7-2020

COVID-19 Peak Immunity Values Seem to Follow Lognormal Distribution

Julio Urenda

Olga Kosheleva

Vladik Kreinovich

Tonghui Wang

Comments:
Technical Report: UTEP-CS-20-79

COVID-19 Peak Immunity Values Seem to Follow Lognormal Distribution

Julio Urenda[1,2], Olga Kosheleva[3],
Vladik Kreinovich[2], and Tonghui Wang[4]
[1]Department of Mathematical Sciences
[2]Department of Computer Science
[3]Department of Teacher Education
University of Texas at El Paso
El Passo, TX 79968, USA
jcurenda@utep.edu. olgak@utep.edu, vladik@utep.edu
[4]Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003, USA
twang@nmsu.edu

Abstract

For the current pandemic, an important open problem is immunity: do people who had this disease become immune against further infections? In the immunity study, it is important to know how frequent are different levels of immunity, i.e., what is the probability distribution of the immunity levels. Different people have different rates of immunity dynamics: for some, immunity gets to the level faster, for others the immunity effect is slower. Similarly, in some patients, immunity stays longer, it others, it decreases faster. In view of this, an important characteristic is peak immunity. A recent study provides some statistics on the peak immunity. There is not enough data to provide a statistically guaranteed selection of a probability distribution, but we can already make some preliminary conclusions. Specifically, based on the available data, we argue that the COVID-19 peak immunity values follow lognormal distribution.

1 Formulation of the Problem

Immunity studies are important. In an epidemic, some people have a mild version of the disease, some have a strong (or even deadly version). Usually, recovered people gain an immunity, meaning that they cannot be infected again – at least for some time.

For some diseases, vaccines are available that induce immunity.

Once the majority of people gain immunity, the epidemic slows down and stops.

Immunity question is especially important for COVID-19. For the current epidemic, no vaccine is available yet. So, to predict the further dynamics of the epidemic – and to develop methods to decrease its spread – it is very important to study the levels of immunity of people who recovered from this disease.

COVID-19 immunity study. The results of studying immunity of patients who recovered from COVID-19 have recently appeared in the paper [1].

The resulting data. In this study, the immunity level of each patient was studied for a 3-month period. For each patient, an important characteristic is the peak immunity level. After the peak, this level starts decreasing, but the peak level usually indicates how strong will the immunity response become if the body encounters the same disease.

Different patients have different peak immunity level. The paper [1] provides the following information about the number of patients with different peak immunity levels (as measured in appropriate units):

- 3.1% of the patients have peak immunity level below 50;
- 7.7% of the patients have peak immunity level between 50 and 200;
- 10.8% of the patients have peak immunity level between 200 and 500;
- 18.5% of the patients have peak immunity level between 500 and 2000; and
- 60% of the patients have peak immunity level above 2000.

So what is the corresponding distribution? To study the disease, it is important to know how many people have different immunity level, i.e., it is important to know the probability distribution of the peak immunity level.

What we do in this paper. The above data provides, in effect, four independent observations – the fifth number is simply 100% minus the sum of the first four values. Of course, it is not enough to have four observations to determine the probability distribution with a statistical guarantee. To get a guaranteed conclusion, we need to wait until more studies are done.

However, some answers are needed right away. So, what we do is provide a reasonable hypothesis about this distribution – a hypothesis that lead to lognormal distribution, and then we show that the above empirical data is indeed in very good accordance with this distribution.

Comment. It is important to notice that for each moment of time, the actual distribution of immunity is more complex. For example, the distribution for days 40-50 after the onset of symptoms is clearly bimodal [1]. This complexity can be explained by the fact that different people have different rates of immunity change. So:

- when we compare the peak immunity values, we get a meaningful distributions, but
- when we compare values at a given intermediate moment of time, we bring together patients at different stages of the immunity dynamics, as a result we get a more complex distribution – probably a convex combination of several log-normal distributions corresponding to different rates.

2 Analysis of the Problem and the Resulting Hypothesis

Analysis of the problem. Usually, immunity to a disease is affected by many different factors. Factors such as general physical fitness, body-mass index, age, etc., can drastically increase or decrease immunity level. This increase of decrease is usually expressed in percent terms: e.g., a typical statement is "regular physical activities increase immunity by 20%". This means that each of the immunity-affecting factors multiplies the average immunity level by a certain constant – a constant which is larger than 1 is it is an increase, and a constant which is smaller than 1 if it is a decrease.

Thus, to find the joint effect of several factors, we need to multiply the average immunity level by all these factors. Thus, the resulting immunity level is a product of several constants corresponding to different factors.

Each of these constants varies from one person to another, so it can be viewed as a random variable. Constants corresponding to different factors are usually independent: e.g., some young people are physically active some are not, same with older people. So, the immunity level is the product of many independent random variables.

What is the resulting distribution? Situations when we have a product of many independent random variables may not be standard in statistics, but we can easily reduce this situation to a more convenient one if we take the logarithm of the immunity level. It is well known that the logarithm of the product is equal to the sum of the logarithm. Thus, the logarithm of the peak immunity level is the sum of a large number of independent random variables – logarithms of the corresponding constants.

It is known that under reasonable conditions, the sum of a large number n of independent random variables tends to the normal (Gaussian) distribution; this is known as the Central Limit Theorem; see, e.g., [2]. This means that when the value n is large, the corresponding distribution is close to Gaussian. This is the reason why normal distributions are ubiquitous.

In our case, we can therefore conclude that the logarithm $X = \ln(Y)$ of the peak immunity level Y is normally distributed – and thus, that the distribution of the peak immunity level Y itself is log-normal.

Let us see if this conclusion is consistent with the above data.

3 Empirical Data Is in Good Accordance with the Log-Normal Distribution

Let us convert the data into cumulative distribution function values. Probability distributions are usually described in terms of cumulative distribution function (cdf) or in terms of the probability density function (pdf). The above distribution is the easiest to describe in terms of the cdf $F_Y(y) \stackrel{\text{def}}{=} \text{Prob}\,(Y \le y)$:

- for $y = 50$, the probability is

$$F_Y(50) = 3.1\%;$$

- for $y = 200$, the probability is

$$F_Y(200) = 3.1\% + 7.7\% = 10.8\%;$$

- for $y = 500$, the probability is

$$F_Y(500) = 3.1\% + 7.7\% + 10.8\% = 21.6\%;$$

and

- for $y = 2000$, the probability is

$$F_Y(2000) = 3.1\% + 7.7\% + 10.8\% + 18.5\% = 40.1\%.$$

Log-normal distribution: a brief reminder. A random variable Y is said to have a log-normal distribution with parameters μ and σ if $X \stackrel{\text{def}}{=} \ln(Y)$ is normally distributed with mean μ and standard deviation σ; this is usually denoted as $X \sim N(\mu, \sigma^2)$. Let $\Phi(t)$ denote the cdf and the pdf of the $N(0,1)$ distribution. Then the pdf of Y is given by

$$f_Y(y) = \frac{1}{y} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right), \tag{1}$$

where $\ln(y)$ is the natural logarithm of y, and the cdf of Y has the form:

$$F_Y(y) \stackrel{\text{def}}{=} \text{Prob}\,(Y \le y) = \Phi\left(\frac{\ln(y) - \mu}{\sigma}\right). \tag{2}$$

Let us go to logarithms. In view of the definition of log-normal distribution – and of the fact that normal distributions are ubiquitous and, as a result, there exists many tests of normality – it is reasonable to test that $\ln(Y)$ is normally distributed.

The original definition of the log-normal distribution says that the natural logarithm of Y is normal. For our data, as we will see, it is more convenient to use decimal logarithms $\log_{10}(x)$. Since $\log_{10}(Y) = \dfrac{\ln(Y)}{\ln(10)}$, natural logarithm is normally distributed if and only the decimal logarithm is.

So, let us use decimal logarithms. It is well know that

$$2^{10} = 1024 \approx 10^3 = 1000,$$

thus, $\log_{10}(2) \approx 0.3$. Hence,

$$\log_{10}(5) = \log_{10}(10/2) = \log_{10}(10) - \log_{10}(2) \approx 0.7,$$

so we conclude that

$$\log_{10}(50) \approx 1.7, \quad \log_{10}(200) \approx 2.3, \quad \log_{10}(500) \approx 2.7, \quad \log_{10}(2000) \approx 3.3.$$

Thus, for $X = \ln(Y)$, the corresponding pdf $F_X(x) = \text{Prob}\,(X \le x)$ has the following form:

- for $x = 1.7$, the probability is $F_X(1.7) = 3.1\%$;
- for $x = 2.3$, the probability is $F_X(2.3) = 10.8\%$;
- for $x = 2.7$, the probability is $F_X(2.7) = 21.6\%$; and
- for $x = 3.3$, the probability is $F_X(3.3) = 40.1\%$.

Let us prepare for checking that the empirical distribution for $X = \log(Y)$ is normal. If we know the value $p = \Phi(t)$, we can determine the corresponding value t as $t = \Phi^{-1}(p)$, where $\Phi^{-1}(p)$ denotes the inverse function of $\Phi(t)$.

For a general normal distribution X, with mean μ and standard deviation σ, the expression $\dfrac{X - \mu}{\sigma}$ is normally distributed with mean 0 and standard deviation 1. So, $F_X(x) = \Phi\left(\dfrac{x - \mu}{\sigma}\right)$. Thus, if we know the value $p = F_X(x)$, we can conclude that $\dfrac{x - \mu}{\sigma} = \Phi^{-1}(p)$ and thus, that $x = \mu + \sigma \cdot \Phi^{-1}(p)$.

Based on the tables of the function $\Phi(t)$, we conclude that:

$$\Phi^{-1}(0.031) \approx -1.86, \qquad \Phi^{-1}(0.108) \approx -1.27,$$

$$\Phi^{-1}(0.216) \approx -0.79, \qquad \Phi^{-1}(0.401) \approx -0.25.$$

Thus, the above four points lead to the following four equations:

$$\mu - 1.86\,\sigma \approx 1.7; \tag{3}$$

$$\mu - 1.27\,\sigma \approx 2.3; \tag{4}$$

$$\mu - 0.79\,\sigma \approx 2.7; \tag{5}$$

$$\mu - 0.25\,\sigma \approx 3.3. \tag{6}$$

How can we check: first idea. We have four equations that bind two parameters μ and σ. In principle, we can determine μ and σ, e.g., from the first two equations, and see if the resulting values fit the other two formulas. One fit may be a coincidence, but two fits would be a good argument that the distribution is actually log-normal.

Checking confirms log-normal character of the distribution of peak immunity. Subtracting the first equation from the second one, we conclude that $0.59\sigma \approx 0.6$, so $\sigma \approx 1.02$. Thus, from the first equation, we can conclude that

$$\mu \approx 1.7 + 1.86 \cdot \sigma \approx 1.7 + 1.86 \cdot 1.02 \approx 3.60.$$

Substituting the values $\mu \approx 3.60$ and $\sigma \approx 1.02$ into the left-hand sides of the formulas (3) and (4), we get:

$$\mu - 0.79\sigma \approx 3.60 - 0.79 \cdot 1.02 \approx 2.79$$

and

$$\mu - 0.25\sigma \approx 3.60 - 0.25 \cdot 1.02 \approx 3.24.$$

These are indeed good approximations to the observed values 2.7 and 3.3. So, the data indeed seems to confirms the log-normal character of the distribution.

How we can check: second idea. Another possibility is to treat Equations (3)–(6) as a regression model; see, e.g., [2]. In this case, the least squares estimates of μ and σ are

$$\widehat{\mu} = 3.588 \qquad \text{and} \qquad \widehat{\sigma} = 1.018,$$

with the R squared value equal to $R^2 = 0.997$.

This also confirms that the data is consistent with log-normal distribution.

Comment. This may seem convincing, but it is worth reminding that we only have four data points, not enough to make a statistically convincing conclusion.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

[1] J. Seow, C. Graham, B. Merrick, S. Acors, K. J. A. Steel, O. Hemmings, A. O'Bryne, N. Kouphou, S. Pickering, R. Galao, G. Betancor, H. D. Wilson, A. W. Signell, H. Winstone, C. Kerridge, N. Temperton, L. Snell, K. Bisnauthsing, A. Moore, A. Green, L. Martinez, B. Stokes, J. Honey, A. Izquierdo-Barras, G. Arbane, A. Patel, L. O'Connell, G. O'Hara, E. MacMahon, S. Douthwaite, G. Nebbia, R. Batra, R. Martinez-Nunez, J. D. Edgeworth, S. J. D. Neil, M. H. Malim, and K. Doores, *Longitudinal evaluation and decline of antibody responses in SARS-COV-2 infection*, medRxiv 2020.07.09.20148429, doi: https://doi.org/10.1101/2020.07.09.20148429.

[2] D. J. Sheskin, *Handbook of Parametric and Non-Parametric Statistical Procedures*, Chapman & Hall/CRC, London, UK, 2011.

[3] World Health Organization (WHO), *Report of the 2019 WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)*, WHO, February 2020, https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf