

RESEARCH

Statistical distribution fitting to the number of COVID-19 deaths in South Africa

Delson Chikobvu^{1,2,3,4}, Caston Sigauke^{1,2,3,4,5}

^{1,2} Correspondence:

^{1,2} Department of Statistics,

Private Bag X5050, 0950

Thohoyandou, South Africa

Full list of author information is available at the end of the article

^{3,4} Equal contributor

Abstract

Background: COVID-19 first detected in Wuhan; China in December 2019, is a disease caused by the coronavirus SARS-CoV-2. It has presented the greatest public health challenges globally since the 1918 influenza which was called the “mother” of all pandemics. South Africa and the rest of Africa are yet to experience the devastating effects of COVID-19.

Methods: In this paper, the reported COVID-19 number of deaths in South Africa, for the period 27 March 2020 to 20 May 2020, is modeled using four statistical distributions which can be grouped under the Generalised Gamma distribution. This exploratory study also uses simple additive models to capture the underlying COVID-19 death rate.

Results: Empirical results show that the Gamma distribution gives the best fit to the data. The hazard rate is still increasing, and the peak number of deaths has not been reached yet despite the lockdown and other measures to try and slow down the progression of the disease.

Conclusions: The exploratory data analysis done in this study is simple and meant to complement the detailed and complex modelling done which is useful in informing policy and decision making.

Keywords: Additive quantile regression; COVID-19; Hazard rate; Generalised Gamma distribution; South Africa

Background

COVID-19 is a highly contagious disease. It is caused by the coronavirus SARS-CoV-2 virus and was first detected in Wuhan, China in December 2019. In South Africa, there are widespread testing campaigns for the coronavirus. Tests are performed widely and as rapidly as possible. To detect the presence of the virus, tests are conducted in South Africa after or before any screening for symptoms. Screening and testing are done through mobile testing clinics, community screening with thousands of deployed community health workers and strict contact tracing protocols to track down possible infected cases.

The number of confirmed COVID-19 cases and COVID-19 deaths are recorded daily. Results sometimes come with a delay. The cumulative number of confirmed cases and the case fatality rate are also reported. Cases are supposed to be detected early and interventions can be targeted, such as quarantine and isolation for confirmed cases to limit transmission. Tests are done when symptoms are present, but other times, mass testing is done. Test samples can only be kept for a few days before samples are discarded.

COVID-19 Mortality data is a reliable indicator of the extent of this pandemic in South Africa. One of the assumptions is that those who have died are tested and the cause of death is captured accurately. However, mortality data comes with at least a 3 weeks' time lag from the time of infection. Its usefulness as a predictor of the current number of positive COVID-19 cases is limited in that sense.

Mortality data can be useful for epidemiology model fitting and testing once a certain number of deaths have occurred. In this study, we model COVID-19 mortality using the three-parameter Generalised Gamma distribution which is a flexible family comprising the Weibull, Gamma, Exponential and the Lognormal statistical distributions. However, COVID-19 mortality data is often under reported for a variety of reasons in South Africa. Reasons include access to facilities and resources needed to do relevant tests especially when death occurs outside of Health facilities.

The COVID-19 mortality modeled in this study reflects interventions by the South African government in terms of its risk-based strategy including the social distancing, wearing of face masks and lockdown including school closures and travel restrictions. The lockdown in South Africa came into effect on midnight of the 26th of March 2020. Since then, a range of various levels of the lockdown has been announced. As the pandemic progresses, more information becomes available around the COVID-19 pandemic and we will learn more around the data and the pandemic.

The rest of the paper is organised as follows: the models are discussed in Section 2. Empirical results are presented and discussed in Section 3, while Section 4 concludes.

Methods

Distribution fitting using the Generalised Gamma distribution

The distribution of deaths due to COVID-19 can be modeled with the Gamma density ([?, ?]; among others). In this study, we consider the use of the Generalised Gamma distribution (GGD). The distribution was introduced by [?] and is a flexible family comprising the Gamma, Weibull, Exponential and includes the Lognormal as a limiting distribution. Let Y to be a random variable representing the reported COVID-19 deaths. Assuming Y follows the GGD then the density of the GGD is given by

$$f(y) = \frac{\alpha^k}{\beta\Gamma(k)} \exp\left(-\alpha y^{\frac{1}{\beta}}\right) y^{\frac{k}{\beta}-1}, y > 0, \quad (1)$$

where k is a nuisance (shape) parameter, α the scale parameter and β the tail index (shape parameter). When $\beta = k = 1$, we get the exponential distribution, whose density is

$$f(y) = \alpha e^{-\alpha y}, y > 0. \quad (2)$$

The exponential distribution has a constant hazard rate.

For $\beta = 1$, we get the Gamma distribution with the following density

$$f(y) = \frac{\alpha^k}{\Gamma(k)} \exp(-\alpha y) y^{k-1}, y > 0. \quad (3)$$

For the case, $k > 1$ the hazard rate is concave and increasing while for $k < 1$ it is convex and decreasing.

And for $k = 1$, we have the Weibull distribution with the following density

$$f(y) = \frac{\alpha}{\beta} \exp\left(-\alpha y^{\frac{1}{\beta}}\right) y^{\frac{1}{\beta}-1}, y > 0 \quad (4)$$

When $\beta > 1$, the hazard rate function increases and decreases when $\beta < 1$. The increasing hazard model could reflect the hazard for a population afflicted by a pandemic and the number of deaths is getting out of control. The level of hazard is expected to increase with time since the pandemic increases. The decreasing hazard model could reflect the hazard for the same population when the pandemic is getting under control or when the pandemic is on its way out. The level of hazard is expected to fall as the time since the pandemic peak increases.

When $k \rightarrow \infty$, this leads to the Lognormal as the limiting distribution. The hazard rate is hump-shaped. Initially, it increases, then reaches a maximum before decreasing towards zero.

A smooth additive model for the death rate series

The use of smooth additive models can assist in capturing the death rate behaviour with time ([?]). This study uses the quantile generalised additive model (quantGAM) discussed by [?] and extended by [?] and defined as

$$q_{Y|X}(\tau) = \sum_{t=1}^n \rho_{\tau}\left(y_t - s(t)\right), \quad \tau \in (0, 1), \quad (5)$$

where y_t denotes the COVID-19 deaths, which are reported on day $t, t = 1, \dots, n, s$ is a smooth function of time t , $q_{Y|X}(\tau)$ is the extreme conditional quantile function and

$$\rho_{\tau}(u) = [\tau \mathbf{I}(u \geq 0) + (1 - \tau) \mathbf{I}(u < 0)] |u|$$

is the pinball loss function with \mathbf{I} denoting an indicator function. The function $s(t)$ in this context represents the underlying COVID-19 death rate.

Results

The data used in this study is from <https://health.hydra.africa/>. The data consists of daily reported COVID-19 positive cases, deaths, recoveries and active cases. The trend is obscured as there are too many unknowns. Confirmed COVID-19 positive cases sometimes take a few days to be reported to the officials for a variety of reasons (there is data lagging in reported cases). There are backlogs in the testing itself. South Africa has differing levels of standards of living which may affect rates of transmission of the disease in various communities and the level of hospital care accessible to individuals.

The number of positive cases relates to the date of release of results by government officials and not the date of infection nor the date of the test. A three-point moving average is calculated on the original COVID-19 deaths to smooth some of these effects. The mortality distribution modeled here will reflect mortality from different age profiles of infected cases. Since mortality rates vary by age with higher chances of dying for older ages, mortality cases are mainly from older patients on average and hence reflects mortality at those older ages. South Africa's COVID-19 Case Fatality Rate (CFR) now stands at around 1.9%. This is low by international standards. As of 20 May 2020, the number of positive cases was 18 003 and the number of COVID-19 death stood at 339.

Exploratory data analysis

Summary statistics of the smoothed series on the number of COVID-19 death (reported 3-point moving average deaths) for the period 27 March 2020 to 20 May 2020 is given in Table ???. The skewness and kurtosis are presented in Table ??? and show that the distribution of the reported deaths is non-normal. This is consistent with the density plot in Figure ??(b) where the bulk of the deaths is to the right of the median deaths. The non-normality of the data is confirmed by both the QQ plot and box plot given in Figure ??(c) and (d), respectively.

Table 1 Summary statistics of the reported COVID-19 deaths (3-point moving average series).

Var	Mean	St Dev	Min	Q1	Median	Q3	Max	Skew	Kurt
y_t	5.881	4.873	0.667	2	5.333	8.333	25	1.474	6.342

From Table ??, y_t denotes smoothed reported deaths using the 3-point moving average, St dev is the Standard deviation, Min represents the Minimum, Max is the Maximum, Q1 represents the Lower quartile, Q3 represents the Upper Quartile, Skew denotes the Skewness, Kurt denotes the Kurtosis.

A plot of the reported COVID-19 deaths, together with the density plot, normal QQ-plot and the box plot, is given in Figure ?. Figure ??(a) shows that there is an upward trend. In the first 12 days it was almost linear with a constant rate of increase. From day 13, the reported death rates fluctuate a lot and there is a steep increase.

Figure 1 Reported COVID-19 deaths (3-point moving average series) for the period 27 March 2020 to 20 May 2020.

Distribution fitting

Figure ?? top panel shows the histogram of reported new COVID-19 deaths superimposed with an empirical density (left panel) and the corresponding cumulative distribution (right panel).

Figure 2 Histogram of reported new COVID-19 deaths superimposed with an empirical density (left panel) and the corresponding cumulative distribution (right panel).

The ‘EasyFit’ open-source software (<http://www.mathwave.com/help/easyfit/index.html>) is used in this study to fit the Generalised Gamma distribution. The parameter estimates are: $k = 0.9935$, $\alpha = 1.452$, $\beta = 4.0377$ and the goodness of fit statistics were 0.1197 and 0.6398 for Kolmogorov-Smirnov and Anderson-Darling, respectively. The value of $k \approx 1$, implying that the Weibull distribution is the best fitting distribution.

The four distributions of the three-parameter Generalised Gamma distribution are then fitted using the R package ‘fitdistrplus’ developed by [?]. The parameters are estimated using the maximum likelihood estimation method. Table ?? presents a summary of the goodness of fit statistics and the goodness of fit criteria for the four distributions. The goodness of fit measures (AIC, BIC) suggest that the Gamma

Table 2 Distribution fitting to COVID-19 death.

	Weibull	Gamma	Lognormal	Exponential
Goodness-of-fit statistics				
Kolmogorov-Smirnov statistic	0.1096	0.1199	0.1565	0.1310
Cramer-von Mises statistic	0.0989	0.1128	0.1812	0.1485
Anderson-Darling statistic	0.5829	0.6436	1.0614	0.9923
	Weibull	Gamma	Lognormal	Exponential
Goodness-of-fit criteria				
Akaike’s Information Criterion	294.15	293.90	296.48	295.79
Bayesian Information Criterion	298.09	297.84	300.42	297.76

distribution as the best distribution to describe COVID-19 death. The Weibull distribution is a very close competitor and is suggested as the best by the other goodness of fit measures.

Fitting the Gamma distribution

The Gamma parameter maximum likelihood (ML) and parametric bootstrap estimates are given in Table ?. The standard errors for the ML estimates are given in parentheses while for the bootstrap medians, the 95% confidence intervals are given in brackets. The correlation between the shape and scale parameter for the gamma distribution was found to be 0.8388.

Table 3 Gamma parameter ML estimates and bootstrap estimates.

	Shape	Scale (Rate)
Maximum likelihood estimate	1.4413(0.2541)	0.2451(0.0515)
Bootstrap median estimate	1.4863	0.2552
95% confidence interval	(1.0688, 2.1364)	(0.1711, 0.3858)

Figure ?? shows the histogram of reported COVID-19 deaths (3-point moving average) superimposed with the Gamma distribution (left panel) and the QQ plot (right panel). The QQ plot shows that the two-parameter Gamma distribution is a good fit to the data.

Figure 3 Histogram of reported COVID-19 deaths (3-point moving average) superimposed with the Gamma distribution (left panel) and the QQ plot (right panel).

The level of the hazard is increasing since the beginning of the pandemic. South Africa is still to attain its peak. The rate of increase however is slowing down. This

Figure 4 Plot of the hazard rate (Gamma).

may be attributed to the success of the various intervention measures taken by the government during this period including the lockdown, social distancing and the wearing of face masks.

A smooth additive model for the death rate series

The use of a simple additive quantile regression model gives us some useful insight into the behaviour of the COVID-19 data. This complements to the detailed epidemic models.

$$q_{Y|X}(\tau) = \sum_{t=1}^n \rho_{\tau} \left(y_t - s(t) \right), \quad \tau \in (0, 1). \tag{6}$$

The basic summary statistics of the raw data is minimum value is 0, first quartile 1, median deaths 4, an average of 6 deaths, the third quartile is 9 with a maximum value of 27 deaths. Reported deaths (raw data) from the COVID-19 in South Africa (SA) for the period 27 March 2020 to 20 May 2020 superimposed with the smoothed variable ‘day’ is given in Figure ??.

Figure 5 Reported deaths (raw data) with COVID-19 in South Africa (SA) for the period 27 March 2020 to 20 May 2020.

The model fits the reported COVID-19 deaths in South Africa together with some basic model checking plots of the residuals, the ACF and PACF of the residuals are as given in Figure ?. A plot of the smoothed ‘day’ variable in Figure ?? shows that the reported COVID-19 deaths are increasing, again an indication that the peak has not yet been reached.

Figure 6 Results of fitting the additive quantile regression model (smoothed variable ‘day’) to the reported death data (top panel), plot of residuals from the fitted model (middle panel) and the ACF PACF of the residuals (bottom panel). Figures 6a and 6b INSERT HERE.

Discussion

This paper gives an indication of the spread of the COVID-19 disease in South Africa. The Gamma distribution with an increasing but concave hazard rate best describes statistically how the disease is spreading. The concave shape is encouraging in the sense that that the increase is happening at a decreasing rate. This points to a some what success in the interventions by government. The disappointment is the fact that the data suggests that the peak has not yet been reached. As the country moves to less strict levels of the lockdown, there is also the added fear of a second wave of the disease taking root in the country. The delay in the onset of the influenza (flu) season in the year 2020 suggests that the social distancing seems to have the desired effect. At the time of writing this paper, the influenza had at least a 2 month delay.

Conclusions

The Gamma distribution is a good fit for the South African COVID-19 data. The hazard rate is still increasing, and the peak number of deaths has not been reached yet despite the lockdown and other measures to try and slow down the progression of the disease. The lockdown and other measures to try and slow down the progression of the disease seems to have slowed down the rate of increase in the hazard rate, but the cases are not yet on a downward trend.

Abbreviations

Competing interests

The authors declare that they have no competing interests.

Author's contributions

The authors equally contributed to this article.

Acknowledgements

The authors would like to thank the reviewers for their comments.

Availability of data and materials

All data generated or analysed during this study are included in this published article including the R codes.

Author details

?? Department of Mathematical Statistics and Actuarial Science, P.O. Box 339, 9300 Bloemfontein, South Africa.

?? Department of Statistics, Private Bag X5050, 0950 Thohoyandou, South Africa.

References

1. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, Cuomo-Dannenburg G, Thompson H, Walker PG, Fu H, Dighe A, Griffin JT, Baguelin M, Bhatia S, Boonyasiri A, Cori A, Cucunubá Z, FitzJohn R, Gaythorpe K, Green W, Hamlet A, Hinsley W, Laydon D, Nedjati-Gilani G, Riley S, van Elsland S, Volz E, Wang H, Wang Y, Xi X, Donnelly CA, Ghani AC, Ferguson NM. Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases* 2020; 1-9. [https://www.thelancet.com/pdfs/journals/laninf/PIIS1473-3099\(20\)30243-7.pdf](https://www.thelancet.com/pdfs/journals/laninf/PIIS1473-3099(20)30243-7.pdf)
2. Wood SN. Simple models for COVID-19 death and fatal infection profiles 2020; 1-9. <https://arxiv.org/abs/2005.02090> Accessed on 8 May 2020.
3. Stacey EW. A Generalization of the Gamma Distribution. *Annals of Mathematical Statistics* 1962; 33(3): 1187-1192. <https://www.jstor.org/stable/2237889>
4. Gaillard P, Goude Y, Nedellec R. Additive models and robust aggregation for GEFcom2014 probabilistic electric load and electricity price forecasting. *Int. J. Forecast.* 2016; 32: 1038-1050.
5. Fasiolo M, Goude Y, Nedellec R, Wood SN. Fast Calibrated Additive Quantile Regression 2020; 1-39. <https://arxiv.org/pdf/1707.03307v4.pdf> (accessed on 31 March 2020).
6. Delignette-Muller ML, Dutang C, Siberchicot A. *fitdistrPlus: An R Package for Fitting Distributions*, 2020 <https://cran.r-project.org/web/packages/fitdistrplus/index.html> (accessed on 20 May 2020).

Appendix A: Figures

Figure 7 Figure A1: Plots of raw and smoothed reported deaths in South Africa (27 March 2020 to 15 May 2020). Figure legend text.

Figure 8 Figure A2: Box plot of reported deaths in South Africa (27 March 2020 to 15 May 2020).

Appendix B: R code