ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1, Issue 1, April 2021, pp. 13-19

Risk Analysis on the Growth Rate of Covid-19 Cases in Indonesia Using Statistical Distribution Model

Dina Tri Utari [a,1,*], Andrie Pasca Hendradewa [b,2,*]

[a] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia
[b] Department of Industrial Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia
[1] dina.t.utari@uii.ac.id*; [2] andrie.pasca@uii.ac.id*
* corresponding author

ARTICLE INFO

Article history
Received February 12, 2021
Revised March 25, 2021
Accepted April 15, 2021

Keywords
covid-19
growth rate
risk analysis
statistical model
geometric distribution

ABSTRACT

Coronavirus or Covid-19 outbreak has been declared as a pandemic and many countries were not ready to deal with such an eventuality. The highly rapid rate of transmission is one reason for the need to take mitigation measures, since healthcare system has limited capacity. Indonesia is one of the countries that has lost medical resources to the pandemic. In order to provide more comprehensive information about the characteristics of Covid-19 in Indonesia, risk analysis of the occurrence of new cases was needed. This study proposes a related overview about risk occurrence of new Covid-19 cases per daily basis by performing distribution fitting technique to form a statistical distribution model. Among the available alternative models, Geometric distribution is the most suitable to describe the growth of new cases in Indonesia.

1. Introduction

In mid-March 2020, the World Health Organization (WHO) officially declared the coronavirus disease (Covid-19) as a pandemic [2]. The disease has passed through the phases of an outbreak and epidemic, as that occurred for H1N1 or swine flu in 2009. Starting from its outbreak in the city of Wuhan, Hubei Province, China in December 2019, the virus then spread to many regions and 114 countries across the globe, and shocked many people for being unprepared in mitigating the impacts of this pandemic. By September 20, 2020, the overall cases worldwide had reached 30.9 million recorded cases with a death toll of 3% [9]. Based on statistical data, although the cure rate for this disease reaches 72%, the high transmission rate leads to the surge of new confirmed cases, thus making the medical resources overwhelmed for lacking capacity.

The skyrocketing infirmed cases of Covid-19 that exceed the capacity of medical resources leads to longer case handling. Until August 28, 2020, it was noted that the occupancy rates of the isolation rooms and ICU bed of 67 referral hospitals in Jakarta reached 69 and 77 percent, respectively [7]. However, the medical resources are not only experiencing limited capacity in terms of medical equipment and isolation or inpatient rooms, but also in terms of medical personnel and doctors. About 92 health workers died for every 100,000 Covid-19 cases identified in Indonesia, according to the

p-ISSN 2798-253X
e-ISSN 2798-3153

Jakarta Globe calculation based on data on 12 September 2020. This number is the fourth-highest fatality rate for health workers across the world after Mexico, Egypt, and the United Kingdom [4].

Various efforts have been made by the Indonesian Government through the National Disaster Mitigation Agency (BNPB) to tackle the Covid-19 problem, one of which is importing medical equipment, in-vitro diagnostic medical devices and household health supplies for mitigation [5]. As seen from a risk analysis, the Covid-19 pandemic case is a risk with a large impact, so it requires proper mitigation to minimize losses.

The above description underlines that the root cause of this pandemic is the massive transmission rate and the rapidly escalating number of new cases in the community. Therefore, minimizing the occurrence of new cases can be considered as a preventive solution to reduce the impact of potential losses incurred. To describe the characteristics of the emergence of Covid-19 cases in Indonesia, a statistical distribution model can be used since almost all events or phenomena can be converted into a model. Although there is no model that can perfectly replicate actual events, the information of a case or phenomenon can be better understood by simplifying the problem into a statistical model. Through a statistical distribution model, this study provides a risk analysis that can be taken into consideration by the public in their activities during a pandemic based on the characteristics of Covid-19 cases growth in Indonesia.

2. Method

2.1. Random Variable

A random variable, say X, is a function defined over a sample space, *S*, that associates a real number, X(e)=x, with each possible outcome e in S [1].

1) Discrete Random Variable
   If the set of all possible values of a random variable, X, is a countable set, $x_1, x_2, \ldots, x_n$, or $x_1, x_2, \ldots$, X is called a discrete random variable. The function
   $f(x)=P[X=x] \qquad x=x_1, x_2, \ldots$ (1)
   That assigns the probability to each possible value x will be called the discrete Probability Density Function (PDF).
   The Cumulative Distribution Function (CDF) of a random variable X is defined for any real x by
   $F(x)=P[X\leq x]$ (2)
2) Continuous Random Variable
   A random variable X is called a continuous random variable if there is a function f(x), called the PDF of X, such as the CDF that can be represented as
   $F(x)=\int_{-\infty}^{x} f(t)dt$ (3)

2.2. Cumulative Distribution Function (CDF)

The CDF of a random variable X is $F(x):=P[X\leq x]$. When an independent and identically distributed (iid) sample $X_1, \ldots, X_n$ is given, the CDF can be estimated by the Empirical Distribution Function (ECDF) [6]

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i\leq x\}} \quad (4)$$

where $1_A:=\begin{cases}1, \text{ A is true}\\0, \text{ A is false}\end{cases}$ is an indicator function.

Continuous random variables are either characterized by the CDF F or the PDF $f=F'$, which represents the infinitesimal relative probability of *X* per unit of length. We write X~F (or X~f) to denote that X has a CDF F (or a PDF f). If two random variables, X and Y, have the same distribution, we write X=Y.

p-ISSN 2798-253X
e-ISSN 2798-3153

2.3. Fitting Distribution

The purpose of fitting a distribution is to predict the probability or forecast the frequency of occurrence of events in a certain interval. By ordering the goodness of fit of various distributions, a decision can be made about which distribution is acceptable and matches the data used.

Assume that an iid sample $X_1, \ldots, X_n$ from the distribution F is given. Tests for the null hypothesis

$$H_0 : F=F_0$$

are against the most general alternative

$$H_1 : F\neq F_0$$

where F_0 is a pre-specified, not-data-dependent, distribution model. If some parameters of F_0 are estimated from the sample, the presented tests will not respect the significance level (α) for which they are constructed, and as a consequence they will be highly conservative.

The next two well-known goodness of fit tests are as follows [6].

1) Kolmogorov-Smirnov Test
   a. Test purpose. Given $X_1, \ldots, X_n \sim F$, test $H_0 : F=F_0$ vs. $H_1 : F\neq F_0$ consistently go against all the alternative to F_0.
   b. Statistic definition. The test statistic uses the supremum distance between F_n and F_0:
   $$D_n := \sqrt{n} \sup_{x\in R} |F_n(x)-F_0(x)| \quad (5)$$
   If $H_0 : F=F_0$ holds, D_n tends to be small. Conversely, when $F\neq F_0$, larger values of D_n are expected, and the test is rejected when D_n is large.
   c. Statistic computation. The computation of D_n can be efficiently achieved by realizing that the maximum difference between F_n and F_0 happens at $x=X_i$, for a certain X_i. From here, sorting the sample and applying the probability transformation F_0 gives the following function:
   $$D_n := \max\left(D_n^+, D_n^-\right), \quad (6)$$
   $$D_n^+ := \sqrt{n} \max_{1\leq i\leq n} \left\{\frac{i}{n} - U_{(i)}\right\}, \; D_n^- := \sqrt{n} \max_{1\leq i\leq n} \left\{U_{(i)} - \frac{i-1}{n}\right\}$$
   where $U_{(j)}$ stands for the j-th sorted $U_i := F_0(X_i)$, $i = 1, \ldots, n$.
   d. Distribution under H_0. If H_0 holds and F_0 is continuous, D_n has an asymptotic CDF given by the Kolmogorov–Smirnov's K function:
   $$\lim_{n\to\infty} P[D_n\leq x] = K(x) := 1-2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2x^2} \quad (7)$$
   e. Highlights and caveats. The Kolmogorov–Smirnov test is a distribution-free test because its distribution under H_0 does not depend on F_0, but only if F_0 is continuous and the sample $X_1, \ldots, X_n$ is also continuous, i.e., the sample has no ties. If these assumptions are met, the iid sample $X_1, \ldots, X_n \sim F_0$ generates the iid sample $U_1, \ldots, U_n \sim U(0, 1)$. As a consequence, the distribution of (6) does not depend on F_0. If F_0 is not continuous or there are ties on the sample, the K function is not the true asymptotic distribution. A possibility if there are ties on the sample is to perturb the sample slightly in order to remove them.
2) Anderson-Darling Test
   a. Test purpose. Given $X_1, \ldots, X_n \sim F$, test $H_0 : F=F_0$ vs. $H_1 : F\neq F_0$ consistently go against all the alternative to F_0.
   b. Statistic definition. The test statistic uses a quadratic distance between F_n and F_0 weighted by $w(x)=\left(F_0(x)(1-F_0(x))\right)^{-1}$:
   $$A_n^2 := n \int \frac{\left(F_n(x)-F_0(x)\right)^2}{F_0(x)\left(1-F_0(x)\right)} dF_0(x) \quad (8)$$
   If H_0 holds, A_n^2 tends to be small (because of the denominator), so rejection happens for large values of A_n^2. Note that, compared with W_n^2, A_n^2 places more weight in the deviations that happen on the tails, that is, when $F_0(x)\approx 0$ or $F_0(x)\approx 1$.

   c. Statistic computation. The computation of A_n^2 can be significantly simplified as:

p-ISSN 2798-253X
e-ISSN 2798-3153

$$A_n^2 := -n - \frac{1}{n}\sum_{i=1}^{n}\{(2i-1)\log(U_{(i)}) + (2n+1-2i)\log(1-U_{(i)})\} \quad (9)$$

d. Distribution under H_0. The asymptotic null distribution of A_n^2 when F_0 is continuous is the CDF of the random variable

$$\sum_{k=1}^{\infty}\frac{Y_j}{j(j+1)}, \text{ where } Y_j \sim \chi_1^2, \text{ j=1, ... are iid.} \quad (10)$$

e. Highlights and caveats. As with the previous tests, the Anderson–Darling test is also distribution-free if F_0 is continuous and there are no ties in the sample. Otherwise, the null asymptotic distribution is different from the one of (10). The Anderson–Darling test also presents empirical evidence pointing out that it is more powerful than the Kolmogorov–Smirnov test for a broad class of alternative hypotheses. In addition, due to its construction, the Anderson–Darling test can better detect the situations in which F_0 and F differ on the tails.

3. Results

The number of Covid-19 cases in Indonesia according to Indonesia Covid-19 Task Force (2020) [3] from the first confirmed case until 10th September 2020 is shown in Figure 1 below. The graph shows the number of daily new cases for the last 6 months, which is represented by X-axis, while the occurrence probability of daily new cases is represented by Y-axis. Each probability value was identified by calculating the unique number of daily new cases divided by the total data record. The number of new cases may vary from one day to another. Therefore, the number of daily new cases can be assumed as an event which is visualized in Figure 1 that contains event (X-axis) and probability of event (Y-axis).

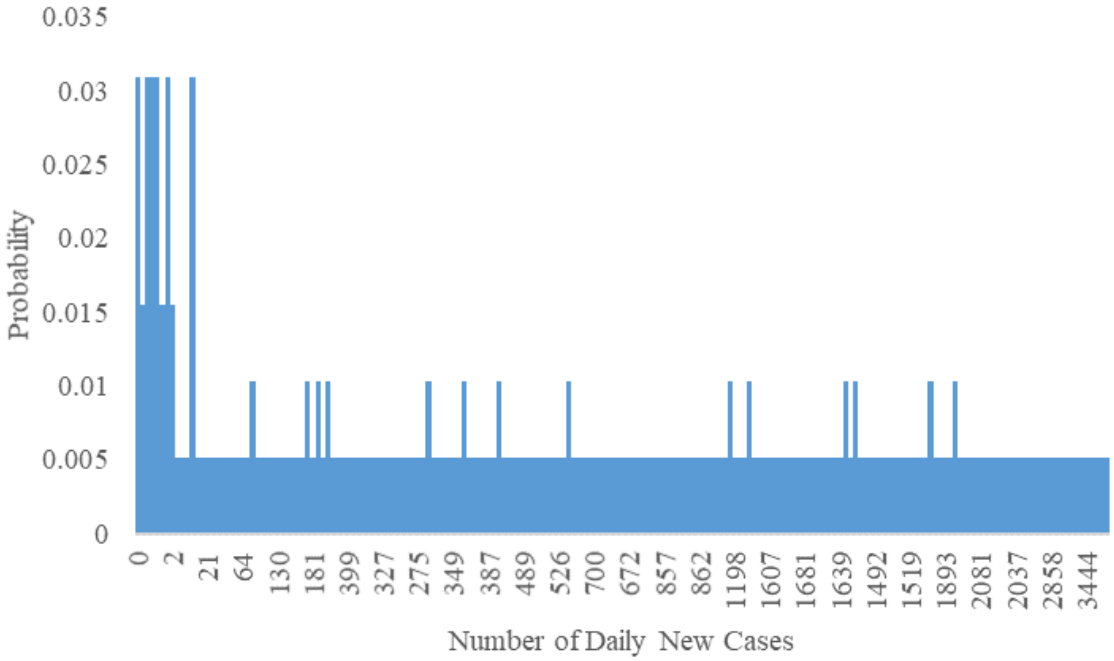

Fig.1. Histogram of daily new Covid-19 cases

The histogram trend as shown in Fig.1 indicates a similar distribution pattern. In order to prove whether the data follows a statistical distribution, a fitting distribution process is required. Since the data is discrete, many possible discrete distributions can be considered as null hypothesis in the fitting distributions process. In order to limit the scope of the fitting process, there are three statistical distribution models chosen as an initial hypothesis in the goodness of fit test. Those three distributions are Geometric, Discrete Uniform, and Negative Binomial. The comparison of goodness-of-fit test results towards the distributions is shown in Table 1 below.

Table 1. Goodness-of-fit test summary

Distributions	Kolmogorov-Smirnov			Anderson-Darling		
	α value			α value		
	0.05	0.02	0.01	0,05	0.02	0.01
Geometric	Fail to Reject	Fail to Reject	Fail to Reject	Reject	Fail to Reject	Fail to Reject
Discrete Uniform	Reject	Reject	Reject	Reject	Reject	Reject
Negative Binomial	Reject	Reject	Reject	Reject	Reject	Reject

Distributions parameter: Geometric (p= 0.0009354), D. Uniform (a=-537, b=2673), Neg. Binomial (n=1, p=0.00124)
Goodness-of-fit details for Geometric distribution: Kolmogorov-Smirnov (Statistic: 0.08878; p-Value: 0.8836), Anderson-Darling (Statistic: 3.2479; p-Value: 1)

According to goodness-of-fit test results using Kolmogorov-Smirnov and Anderson-Darling method in several α values, the null hypothesis of Geometric distribution is almost failed to be rejected. Therefore, the Geometric distribution was then chosen as the fittest statistical model to represent the Covid-19 cases in Indonesia.

4. Discussion

The Geometric distribution has two definitions, such as the number of trials until the first success in a sequence of independent Bernoulli trials, and the number of failures before the first success in a sequence of independent Bernoulli trials. A Bernoulli trial is an experiment that has two results, usually referred to as a "failure" or a "success." The success occurs with probability p and the failure occurs with probability 1-p. "Success" means that a specific event occurred, whereas "failure" indicates that the event did not occur. Because the event can be negative (death, recurrence of cancer, etc) [8], the probability density function (PDF) for Geometric distribution is shown in the following formula:

$$f(x)=p(1-p)^x$$

Meanwhile, the cumulative distribution function (CDF) for Geometric distribution is expressed as the following formula:

$$f(x)=1-(1-p)^{x+1}$$

The PDF represents the probability of getting x failures before the first success, while the CDF represents the probability of getting most of x failures before the first success. In order to analyze the risk of Covid-19 cases growth, the Geometric CDF was then plotted into a graph to give a projection about the probability of an event occurring on or before n^{th} trial. In this case, cumulative probability (p) value can be interpreted as a potential risk of Covid-19 occurrence and the number of trials (n) quantifies the number of activities which are potentially exposed to virus infection. The graph (with additional 5% and 95% potential risk indicator) is shown on Fig.2 below.

p-ISSN 2798-253X
e-ISSN 2798-3153

Fig.2. Probability of Covid-19 occurrence on or before n^{th} potentially exposed activity

The graph shows that the less the activity the less likely the case to occur. Otherwise, the more the activities which are potentially exposed to virus transmission or infection, the more likely the case to occur. It also shows that doing 49 trials/activities can lead to 5% risk of infection, while 95% risk potential can be obtained by doing 3099 potentially exposed activities. The result of research analysis can be considered as the risk analysis of Covid-19 occurrence during the pandemic, only if the parameter (p) does not change significantly. In case, the parameter (p) increases as indicated by the increasing number of new cases, the Geometric distribution curve will get narrower that makes either 5% and 95% risk potential can be reached with fewer number of trials.

5. Conclusion

As the Covid-19 pandemic evolves around the world, there has been an alarming surge of new confirmed cases, especially in Indonesia. The rapid spread is mainly attributed to the highly contagious and easily transmittable characteristics of the Covid-19 virus from one person to another. These extraordinary situations certainly requires an understanding of its risk and the ability to mitigate the risk. This study describes the potential risk of Covid-19 transmission in Indonesia from the initial cases until the period of early September 2020 in order to inform related parties with the proper way to mitigate the impact of the pandemic. The results of the fitting distribution process revealed that the characteristics of the Covid-19 pandemic in Indonesia follows the Geometric distribution with parameter (p) of 0.0009354. The characteristic of an event, which follows the Geometric distribution means that the probability of its occurrence depends on the number of trials mattered. Thus, in the case of Covid-19, the number of trials indicates the amount of activity that potentially leads to the outbreak of new cases. These potentially high-risk activities can be interpreted as activities that are carried out alone or those that involve social interactions. Based on the concept of geometric distribution (which represents Covid-19 cases in Indonesia), the risk mitigation measures to minimize new Covid-19 occurrence are: (1) reducing the probability (p) of occurrence, by way of implementing strict health protocols (e.g: washing hands, wearing face masks, and maintaining body immunity); and (2) reducing the number of trials. (n) by limiting each individual from potentially high-risk activities that may lead to infection (e.g: avoiding the crowd and physical contact, and avoiding touching the face).

p-ISSN 2798-253X
e-ISSN 2798-3153

Acknowledgment

The authors would like to acknowledge the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia for their valuable support in this research.

References

[1] Bain, L. J., and Engelhardt, M.,"Introduction to Probability and Mathematical Statistics", Pacific Grove: Duxbury, 1992.

[2] Ducharme, J., "World Health Organization Declares Covid-19 a 'Pandemic.' Here's What That Means", Retrieved from https://time.com/5791661/who-coronavirus-pandemic-declaration/, on 10 September 2020.

[3] Indonesia Covid-19 Task Force, "Peta Sebaran", Retrieved from https://covid19.go.id/peta-sebaran, on 10 September 2020.

[4] Jakarta Globe, "Indonesia Is Losing Health Workers at an Alarming Rate", Jakarta Globe, Retrieved from https://jakartaglobe.id/news/indonesia-is-losing-health-workers-at-an-alarming-rate, on 20 September 2020.

[5] Mariska, D., "Customs Office Simplifies Import Requirements for Medical Equipment", Jakarta Globe, Retrieved from https://jakartaglobe.id/news/customs-office-simplifies-import-requirements-for-medical-equipment/, on 20 September 20200.

[6] Portugues, E. G., "Notes for Nonparametric Statistics", Madrid: University of Madrid, 2020.

[7] Syakriah, A., Atika, S., "Patients crowd hospitals as Indonesia loses 183 'priceless' medical workers", The Jakarta Post, Retrieved from https://www.thejakartapost.com/news/2020/09/02/patients-crowd-hospitals-as-indonesia-loses-183-priceless-medical-workers.html, on 10 September 2020.

[8] Wicklin, R., "The Geometric Distribution in SAS", SAS: Analytics, Artificial Intelligence, and Data Management, Retrieved from https://blogs.sas.com/content/iml/2020/04/06/geometric-distribution-sas.html, on 15 September 2020 2020.

[9] Worldometer, "Covid-19 Coronavirus Pandemic", Retrieved from, https://www.worldometers.info/coronavirus/, on 20 September 2020.

p-ISSN 2798-253X
e-ISSN 2798-3153