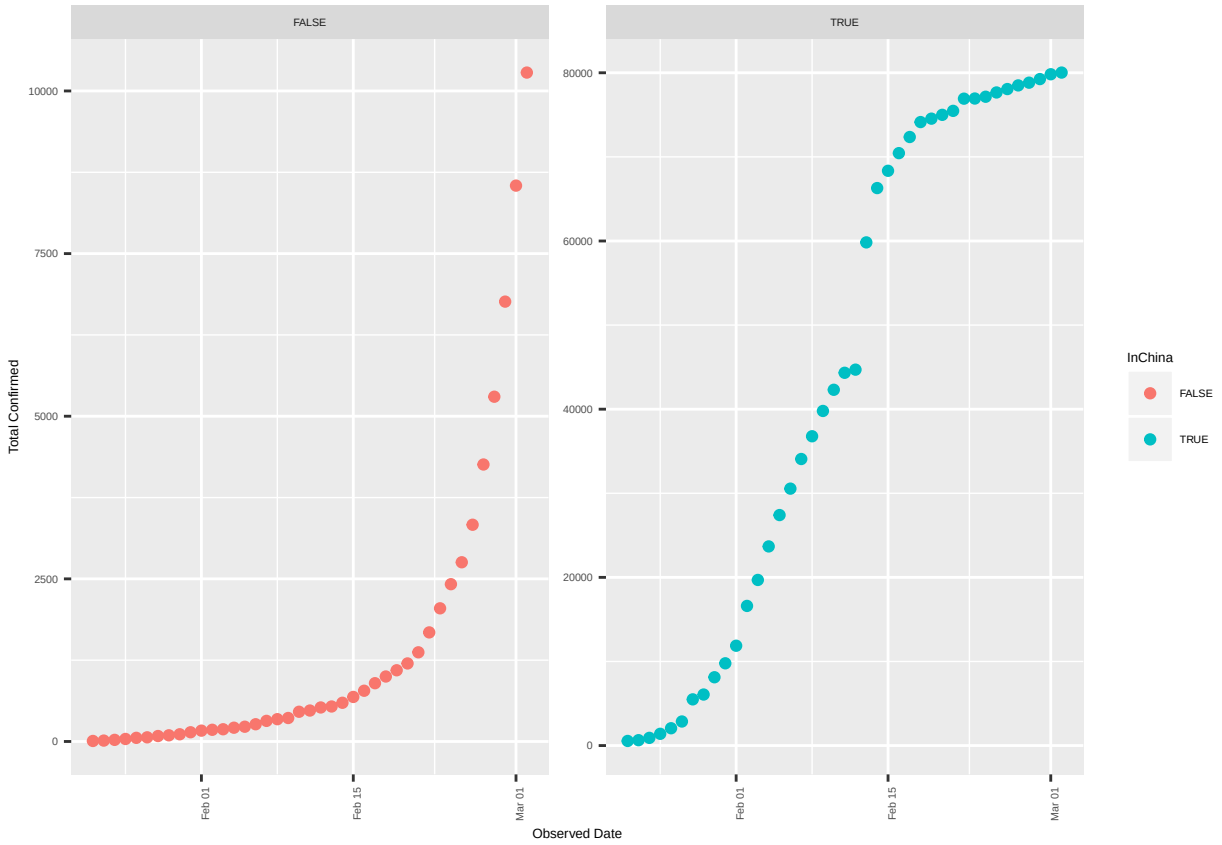
Table 1: Observations Grouped by Inside /Outside China

YEARMD	OutChina	InChina
2020-01-22	8	547
2020-01-23	14	639
2020-01-24	25	916
2020-01-25	39	1399
2020-01-26	56	2062
2020-01-27	64	2863
2020-01-28	84	5494
2020-01-29	95	6070
2020-01-30	111	8124
2020-01-31	142	9783
2020-02-01	167	11871
2020-02-02	180	16607
2020-02-03	188	19693
2020-02-04	212	23680
2020-02-05	227	27409
2020-02-06	265	30553
2020-02-07	317	34075
2020-02-08	343	36778
2020-02-09	361	39790
2020-02-10	457	42306
2020-02-11	476	44327
2020-02-12	523	44699
2020-02-13	538	59832
2020-02-14	595	66292
2020-02-15	685	68347
2020-02-16	780	70446
2020-02-17	896	72364
2020-02-18	999	74139
2020-02-19	1095	74546
2020-02-20	1200	74999
2020-02-21	1371	75472
2020-02-22	1677	76922
2020-02-23	2047	76938
2020-02-24	2418	77152
2020-02-25	2755	77660
2020-02-26	3332	78065
2020-02-27	4258	78498

YEARMD	OutChina	InChina
2020-02-28	5300	78824
2020-02-29	6762	79251
2020-03-01	8545	79826
2020-03-02	10283	80026

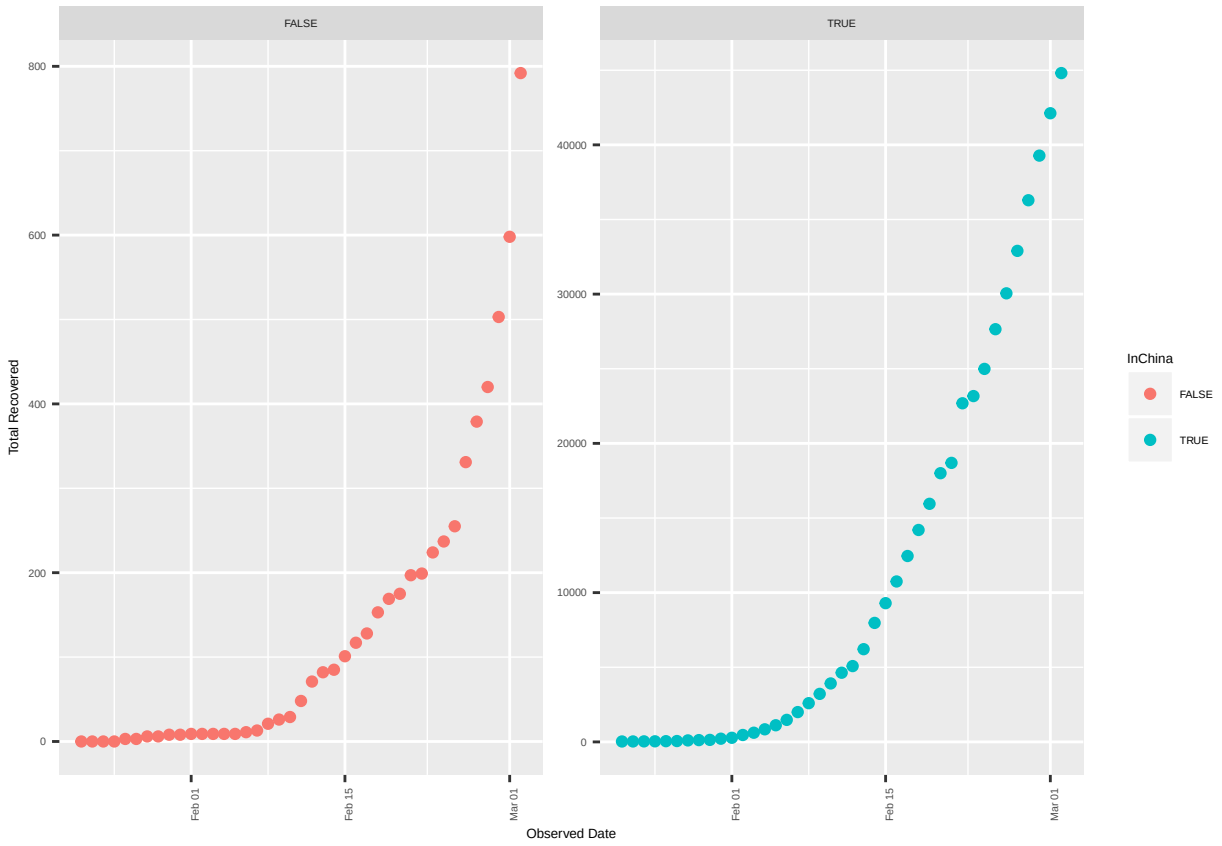
Confirmed Cases

```
## Saving 6.5 x 4.5 in image
```



Recovered Cases

```
## Saving 6.5 x 4.5 in image
```



Confirmed Cases Group by Inside|Outside China

Note: We have removed "Other" and "China" from this plot to show countries in the same scale. "Other" referred to patients on cruise ships who came from multiple countries. "China" Totals are too high to plot on the same scale as other countries.

We can see that the spread of nCoV follows the same curve inside and outside of China. This also show that the curve is increasing slope during this period of spread. A typical virus will spread increasingly and then slow down before fading away. We can see that cases in China are starting to decrease in slope but outside are still on the rise.

```
## [1] "Countries with the highest number of cases outside China are currently:"
## [2] "South Korea"
## [3] "Italy"
## [4] "Iran"
```

This data shows that countries geographicaly close to China have the most cases so far. This is to be expected since travel volumes are high between Wuhan and Hong Kong, Singapore, and other neighbor countries.

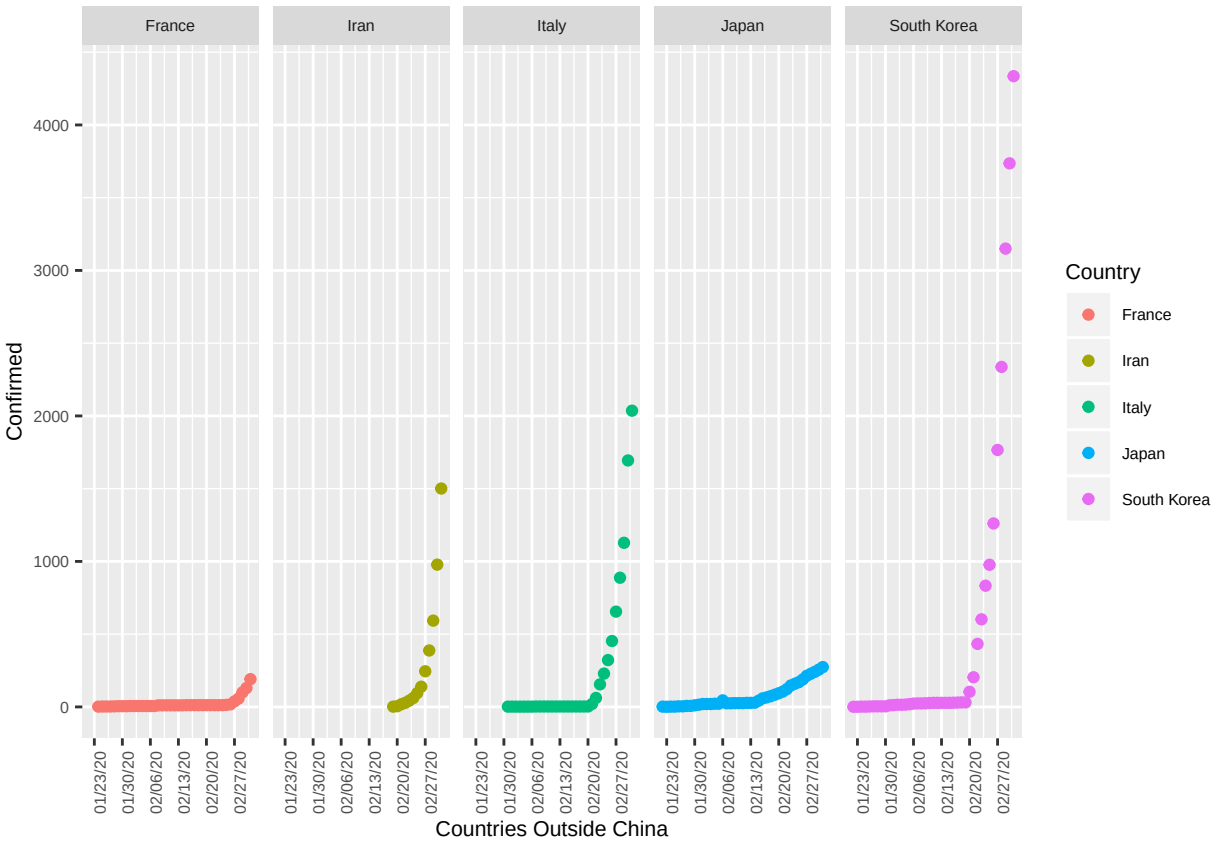
Table 2: Countries >10 Cases

Country	Total
South Korea	4335
Italy	2036
Iran	1501
Japan	274
France	191

Country	Total
Germany	159
Spain	120
Singapore	108
US	101
Hong Kong	100
Kuwait	56
Bahrain	49
Thailand	43
Switzerland	42
Taiwan	41
UK	40
Australia	30
Malaysia	29
Canada	27
Iraq	26
Norway	25
United Arab Emirates	21
Austria	18
Netherlands	18
Vietnam	16
Sweden	15
Lebanon	13

A faceted plot of Top 5 Countries outside of China

This plot shows that the increase in each of these countries is following the same pattern during this early stage of the virus spread. It appears from these graphs that the nCoV virus spreads at an exponential rate at first.



Comparing Data to a R-normal bell curve

R has four in built functions to generate normal distribution. They are described below.

dnorm(x, mean, sd)

Following is the description of the parameters used in above functions x is a vector of numbers.

mean is the mean value of the sample data. It's default value is zero. sd is the standard deviation. It's default value is 1.

dnorm() This function gives height of the probability distribution at each point for a given mean and standard deviation.

```
# Add some rows for future data...to predict
    lastrow<-nrow(Tidy_df)
    lastdate<-max(Tidy_df$YEARMD)
    days_till_olympics<- as.numeric( Olympics - lastdate )

    sprintf("%s Days Till 2020 Olympics in Tokyo", days_till_olympics)
```

```
## [1] "144 Days Till 2020 Olympics in Tokyo"
```

```
    g<-ncov_df %>% filter(InChina==TRUE) %>%
      group_by(YEARMD, InChina) %>%
      summarise(Total=sum(Confirmed)-sum(Recovered))

    ## save to a file for the slideshow
    png(filename="rnorm.png")
```

```
    ## put two plots on one slide
    par(mfrow=c(1,2))
      plot(x=g$YEARMD,y=g$Total,
      xlab = "Observed Date", ylab = "Total = Confirmed - Recovered")

    inchina_sd<-sd(Tidy_df$InChina)
    outchina_sd<-sd(Tidy_df$OutChina)
    sprintf("InChina Standard Deviation is: %s",inchina_sd)
```

```
## [1] "InChina Standard Deviation is: 30860.0323699109"
```

```
    sprintf("OutChina Standard Deviation is: %s",outchina_sd)
```

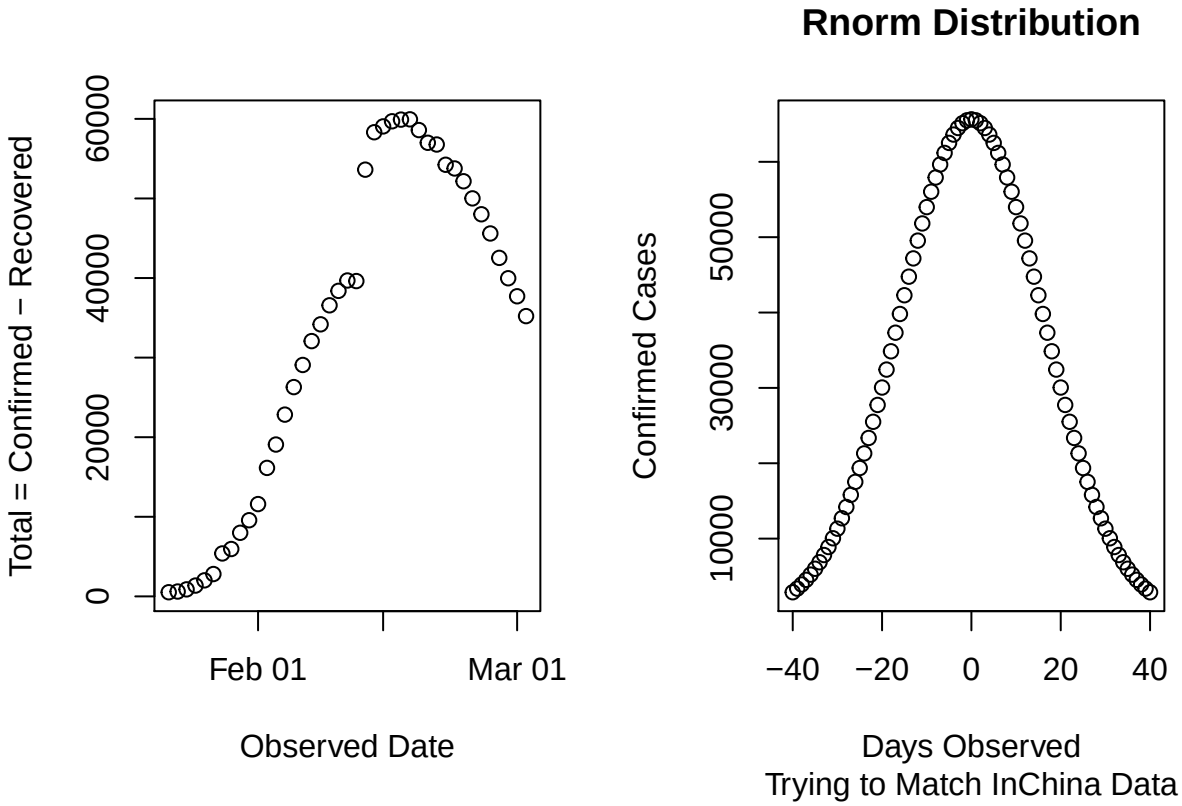
```
## [1] "OutChina Standard Deviation is: 2367.37795909726"
```

```
    x=seq(-40,40 ,by=1)
    y=dnorm((x)/16, mean=0, sd=1)*OutChinaMax*16
    z=dnorm(x/80, mean=0, sd=outchina_sd)*InChinaMax


    plot(x,y, main="R-norm Distribution",
        sub="Trying to Match InChina Data",
        xlab="Days Observed", ylab="Confirmed Cases")
    dev.off()
```

```
## pdf
##   2
```

```
    ## put two plots side-by-side
    par(mfrow=c(1,2))
      plot(x=g$YEARMD,y=g$Total,
      xlab = "Observed Date", ylab = "Total = Confirmed - Recovered")
    ## plot to the screen
    plot(x,y, main="Rnorm Distribution",
        sub="Trying to Match InChina Data",
        xlab="Days Observed", ylab="Confirmed Cases")
```




```
    sprintf("%s Days Till 2020 Olympics in Tokyo", days_till_olympics)
```

```
## [1] "144 Days Till 2020 Olympics in Tokyo"
```

Conclusion

We try to construct an R-normal bell curve that matches the nCoV pattern. Virus such as SARS and the Flu typically follow a bell curve. People with the virus recover or die at the same rate that they caught the virus and the curve slopes down slowly at first, and then steeply.

The -40 on the x-axis represents 40 days ago, when the first sample is plotted. The 0 on our curve represents today, the curve predicts that 40 days from now, the number of cases will decrease greatly in China, 100 days before the Olympics, July 24, 2020.