International Journal on Infectious Disease and Epidemiology

Bayesian Inference on Numbers of Exposed Hosts at Final Stage of Infectious Diseases: A Case Study on COVID-19 in South Korea

Research Article

Volume 3 Issue 2- 2022

Author Details

K B Lee[1,2], Seung Nam Park[1,2] and Hyong Ha Kim[1]*

[1]Korea Research Institute of Standards and Science, Republic of Korea
[2]University of Science and Technology, Republic of Korea

*Corresponding author

Seung Nam Park, Korea Research Institute of Standards and Science, University of Science and Technology, Republic of Korea

Article History

Received: June 02, 2022 Accepted: June 04, 2022 Published: June 15, 2022

Abstract

As South Korea was approaching the final stage of the coronavirus disease (COVID-19) pandemic, the number of existing exposed hosts in a region, regardless of their symptoms, is curiosity-striking. This paper demonstrates how to statistically infer this number using the daily reported data of confirmed new cases. The number of exposed hosts in a compartment model, including an incubation period, was distributed statistically following a binomial distribution. Using a prior distribution with weak information on the number and a binominal likelihood with the progress rate and observed data, the Markov chain Monte Carlo (MCMC) method enables a sampler to generate a posterior distribution and to create a Bayesian inference on the numbers. The regional inference for a month, from July 11 to August 11, 2020, suggests that the number was between 29 and 35 in Seoul and 2 in Daejeon for the Highest Posterior Density (HPD) interval (3%–97%). The inferences on the temporal drift over the five sub-periods showed a significant reduction in Seoul, Gwangju, and Daejeon, between July 26 and August 6, 2020. Daejeon temporarily approached a regional eradication with a confidence level of 95%. The MCMC method was useful for Bayesian inferences on the number of exposed hosts, in terms of probability density distribution. Any significant temporal change in the number can be confirmed based on the confidence level of the HPD interval.

Keywords: COVID-19; Epidemiology; Mathematical model; Bayesian inference

Introduction

Since the World Health Organization announced coronavirus disease (COVID-19) as a pandemic, the global number of Confirmed New Cases (CNCs) of COVID-19 were hitting record as of mid-August 2020 [1,2]. Amid such a helpless circumstance, its spread in South Korea seemed to slow down in terms of the CNCs. During the initial stage, the highly infectious nature of this novel virus caused panic among both the general public and health authorities, which remains to cause a lot of concern regarding the eradication of this disease. At the initial stage of spread, several quick reports had predicted the size of the epidemic and temporal evolution of the disease based on mathematical models [3-5].

Ever since the outbreak of the COVID-19 pandemic, health authorities have reported the daily accumulated number of CNCs to the public, and it appeared that the pandemic curve was steadily saturating, with the number of CNCs randomly scattered. As the disease seems to be approaching an ending stage in certain communities, it is important for the health authorities to be capable of statistically estimating the number of exposed hosts regardless of the disease's symptoms. A statistical inference on the number of cases in a given city or province can give clues to when they could statistically declare an eradication of the disease in the community, within a certain level of confidence.

Traditionally, there have been two approaches to modeling the evolution of infectious diseases: deterministic and stochastic. The deterministic approach is based on a system of Ordinary Differential Equations (ODEs). Dependent variables such as the number of susceptible, exposed, and infected ODEs evolve in a deterministic pattern from a set of initial values. The deterministic methods are not realistic in that all variables are treated as real numbers. The stochastic approach can supplement them by considering observation noises and process noises of infectious diseases by chance [6]. The result of the stochastic method asymptotically converges to that of the deterministic method as the number of computational trials increase. Moreover, it is the only method that can uniquely demonstrate the extinction of the disease at a final stage [6]. In the final stage, where the numbers of CNCs are randomly scattered, the stochastic approach can statistically predict the number of individuals in each state of a disease progress. The number jumping from one state to the next follows a probability distribution, parameters of which are described by the transmission rates.

In this study, we first verified a Discrete-Time Stochastic Method (DSM) obeying binomial distribution in comparison with the other method. Following which, the Markov Chain Monte Carlo (MCMC) method was applied to create Bayesian inferences on the number of exposed hosts in several regions of South Korea, where we used a prior distribution with less information and a binomial likelihood, and the daily reported data of CNCs. The regions include five metropolitan cites and a province in South Korea, where the numbers of CNCs remained significant as of August 2020. The inferences were conducted over a period of an entire month and over five sub-periods to depict the temporal evolution of the whole period. The inferences were reported in terms of probability density distributions. Finally, we performed a simple temporospatial correlation analysis between the regions at different sub-periods to show evidence of inter-regional transmissions.

Data

The Korean Center for Disease Control and Prevention (KCDC) has been reporting the spread of the COVID-19 pandemic in South Korea every day on their website. We collected data on the number of CNCs daily of five metropolitan cities and the Gyeonggi province surrounding Seoul [7]. The numbers of CNCs excluded quarantined individuals entering South Korea. Figure1 demonstrates the dataset collected between July 11, 2020 and August 11, 2020. The population size of the cities and provinces, as of July 2020, were collected from the website of the Ministry of Public Administration and Security [8]. To infer the number of exposed hosts, it is essential to know the progress rate, κ = 1/4 (1/day), the inverse of which is the incubation period of the disease. To verify the DSM, we adopted the following model parameters: transmission rate β= 0.5 and isolation rate α= 1/4 (1/day). These parameters were applied to predict the time evolution of COVID-19 in South Korea [3]. The transmission rate was obtained from the reproduction number ratio in COVID-19 cases of Daegu and North Kyongsang Province during early stages of its rapid spread [4].

Methods

Stochastic SEIHR Models

In stochastic compartment models, the transitions between the compartments or states are determined by their transition rates. In the SEIHR model, the compartments consist of susceptible (S), exposed (E), infected (I), hospitalized (H), and recovered (R) states. The deterministic models share transition rates with stochastic SEIHR models. The stochastic models can be simulated by Event-Driven Method (EDM) and DSM. EDM is the most realistic in mimicking the transition between discrete states [6], which can be implemented using Gillespie's algorithm [9]. However, DSM is faster in computation and simpler in implementation than EDM. Thus, to replace EDM with DSM, it is necessary to confirm the computational accuracy of DSM in comparison with EDM.

DSM was applied to study a behavior-disease model by [10] and recently, COVID-19 by [11], where it was assumed that a set of chains through an infection progress is generated following binomial distributions of the transitions of individuals within a discrete unit of time. Each member in a state at time t has a probability with the relevant transition rate multiplied by a time interval Δt, during the interval between t and t+Δt, and jumps to the next state at time t+Δt. The number of newly exposed individuals E+, transiting from the susceptible state to the exposed state during interval Δt, is a random variable obeying a binomial distribution of Bin[S(t), βΔtI(t)/N], where S(t) is the number of susceptible hosts at time t. Similarly, the number of newly confirmed individuals I+, at time t+Δt, is a random variable following a binomial distribution of Bin[E, κΔt] [11]. Considering all possible transition rates, a Markov chain with the net numbers in the states {S(t), E(t), I(t), H(t), R(t)| t = 0, Δt, 2Δt, …} is formed. The next state of the system is determined by the current state through a part of the Markov chain relations:

$$S\,(t+\Delta t) = S(t) - E+$$

$$= S(t) - \mathrm{Bin}\,[S(t),\beta\Delta t I(t)/N], \qquad (1)$$

$$E\,(t+\Delta t) = E(t) + E+ - I+$$

$$= E(t) + \mathrm{Bin}\,[S(t),\, \beta\Delta t I(t)/N] - \mathrm{Bin}\,[E(t),\kappa\Delta t], \qquad (2)$$

$$I\,(t+\Delta t) = I(t) + I+ - H+$$

$$= I(t) + \mathrm{Bin}\,[E(t),\, \kappa\Delta t] - \mathrm{Bin}\,[I(t),\, \alpha\Delta t] \qquad (3)$$

Markov Chain Monte Carlo (MCMC) Method

MCMC refers to a class of methods for sampling from a probability distribution to construct a distribution with the most likelihood. This method consists of two techniques: Monte Carlo and Markov chain. Monte Carlo is a general technique that relies on a series of continuous random sampling to obtain a numerical solution. When the sampling forms a Markov chain through the Markov process, it is not necessary to know the entire history of the sampling process to make the next inferences [12]. Owing to this advantage, MCMC method has been used for Bayesian inferences. Among several MCMC packages, PyMC3 is the most recent open-source package written in Python. Computational models can be specified directly in Python code for automatic Bayesian inferences [13].

Assuming that the number of newly confirmed individuals I+ obeys a binomial distribution, as shown in Section 5.1, the computational model for this study is proposed by defining a prior on the number of the exposed as a discrete uniform distribution with both lower and upper limits, and a likelihood distribution as binomial distribution, as such

$$\mathrm{Prior}\ (E) \sim \mathrm{Uniform}\ (\mathrm{lower} = 0,\ \mathrm{upper}), \qquad (4)$$

$$\mathrm{Likelihood} \sim \mathrm{Binomial}\ (E,\ \kappa\Delta t;\ X), \qquad (5)$$

where E is the number of exposed hosts and X is a set of n observed data points. Using the number of daily CNCs as the observed data and the prior number of exposed, this model allows the MCMC method to generate the samples of a target posterior distribution on a number of exposed hosts. The prior uniform distributions were selected so that its upper limit was sufficiently high to provide less prior information on the number of exposed hosts. In all the inferences, two chains of sampling were obtained by the Metropolis sampler with a sample number of 2,000 or 5,000, following a tuning process with 1,000 samples.

Results

Validation of DSM in Comparison with EDA

To validate DSM in terms of computational accuracy, the progress of COVID-19 was repeatedly calculated by both stochastic models of DSM and EDA using the model parameters described in Section 2 (DATA). (Figure 2) shows 500 traces of the calculation using both methods, of which the mean is depicted in solid black lines. The upper plots were obtained using DSM and the lower plots by EDA using Gillespie's algorithm. At a glance, the difference in scatters between both methods was insignificant. Comparing the mean of traces from both methods, the peak positions of the infected population agrees within 5% in x-axis and 10% in y-axis, respectively. These computations in Python 3.0 with a laptop computer took 1.4 s and 57 s for DSM and EDA, respectively.

The Numbers of Exposed Hosts in the Regions for a Month

As shown in (Figure 1), which represents the data set for a month, the daily reported number of CNCs in the regions appeared randomly scattered. We created Bayesian inferences on a number of exposed hosts in each region for the entire period using the MCMC method, as shown in (Figure 3). The upper limits of the discrete uniform distribution of the prior were selected as 100 for Seoul and Gyeonggi, and as 50 for other regions. Reducing such upper limits to two-thirds does not significantly change the posterior. This confirms that the information from the prior was weak and the inference mainly depended on the observed data through the most likelihood. The plots on the right-hand side depict sampled values (y-axis) from the posterior distribution for the sampling numbers (x-axis, 5000 times). The plots on the left-hand side represent Probability Density Functions (PDFs) obtained by the sampling results on the right-hand side. The x-axis and y-axis represent the number of exposed hosts and probability density, respectively. Each PDF with a color on the left-hand side corresponds to the Markov chain with the same color on the right-hand side.

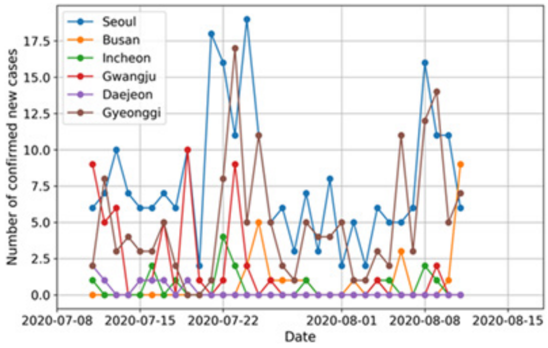

Figure 1: Daily reported number of confirmed new cases of COVID-19 in 5 metropolitan cites and Gyeonggi province of South Korea between July 11 and August 11, 2020.

In general, a PDF provides much more information than point estimations of the mother population, such as its mean, Standard Deviation (SD), and confidence interval of the mean. If the distribution does not appear normal, as shown in (Figure 3), it is highly recommended to represent the statistics in the PDFs. The repeatable PDFs, given by two independent samplings, ensured that the samplings and the resultant inferences were acceptable. In (Table 1), the PDFs are summarized in terms of the mean, SD, and Highest Posterior Density (HPD) interval between 3% and 97%. Note that the HPD intervals in (Table 1) are expressed in natural numbers because the priors were discrete uniform distributions and the sampling outputs were integers.

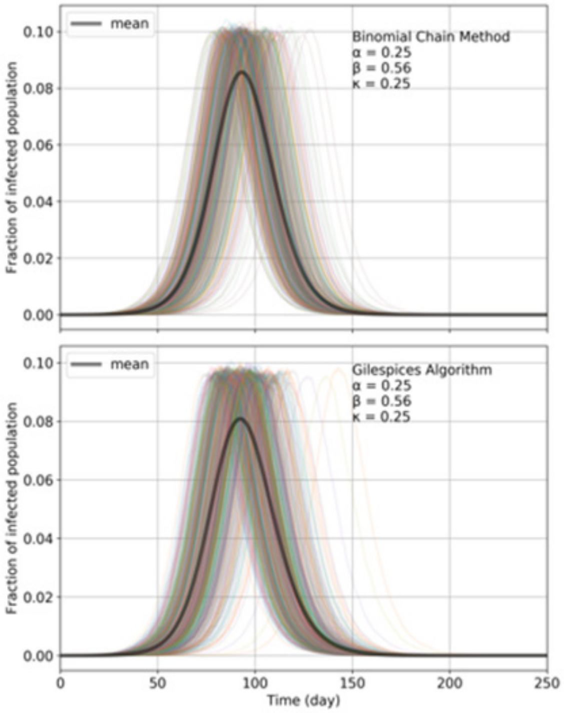

Figure 2: Comparison of two stochastic methods: the binomial chain method (BCM) and event-driven approach (EDA) in predictions of time evolutions of the pandemic, of which the model parameters are relevant to those of an early stage of COVID-19 in South Korea. The computation trials were 500 times. In addition, EDA applied Gillespie's algorithm.

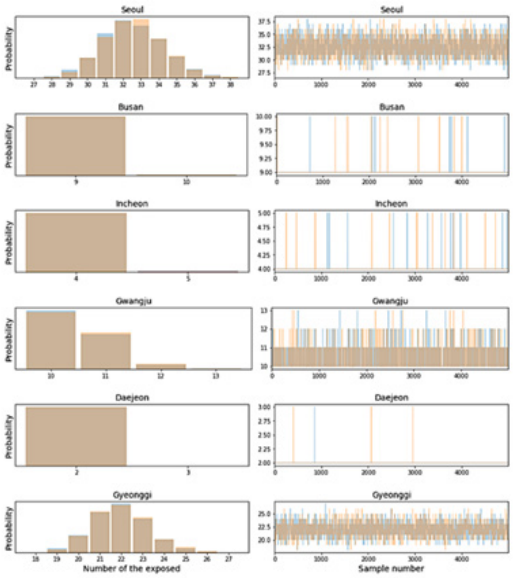

Figure 3: The inferences on the regional number of exposed hosts represented by the samples (the right-hand side) obtained by the MCMC method and its probability density distributions (the left-hand side). These inferences used observed data reported daily for a month between July 11 and August 11, 2020.

The Numbers of Exposed Hosts in the Regions for Five Sub-Periods

We repeated the Bayesian inference on the number of exposed hosts in the regions while moving an observation window with a width of 10 days over the entire period. The period was divided into 5 sub-periods with an interval of 10 days for this study to infer the numbers of exposed hosts in all regions as a function of the sub-periods. The halves of the intervals overlapped with the adjacent (next and/or previous) intervals to reveal moving averages as the sub-period changes. The sub-periods are denoted by titles on all plots in (Figure 4), which can be interpreted similarly as shown in (Figure 3), except that all regional distributions are simultaneously plotted on a plot. Again, each plot on the right-hand side shows 2,000 sampled values, while each plot on the left-hand side represents PDFs. Some PDFs clearly overlapped: Gyeonggi/Gwangju in Period 0, Busan/Incheon in Period 1 and 2, Gwangju/Incheon and Seoul/Gyeonggi in Periods 3 and 4. (Table 2) is included to help readers discriminate between the overlapped PDFs. (Table 2) summarizes the PDFs in (Figure 4) in terms of the mean, SD, and HPD interval between 3% and 97%. The HPD interval in Bayesian statistics roughly corresponds to the confidence interval of classical statistics. For instance, we can statistically inform the relevant health authority that the number of exposed hosts between 30 and 39 existed with a confidence level of 95% in Seoul between 11 and 22 July 2020.

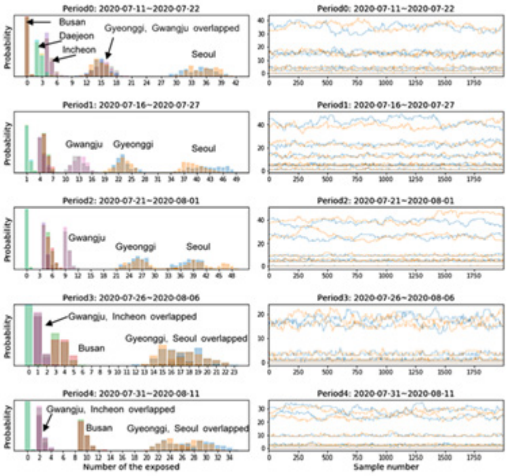

Figure 4: Posterior probability density distributions (the left-hand side) of regional number of exposed host and samples (the right-hand side) for the 5 sub-periods, in which the entire period was divided, with 5 days overlapping, to observe temporal drifts of the probability distributions.

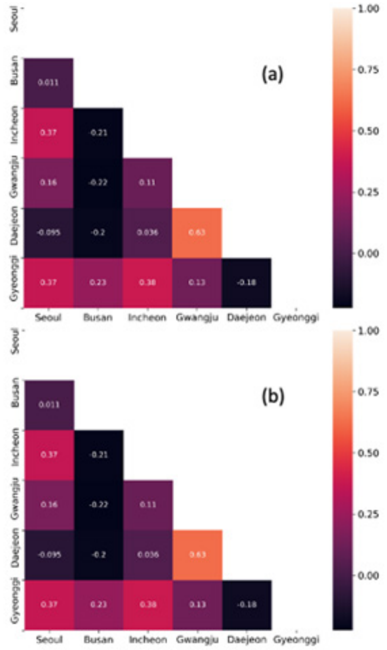

Figure 5: Temporospatial correlation analysis charts between different regions and two time periods. The upper chart (a) is for the period between July 11 and August 11 and the lower (b) for the period between July 11 and August 16, 2020.

Temporospatial Correlations of Numbers of the Confirmed New Cases (CNCs)

When deciding whether it is necessary to include an inter-regional transmission in a mathematical model of infectious diseases, a simple temporospatial correlation analysis may be useful. During the period between July 11 and August 11, 2020, a maximum of 10 CNCs in a day were reported in Gwangju, while zero to two CNCs were intermittently reported in Daejeon. The KCDC confirmed that cases in these two cities were connected due to a mutual gathering of some of their residents. We constructed a correlation coefficient chart, as shown in (Figure 5(a)), which shows the highest correlation coefficient (0.63) between the two cities. Considering negative coefficients as a noise level, the correlation between Gwangju and Daejeon is temporally the most significant in all different combinations.

A few days after August 12, rapid increases in CNCs were reported in Seoul and regions (Incheon and Gyeonggi) around Seoul. (Figure 5(b)) shows the correlation coefficient chart reconstructed for the period between July 11 and August 16, 2020. It shows the strongest correlation of 0.95 between Seoul and Gyeonggi and the second strongest correlation of 0.86 between Seoul and Incheon. It is also worth noting that the correlation between Seoul and Busan has grown to a similar level of correlation (0.54) between Gwangju and Daejeon.

Discussion

We have validated DSM in comparison with EDA, which is more fundamental in the stochastic methods. With a trial computation number of 500, the discrepancy between the mean of the traces was less than 10%. Even if we expect the discrepancy to decrease as the number of trials increases, we can accept 10% as the computational accuracy of DSM for the inferences. Considering the simplicity of implementing DSM and a high computational speed, 40 times faster than EDA, we may consider the inaccuracy of DSM to be compensated by these advantages.

Owing to the randomness in the daily reported numbers of CNCs during the entire period under investigation, as shown in (Figure 2), it is not easy to distinguish a significant trend or pattern in each region. Therefore, using the MCMC method, we created Bayesian inferences of the number of exposed hosts, which was a source of the new cases (Figure 3) shows the posterior distributions of the number in each region. The peak positions and shapes of the PDFs differ region to region. The distributions of Seoul and Gyeonggi have more normal distributions than the other regions. This is because the binominal distribution converges to a normal distribution as the trial number (in this case, the trial number is the number of exposed hosts) increases. As shown in (Table 1), the normality of the distributions of Seoul and Gyeonggi reflects the fact that the confidence interval (mean ± 2SD) almost agrees with the interval between PHD 3% and PHD 97%. On the other hand, because of the small numbers of exposed hosts, the other regions show different uncertain behaviors. Regions with numbers less than 10 are expected to have deterministic numbers (9 in Busan, 4 in Incheon, and 2 in Daejeon) with a confidence level of 95%.

Looking into (Table 2) and (Figure 4) allows us to identify significant temporal changes in the mean for few regions such as Seoul and Gyeonggi. In the case of Seoul, the mean over Period 3 is significantly less than that over Period 0 and 1. Gyeonggi shows a steady increase until Period 2 following a slight decrease, and finally results in an overlap with Seoul over Period 3. Then, both regions show increases in the mean during Period 4. In the case of Gwangju, where an inter-regional transmission with Daejeon was reported, it steadily decreased until Period 3. This decrease was synchronized with that of Daejeon, where regional eradication was expected after Period 1. Since a significant appearance of exposed hosts has not been reported at Period 0, Busan has shown a steady increase in the number until Period 4 after no significant appearance of exposed hosts during Period 0. This increase is contradictory to other regions, which might be reflected in some of the negative correlation coefficients, as shown in (Figure 4). It has been confirmed that the cases in Busan had a separate transmission chain triggered by an international vessel with its sailors anchored to the Busan International Harbor.

The arguments from (Table 2) and (Figure 4) are consistent with the regional correlation coefficient chart for a certain period between July 11, 2020 and August 11, 2020, as shown in (Figure 5(a)). The negative correlation coefficient of Busan with other regions, except Gyeonggi, did not seem to be of noise; however, it was related to a different transmission channel. As shown in (Figure 5(b)), the regional correlation chart, including more data for several days after August 12, shows the strongest correlation of 0.95 between Seoul and Gyeonggi, which seems to be led by a rapid increase of CNCs in both regions. This abrupt increase may have originated from Period 4, as expected in previous arguments. Such an increase in Seoul and Gyeonggi also made it reasonable to raise concerns that it could have affected and refired a mass infection in Daejeon, where at least a temporary local eradication could have been declared according to previous statistical inference.

In summary, when approaching an ending stage of infectious diseases such as the COVID-19 pandemic, the daily reported number of CNCs may appear random, and it is natural to be curious about how many exposed hosts statistically exist in a certain community. Noting that the number of CNCs follows a binomial distribution with an unknown number of exposed hosts and the probability of a successful new case generation with the progress rate κ, the MCMC method enables us to infer the numbers of exposed hosts in the community. Using the data of CNCs reported daily by the KCDC, we demonstrated how to create a Bayesian inference on the number of the exposed hosts for a certain period of time. Dividing the entire period into sub-periods allows for additional inferences regarding the temporal evolution of results. Additionally, we demonstrated that the additional temporo-spatial correlation analysis can substantially deepen the knowledge obtained by the Bayesian inference alone, using the MCMC method.

Acknowledgements

This study was conducted to express our strong solidarity with experts from medical sectors and quarantine authorities and to convey our endless respect to them. We strongly appreciate Dr. Hyojung Lee from the National Institute of Mathematical Science, who read our manuscript and provided valuable comments.

References

1. World Health Organization (2020) WHO Statement regarding cluster of pneumonia cases in Wuhan, China.
2. World Health Organization (2020) WHO Director General's opening remarks at the media briefing on COVID-19.
3. Ki M (2020) Epidemiologic characteristics of early cases with 2019 novel coronavirus (2019-nCoV) disease in Korea. Epidemiol Health 42: e2020007.
4. Choi S, Ki M (2020) Estimating the reproductive number and the outbreak size of COVID-19 in Korea. Epidemiol Health 42: e2020011.
5. Kim S, Yu B, Jung E (2020) Prediction of COVID-19 transmission dynamics using a mathematical model considering behavior changes in Korea. Epidemiol Health 42: e2020026.
6. Keeling M J, Rohani P (2008) Modeling infectious diseases in humans and animals. 200-205.
7. Korea Center for Disease Control and Prevention: Daily report on COVID-19 Daily Confirmed New Cases in South Korea.
8. Ministry of Public Administration and Security: Regional population statistics of the Republic of Korea.
9. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 22(4): 403-434.
10. Perra N, Balcan D, Goncalves B, Vespignani A (2011) Towards a characterization of behavior-disease models. PLoS ONE 6(8): e23084.
11. He S, Tang S, Rong L (2020) A discrete stochastic model of the COVID-19 outbreak: Forecast and control. Mathematical Bioscience and Engineering 17(4): 2792–2804.
12. Park S N, Shin H S, Cho H, Kim M S (2020) Long-term drift analysis of Zener voltage standards and proposal of an initial calibration interval using calibration records accumulated for 15 years. Metrologia 57(6): 065007.
13. Salvatier J, Wiecki T V, Fonnesbeck C (2016) Probabilistic programming in Python using PyMC3. PeerJ Computer Science 2: e55.

Citation: Lee K Park SN, Kim HH. Bayesian Inference on Numbers of Exposed Hosts at Final Stage of Infectious Diseases: A Case Study on COVID-19 in South Korea.*Int Jr Infect Dis & Epidemlgy*. 2022;3(2):1–5. DOI: 10.51626/ijide.2022.03.00023