# Appendix A

## Virus variants and prime numbers

In this Appendix we introduce a suggestive hypothesis to support the choice of the function exploited to fit the WHO data on relevant SARS-CoV-2 variants. The connection discussed here between virus mutations, quantum states and prime numbers yields a heuristic justification of the analytic form of the function used for the fit, although a precise theoretical framework is still lacking.

**Prime numbers**

As discussed in the Methods section, the fit of WHO data was obtained by means of the function $v(N) = k \cdot \frac{N}{\log N}$, where $k$ is the constant of the numerical fit.

In 1801 Gauss [1] found that the function $\pi(x) \sim \frac{x}{\log x}$ yields asymptotically (i.e. for $x$ sufficiently large) the cumulative number of primes less than a given number *x*. By comparing the Gauss function $\pi(x)$ and the function exploited for the fit, it turns out that the cumulative number of relevant variants $v$ for $N$ infected cases in the world is proportional to the cumulative number $\pi(N)$ of primes less than $N$, i.e. $v(N) = k \cdot \pi(N)$.

A more accurate expression of the number of primes less than $x$ is given by the logarithmic integral function $Li(x)$, defined as $Li(x) = \int_0^x \frac{dt}{\log t}$ . By exploiting the logarithmic integral function for the fit of WHO data, the cumulative number $\hat{v}$ of relevant SARS-CoV-2 variants for $N$ infected cases in the world becomes: $\hat{v}(N) = h \cdot Li(N) = h \cdot \int_0^N \frac{dt}{\log t}$ with the constant $h$ given by $h = 3.16 \cdot 10^{-6}$ and 95% CI = $(2.74 - 3.57) \cdot 10^{-6}$. The adjusted $R$-squared, measuring the goodness of fit, is $R^2 = 0.97$ both with the Gauss function and the logarithmic integral function. The difference between the fits $v(N)$ and $\hat{v}(N)$ is less than 1 for $N \leq 9.6 \cdot 10^8$, therefore in the current range of $N$ values we can use the simpler function $v(N)$.

The Gauss function provides an asymptotic approximation of the number of primes less than a sufficiently large quantity. For this reason, in our fit we considered the large number of cases in the world instead of focusing separately on single countries or geographical areas, where the cases are fewer.

**Zeta function and quantum states**

In 1859 Riemann [2] found an exact expression for the number of primes less than a given quantity. Riemann's formula involves a sum over the zeroes of the so-called zeta function $\zeta(s)$, whose relevant zeroes would have all real part ½ according to Riemann's Hypothesis [3]. The distribution of the zeroes of $\zeta(s)$ along the critical line z = ½ determines the distribution of the primes, as well established in number theory [4].

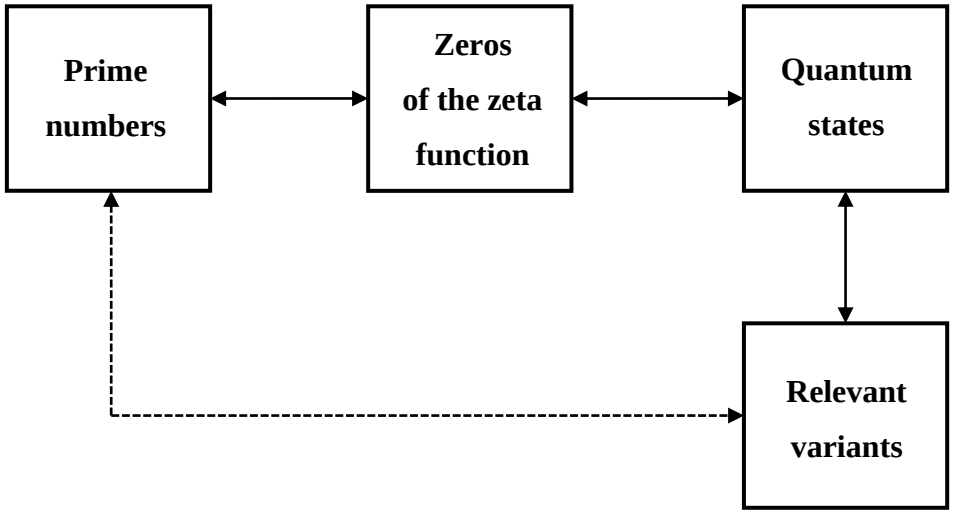
In 1977 Montgomery [5] found the function which describes the spacing between the zeroes of $\zeta(s)$. As spotted by the physicist Dyson, such function is the same as that describing the spacing between the energy levels of heavy atomic nucleus. The link between zeroes of $\zeta(s)$ and quantum states also extends to chaotic systems, as pointed out by Berry [6] and confirmed numerically by Odlyzko [7].

In conclusion, a connection seems to exist between the distribution of the prime numbers and the quantum states of a physical system.

**Virus mutations**

As suggested in 1944 by Schrödinger [8] in his renowned essay "What is life?", a genetic mutation can be considered a sort of "quantum jump", i.e. a transition between two different states of a quantum system. In particular, a virus variant is a genetic mutation due to a quantum transition between two different configurations in the structure of the virus.

The following figure schematises the connection between the cumulative number of primes less than a given quantity and the number of relevant variants for a given number of cases. The intermediate links in the scheme are the distribution of the zeroes of Riemann's zeta function and the spacing between the quantum states of a physical system.




The numerical fit of WHO data could be performed by exploiting a great variety of functions different from the one we used. However, our choice is supported by the connection between virus variants and prime numbers inspired by quantum physics and number theory.

**References of Appendix A**

[1] Gauss CF. Disquisitiones Arithmeticae. Leipzig: Fleischer; 1801.

[2] Riemann B. On the Number of Prime Numbers less than a Given Quantity. Berlin: Monatsberichte der Berliner Akademie; 1859.

[3] Edwards HM. Riemann's Zeta Function. Mineola, New York: Dover Publications, Inc; 1974.

[4] Hardy GH, Littlewood JE. Contributions to the theory of the Riemann zeta-function and the theory of the distribution of primes. Acta Math 1916; 41: 119-96. https://doi.org/10.1007/BF02422942.

[5] Montgomery HL. Extreme values of the Riemann zeta function. Comm. Math Helv 1977; 52:511-18.

[6] Berry MV. Riemann's Zeta function: A model for quantum chaos? In: Seligman TH, Nishioka H (editors). Quantum Chaos and Statistical Nuclear Physics. Lecture Notes in Physics, vol 263. Berlin, Heidelberg: Springer; 1986. https://doi.org/10.1007/3-540-17171-1_1.

[7] Odlyzko AM. On the distribution of spacings between zeroes of the zeta function. Mathematics of Computation 1987; 48 (177): 273-308. https://doi.org/10.2307/2007890.

[8] Schrödinger E. What is Life? The Physical Aspects of the Living Cell. Cambridge: Cambridge University Press; 1944.

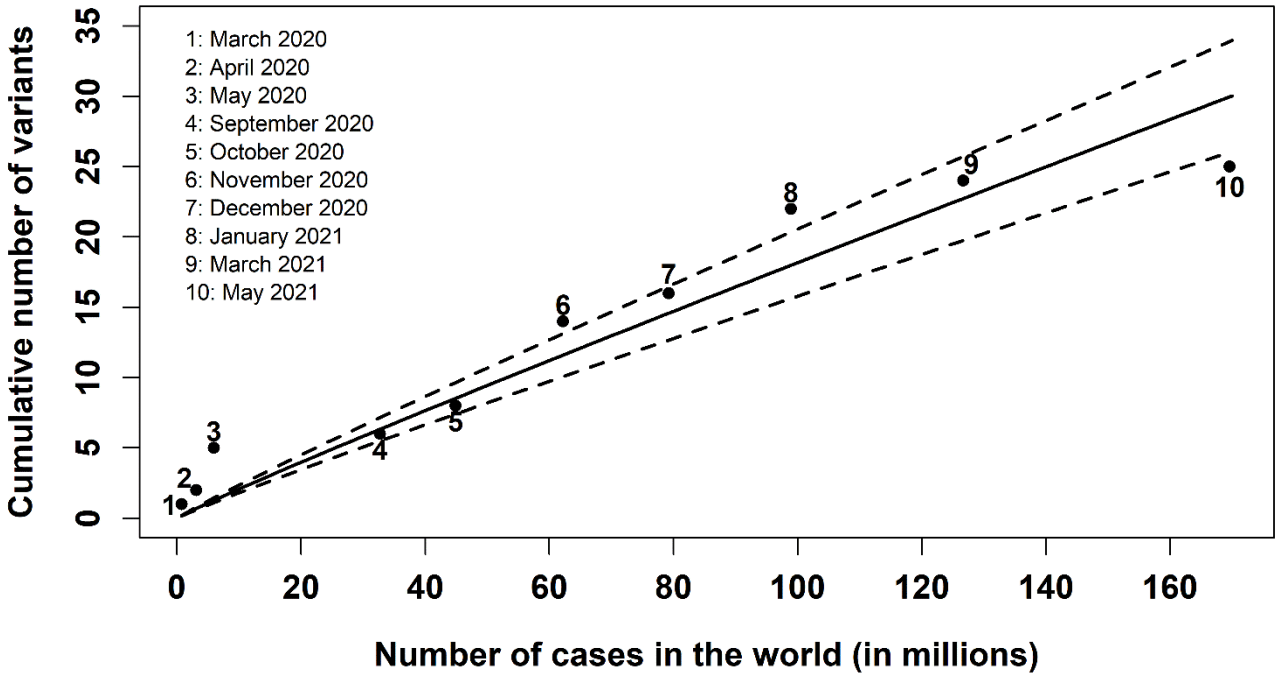
# Appendix B

## Details on the numerical fit

In this appendix we discuss some details on the fit of WHO data (confidence interval, residuals and numerical derivatives) and consider the approximation given by the linear regression.

### Confidence interval and residuals

The 95% confidence interval (CI) of the constant $k = 3.35 \cdot 10^{-6}$ of the numerical fit of WHO data on relevant SARS-CoV-2 variants is $95\%\ \mathrm{CI} = (2.91 - 3.79) \cdot 10^{-6}$. In the figure below the dashed lines correspond to the upper and lower limits of the 95% CI of the constant $k$.




The residual $r_i$ of the $i$-th value in a set of $n$ data is the difference between the observed value $y_i$ and the corresponding value $\hat{y}_i$ predicted by the fit: $r_i = y_i - \hat{y}_i$, with $i = 1, 2, \dots, n$. In our fit the maximum absolute value $r_{max}$ of the residuals $r_i$ is given by

$$r_{max} = \max_i |y_i - \hat{y}_i| = 4.96$$

corresponding to the last observation ($i = 10$) shown in the previous figure.

The residual standard deviation $\sigma_r$ is defined as:

$$\sigma_r = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{df}}$$

where $df$ is the degree of freedom: $df = n - p$ ($n$ is the number of observed data and $p$ the number of parameters in the fit: $p = 1$ in our case). The residual standard deviation for our fit is

$$\sigma_r = \sqrt{\frac{\sum_{i=1}^{10}(y_i - \hat{y}_i)^2}{10 - 1}} = 2.72$$

**Numerical derivatives**

In order to obtain the derivatives of the function underlying the WHO data on relevant SARS-CoV-2 variants, we used Wolfram Mathematica to build and derive the Lagrange polynomial and the B-splines interpolating the observed data. Moreover, we computed the so-called three-points and five-points formulas discussed e.g. by Burden RL and Faires JD in "Numerical analysis" (7th edition. Pacific Grove: Brooks/Cole; 2001).

If $x_0, x_1, \ldots, x_n$ are $n + 1$ distinct numbers in an interval $I$ and $f(x)$ is a function whose values are known in these points, the $(n + 1)$-point formula expressing the derivative $f'$ of the function $f$ in a point $x_j$, with $j = 0, 1, \ldots, n$, is given by:

$$f'(x_j) = \sum_{k=0}^{n} f(x_k) \cdot L'_{n;k}(x_j) + \frac{f^{(n+1)}\left(\xi(x_j)\right)}{(n+1)!} \prod_{\substack{k=0 \\ k \neq j}}^{n} (x_j - x_k)$$

where $\xi$ is a point in $I$, depending on $x_j$, and $L'_{n;k}(x)$ is the derivate of the $k$-th coefficient $L_{n;k}(x)$ of the $n$-th Lagrange interpolating polynomial:

$$L_{n;k}(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i}$$

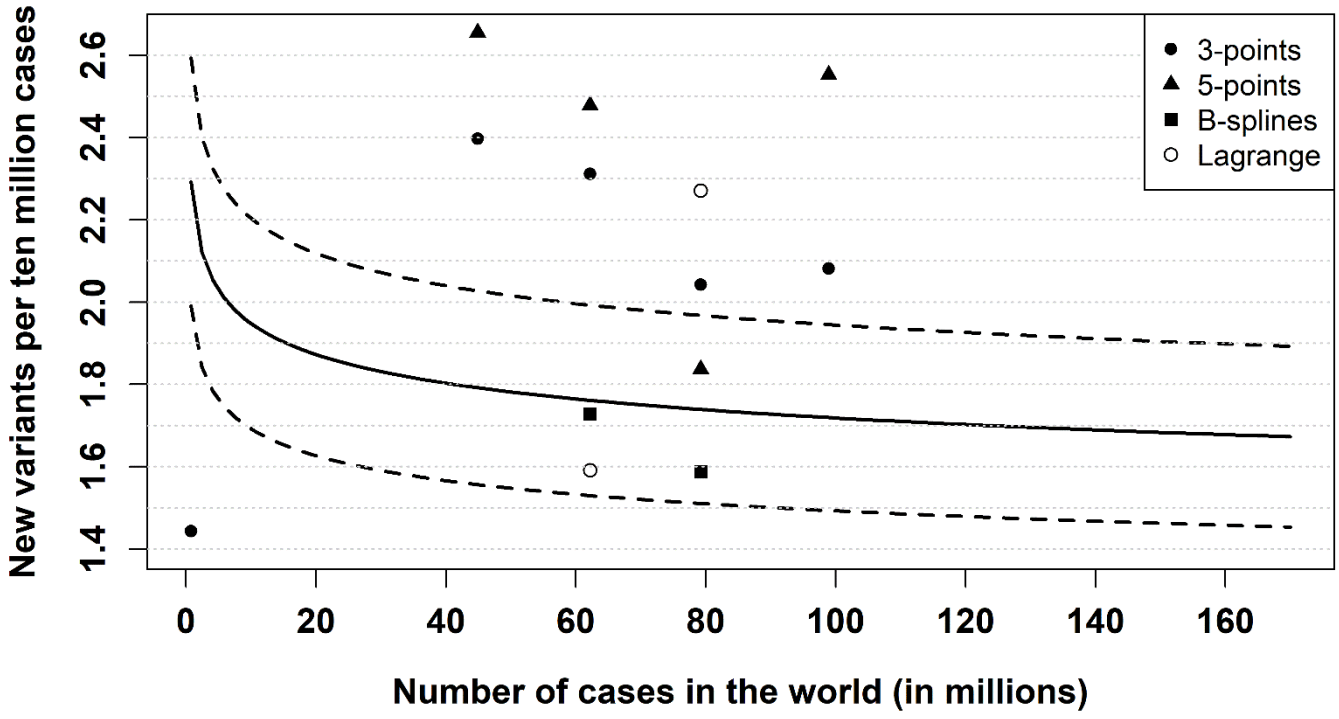
The most common formulas are those involving three and five evaluating points:

$$f'(x_j) \cong \sum_{k=j-h}^{j+h} f(x_k) \cdot L'_{n;k}(x_j)$$

where $h = 1$ or $h = 2$ for the three-points or five-points formula, respectively.

The sets of three points $\{x_{j-1}, x_j, x_{j+1}\}$ and five points $\{x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2}\}$, belonging to the complete set $\{x_0, x_1, \dots, x_n\}$, are chosen so that the point $x_j$ where the derivative must be computed is in central position (or as most central as possible), since in this case the approximation error is minimum.

In the figure below the solid line represents the new variants per ten million cases $n = k \cdot 10^7 \frac{\log N - 1}{(\log N)^2}$, while the dashed lines correspond to the upper and lower limits of the 95% CI $= (2.91 - 3.79) \cdot 10^{-6}$ of the constant $k = 3.35 \cdot 10^{-6}$. The dots represent the new variants per ten million cases computed through the formula $n \cong f'(N) \cdot \Delta N$, where $\Delta N = 10^7$ and the numerical derivative $f'$ of the function $f$ underlying the observed data is obtained with the methods reported in the legend (three or five points formulas, B-splines and Lagrange interpolation).

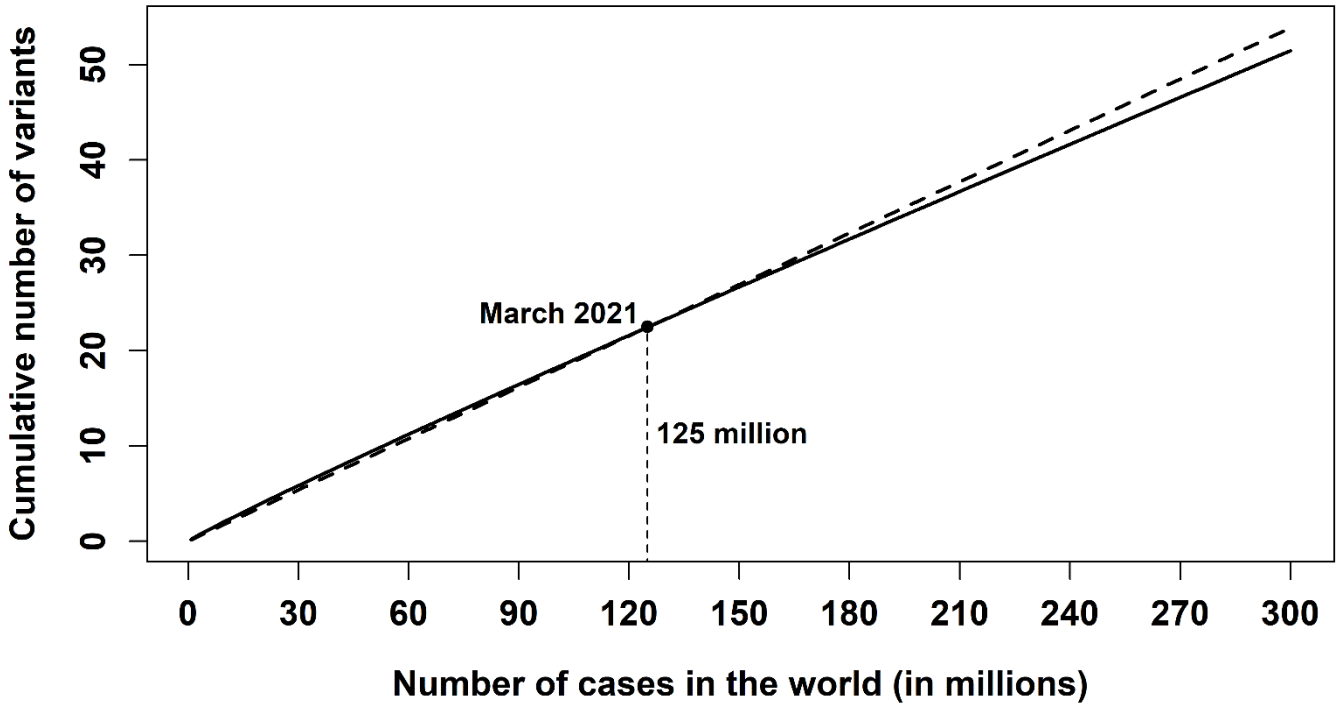



**Linear regression**

The numerical fit $v = k \cdot N/\log N$, where $v$ is the cumulative number of relevant SARS-CoV-2 variants and $N$ is the cumulative number of cases, can be approximated by the linear regression $\tilde{v} = h \cdot N$, with $h = 1.80 \cdot 10^{-7}$ and 95% CI = $(1.54 - 2.05) \cdot 10^{-7}$.

The linear model does not satisfy the third condition listed in the Methods section. However, it is close to the logarithmic fit up to large numbers of cases. Specifically, the difference between the number of relevant SARS-CoV-2 variants predicted by the two fits is zero for $N = 125$ million cases (March 2021).

The difference between the number of variants predicted by the two fits raises to 12 variants for $N = 760$ million cases in the world (corresponding to about three times the total cases in November 2021).




In the logarithmic fit $v = k \cdot \frac{N}{\log N}$ the number $n$ of new relevant variants per ten million cases decreases as the number $N$ of cases increases, as discussed in the Results section: $n = 33.5 \cdot \frac{\log N - 1}{(\log N)^2}$.

On the contrary, in the linear fit $\tilde{v} = h \cdot N$ the number $n$ is constant: $n \cong \tilde{v}'(N) \cdot \Delta N = 1.80$, being $\tilde{v}'(N) = h = 1.80 \cdot 10^{-7}$ and $\Delta N = 10^7$. This different behaviour between the logarithmic and linear fits is represented in the figure below.