Journal of Epidemiology and Global Health
Vol. **11**(**2**); June (2021), *pp.* 146–149
DOI: https://doi.org/10.2991/jegh.k.210108.001; ISSN 2210-6006; eISSN 2210-6014
https://www.atlantis-press.com/journals/jegh

Perspective

COVID-19: Nothing is Normal in this Pandemic

Luzia Gonçalves[1,2,*], Maria Antónia Amaral Turkman[2], Carlos Geraldes[2,3], Tiago A. Marques[2,4,5], Lisete Sousa[2,6]

[1]*Global Health and Tropical Medicine, Unidade de Saúde Pública Internacional e Bioestatística, Instituto de Higiene e Medicina Tropical, Universidade NOVA de Lisboa, Rua da Junqueira 100, Lisboa 1349-008, Portugal*
[2]*CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3]*ISEL - Instituto Superior de Engenharia de Lisboa – Instituto Politécnico de Lisboa, Portugal*
[4]*Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[5]*Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, Scotland*
[6]*Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal*

ARTICLE INFO

Article History

Received 31 August 2020
Accepted 12 December 2020

Keywords

Epidemic curve
normal distribution
log-normal distribution
Gaussian curve
COVID-19

ABSTRACT

This manuscript brings attention to inaccurate epidemiological concepts that emerged during the COVID-19 pandemic. In social media and scientific journals, some wrong references were given to a "normal epidemic curve" and also to a "log-normal curve/distribution". For many years, textbooks and courses of reputable institutions and scientific journals have disseminated misleading concepts. For example, calling histogram to plots of epidemic curves or using epidemic data to introduce the concept of a Gaussian distribution, ignoring its temporal indexing. Although an epidemic curve may look like a Gaussian curve and be eventually modelled by a Gauss function, it is not a normal distribution or a log-normal, as some authors claim. A pandemic produces highly-complex data and to tackle it effectively statistical and mathematical modelling need to go beyond the "one-size-fits-all solution". Classical textbooks need to be updated since pandemics happen and epidemiology needs to provide reliable information to policy recommendations and actions.

1. COVID-19 PANDEMIC AND MODELS BASED ON PREVIOUS LITERATURE

The COVID-19 pandemic brings a new update to the famous George E. P. Box quote "all models are wrong, but some are useful" [1]. Martin Goodson wrote "All models are wrong, but some are completely wrong" [2]. Many models were produced in a short period by the scientific community and, contrary to what is usual, with vast dissemination to a broad audience by TV, newspapers, and social networks. Unfortunately, bad things seem to outmanoeuvre good ones in spreading speed.

In Portugal, like in other countries, many models appeared in an earlier phase of the pandemic with catastrophic numbers of deaths and infected cases, often without a clear distinction among suspected, symptomatic, asymptomatic, confirmed, and reported cases. Some mathematicians were in the front line, but few statisticians appeared in this phase. Undoubtedly well-intentioned, many non-statisticians and non-mathematicians gave their contributions to help in this hard situation by modelling "something", but findings were involved in controversy. In fact, without enough and reliable data, it is impossible to establish good models or reliable predictions.

Worldwide, the motto "Let's flatten the curve" produced a competition between "the best models" to express the number of cases with and without containment measures. In a second phase, a new competition was guessing when would be the peak of the epidemic curve and latter for an end date for this pandemic or the second waves in each country. In Portugal, some serious models were produced, but in newspapers and social networks, some inaccurate references were given to a "normal epidemic curve" and also to a "log-normal curve/distribution", even using classical sampling statistics to describe properties of the epidemic curve. By then, some statisticians intervened to clarify that there was a confusion between "epidemic curve and probability distribution" and the Portuguese Statistical Society took a position to avoid the dissemination of these wrong concepts. In other countries, we found similar approaches. Rashed et al. [3] studied 16 prefectures in Japan and they stated that the number of daily COVID-19 confirmed cases follow bell-shape or log-normal distribution in most prefectures. At the beginning, some of these curves seemed symmetric, but the bell-shape does not represent a normal distribution or a log-normal distribution.

Looking at different time windows, many curves may emerge, taking different forms over time intervals. Figure 1 expresses the dynamics of new reported cases until the 31 July 2020, according to the European Centre for Disease Prevention and Control, in Portugal and in the Netherlands, Gaussian and lognormal functions were fitted to daily cases from first day to the 10 May 2020 in Portugal and 17 May 2020 in the Netherlands. Dashed lines show

*Corresponding author. Email: luziag@ihmt.unl.pt

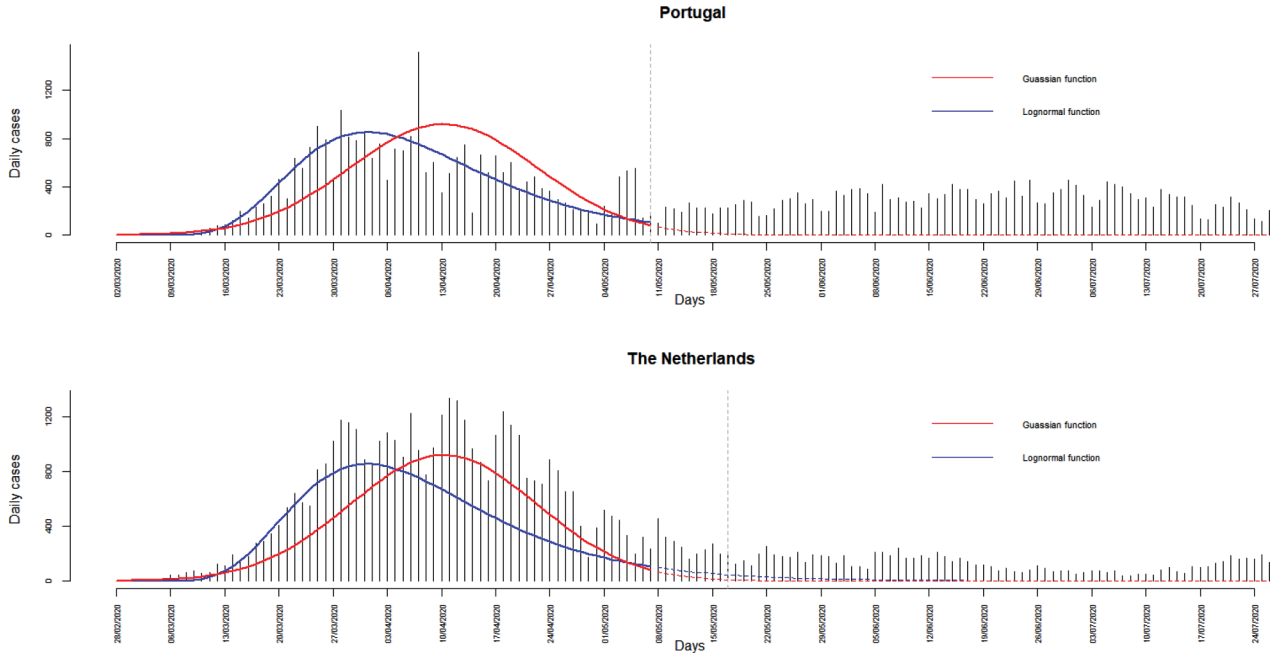

Figure 1 | COVID-19 daily cases in Portugal and in the Netherlands until the 31 July 2020. Fitted models (solid lines) and predicted cases (dashed lines) correspond to Gaussian (red) and lognormal (blue) curves. Green vertical lines correspond to the 10 May 2020 in Portugal and to the 17 May 2020 in the Netherlands.

discrepancies between predicted and observed daily cases, highlighting how bad these popular models are. Section 3 gives some mathematical details and assumptions about the Gaussian curve.

The importance of temporal indexation needs to be present in this type of data and unfortunately this key element has often been ignored. During this debate, advocates of "a normal/log-normal curve" have argued that this is a basic concept explained in first lectures on epidemics and there are many introductory textbooks of epidemiology, describing the epidemic curve as a normal/log-normal distribution. In fact, this confusion appears in several old and recent documents of reputable international institutions, books and important scientific journals [3–8]. For many decades, applied epidemiology training programs of the Centers for Disease Control and Prevention (CDC) have been a crucial impact worldwide and particularly, in African and Asian countries (e.g., Reddy et al. [9]), where outbreaks and epidemics are more frequent. In many low and middle-income countries, teaching materials from several CDC courses are very important references and often unique documents for health professionals. Thus, these crucial documents need to be urgently updated.

2. TIME IS AN IMPORTANT KEY IN EPIDEMIOLOGY AND IN THE EPIDEMIC CURVE

In epidemiology, there are many other curves (e.g., survival curve) and in statistics, there are countless probability distributions (e.g., Weibull, Gamma, and Beta) [10]. These concepts - epidemic curve and probability distribution - are entirely different. In an epidemic situation, we are interested in the epidemic curve that has a particular and fundamental characteristic – cases are indexed by time (e.g., day or week). To visualize an epidemic curve, we put on the vertical axis the number of cases (discrete nature) and the horizontal axis the time unit. The temporal correlation is a key characteristic since new cases are strongly correlated with past cases. The normal distribution is widely used to describe several continuous variables (e.g., body weight). In a probability distribution the values (x) of a random variable (X), are represented on the horizontal axis, and on the vertical axis, it is displayed the density or the probability mass function, according to whether the variable is continuous or discrete. Thus, if we explore the new cases according to a common distribution used to describe the frequency distribution of a random sample, we lose the temporal reference. In fact, the desirable property of independent and identically distributed (i.i.d.) variables, common to most of the biomedical applications, fails in an epidemic curve situation. Moreover, although the epidemic curve may look like a Gaussian curve and be eventually modelled by a Gauss function, it is not a Gaussian (normal) distribution or a log-normal distribution, as some claim [3–7].

3. STATISTICAL RATIONAL

To explain why this confusion may arise, we will consider a toy example with the new cases of an epidemic in T days, denoted by $\{y_t, t = 1, 2, \ldots, T\}$. It may be reasonable to assume that each $\ln(y_t)$ follows a quadratic regression in time. Thus, if so, we can write:

$$\ln(y_t) = a_0 + a_1 t + a_2 t^2 + \varepsilon_t, \quad t = 1, 2, \ldots, T,$$

where ε_t are random variables i.i.d. with a normal distribution with mean 0 and small variance $\sigma^2 < 1$ (although the assumption of independent errors is not very credible in an epidemic situation). This means that,

$$y_t = \exp(a_0 + a_1 t + a_2 t^2 + \varepsilon_t), \quad t = 1, 2, \ldots, T. \tag{1}$$

Considering,

$$a_0 = \ln(a) - \frac{b^2}{2s^2}$$
$$a_1 = \frac{b}{s^2}$$
$$a_2 = -\frac{1}{2s^2},$$

the expression (1) can be written as,

$$y_t = e^{\ln(a)} e^{-\frac{b^2}{2s^2} + \frac{2bt}{2s^2} - \frac{t^2}{2s^2}} e^{\varepsilon_t} = a e^{-\frac{(t-b)^2}{2s^2}} e^{\varepsilon_t}.$$

This last expression, ignoring the error term e^{ε_t} which in principle is small, since it was assumed that $\sigma^2 < 1$, shows that y_t, as a function of time, can be described, approximately, by a Gauss function. Moreover, assuming ε_t i.i.d with $N(0, \sigma^2)$, the random variables $\ln(y_t)$, $t = 1, 2, \ldots, T$, for each t follow a normal distribution with mean $a_0 + a_1 t + a_2 t^2$ (clearly depends on time) and variance σ^2. Consequently, Y_t, $t = 1, 2, \ldots, T$, follows for each t a log-normal distribution with mean

$$e^{\left(a_0 + a_1 t + a_2 t^2 + \frac{\sigma^2}{2}\right)} = e^{\ln(a) - \frac{(t-b)^2}{2s^2} + \frac{\sigma^2}{2}} = a e^{-\frac{(t-b)^2}{2s^2} + \frac{\sigma^2}{2}},$$

and variance

$$e^{2\left(a_0 + a_1 t + a_2 t^2 + \frac{\sigma^2}{2}\right)} \cdot (e^{\sigma^2} - 1) = e^{2\left(\ln(a) - \frac{(t-b)^2}{2s^2} + \frac{\sigma^2}{2}\right)} (e^{\sigma^2} - 1) = a^2 e^{-\left(\frac{t-b}{s}\right)^2} e^{\sigma^2} (e^{\sigma^2} - 1).$$

A similar argument applies when, instead of a lognormal situation as above, it is assumed a model of the type

$$y_t = \exp(a_0 + a_1 t + a_2 t^2) + \varepsilon_t, \quad t = 1, 2, \ldots, T$$

with ε_t independent normal with mean 0 and variance $\sigma^2(t)$ dependent on time. In this case Y_t, $t = 1, 2, \ldots, T$, follows, for each t, a normal distribution with mean $e^{(a_0 + a_1 t + a_2 t^2)} = a e^{-\frac{(t-b)^2}{2s^2}}$.

We stress however that these type of models are not adequate in an epidemic situation since they ignore the dependence inherent among the daily cases.

4. IMPLICATIONS FOR TEACHING, RESEARCH AND PUBLIC HEALTH POLICIES

The assumptions leading to the previous results will seldom be adequate in an epidemic situation and hence it is very unlikely that a Gaussian function will describe properly an epidemic curve. Moreover, in practice, errors are expected to be correlated, not independent. These random variables, Y_t or $\ln(Y_t)$, are not i.i.d. since they have a different distribution for each time t. This makes all the difference, because since Y_t ($t = 1, 2, \ldots, T$) are not identically distributed random variables, common summary statistics computed for random samples (i.i.d. random variables), such as mean, median, symmetry, kurtosis, etc., lose their meaning in this case. Also, since a histogram is a plot used for i.i.d. random variables, it has also no meaning in an epidemic situation in which time is an essential component. Singh [11] fell into this trap and the arguments put forward for the new COVID 19 cases in India have, consequently, no theoretical support. Also, the usual plot displaying new cases against time is not a histogram as some authors call it [4,12]. It is simply a time series plot.

These points are crucial for the analysis of outbreaks, epidemic or pandemic situations and also for teaching purposes. In an emergency situation, to provide a quick response, there is a tendency to use basic models described in the existing literature. Thus, this literature must be trusted. Statistical and mathematical backgrounds need to be always present to avoid severe consequences in modelling infection diseases with wrong models and summary descriptions, with public health policy implications as in the COVID-19 pandemic. Simple and understandable concepts for all do not represent the best information as a decision support tool.

Mathematicians and epidemiologists have devised models to describe the behaviour of an epidemic based on sound theoretical grounds and these are the ones that should be taught and used in practice. Despite their solid theoretical foundations, these models have failed in a short- or long-term COVID-19 forecasting in several settings worldwide [13,14]. From the mathematical and statistical point of view, several criticisms and controversial, around the COVID-19 forecasting, are natural. The response to the COVID-19 pandemic has so many dimensions that it is very difficult to include all dimensions in a single robust model. According to Ioannidis et al. [13], some potential reasons for the failure of several models are poor data sources, wrong assumptions in the modelling, lack of incorporation of epidemiological features, selective reporting, etc. Social media tend to report extreme forecasting and this selective reporting may have serious consequences within the general public in terms of public health measures [13]. Decision-makers and the general public may also be affected by these high-criticism environments because some models are complex and without understanding their uncertainty, assumptions and limitations, it is difficult to trust in some of them.

In many European countries, also due to the epidemiological profile, based on non-transmission diseases, it became clear that the mathematical background is crucial to tackling infectious diseases. Brownson et al. [15] analysed and reflected about training in epidemiology and they stated: "future epidemiologists need strong quantitative backgrounds". This pandemic showed that the "future" is "now". Epidemiologists are under enormous pressure and are required to have "superpowers". Some of them are public health specialists, accumulating also teaching and public health research at universities, institutional and political positions. In fact, epidemiology needs to integrate multiple perspectives and multiple disciplines, ranging from social sciences to "hard" science. Epidemiologists have a central role in the interconnection of several scientific domains. Certainly, they need a quantitative background, but the "strong" investment in new theoretical developments, new models, computational algorithms and software to handle and to analyse data needs to be done by computer scientists, statisticians, and mathematicians. In terms of communication with policymakers, communities and media, epidemiologists have a clear advantage over mathematicians, statisticians and computer scientists. These groups, with some good exceptions, have also failed in science communication with scientists from other backgrounds, the general public, and decision-makers, during this pandemic.

To sum up, this ongoing pandemic brings a critical debate around past and current modelling, under enormous pressure, sometimes without serious peer review process and a good science communication strategy. These issues take a long time, and we are at the

beginning of a long screening, but certainly, best practices are emerging from the current collaborative work environment. The multi-disciplinary expertise available through established networks and independent scientific assessments are certainly powerful ways of bringing more scientific rigour to prepare the next global health emergency. However, it is worth revisiting the foundation of an "outbreak science" proposed by Rivers et al. [16], to prepare the next pandemic and avoid reactive mobilizations based on existing theory and practice of public health and classic epidemiology.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

AUTHORS' CONTRIBUTION

This document arises from discussions within a CEAUL working group, of which all authors are members. LG and MAT wrote the first draft of the manuscript. All authors accompanied and discussed the news and documents produced during and about this pandemic, seeking the sources of same concepts. All authors revised the manuscript and approved the final version.

FUNDING

This work was partially support by CEAUL (funded by FCT – Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020).

REFERENCES

[1] Box GEP. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN, editors. Robustness in statistics. New York, NY: Academic Press; 1979, pp. 201–36.

[2] Goodson M. All models are wrong, but some are completely wrong. Royal Statistical Society Data Science Section. 2020. Available from: https://rssdss.design.blog/2020/03/31/all-models-are-wrong-but-some-are-completely-wrong/.

[3] Rashed EA, Kodera S, Gomez-Tames J, Hirata A. Influence of absolute humidity, temperature and population density on COVID-19 spread and decay durations: multi-prefecture study in Japan. Int J Environ Res Public Health 2020;17;5354.

[4] Centers for Disease Control and Prevention. Lesson 1: Introduction to Epidemiology. Section 11: Epidemic Disease Occurrence. 2012. Available from: https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section11.html.

[5] Dicker R, Coronado F, Koo D, Parrish RG. Principles of epidemiology in public health, practice. An introduction to applied epidemiology and biostatistics. Atlanta, GA: Centers for Disease Control and Prevention (CDC) Office of Workforce and Career Development; 2012.

[6] Rothman KJ. Epidemiology: an introduction. New York, USA: Oxford University Press; 2012.

[7] Greenstone M, Nigam V. Does social distancing matter? University of Chicago, Becker Friedman Institute for Economics. Working Paper. Chicago, USA: SSRN; 2020.

[8] Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, et al. Propagation analysis and prediction of the COVID-19. Infect Dis Modell 2020;5;282–92.

[9] Reddy C, Kuonza L, Ngobeni H, Mayet NT, Doyle TJ, Williams S. South Africa field epidemiology training program: developing and building applied epidemiology capacity, 2007–2016. BMC Public Health 2019;19;469.

[10] Johnson NL, Kotz S, Balakrishnan N. Continuous univariate distributions. Vol. 1, 2nd ed. New York, NY: John Wiley & Sons; 1994.

[11] Singh A. India Covid19: one histogram told 3 important stories. 2020. Available from: https://rpubs.com/anupamsingh1.

[12] Fontaine RE. Describing epidemiologic data. In: Rasmussen SA, Goodman RA, editors. The CDC Field Epidemiology Manual. New York, NY: Oxford University Press; 2019.

[13] Ioannidis JPA, Cripps S, Tanner MA. Forecasting for COVID-19 has failed. 2020. Available from: https://forecasters.org/blog/2020/06/14/forecasting-for-covid-19-has-failed/.

[14] Holmdahl I, Buckee C. Wrong but useful — what Covid-19 epidemiologic models can and cannot tell us. N Engl J Med 2020;383;303–5.

[15] Brownson RC, Samet JM, Bensyl DM. Applied epidemiology and public health: are we training the future generations appropriately? Ann Epidemiol 2017;27;77–82.

[16] Rivers C, Chretien JP, Riley S, Pavlin JA, Woodward A, Brett-Major D, et al. Using "outbreak science" to strengthen the use of models during epidemics. Nat Commun 2019;10;3102.