Simple method for estimating daily and total COVID-19 deaths using a Gumbel model

Hiroshi Furutani (✉ nqj37715@nifty.com)

Doshisha University

Tomoyuki Hiroyasu

Doshisha University

Yoshiyasu Okuhara

Kōchi University

Research Article

Keywords: COVID-19, mortality, forecast, extreme values, Gumbel distribution

Posted Date: December 15th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-120984/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Simple method for estimating daily and total COVID-19 deaths using a Gumbel model

Hiroshi Furutani^{1,*}, Tomoyuki Hiroyasu^{1,2}, and Yoshiyasu Okuhara³

¹*Innovative Computing Research Center, Doshisha University, Kyo-Tanabe City, Kyoto, Japan*

²*Department of Biomedical Sciences and Informatics, Doshisha University, Kyo-Tanabe City, Kyoto, Japan*

³*Medical School, Kochi University, Nankoku City, Japan*

*corresponding author: nqj37715@nifty.com

Abstract

The purpose of the present paper is to introduce a method for forecasting daily and total numbers of COVID-19-associated deaths. We apply the Gumbel distribution function for the analysis of time series data of the first wave. The Gumbel distribution function $F(t)$ has a notable property of $F(t) = 1/2.718$ at the node (peak) point of the distribution. Therefore, we can forecast the number of total deaths N . In the present study, the Gumbel model gives the estimate $N \approx 2.718N_1$, where N_1 is the partial sum of the daily numbers of deaths until the day of the peak. The proposed model can also forecast the daily numbers after the peak. We investigated the data of New York City, Belgium, Switzerland, Sweden, and the United Kingdom. The Gumbel model gives reasonable results for New York City, Belgium, and Switzerland. On the other hand, the proposed method underestimates N for Sweden and the United Kingdom. The proposed approach is very simple, and carrying out the analysis is easy. This method uses spreadsheet software for most of the calculations, and no special program is needed.

keywords: COVID-19, mortality, forecast, extreme values, Gumbel distribution

Coronaviruses are large, enveloped RNA viruses, the importance of which has been recognized through the identification of a newly emerged coronavirus as the causative agent of severe acute respiratory syndrome (SARS).¹ The worldwide pandemic called COVID-19 is caused by the most recently discovered coronavirus, SARS-CoV-2, and is “the third documented spillover of an animal coronavirus to humans in only three decades that has resulted in a major epidemic”.² The 2003 outbreak of SARS had a case-fatality rate of approximately 10% (774 deaths), and MERS killed 34% of people with illness between 2012 and 2019 (2494 deaths). While the case-fatality rate of COVID-19 is a few percent, the associated pandemic has caused many more deaths worldwide.³ The international research community is currently fighting this global health crisis. There are numerous efforts to develop methods of predicting the time evolution of a pandemic for the purpose of science-based policy making.⁴⁻⁶

In the first wave of a pandemic in many regions, the daily plot of reported deaths is single-peaked and skewed to the right. The time series data have three phases: (A) an exponential increase in the early stage, (B) the intermediate stage changing from increasing to decreasing, and (C) a slowly decreasing final stage. Taking this characteristic into account, the present study makes use of extreme value (EV) statistics to estimate the daily numbers in (C) and the total number of deaths from the data of (A) and (B).

EV theory provides methods to analyze the maximum (or minimum) value samples derived from distributions, such as normal or exponential distributions. EV theory has been used in many fields to estimate the probability of rare events from observed data.⁹ The Fisher-Tippett type I distribution,¹⁰ i.e., the Gumbel distribution, is a distribution of maximum (or minimum) values from the data of several distributions. The cumulative distribution function of the Gumbel distribution for maximum values is given as

$$F(t) = \exp\{-e^{-y(t)}\}, \quad (1)$$

where $y(t)$ is a linear function of t ,

$$y(t) = \frac{(t - \beta)}{\alpha}, \quad \alpha > 0. \quad (2)$$

Here, α and β are parameters that decide the shape and position of the distribution, respectively. The parameter β corresponds to the position of the node, i.e., the peak position, and is rather easy to estimate from the distribution curve. A notable property of the Gumbel model is that we can estimate the total number N from the data of $t \leq \beta$. To this end, we make use of the value of $F(t)$ at the node:

$$F(\beta) = e^{-1} = 1/2.718. \quad (3)$$

	NYC	BEL	CHE	SWE	UK
N	17413	9236	1637	5668	39339
N_e	17601	10892	1587	3928	29393
N_1	6475	4007	584	1445	10813
N/N_1	2.689	2.305	2.803	3.922	3.638
date of $t = 1$	2020/3/22	2020/3/21	2020/3/18	2020/3/22	2020/3/24
t_{max}	15	15	15	20	20
t_f	63	67	66	115	83
date of $t = t_f$	2020/5/23	2020/5/26	2020/5/22	2020/7/14	2020/6/14
a	11.417	14.246	12.822	15.148	12.322
b	18.264	22.319	20.716	24.148	19.611
t_b	18	22	21	24	19
$n(t_b)$	566.714	285.429	47.444	98.857	942.429
$\hat{\alpha}$	11.426	14.039	12.309	14.617	11.474
t_c	33.911	42.313	38.084	44.993	36.599
$\hat{t}_c$	33.936	41.697	36.561	43.416	34.079

Table 1: Summary of Gumbel analysis. (1) N : total number of deaths, (2) N_e : estimate of N , (3) N_1 : partial sum of daily data, (4) N/N_1 : ratio of N to N_1 , (5) date of $t = 1$: date of $t = 1$ used in the analysis, (6) t_{max} : number of data points used in the regression analysis, (7) t_f : maximum t used in the calculation of N , (8) date of t_f : date of t_f used in the analysis, (9) a : estimate of parameter α in eq. (2), (10) b : estimate of parameter β in eq. (2), (11) t_b : t of the peak position, (12) $n(t_b)$: daily number of deaths for $t = t_b$, (13) $\hat{\alpha}$: estimate of parameter α in eq. (12), (14) t_c : convergence time, (15) $\hat{t}_c$: estimate of t_c using $\hat{\alpha}$.

Since the partial sum of daily reports N_1 with $t \leq \beta$ is

$$N_1 = NF(\beta),$$

we can forecast N by

$$N = N_1/F(\beta) = eN_1 = 2.718N_1. \quad (4)$$

This means that the total number N can be estimated from the early stage of daily data $t \leq \beta$. Using the relation $n(\beta) = N/(\alpha e)$, where $n(\beta)$ is the maximum daily number at $t = \beta$, we can estimate α by

$$\alpha = \frac{N}{en(\beta)}.$$

From this, we define the convergence time t_c by

$$t_c = 2.9702\alpha,$$

using the relation that 95th percentile of the Gumbel distribution is $\beta + 2.9702\alpha$.

The proposed method is simple and easy to use. We have performed the main part of the calculations by using spreadsheet software.

Results

The present study analyzes the daily numbers and the total number of deaths in the first wave of the COVID-19 pandemic. We have selected the data for one city in the United States, New York City, and four European countries, Belgium, Switzerland, Sweden, and the United Kingdom. All plots of time series data for this city and these countries are single-peaked and skewed to the right. New York City reports two types of data: confirmed deaths and probable deaths. Confirmed deaths counts the number of people who had laboratory-confirmed COVID-19. We analyzed the time series data for confirmed deaths in New York City. Moreover, we chose Sweden because of the herd immunity strategy taken by this country.

Table 1 shows the summary of our results. The total populations of this city and these countries are approximately 83.37×10^5 (NYC), 115.90×10^5 (BEL), 85.70×10^5 (CHE), 102.84×10^5 (SWE), and 668.34×10^5 (UK). In the period of the first wave of the pandemic, the total numbers of deaths per 100,000 people are 208.86 in NYC, 79.69 in BEL, 19.10 in CHE, 55.11 in SWE, and 58.86 in UK. These values are divided into three groups: $NYC > \{BEL, SWE, UK\} > CHE$. The Gumbel estimate of the total number N suggests a different grouping. Whereas the estimate N_e for N is reasonable for NYC, BEL, and CHE, the estimated values are unsatisfactory for SWE and UK. This indicates that the mortality of a region does not affect the applicability of the Gumbel distribution for the analysis of a pandemic.

Figure 1: Gumbel model estimation based on the time series data for New York City. The observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $y_e(t) = -1.5998 + 0.08759t$.

New York City

In Figure 1, we report the results of our analysis of the daily data for New York City as function of time. The solid line indicates the theoretical plot of daily numbers $N_e \cdot f(t)$ using the Gumbel distribution. Here, $f(t)$ is the probability density function of the Gumbel distribution. The dashed curve indicates the 7-day moving average of the reported data.

The number of data points used for the regression analysis is 15. The output of the linear regression analysis is presented in the form of $c_0 + c_1t$ with constants c_0 and c_1 . The 95% confidence interval of c_0 is from -1.6130 to -1.5865 , and that of c_1 is from 0.08613 to 0.08905 . The coefficient of correlation is $R(T, Y) = 0.9996$, and the coefficient of determination is $R(T, Y)^2 = 0.9992$. As this large value of the correlation coefficient suggests, the theoretical curve almost perfectly fits the reported data in the exponentially increasing period. The theoretical curve for forecasting in the decreasing phase also shows a reasonable fit to the observation.

Belgium

In Figure 2, we report the results of our analysis of the daily data for Belgium as a function of time. The solid line indicates the theoretical result of the Gumbel model. The dashed curve indicates the 7-day moving average of the reported data. The number of data points used in the regression analysis is 15. The 95% confidence interval of c_0 is from -1.5731 to -1.5602 , and that of c_1 is from 0.06948 to 0.07091 . The coefficient of correlation is $R(T, Y) = 0.9999$, and the coefficient of determination is $R(T, Y)^2 = 0.9997$.

Switzerland

In Figure 3, we show the results for Switzerland. The solid line indicates the theoretical curve of the daily numbers. The dashed curve indicates the 9-day moving average of the reported data. The number of data points used in the regression analysis is 15. The 95% confidence interval of c_0 is from -1.6779 to -1.5535 , and that of c_1 is from 0.07115 to 0.08483 . The coefficient of correlation is $R(T, Y) = 0.990$, and the coefficient of determination is $R(T, Y)^2 = 0.979$.

Sweden

In Figure 4, we report the results of our analysis of the daily data for Sweden as a function of time. The solid line indicates the theoretical curve of the daily numbers. The dashed curve indicates the 7-day moving average of the reported data. The number of data points used in the regression analysis is 20. The 95% confidence interval of c_0 is from -1.6199 to -1.5900 , and that of c_1 is from 0.06477 to 0.06726 . The coefficient of correlation is $R(T, Y) = 0.990$, and the coefficient of determination is $R(T, Y)^2 = 0.979$.

The large discrepancy between the observed data and theoretical prediction seems to suggest that this discrepancy is the result of the herd immunity policy taken by Sweden. Although there has been much discussion on the effect of herd immunity,¹² scientific investigations on this issue are still limited. Therefore, it is too early to make such a conclusion at this stage of our knowledge of COVID-19.

Figure 2: Gumbel model estimation based on the time series data for Belgium. The observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $y_e(t) = -1.5667 + 0.07019t$.

Figure 3: Gumbel model estimation based on the time series data for Switzerland. The observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $y_e(t) = -1.6157 + 0.07799t$.

Figure 4: Gumbel model estimation based on the time series data for Sweden. Observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $y_e(t) = -1.6049 + 0.06602t$.

Figure 5: Gumbel model estimation based on the time series data for the United Kingdom. The observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $y_e(t) = -1.5916 + 0.08116t$.

United Kingdom

The fit for the United Kingdom is reported in Figure 5. The solid line indicates the Gumbel model curve of the daily numbers. The dashed curve indicates the 7-day moving average of the daily numbers. The number of data points used in the regression analysis is 20. The 95% confidence interval of c_0 is from -1.6205 to -1.5617 , and that of c_1 is from 0.07866 to 0.08366 . The coefficient of correlation is $R(T, Y) = 0.998$, and the coefficient of determination is $R(T, Y)^2 = 0.996$.

The output of our analysis for the United Kingdom is $N/N_e = 1.338$, and this ratio for Sweden is 1.443 . The theoretical curve also deviates from the reported data in the same manner as Sweden.

Discussion

The present study demonstrates the advantage of the EV theory for investigating time series data of deaths in pandemic outbreaks. We make use of one class of Gumbel distribution for the extremes of maximum values. The Gumbel distribution provides a means of extrapolation in time with theoretical justification by the EV theory and can be used to estimate the data in the decreasing phase from the data in the exponentially increasing phase.

Gompertz (1825) proposed a formula for describing the mortality rates of the elderly. This law states that death rates increase exponentially with age.¹³ Gumbel discussed the Gompertz's model in one section (6.3.5 Oldest Ages) of his

book.⁷ He applied the Gumbel model of minimum values for the analysis of mortality. The cumulative distribution function is given by

$$F_1(t) = 1 - \exp\{-e^{(t-\beta)/\alpha}\}.$$

The survival function of the Gompertz model $S(t)$ is defined for $t \geq 0$. It can be shown¹⁴ that $S(t)$ is approximated as the survival function of the Gumbel distribution for $-\infty < t < \infty$,

$$S(t) \approx 1 - F_1(t) = \exp\{-e^{(t-\beta)/\alpha}\}.$$

There are theoretical efforts to derive the Gumbel distribution of the smallest values to explain the phenomena of exponentially increasing mortality. Abernethy (1979) applied EV theory to this problem.¹⁵ This previous study used an analogy between death of the organism and the failure of a multicomponent system and mathematically derived this type of Gumbel distribution. We hope that this line of approach can explain the applicability of the Gumbel distribution of maximum values to the pandemic data.

In many regions, it is reported that the outbreak of a second wave of the pandemic began before the first wave of COVID-19 converged to a final stage.¹⁶ We have carried out preliminary calculations for the first wave data of the United States. The dataset is that used in this analysis for European countries. The estimated number of total deaths N_e is approximately 100,000. However, the second wave of the pandemic began in the decreasing phase of the first wave, so we cannot calculate the total number of deaths N from the reported data. Therefore, we are now trying to design a more general tool for analyzing the first and second waves of a pandemic simultaneously.

Methods

Data extraction

From the website for New York City, we downloaded the dataset of the daily number of COVID-19 deaths: <https://www1.nyc.gov/site/doh/covid/covid-19-data-deaths.page>. This dataset includes two types of data: confirmed deaths and probable deaths. This analysis uses the data of confirmed deaths.

For the analysis of the data for Belgium (BEL), Switzerland (CHE), Sweden (SWE), and the United Kingdom (UK), we downloaded datasets from the website of the European Center for Disease Prevention and Control: <https://www.ecdc.europa.eu>. (The filename is COVID-19-geographic-distribution-worldwide-2020-09-12.xlsx.)

Gumbel model

The probability density function $f(t)$ of the Gumbel distribution is given by

$$f(t) = \frac{1}{\alpha} e^{-y(t)} F(t). \quad (5)$$

The mean $\bar{t}$ and variance V of the distribution are

$$\bar{t} = \beta + 0.5772\alpha, \quad V = \frac{\pi^2}{6}\alpha^2,$$

where the mode, which is the peak position of $f(t)$, is β . Useful information for the parameter estimation is obtained for $t = \beta$:

$$F(\beta) = \frac{1}{e}, \quad f(\beta) = \frac{1}{\alpha e}. \quad (6)$$

For the Gumbel distribution, it is easy to calculate the time $t = q(P)$ that satisfies the relation $F(t) = P$ with $0 < P < 1$. Using eq. (1), we obtain the following quantile:

$$q(P) = \beta - \alpha \ln \ln \frac{1}{P}.$$

We present quantiles for $P = 0.5, 0.90$, and 0.95 :

$$q(0.5) = \beta - \alpha \ln \ln 2 = \beta + 0.3665\alpha, \quad q(0.90) = \beta + 2.2504\alpha, \quad q(0.95) = \beta + 2.9702\alpha.$$

Here, $q(0.5)$ is the median of distribution. We define the convergence time t_c measured from the peak position by the quantile of $P = 0.95$ as

$$t_c = 2.9702\alpha. \quad (7)$$

Estimation of daily and total numbers

This subsection describes the process for estimating the daily data and the total number. The date is indexed as $t = 0$ when the first case was reported. Later we will redefine the starting date for the convenience of comparing the data of different regions.

We use the 7-day moving average $n(t)$ as the daily data in this analysis:

$$n(t) = (d_{t-3} + d_{t-2} + \dots + d_{t+3})/7,$$

where d_t denotes the reported daily number. Note that $n(t)$ includes the information of d_{t+1} , d_{t+2} , and d_{t+3} . In the process of calculations, note that, for the data of Switzerland, it is necessary to use the 9-day average because even data of the seven-day average fluctuate considerably around the trend line.

The accumulated number $U(t)$ is calculated using d_i :

$$U(t) = \sum_{i=0}^t d_i.$$

From $U(t)$, we estimate the distribution function $F(t)$.

The first step is the estimation of total number $N = \sum_t n(t)$. To this end, we used eq. (3). Therefore, it is necessary to estimate mode b . We define the estimated mode t_b , which takes an integer value recursively defined as

$$t_b = \arg \max_t f(t) \ (t \leq t_b) \quad \text{and} \quad f(t_b) = \max\{f(t_b), f(t_b + 1), \dots, f(t_b + 5)\}.$$

We define the number N_1 as the partial sum of daily data $\{d_t, t \leq t_b\}$. The estimate of the total number N_e can be calculated by

$$N_1 = U(t_b), \quad N_e = eN_1 = 2.718N_1. \quad (8)$$

The next step is the estimation of daily number $n(t)$. The distribution function $F(t)$ is approximated using $U(t)$ and N_e :

$$\tilde{F}(t) = U(t)/N_e. \quad (9)$$

The linear function $y(t) = (t - \beta)/\alpha$ is estimated by the regression analysis of the data of intermediate estimate $\tilde{y}$:

$$\tilde{y}(t) = -\ln\{-\ln\tilde{F}(t)\}.$$

Next, we rewrite the definition of t using $\tilde{F}(t)$ of eq. (9). We set the first time satisfying the condition $\tilde{F}(t) > 0.01$ as $t = 1$. Setting the final time t_f as the smallest t satisfying

$$t > t_b, \quad \text{and} \quad n(t) < 0.1n(t_b),$$

and we define the “observed” total number of deaths by

$$N = U(t_f).$$

Forecasting of time series data

We calculate the estimate of $y(t)$ from the intermediate estimate $\tilde{y}$ using the regression analysis program. After performing a linear regression of the dependent variable $\tilde{y}$ and the independent variable t , we obtain two estimated parameters a and b for α and β , respectively. The present study uses t_{max} points for the regression :

$$T: 1, 2, \dots, t_{max} \quad \text{and} \quad Y: \tilde{y}(1), \tilde{y}(2), \dots, \tilde{y}(t_{max}).$$

Here, time (day) starts from $t = 1$. We define the estimate of $y(t)$ as

$$y_e(t) = \frac{t - b}{a}. \quad (10)$$

The goal is to calculate the estimate of $F(t)$:

$$F_e(t) = \exp\{-\exp[-y_e(t)]\} = \exp\{-\exp[-(t - b)/a]\}. \quad (11)$$

The daily moving average $n(t)$ is estimated by

$$n_e(t) = \{F_e(t) - F_e(t - 1)\}N_e.$$

There is another way to estimate α . From eq. (6), we obtain an estimate of α as

$$\hat{\alpha} = \frac{N_e}{en(t_b)} = \frac{N_1}{n(t_b)}, \quad (12)$$

and a direct estimate of t_c is calculated by

$$\hat{t}_c = 2.9702\hat{\alpha}. \quad (13)$$

Note that this estimation uses N_1 and $n(t_b)$, both obtainable directly from observed data.

References

- [1] P. S. Masters (2006), The molecular biology of coronaviruses, *Adv. Virus Res.* **66**, 193–292.
- [2] Coronavirus Study Group of the International Committee on Taxonomy of Viruses (2020), The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and named it SARS-CoV-2, *Nat. Microbiol.* **5**, 536–544.
- [3] E. Mahase (2020), Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality, *BMJ* **368**, m641.
- [4] J. C. Miller (2012), A note on the derivation of epidemic final sizes, *Bull. Math. Biol.* **74**, 2125–2141.
- [5] F. Brauer (2019), Early estimates of epidemic final sizes, *J. Biol. Dyn.* **13**, 23–30.
- [6] D. Fanelli and F. Piazza (2020), Analysis and forecast of COVID-19 spreading in China, Italy and France, *Chaos, Solitons and Fractals* **134**, 109761.
- [7] E. J. Gumbel (1958), *Statistics of Extremes*, Columbia University Press, New York.
- [8] S. Coles (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, London.
- [9] Z. S. Asadi and R. E. Melchers (2017), Extreme value statistics for pitting corrosion of old underground cast iron pipes, *Reliab. Eng. Syst. Saf.* **162**, 64–71.
- [10] R. A. Fisher and L. H. C. Tippett (1928), Limiting forms of the frequency distributions of the largest or smallest numbers of a sample, *Proc. Cambridge Phil. Soc.* **24**, 180–190.
- [11] D. J. T. Carter and P. G. Challenor (1983), Methods of fitting the Fisher-Tippett type 1 extreme value distribution, *Ocean Engng.* **10**, 191–199.
- [12] D. Jones and S. Helmreich (2020), A history of herd immunity, *Lancet* **396**, 810–811.
- [13] B. Gompertz (1825), On the nature of the function expressive of the law of human mortality, and on a new method of determining the value of life contingencies, *Phil. Trans. R. Soc.* **115**, 513–585.
- [14] P. Pflaumer (2007) Life table forecasting with the Gompertz distribution, *JSM Proc.* Alexandria, VA, 3564–3571.
- [15] J. D. Abernethy (1979), The exponential increase in mortality rate with age attributed to wearing-out of biological components, *J. Theor. Biol.* **80**, 333–354.
- [16] G. Cacciapaglia, C. Cot and F. Sannino (2020), Second wave COVID-19 pandemic in Europe: a temporal playbook. *Sci. Rep.* **10**, 15514.

Author contributions

This work has been performed equally by the authors. H.F. wrote the initial manuscript.

Competing interests

The authors declare no competing interests.

Figures

Figure 1

Gumbel model estimation based on the time series data for New York City. The observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $ye(t) = -1.5998 + 0.08759t$.

Figure 2

Gumbel model estimation based on the time series data for Belgium. The observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $ye(t) = -1.5667 + 0.07019t$.

Figure 3

Gumbel model estimation based on the time series data for Switzerland. The observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $ye(t) = -1.6157 + 0.07799t$.

Figure 4

Gumbel model estimation based on the time series data for Sweden. Observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $ye(t) = -1.6049 + 0.06602t$.

Figure 5

Gumbel model estimation based on the time series data for the United Kingdom. The observed data points are indicated by the dashed line, and the theoretical estimation is indicated by the solid line. The linear function used to plot the Gumbel distribution is $ye(t) = -1.5916 + 0.08116t$.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.zip](#)
- [SupplementaryInformation.pdf](#)