Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 233-263
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Probability Distribution of SARS-Cov-2 (COVID) Infectivity Following Onset of Symptoms: Analysis from First Principles

Mark P. Silverman

Department of Physics, Trinity College, Hartford, USA
Email: mark.silverman@trincoll.edu

How to cite this paper: Silverman, M.P. (2023) Probability Distribution of SARS-Cov-2 (COVID) Infectivity Following Onset of Symptoms: Analysis from First Principles. *Open Journal of Statistics*, **13**, 233-263.
https://doi.org/10.4236/ojs.2023.132013

Received: March 18, 2023
Accepted: April 24, 2023
Published: April 27, 2023

Abstract

The phasing out of protective measures by governments and public health agencies, despite continued seriousness of the coronavirus pandemic, leaves individuals who are concerned for their health with two basic options over which they have control: 1) minimize risk of infection by being vaccinated and by wearing a face mask when appropriate, and 2) minimize risk of transmission upon infection by self-isolating. For the latter to be effective, it is essential to have an accurate sense of the probability of infectivity as a function of time following the onset of symptoms. Epidemiological considerations suggest that the period of infectivity follows a lognormal distribution. This proposition is tested empirically by construction of the lognormal probability density function and cumulative distribution function based on quantiles of infectivity reported by several independent investigations. A comprehensive examination of a prototypical ideal clinical study, based on general statistical principles (the Principle of Maximum Entropy and the Central Limit Theorem) reveals that the probability of infectivity is a lognormal random variable. Subsequent evolution of new variants may change the parameters of the distribution, which can be updated by the methods in this paper, but the form of the probability function is expected to remain lognormal as this is the most probable distribution consistent with mathematical requirements and available information.

Keywords

COVID, SARS-Cov-2, Period of Infectivity, Probability of Infectivity, Viral Shedding, Infectiousness, Kaplan-Meier Curve, Principle of Maximum Entropy

1. Introduction

Although the SARS-Cov-2 virus presents many biochemical and biophysical

mysteries for scientists to research, the question of greatest import for the general public, to judge from numerous queries online and in the print news media, is a practical one: "If I get coronavirus disease (COVID), how long must I isolate". This question is of critical significance to infected individuals and to those who determine public health policy, since the answer can have serious repercussions for the family, co-workers, and, in general, community of a sick person who does not self-quarantine long enough.

Responses to the preceding question are given by numerous websites whose information often comes from secondary or tertiary news sources and rarely directly from scientists and clinicians who study or treat the disease; see for example [1] [2] [3] [4] [5]. Reports such as these (which are but a small sample of many comparable sources) can be helpful, but the information provided is usually broadly general, and possibly mutually conflicting, with only qualitative assertions of infectivity ("very likely", "not very likely", etc.), rather than a quantitative measure of probability. Part of the problem is the complexity, uncertainty, and disagreement reflected in the medical literature itself [6] [7] [8] [9]. For example, the authors of Ref [9] take issue with the US Centers for Disease Control and Prevention (CDC), which decreased the recommended isolation period after a positive rapid antigen test from 10 days to 5 days.

In confronting the question of isolation, it is essential to understand (as the general public, news media, and even medical professionals often do not) that the relevant answer is *not* a number, but a *probability distribution*. The period of isolation is a random variable, *i.e.* an uncertain quantity by virtue of many uncontrollable conditions such as the health of the person prior to COVID infection, the severity of the disease produced by infection, the specific variant of the infecting virus, the incubation period of the virus [10], and other variables.

In this paper I propose that the conditional probability that a person with a positive rapid antigen test on day 0 will still test positive (and therefore potentially shed virus) t days afterward follows a lognormal distribution. This proposition is supported empirically by constructing the lognormal probability density function (pdf) and cumulative distribution function (cdf) from data provided by several published medical studies. Further support is provided by an objective (*i.e.* model-free) theoretical analysis of the probability of infectivity, employing the Principle of Maximum Entropy (PME) and the Central Limit Theorem (CLT).

Given the mathematical form of the lognormal distribution, the complete pdf and cdf are obtained from two empirical data from the medical literature, e.g. the median number of days of infectivity (*i.e.* the 50% quantile) as inferred from a positive antigen test, and the 95% quantile (*i.e.* the number of days after which the probability of a positive test is less than or equal to 5%). In statistical practice as applied to science and medicine, the 5% level is often set as the threshold of statistical significance.

The utility of knowing the actual distribution function (in contrast to a few isolated qualitative guesses or estimates) is that the pdf provides 1) a self-con-

sistent measure of risk of infectivity as a function of time, 2) a mathematical structure for calculating statistical moments, quantiles, and uncertainties, as well as graphical output analogous to Kaplan-Meier type survival curves [11], 3) predictions against which new empirical observations can be tested, and 4) a systematic framework for establishing whether different estimates of risk or different isolation guidelines are mutually consistent. Moreover, social science studies have indicated that people seriously underestimate the probability of having and spreading the coronavirus [12]. It is all the more necessary, therefore, that reliable probabilities concerning self-isolation upon infection be available to the public and to public health agencies.

The remainder of the paper is organized as follows:

In Section 2 the lognormal pdf and cdf of the period of COVID infectivity are determined from several published clinical studies and tested empirically for consistency. Graphical displays of the complementary cumulative distribution function (ccdf) show directly the probability of viral shedding (as determined by positive antigen or PCR tests) as a function of time following the onset of COVID symptoms.

Because relatively few published reports of COVID clinical studies were found to provide adequate information for construction of the desired distribution functions, the author describes in Section 3 a prototypical procedure for obtaining the pertinent statistics. These statistics comprise the mean numbers or quantiles (percentages) of daily positive tests within a tracked cohort. As statistics, these data are variates (realizations) of random variables of unknown distributions. An objective analysis employing the PME and CLT shows that the most probable distribution of the number or fraction of positive COVID tests within a cohort, given the constraining information, is multinomial in form in which the probability of infectivity p_t at day t is itself a lognormal random variable (RV). It is then shown that the overall probability function of a multinomial distribution of a set of powers of products of lognormally distributed variables $\{p_t\}$ very closely approximates a lognormal pdf. The full demonstration comprises a sequence of steps, each one clearly explained and implemented in a separate subsection of Section 3.

Section 4 examines the constraints of application of the PME and addresses questions of whether the derived distributions uniquely describe the period of infectivity and probability of infectivity.

Conclusions and implications of this paper are summarized in the final Section 5.

2. Period of Infectivity Following COVID Infection

It is widely, yet erroneously, believed that anthropometric features like height, weight, body mass index (BMI) and other variable human attributes follow a Gaussian distribution, which, after all, is called the "normal" distribution. However, as shown recently, height and weight are correlated bivariate lognormal

random variables [13], and BMI is a univariate lognormal variable [14]. Indeed, any non-negative distributed anthropometric quantity is more likely to follow a lognormal distribution than a normal distribution for at least two reasons. First, the range of a normal distribution spans the entire real domain and therefore theoretically includes unphysical negative variates, which lead to spurious results unless the ratio of mean to standard deviation is sufficiently high. Second, the normal distribution is symmetric about the mean, whereas non-negative physical variables often display a positive skewness in view of the fact that the lower bound is fixed and the upper bound is unlimited.

Lognormal variables occur widely in medicine and biosciences, physical sciences, and social sciences [15] and engineering [16]. Of especial relevance to this paper is the comprehensive investigation of Sartwell [17] more than 60 years ago, showing that the incubation periods of approximately 20 diseases followed a lognormal distribution. Wide variations in the incubation periods were attributed to differences in the strains of the pathogens and their modes of infection. Subsequent studies have also borne out Sartwell's conclusion regarding lognormal incubation periods in the case of coronavirus [18].

The incubation period is, in the words of Sartwell, "the time required for multiplication of the parasitic organism within the host organism up to the threshold point at which the parasite population is large enough to produce symptoms of the host". With regard to coronavirus and the current analysis, this can be considered the latent period between exposure to an infected agent and the first occurrence of a positive antigen test on day 0. The interval of concern in this paper, to be referred to as the infectivity period, is the interval between day 0 and the first occurrence of a negative test on day $t+1$, under the condition that the patient is isolated from further exposure and subject to antigen testing once a day. In analogy to the incubation period, the infectivity period is likewise a time period during which coronavirus levels grow and eventually subside in response to the body's immune system.

From a statistical perspective, incubation and infectivity periods represent stochastic birth-and-death processes [19], whereupon one might anticipate that the statistics of the coronavirus infectivity period would be lognormal, if the incubation period is lognormal. Besides this epidemiological expectation, the Principle of Maximum entropy, discussed in more detail in Section 3, lends support to a lognormally distributed period of infectivity.

2.1. Properties of the Lognormal Distribution

The lognormal distribution, represented symbolically by $\Lambda\left(m,s^2\right)$, is a two-parameter distribution defined by the mean m and variance s^2 of the *parent* normal distribution, symbolically represented by $N\left(m,s^2\right)$. It is to be recalled that a random variable X is lognormal if the variable $Y=\ln X$ is normal. The preceding symbolism implies that the corresponding sets of variates (*i.e.* realizations) $\left\{x_t\right\}$, $\left\{y_t\right\}$ $t=1,\cdots,T$, of the variables are related by $y_t=\ln\left(x_t\right)$.

The pdf of a lognormal variable X

$$p_X(x) = \frac{1}{\sqrt{2\pi s^2}\,x}\exp\left(-\left(\ln(x)-m\right)^2/2s^2\right), \tag{1}$$

readily follows from the well known pdf of a normal variable Y

$$p_Y(y) = \frac{1}{\sqrt{2\pi s^2}}\exp\left(-(y-m)^2/2s^2\right) \tag{2}$$

by the transformation $y=\ln(x)$. The k^{th} statistical moment of $\Lambda\left(m,s^2\right)$ is defined by the expectation

$$M_k \equiv \left\langle X^k \right\rangle = \frac{1}{\sqrt{2\pi s^2}}\int_0^\infty x^{k-1}\exp\left(-\left(\ln(x)-m\right)^2/2s^2\right)\mathrm{d}x \tag{3}$$

where the quantity $M_0 = 1$ establishes the normalization required of a probability distribution.

Moments and statistics relevant to this paper are the lognormal mean

$$\mu \equiv M_1 = \exp\left(m+\frac{1}{2}s^2\right), \tag{4}$$

mean square

$$M_2 = \exp\left(2m+2s^2\right), \tag{5}$$

variance

$$\sigma^2 \equiv M_2 - M_1^2 = \exp(2m)\left[\exp\left(2s^2\right)-\exp\left(s^2\right)\right], \tag{6}$$

and skewness (a measure of asymmetry about the mean)

$$Sk \equiv \left\langle\left(\frac{X-\mu}{\sigma}\right)^3\right\rangle = \left(\exp\left(s^2\right)+2\right)\sqrt{\exp\left(s^2\right)-1}\,. \tag{7}$$

The inverse expressions by which m and s^2 are obtained from μ and σ^2 are

$$m = \ln\left(\mu^2/\sqrt{\mu^2+\sigma^2}\right) = \ln\left(M_1^2/\sqrt{M_2}\right) \tag{8}$$

$$s^2 = 2\ln\left(\sqrt{\mu^2+\sigma^2}/\mu\right) = \ln\left(1+\frac{\sigma^2}{\mu^2}\right) = 2\ln\left(\sqrt{M_2}/M_1\right). \tag{9}$$

Analyses of the lognormal distribution in greater detail are given in References [13] [14] [20] in conjunction with specific applications relating to the distribution of anthropometric attributes.

The lognormal cumulative distribution function (cdf),

$$F_X(q_k) \equiv \int_0^{q_k} p_X(x)\mathrm{d}x = \frac{1}{2}\left(1+\operatorname{erf}\left(\frac{\ln(q_k)-m}{\sqrt{2}\,s}\right)\right) = k, \tag{10}$$

where $1 \ge k \ge 0$, leads to the quantiles q_k, which in most cases must be solved numerically for q_k. The exception is the median, $F_X(q_{0.5}) = 0.5$, for which

$$q_{0.5} = \exp(m). \tag{11}$$

The first and third quartiles, for which $F_X(q_{0.25}) = 0.25$ and $F_X(q_{0.75}) = 0.75$, are respectively

$$q_{0.25} = \exp(m - 0.67448975\,s) \tag{12}$$

$$q_{0.75} = \exp(m + 0.67448975\,s), \tag{13}$$

and the 95% quantile, for which $F_X(q_{0.95}) = 0.95$, is

$$q_{0.95} = \exp(m + 1.64485363\,s). \tag{14}$$

It is also relevant to include the complementary cumulative distribution function (ccdf)

$$\tilde{F}_X(q_k) \equiv \int_{q_k}^{\infty} p_X(x)\,\mathrm{d}x = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{\ln(q_k) - m}{\sqrt{2}\,s}\right)\right) = 1 - k \tag{15}$$

which, in the present context, would correspond to a Kaplan-Meier type survival curve [21]. Thus, the 95% quantile (14) is also the 5% quantile of $\tilde{F}_X(q_{0.95}) \equiv 0.05$.

2.2. Lognormal pdf and ccdf of COVID Period of Infectivity

From a report of the Erasmus Medical Center (Rotterdam NL) trial [22] that the median time period of positive tests to a first negative test was $q_{0.5} = 8$ days following the onset of symptoms at day 0, and that on average 5% of people infected with COVID remained infected at $q_{0.95} = 15.2$ days, there follows from Equation (11) the parameter

$$m_0 = \ln(8) = 2.07944 \tag{16}$$

and subsequently from Equation (14) or (15) the parameter

$$s_0 = 0.39022\,. \tag{17}$$

Empirical parameters (16) and (17), marked by a subscript 0, uniquely determine the lognormal pdf (1), cdf (10), and ccdf (15) of the infectivity period characteristic of the investigation reported in Ref. [22]. A plot of the pdf of the distribution $\Lambda(m_0, s_0^2)$ is shown in **Figure 1**. The mean, median, standard deviation, and skewness of the curve, calculated from Equations (4), (11), (6), and (7) are respectively $\bar{t} = 8.6329$, $\hat{t} = 8.0000$, $\sigma_t = 3.5011$, and $Sk_t = 1.2834$.

Figure 2 shows a plot of the corresponding ccdf. Dashed horizontal black lines mark the 50% and 5% quantiles with associated time periods starting at the vertical dashed black lines. These are the two data points from [22] from which the lognormal parameters m_0, s_0 were determined. However, **Figure 2** also provides a means of testing the consistency of the lognormal distribution with other published quantile information. In an independent investigation by the UMass Chan Medical School [2], researchers reported the 30% quantile at period $t \geq 10$ days. As seen in the figure, the intersection of the horizontal and vertical dashed blue lines corresponding to a 30% infectivity at $t \geq 10$ days falls very nearly exactly on the curve in **Figure 2**. To this extent, therefore, the reports of infectivity from the two cited independent investigations appear to be mutually consistent and support the lognormal hypothesis.

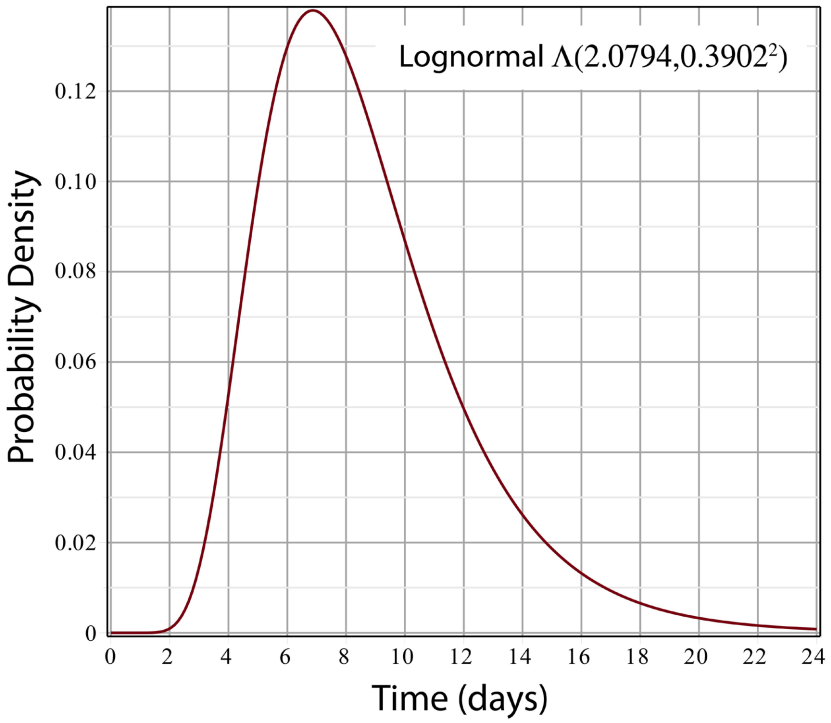


Figure 1. Lognormal probability density of period of infectivity following onset of COVID symptoms on day 0.

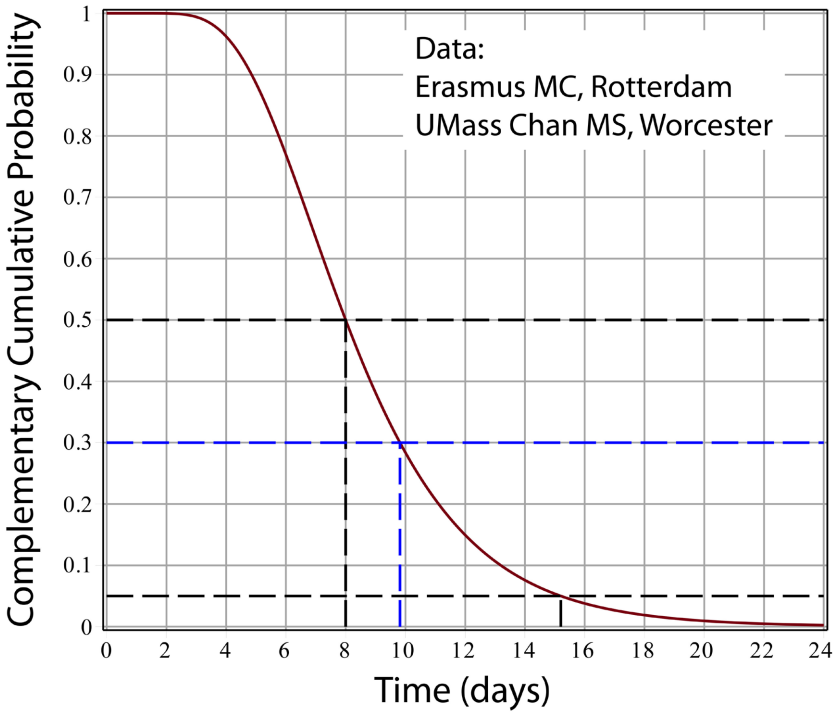


Figure 2. Ccdf of the probability density in **Figure 1**. Dashed black lines mark the 50% and 5% quantile points from Erasmus Medical Center used to construct the lognormal pdf. Dashed blue lines mark the 30% quantile point reported by the UMass Chan Medical School. Consistency of the two sets of measurements support a lognormal distribution.

In judging the consistency of outcomes, it is important to be aware of significant differences in the protocols and sample sizes of different investigations. Consider the following examples.

A third independent investigation (University of Exeter UK) [23] reported that 26% and 28% of individuals in two cohorts tested positive at and beyond day 11 of a total trial period lasting at least 31 days. Time was measured from the onset of symptoms to the last positive test. These results appear consistent with the 30% quantile at $t \geq 10$ days depicted in **Figure 2**, since the standard error of the mean outcome (27%) of the two cohorts is $\sigma_t / \sqrt{2} = 2.48$ days, which exceeds 1 day. However, the tests employed were polymerase chain reaction (PCR) tests rather than antigen tests. PCR tests measure viral genetic material, whereas antigen tests detect substances that cause the body to produce an immune response (*i.e.* create antibodies). Antigen tests may be less accurate than PCR tests, but the latter can detect residual viral genetic material and generate a positive result after an active infection is over [24]. That is why antigen tests are ordinarily used to ascertain current infectivity, apart from the other advantage of providing results rapidly. Nevertheless, infectivity measured by PCR tests also appear consistent with the lognormal distribution in **Figure 2**.

A fourth independent investigation (Imperial College London) [25] reported that the overall median amount of time that people were infectious, as determined by PCR tests, was 5 days [26]. The sample size of this study, however, was particularly small. From an initial total of 57 people in the sample, the date of onset of symptoms was known with certainty for only 38 of whom 34 contributed the following data pertinent to this paper: (a) 22 of 34—*i.e.* 64.7%—remained infectious at $t \geq 5$ days, and (b) 8 of 34—*i.e.* 23.5%—remained infectious at $t \geq 7$ days. From the quantiles specified in (a) and (b), one can construct the ccdf shown in **Figure 3**. Horizontal and vertical dashed black lines mark the two data points from which the lognormal parameters

$$\begin{aligned} m_1 &= 1.7250 \\ s_1 &= 0.3062 \end{aligned} \tag{18}$$

were determined. The mean, median, standard deviation, and skewness of the curve, calculated from Equations (4), (11), (6), and (7) are respectively $\bar{t} = 5.8818$, $\hat{t} = 5.6125$, $\sigma_t = 1.8440$, and $Sk_t = 0.9713$. The cdf curve in **Figure 3** is not consistent with the cdf curve in **Figure 2**. However, the three sample points *are* consistent with a lognormal distribution. The horizontal and vertical dashed blue lines mark the lognormal predicted median at 5.61 days, which is within 1 standard deviation (1.84 days; confidence interval of 29.1%) of the reported sample median (5 days).

It should be stressed that inconsistency between two independent studies does not necessarily mean that either of them was flawed. Differences among trial outcomes can arise because of different protocols (e.g. antigen vs. PCR testing), demographics (a cohort with predominantly young, previously healthy participants vs. elderly, less healthy patients), sample size (affecting statistical uncertainties),

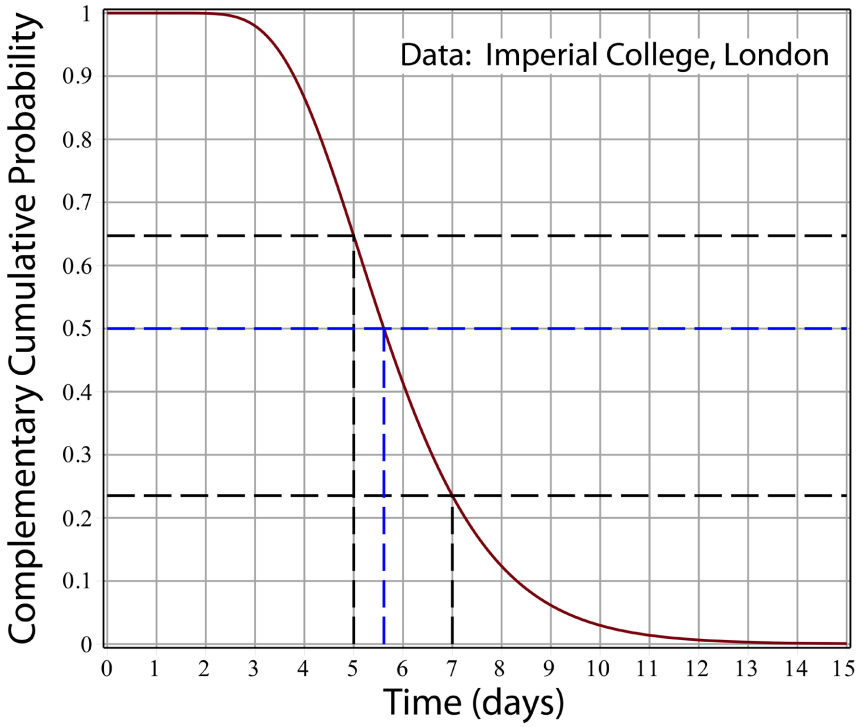


Figure 3. Ccdf constructed from the 64.7% and 23.5% quantile points (dashed black lines) reported by the Imperial College group. The same report claimed a 50% quantile (median) of 5 days in comparison with the lognormal predicted median of 5.6 days.

coronavirus variants (affecting severity of infection), among other reasons. Indeed, daily sampling of COVID infectivity has revealed significant heterogeneity in infectiousness [27]. However, in the author's attempts to find reliable statistical information regarding COVID infectivity, a pervasive problem was that few accessible published reports contained the kind of quantitative information needed to construct a pdf or cdf of the infectivity period. In the following section the fundamentals of an ideal clinical trial is outlined that would provide the needed information. The analysis, based on general statistical principles rather than any detailed dynamical model, leads to lognormal distributions for the probability of infectivity throughout the period of viral shedding.

3. Analysis of an Ideal Clinical Trial

The reasoning supporting the proposition that coronavirus infectivity, and perhaps the infectivity of other diseases as well, follows a lognormal distribution makes use of the Principle of Maximum Entropy (PME), the Central Limit Theorem (CLT), and the relation between the lognormal distribution and the distribution of quotients of two random variables. The analysis begins with an examination of the fundamentals of an ideal clinical trial. Although different investigations to study infectivity as a function of time may employ different procedures, they all must in some way be comparable to the archetype examined in this section. If this were not the case, then the data derived from the outcome of

such a clinical trial would be insufficient to determine the sought-for statistical distributions.

Consider a clinical trial comprising G groups of patients displaying symptoms of COVID on day 0. The trial runs for T days. The total number of patients in each group is $n_{0,g}$ $(g=1,\cdots,G)$. An antigen test is administered daily, and the number of patients in each group testing positive is recorded over the time period t from day 0 to day T, as illustrated in Figure 4 in which $n_{t,g}$ $(t=0,\cdots,T)$ is the number testing positive in group g on day t. The sample mean of positive tests on day t is then

$$n_t = \frac{1}{G}\sum_{g=1}^{G} n_{t,g} \,. \tag{19}$$

The set of numbers $\{n_{t,g}\}$ are all random variables (RVs) of unknown distribution. Moreover, the information reported by a clinical trial ordinarily comprises just the sample means, or more likely just a subset of the sample ratios n_t/n_0 from which quantiles of infectivity (such as plotted in Figure 2 and Figure 3) are deducible, rather than the entire record. However, these sample means and ratios are also random variables. For such numbers to be predictively useful, what is desired is an objective estimate of their statistical distributions on the basis of this incomplete information. From these distributions are obtained the probability of infectiousness (*i.e.* testing positive) on day t, given that symptoms began (or the first positive test occurred) on day 0. By "objective" is meant that the resulting solution is free of extraneous assumptions and is determined only by the information that one has.

Time Units	Group No.	1	2	⋯	G	Sample Mean
0		$n_{0,1}$	$n_{0,2}$	⋯	$n_{0,G}$	n_0
1		$n_{1,1}$	$n_{1,2}$	⋯	$n_{1,G}$	n_1
2		$n_{2,1}$	$n_{2,2}$	⋯	$n_{2,G}$	n_2
⋮		⋮	⋮	⋮	⋮	⋮
T		$n_{T,1}$	$n_{T,2}$	⋯	$n_{T,G}$	n_T

Figure 4. Hypothetical statistical data from an ideal clinical trial for determining the probability of infectivity at day t following onset of symptoms on day 0. The trial comprises G groups with a total of $n_{0,g}$ participants in group $g=1,\cdots,G$. The number of participants testing positive in group g on day t is $n_{t,g}$. Information made available for analysis is a set of all or part of the group-averaged sample means $\{n_t\}$ or ratios $\{n_t/n_0\}$.

3.1. Maximum Entropy Distribution

The most objective distribution compatible with known information can be found by using two fundamental statistical principles: the Principle of Maximum Entropy (PME) and the Central Limit Theorem (CLT).

A) The PME:

It is beyond the scope of this paper to provide a detailed discussion of the meaning and scope of entropy in physics. For the present purposes, let it be sufficient to say that entropy is a measure of the total number of unobserved degrees of freedom (e.g. particle coordinates) of a system that manifests some macroscopic state (e.g. temperature, pressure, and density). The greater the entropy, the more probable the state. In brief, entropy is a measure of information, and the PME produces the most probable solution, given the information available. In many of the problems to which the PME has been applied, the number of unobserved ways the PME solution can be realized is *astronomically* larger than for any other proposed distribution [28] that satisfies the initial constraints.

The PME was originally employed by Jaynes [29] [30], using the Shannon expression for entropy [31], to derive the fundamental principles and relations of equilibrium statistical mechanics. However, it can be employed to solve many problems outside the domain of physics. One such example, which illustrates the method in more detail than can be presented here, and which may prove useful especially in science and medicine, is to ascertain the likelihood that a submitted work was plagiarized [32].

The PME is a variational procedure for finding an unknown probability distribution. Given a set of outcomes $\{x_k\}$ $(k=1,\cdots,K)$ with an associated set of unknown probabilities $\{p_k\}$, the Shannon entropy is

$$H(p_1,\cdots,p_K)\equiv -\sum_{k=1}^{K} p_k \ln p_k \,. \tag{20}$$

Suppose all that is known about a system is the mean outcome

$$\mu = \langle X\rangle = \sum_{k=1}^{K} x_k p_k \tag{21}$$

together with the normalization requirement for a set of probabilities

$$1 = \sum_{k=1}^{K} p_k \,. \tag{22}$$

One then constructs the functional

$$\tilde{H}(p_1,\cdots,p_K) = -\sum_{k=1}^{K} p_k \ln p_k + \lambda_0\left(1-\sum_{k=1}^{K} p_k\right) + \lambda_1\left(\mu - \sum_{k=1}^{K} x_k p_k\right), \tag{23}$$

where λ_0 and λ_1 are Lagrange multipliers, and varies it with respect to each p_j

$$\frac{\delta H(p_1,\cdots,p_K)}{\delta p_j} = 0 \tag{24}$$

for $j=1,\cdots,K$.

Under conditions where the sought-for distribution has a finite, nonzero mean, and the number of possible outcomes is finite, the implementation of Equation (24) leads to a binomial distribution [33]. In the present context of ascertaining the probability that m patients out of a sample of size n_0 and mean n_t test positive on day t, the PME solution takes the form

$$P\left(m\left|n_0,n_t\right.\right)=\frac{n_0!}{m!\left(n_0-m\right)!}\left(\frac{n_t}{n_0}\right)^m\left(1-\frac{n_t}{n_0}\right)^{n_0-m}. \tag{25}$$

From the binomial probability function (25), it is seen that the probability of infectivity p_t on day t is

$$p_t=n_t/n_0, \tag{26}$$

the ratio of the sample mean to the total sample size. For example if $n_0=100$ and $n_8=50$, then an individual has a 50% chance of still being infectious at day 8 for this particular set of trials.

If the information available initially included a set of mean values $\{n_t\}$, as shown in the rightmost column of **Figure 4**, then the PME solution, which generalizes Equation (25), would be a multinomial distribution [34]

$$P\left(m_1,\cdots,m_T\left|n_0,p_1,\cdots,p_T\right.\right)\equiv P\left(\left\{m_t\right\}\left|n_0,\left\{p_t\right\}\right.\right)=n_0!\prod_{t=1}^{T}\frac{p_t^{m_t}}{m_t!} \tag{27}$$

where the probability of infectivity p_t $\left(t=1,\cdots,T\right)$ is still defined by Equation (26) and the partition numbers m_t satisfy the relation

$$\sum_{t=1}^{T}m_t=n_0. \tag{28}$$

Whereas each p_t is a parameter of the PME solution, it is, in fact, the ratio of the means of two random variables of unknown distribution. To find the distribution of this ratio, one can employ the CLT.

Before doing so, it is useful to mention that the probability function (25) or (27) leads to the variance for p_t

$$\sigma_t^2=n_0p_t\left(1-p_t\right). \tag{29}$$

In the limit of increasing sample size n_0 and decreasing probability p_t, the variance (29) approaches the mean

$$\sigma_t^2\approx n_0p_t=n_t, \tag{30}$$

which is a characteristic of a Poisson distribution. And, indeed, it can be shown (by means of the moment generating function) that the probability function (25) reduces to the Poisson probability function

$$P\left(m\left|\lambda_t\right.\right)=\frac{\lambda_t^m}{m!}\exp\left(-\lambda_t\right) \tag{31}$$

under conditions of low probability and large sample size, such that product yields the sample mean λ_t. For purposes of illustration, the Poisson condition on variance will be assumed to hold such that relation (30) is valid. This is a convenience, not a requirement.

Finally, it is worth emphasizing that the PME solutions (25), (26), and (27) did not depend on sample size, but only on the specified known information. The functional form of a PME solution would be different if more information than sample means were known at the outset. For example, if the mean and variance were known information for some system whose random variable spanned the entire real axis, then the derived PME solution would be of Gaussian form. A Gaussian solution, however, would not apply in the present case, even if the variance were known, because the initial information, besides relations (21) and (22), required only non-negative outcomes. Under such circumstances the PME solution for a segment of the real axis, given known mean and variance, would be a truncated Gaussian distribution [35] [36]. But this PME solution would also not apply in the present case because implicit in the initial information is the requirement that the pdf or cdf reach zero smoothly at the origin, whereas the pdf of a truncated Gaussian can be nonzero at the origin.

By contrast, the requirements of known mean and variance, together with non-negativity and continuity at the origin are satisfied by the lognormal distribution. Nevertheless, it is an open question whether the lognormal distribution yields the greatest entropy of any continuous distribution consistent with this initial information. However, even without rigorous proof of this point, the validity of a lognormal distribution of the COVID period of infectivity can be tested empirically once clinical investigations provide the necessary data. The constraints on PME-derived distributions are consider further in Section 4.

B) The CLT:

As a broadly applicable principle, the CLT states that the distribution of the mean of G samples from a population of finite mean μ and variance σ^2 approaches the normal distribution $N\left(\mu, \frac{\sigma^2}{G}\right)$ in the limit of increasing sample size *irrespective* of the actual distribution of the individual samples [37] [38]. The CLT does not indicate how rapidly the normal distribution is approached as the sample size is increased. Nevertheless, the power of the CLT is that the type of random variables being sampled is irrelevant provided the sample size is large enough and the first two moments of the distribution are finite. Since the goal here is to find an objective estimate of the *population* statistics, the assumption of a sufficiently large sample size is justified. The significance of the CLT in the present context is that the probability of infectivity (26) is a random variable Z_t whose distribution function is, for all practical purposes, that of the ratio of two normal RVs

$$Z_t \sim N\left(n_t, s_t^2\right) \Big/ N\left(n_0, s_0^2\right) \tag{32}$$

where, depending on the initially known information, the variances can either be assumed Poissonian

$$\begin{aligned} s_t^2 &= n_t / G \\ s_0^2 &= n_0 / G \end{aligned} \tag{33}$$

or else taken as independently measured quantities.

Although there are distributions, such as the Cauchy distribution [39], which have no finite moments and therefore do not meet the criteria for application of the PME or CLT, such exceptions do not occur in clinical trials with finite numbers of non-negative integer variables.

3.2. Distribution of the Ratio of Two Normal Random Variables

Following the results of the preceding section, the next step is to determine the distribution of a variable Z, which is the ratio of two normal RVs. Given independent random variables X and Y with respective pdfs $p_X(x)$ and $p_Y(y)$, the pdf $p_Z(z)$ of the quotient

$$Z = X/Y \tag{34}$$

is readily obtained by use of the Dirac delta function $\delta(z - x/y)$ [40]

$$p_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_X(x)\, p_Y(y)\, \delta(z - x/y)\, \mathrm{d}x\mathrm{d}y \tag{35}$$

to yield [20]

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(zy)\, p_Y(y)|y|\, \mathrm{d}y\,. \tag{36}$$

If X and Y are normal RVs

$$\begin{aligned} X &= N\left(m_1, s_1^2\right) \\ Y &= N\left(m_2, s_2^2\right) \end{aligned} \tag{37}$$

substitution into Equation (36) of Gaussian pdfs (2) with the appropriate means and variances shown in relations (37) leads to the complicated expression [41]

$$\begin{aligned} p_Z(z) = {} & \frac{1}{\sqrt{2\pi}} \frac{m_1 s_2^2 z + m_2 s_1^2}{\left(s_2^2 z^2 + s_1^2\right)^{3/2}} \exp\left(-\frac{\left(m_1 - m_2 z\right)^2}{2\left(s_2^2 z^2 + s_1^2\right)}\right) \operatorname{erf}\left(\frac{m_1 s_2^2 z + m_2 s_1^2}{\sqrt{2}\, s_1 s_2 \left(s_2^2 z^2 + s_1^2\right)^{1/2}}\right) \\ & + \frac{s_1 s_2}{\pi\left(s_2^2 z^2 + s_1^2\right)} \exp\left(-\left(\frac{m_1^2}{s_1^2} + \frac{m_2^2}{s_2^2}\right)\right). \end{aligned} \tag{38}$$

To the author's knowledge, pdf (38) has not been previously associated with any named random variable. However, it has been shown [41] that if X and Y are non-negative with $m_i/s_i \gg 1 \ (i = 1,2)$, then pdf (38) can be closely approximated by the non-Gaussian expression

$$p_Z(z) \approx \frac{1}{\sqrt{2\pi}} \frac{m_1 s_2^2 z + m_2 s_1^2}{\left(s_2^2 z^2 + s_1^2\right)^{3/2}} \exp\left(-\frac{\left(m_1 - m_2 z\right)^2}{2\left(s_2^2 z^2 + s_1^2\right)}\right), \tag{39}$$

which can then be transformed into a standard normal distribution $N(0,1)$ under a change of variable from z to

$$\theta = \left(m_2 z - m_1\right) \Big/ \left(s_2^2 z^2 + s_1^2\right)^{1/2}. \tag{40}$$

The random variable Θ corresponding to the variate θ is then symbolized by

$$\Theta = N(0,1). \tag{41}$$

Inverting relation (40) in order to solve for z in terms of θ, one obtains an expression of which the logarithm is

$$\ln(z) = \ln\left(m_1 m_2 + \sqrt{\left(m_1^2 s_2^2 + m_2^2 s_1^2\right)\theta^2 - s_1^2 s_2^2 \theta^4}\right) - \ln\left(m_2^2 - s_2^2 \theta^2\right). \tag{42}$$

By means of approximations consistent with those leading to Equation (39), one can reduce Equation (42) to the simpler form

$$\ln(z) \approx \ln(m) + \left(\frac{s}{m}\right)\theta \tag{43}$$

in which

$$m = m_1 / m_2 \tag{44}$$

and

$$\frac{s}{m} = \sqrt{\frac{s_1^2}{m_1^2} + \frac{s_2^2}{m_2^2}}. \tag{45}$$

It then follows from relations (41) and (43) that the RV

$$\ln Z \sim \ln(m) + (s/m)N(0,1) = N\left(\ln(m), s^2/m^2\right) \tag{46}$$

is approximately normal, whereupon Z is representable as a lognormal RV

$$Z \equiv X/Y \sim \Lambda\left(\ln(m), s^2/m^2\right) \tag{47}$$

with pdf

$$p_Z(z) = \frac{1}{\sqrt{2\pi\left(s^2/m^2\right)}} \frac{\exp\left(\left(z - \ln(m)\right)^2 / 2\left(s^2/m^2\right)\right)}{z}, \tag{48}$$

provided the foregoing conditions $\left(m_1/s_1 \gg 1, m_2/s_2 \gg 1\right)$ maintain.

The consistency of the preceding analysis becomes apparent once one recognizes that the mean and variance of X/Y are known to be approximately

$$\mu_{X/Y} \approx m_1/m_2 \tag{49}$$

$$\frac{\sigma_{X/Y}^2}{\mu_{X/Y}^2} \approx \frac{s_1^2}{m_1^2} + \frac{s_2^2}{m_2^2} \tag{50}$$

as derived by a method of series expansion *independent* of the exact distributions of the numerator and denominator [42]. The method of approximation is sometimes referred to as Error Propagation Theory (EPT) [43]. Thus, one can re-express relation (46) as

$$\ln(Z) = \ln(X/Y) \sim N\left(\ln\left(\mu_{X/Y}\right), \sigma_{X/Y}^2 / \mu_{X/Y}^2\right). \tag{51}$$

If Z were *exactly* lognormal $\Lambda\left(m_Z, s_Z^2\right)$, then the exact parameters m_Z and s_Z^2 would relate to the *exact* mean μ_Z and variance σ_Z^2 according to relations (8) and (9), respectively. However, under the assumed condition that $\mu_Z/\sigma_Z \gg 1$, these expressions reduce to

$$m_Z \equiv \ln\left(\frac{\mu_Z^2}{\sqrt{\mu_Z^2+\sigma_Z^2}}\right) \approx \ln(\mu_Z) = \ln(m_1/m_2) \tag{52}$$

$$s_Z^2 \equiv \ln\left(1+\frac{\sigma_Z^2}{\mu_Z^2}\right) \approx \frac{\sigma_Z^2}{\mu_Z^2} = \frac{s_1^2}{m_1^2}+\frac{s_2^2}{m_2^2} \tag{53}$$

in agreement with EPT expressions (49) and (50).

In short, the analysis in this section has established that the quotient variable Z reduces to a lognormal RV under the conditions likely to pertain in a clinical investigation to determine the probability p_t of COVID infectivity.

3.3. Test of the Lognormality of the Ratio of Two Normal Random Variables

To ascertain how well a lognormal distribution matches the distribution of the ratio of two normal RVs, the results of a hypothetical clinical trial were recorded in Table 1. The trial consisted of $G=10$ groups of infected participants who were tested daily for a total of 11 days, as marked in column 1, following the onset of symptoms on day 0. Column 2 records the daily mean number of patients with positive tests. Column 3 records the daily standard errors, *i.e.* the standard deviation of the means in column 2. Column 4 is the EPT estimate of the probability of infectivity p_t. Column 5 is the EPT estimate of the standard deviation of p_t. Columns 6 and 7 record the parameters of the corresponding lognormal distribution (47), defined by Equations (52) and (53). The numbers in column 2 most likely do not correspond to any real clinical trial. They were chosen only to provide a set of values of p_t more or less uniformly spanning the full range from 1 to close to 0 for testing how well the predicted relation Equation (47) holds.

Table 1. Distributions of normal ratio (X/Y) and corresponding lognormal (Z).

Time Units	n_t	$s_t=\sqrt{\frac{n_t}{G}}$	$m_{X/Y}=n_t/n_0$ (p_t)	$s_{X/Y}=\sqrt{\frac{s_t^2}{n_0^2}+\frac{n_t^2 s_0^2}{n_0^2}}$	m_Z Lognormal	s_Z Lognormal
0	100	3.1623	1.00	0.04478	0	0.04478
1	90	3.0000	0.90	0.04600	−1.1054	0.04600
2	80	2.8284	0.80	0.04748	−0.2233	0.04748
3	70	2.6457	0.70	0.04933	−0.3569	0.04933
4	60	2.4495	0.60	0.05168	−0.5112	0.05168
5	50	2.2361	0.50	0.05480	−0.6936	0.05480
6	40	2.0000	0.40	0.05918	−0.9170	0.05918
7	30	1.7321	0.30	0.06583	−1.2051	0.06583
8	20	1.4142	0.20	0.07742	−1.6114	0.07742
9	10	1.0000	0.10	0.10506	−2.3071	0.10467
10	5	0.0707	0.05	0.00726	−3.0051	0.14425
11	2	0.0447	0.02	0.00452	−3.9359	0.22315

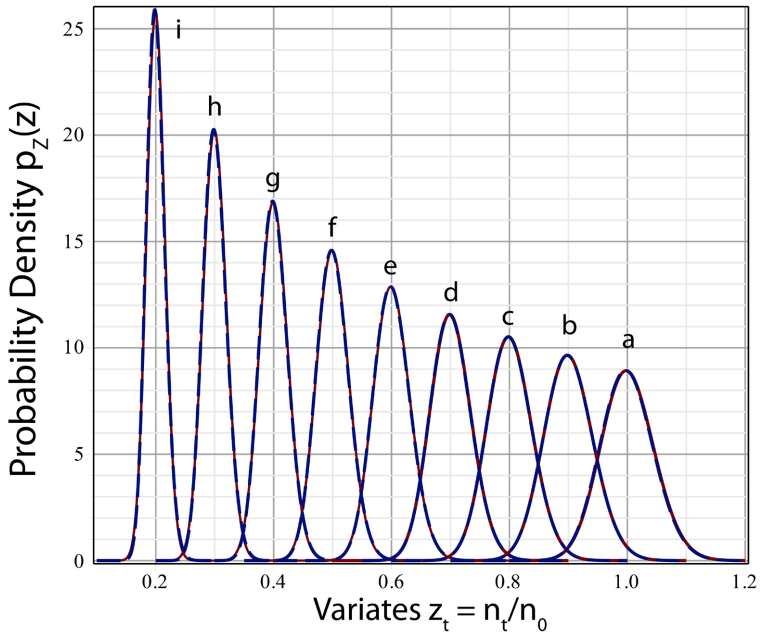
Table 2 compares the exact (to 4 decimal places) numerically calculated first and second moments and the medians of the normal ratio (X/Y) and lognormal (Λ) distributions created in Table 1. As seen from columns (2, 3) and (4, 5), the two sets of moments are identical to 4 decimal places. Columns (6, 7) also show the two sets of medians to differ weakly and progressively only in the 4th decimal place, as p_t decreases with time *t*. The identity of any finite number of moments does not establish the identity of two distributions. Such identity could be established by demonstrating that two distributions have the same moment generating function (mgf) or characteristic function (cf), but the mgf of a lognormal distribution does not exist, and there is no closed form expression for the cf.

An alternative procedure in the present case is simply to compare the two pdfs, Equations (38) and (48), as shown graphically in Figure 5(a) and Figure 5(b) for the full set of distributions of p_t $(t=0,\cdots,11)$ displayed in Table 1. Solid red curves pertain to the normal ratio distribution; dashed blue curves pertain to the corresponding lognormal distribution. Plots for days 0 through 8 are displayed in Figure 5(a). For each of the 9 days (infectivity probability ranging from 100% to 20%) superposed plots of the two pdfs are visually indistinguishable. Plots for days 9 through 11 (infectivity probability ranging from 10% to 2%) are displayed in Figure 5(b). Although the cohorts comprise very few participants who test positive for these days (respective means are 10, 5, and 2), the superposed plots of the two pdfs are barely distinguishable visually.

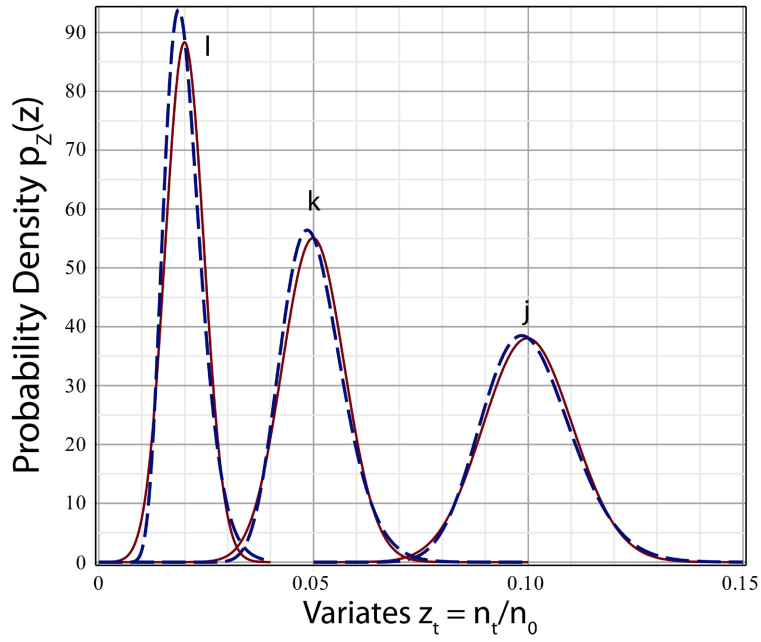
From the preceding analysis and graphical displays it can be concluded that the distribution of the probability of infectivity p_t is for all practical purposes very well represented by a lognormal distribution.

Table 2. Comparison of first and second moments and medians of normal ratio (X/Y) and corresponding lognormal (Z).

Time Units	$(M_1)_{X/Y}$	$(M_1)_\Lambda$	$(M_2)_{X/Y}$	$(M_2)_\Lambda$	$(M_{1/2})_{X/Y}$	$(M_{1/2})_\Lambda$
0	1.0010	1.0010	1.0040	1.0040	1.0000	1.0000
1	0.9009	0.9009	0.8133	0.8133	0.9000	0.9000
2	0.8008	0.8008	0.6247	0.6247	0.8000	0.7999
3	0.7007	0.7007	0.4922	0.4922	0.7000	0.6999
4	0.6006	0.6006	0.3617	0.3617	0.6000	0.5998
5	0.5005	0.5005	0.2513	0.2513	0.5000	0.4998
6	0.4004	0.4004	0.1609	0.1609	0.4000	0.3997
7	0.3003	0.3003	0.0906	0.0906	0.3000	0.2997
8	0.2002	0.2002	0.0403	0.0403	0.2000	0.1996
9	0.1001	0.1001	0.0101	0.0101	0.1000	0.0996
10	0.0501	0.0501	0.0026	0.0026	0.0500	0.0495
11	0.0200	0.0200	0.0004	0.0004	0.0200	0.0195



(a)



(b)

Figure 5. (a) Probability density profiles of the outcomes $z_t = n_t/n_0 \quad (t = 0,\cdots,8)$ where $n_0 = 100$ and n_t = (a) 100, (b) 90, (c) 80, (d) 70, (e) 60, (f) 50, (g) 40, (h) 30, (i) 20 as calculated from the exact relation (38) for the ratio of two normal RVs (solid red curve) and from the equivalent lognormal density (48) (dashed blue curve). The two sets of curves are visually indistinguishable; (b) probability density profiles of the outcomes $z_t = n_t/n_0 \quad (t = 9,10,11)$ where $n_0 = 100$ and n_t = (j) 10, (k) 5, (l) 2 as calculated from the exact relation (38) for the ratio of two normal RVs (solid red curve) and from the equivalent lognormal density (48) (dashed blue curve). The two sets of curves match closely, apart from small deviations near the peaks and wings.

3.4. Probability Distribution of Infectivity

Return now to the PME-derived binomial distribution (25), which can be re-expressed in the form

$$P\left(m\left|z_t,n_0\right.\right)=\frac{n_0!}{m!\left(n_0-m\right)!}\left(z_t\right)^m\left(1-z_t\right)^{n_0-m}, \tag{54}$$

where, in view of the preceding section,

$$z_t\equiv n_t/n_0=p_t \tag{55}$$

is itself a variate of the lognormal distribution (47) with range from 0 to 1 (since it is a probability). Thus, the distribution of the number of positive outcomes m at time t is obtained by integrating relation (54) over the range of z_t as follows

$$\begin{aligned}P\left(m\left|t,n_0\right.\right)&=\int_0^1 P\left(m\left|z_t,n_0\right.\right)p_{Z_t}\left(z_t\right)\mathrm{d}z_t\\&=\frac{n_0!}{\sqrt{\frac{\pi}{2}}\left(1-\operatorname{erf}\left(\frac{m_t}{\sqrt{2}s_t}\right)\right)m!\left(n_0-m!\right)}\int_0^1 z_t^{m-1}\left(1-z_t\right)^{n_0-m}\exp\left(-\frac{\left(\ln\left(z_t\right)-m_t\right)^2}{2s_t^2}\right)\mathrm{d}z_t\end{aligned} \tag{56}$$

where the parenthetical expression with error function in the denominator arises from normalization of p_t whose range spans only a segment (0, 1) of the positive real axis.

There is no closed form expression for the integral (56), but numerical evaluations with graphical display confirm that probability function (56) describes a lognormal distribution as illustrated in **Figure 6** for outcomes of a hypothetical clinical trial with 10 cohorts of 50 patients each. Blue diamonds in the figure mark points of the binomial distribution (25) with arbitrarily chosen fixed probability $z_t=0.5$ for illustration. Red circles mark points of the compound probability function (56) with lognormal distribution of z_t with parameters

$$\begin{aligned}m_t&=\ln\left(0.5\right)=-0.6931\\s_t&=\sqrt{\left|m_t\right|/10}=0.2633\end{aligned} \tag{57}$$

corresponding to a trial with 10 groups and Poissonian variance. The solid blue curve, which closely superposes the trajectory of red circles of the compound distribution is a lognormal distribution with parameters

$$\begin{aligned}m_t^{(\Lambda)}&=\ln\left(25.3\right)=3.231\\s_t^{(\Lambda)}&=0.309\end{aligned} \tag{58}$$

obtained by visual best fit to the plot of Equation (56). The parameters (58) agree closely with theoretically expected lognormal parameters

$$\begin{aligned}m_{theory}^{(\Lambda)}&=3.208\\s_{theory}^{(\Lambda)}&=0.285\end{aligned} \tag{59}$$

obtained from relations (8) and (9) in terms of the mean and variance

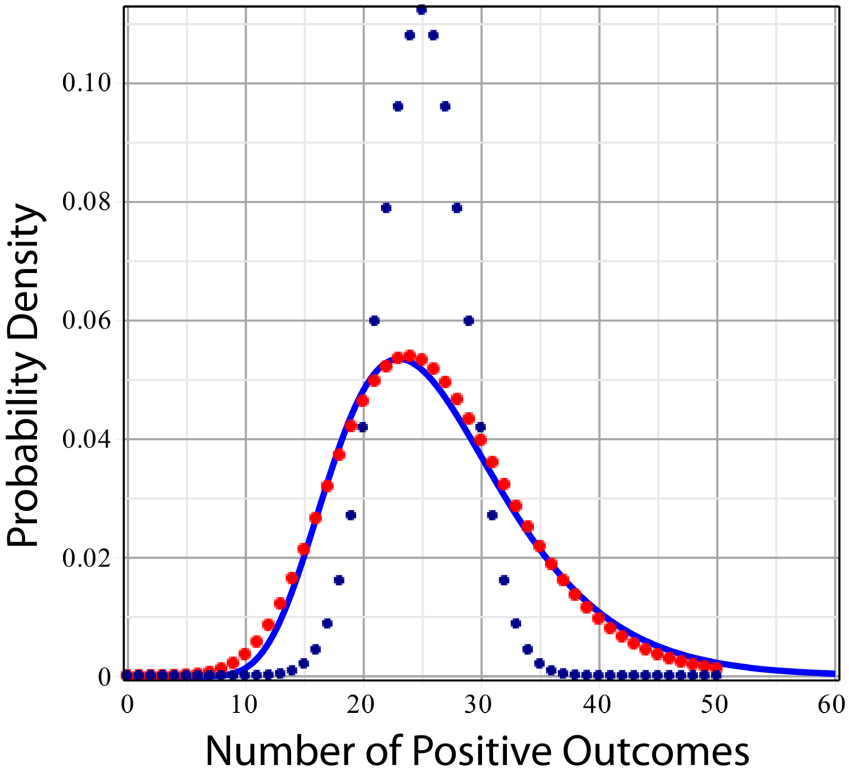


Figure 6. Binomial probability (blue diamonds) of m successes out of 50 trials with probability of success $z = 0.5$. Probability function (red circles) of the binomial distribution compounded with a lognormal distribution of z. Lognormal probability density (solid blue curve) with parameters derived from the mean and variance of the compounded binomial-lognormal distribution.

$$\begin{aligned} \mu_t &= 25.760 \\ \sigma_t^2 &= 56.316 \end{aligned} \tag{60}$$

calculated directly from Equation (56) .

3.5. Distribution of Infectivity as a Function of Time

The PME-derived multinomial probability function (27), compounded with lognormal pdfs $p_{Z_t}(z_t)$ of infectivity for the span of time $t = 1, \cdots, T$ predicts the fraction of patients testing positive each day of the infectivity period. This is the information needed for determining quantiles and other statistics of the period of infectivity. However complicated such a compound probability function may be, it is straightforward to show that the function describes a lognormal distribution.

In structure, the probability function (27) comprises factors of powers of the lognormal variables $z_t = n_t / n_0 = p_t$. In analogy to the CLT, by which sums of random variables approach a normal RV (provided certain conditions are met), products of powers of random variables approach a lognormal RV. Moreover, if the factor variables are themselves lognormal, then a product of powers of these

factors is *exactly* lognormal, as shown by the following argument.

Consider the variable Z defined by the product of powers c_i of $i=1,\cdots,M$ independent lognormal variables $\Lambda\left(a_i,b_i^2\right)$

$$Z=\prod_{i=1}^{M}\Lambda\left(a_i,b_i^2\right)^{c_i}. \tag{61}$$

The natural logarithm of Z yields the expression

$$\ln Z=\sum_{i=1}^{M}c_i\ln\Lambda\left(a_i,b_i^2\right)=\sum_{i=1}^{M}c_iN\left(a_i,b_i^2\right) \tag{62}$$

where, by definition, $\ln\Lambda$ is a normal RV. Since the superposition of independent normal RVs is a normal RV

$$\sum_{i=1}^{M}c_iN\left(a_i,b_i^2\right)=N\left(\sum_{i=1}^{M}c_ia_i,\sum_{i=1}^{M}c_i^2b_i^2\right), \tag{63}$$

it then follows that $\ln Z$ is a normal RV, and therefore Z itself is a lognormal RV

$$Z=\Lambda\left(\sum_{i=1}^{M}c_ia_i,\sum_{i=1}^{M}c_i^2b_i^2\right). \tag{64}$$

In summary, the compound multinomial distribution

$$P\left(\{m_t\}\middle|\{t\},n_0\right)\propto\int_0^1\cdots\int_0^1P\left(\{m_t\}\middle|\{z_t\},n_0\right)\prod_{t=1}^{T}p_{Z_t}\left(z_t\right)\mathrm{d}z_t \tag{65}$$

that generalizes probability function (56) describes a multivariate lognormal distribution of infectivity throughout the trial period.

4. Recapitulation and Constraints

Having derived the pdfs and (c)cdfs descibing COVID infectivity in Sections 2 and 3, it is well to examine briefly what has been achieved and what conditions pertain.

In Section 2 the focus was on the *period* of infectivity following onset of symptoms. *Time* is the random variable here for which a lognormal distribution was adopted empirically on the basis of an epidemiological analogy between the periods of incubation and infectivity. Data from one large medical study were used to deduce the parameters of the lognormal distribution, and data from other independent studies provided tests of the consistency of the derived distribution function.

Subsequently, after discussion of the PME in Section 3, it was conjectured that under conditions in which the initial information included only the mean and variance of a non-negative RV whose pdf vanished continuously at the origin, the PME solution would be, or would closely approximate, a lognormal distribution. To the author's knowledge, this conjecture has never been proven or disproven. However, the investigation in Section 2 raises two questions: (1) is the lognormal the only distribution that might describe the statistics of the period of infectivity? (2) If there are other such distributions, does this non-uniqueness

matter practically? The short answers to the questions are (1) "No" and (2) "Maybe not". Here is the reasoning.

It has been stressed that the distribution to which the PME leads in any given problem depends on the kind of initial information. Consider, for example the lognormal and gamma distributions. The PME leads to the first with pdf (1) if the initial information comprises the mean and variance of $\ln X$ where X is a non-negative RV. However, the PME leads to the second, with pdf (of the same variable X)

$$p_X(x) = \frac{x^{k-1}\exp(-x/\theta)}{\theta^k \Gamma(k)}, \tag{66}$$

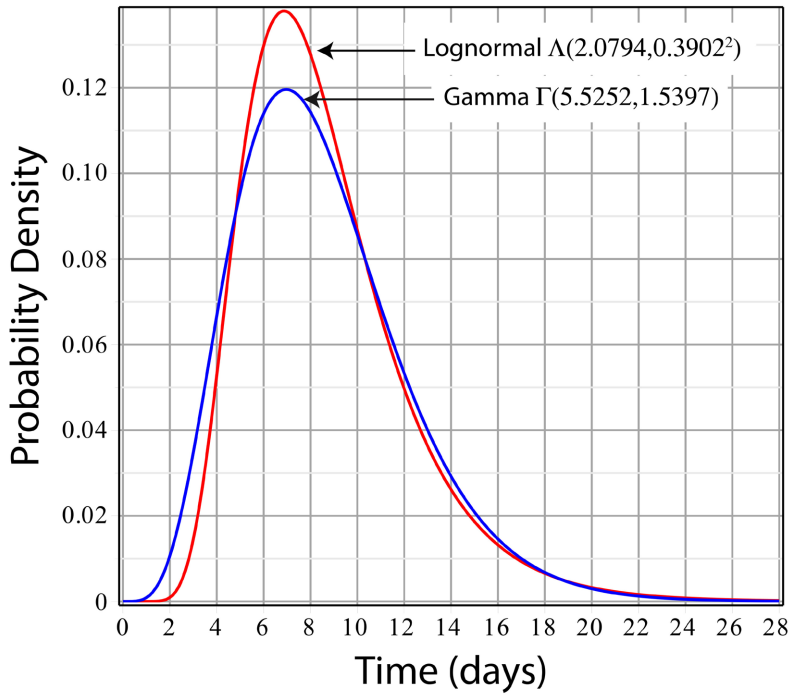
if the initial information comprises the mean of X and the mean of $\ln X$ [44]. In pdf (66) the shape parameter k and scale parameter θ are greater than 0. A plot of the gamma pdf (66) can lead to a profile similar to that of lognormal pdf (1) if the parameters are just right.

In determining the parameters of the lognormal distribution in Section 2, the author made use of available information consisting of two quantiles of the period of infectivity and *not* the constraints required for a PME solution to be either lognormal or gamma. However, the gamma parameters

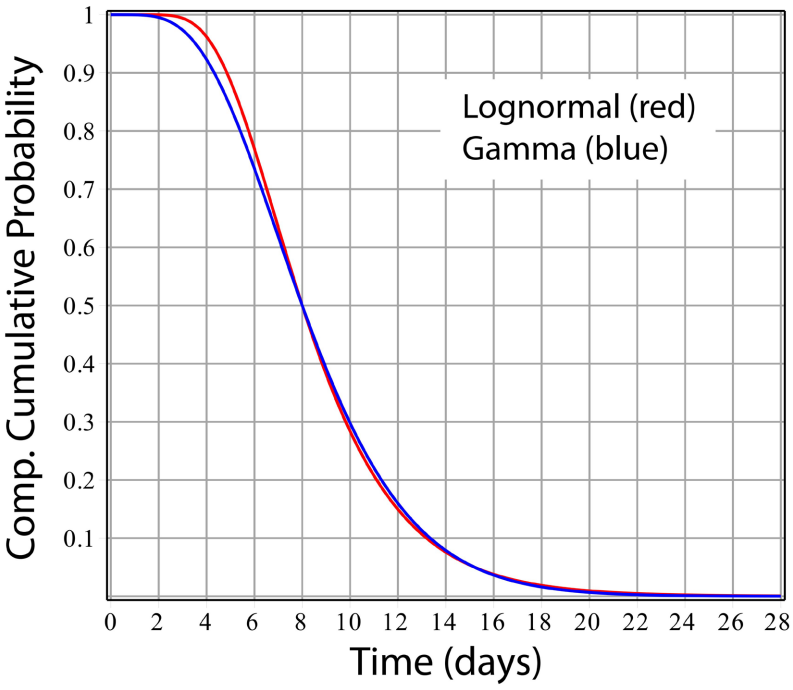
$$\begin{aligned} k_0 &= 5.5252 \\ \theta_0 &= 1.5397 \end{aligned} \tag{67}$$

determined from the same (50%, 95%) quantiles as the lognormal parameters (16) and (17) in Section 2, results in a pdf profile in **Figure 7(a)** and ccdf profile in **Figure 7(b)** that closely match the corresponding lognormal profiles. As seen especially in **Figure 7(b)**, both distributions appear to provide statistically equivalent cumulative probabilities of viral shedding as a function of time. Moreover, the entropies of the two distributions (the calculation of which lies outside the scope of this paper) are close. Up to this point, therefore, a lognormal or gamma ccdf seems to account satisfactorily for the period of COVID infectivity, and more data would be needed from more clinical studies to distinguish which is better.

In Section 3 the focus was on the *mean numbers* of positive samples $\{n_t\}$ in a proposed clinical investigation comprising a finite number of grouped participants over a limited period of time T. These are precisely the constraints for which the PME leads exactly to a multinomial distribution of ratios $\{z_t = n_t/n_0\}$, interpretable as probabilities of infectivity and shown by means of the CLT and some mathematical analysis to closely approximate lognormal random variables. It then follows that the fractions (or quantiles) of virus-shedding participants throughout the period T comprise a multivariate lognormal distribution. Although the random variable in Section 3 was not time, a reasonably complete set of the numbers $\{n_t\}$ would yield quantiles for the period of infectivity and thereby test whether the cdf (or ccdf) is consistent with a lognormal distribution, as predicted.



(a)



(b)

Figure 7. (a) Comparison of lognormal probability density (red) and gamma probability density (blue) of the period of COVID infectivity. The two sets of distribution parameters (m_0, s_0) and (k_0, θ_0) were both determined from the same two quantiles (50%, 95%) upon which **Figure 1** and **Figure 2** are based; (b) comparison of lognormal (red) and gamma (blue) complementary cumulative distribution functions determined from the parameters shown in (a).

Finally, as a matter of practicality, once it has been established by clinical studies of the kind proposed in Section 3 that the probability distribution of the period of infectivity is lognormal (or gamma or some other two-parameter distribution), then all one would need are two quantiles (or the mean and variance) of the period as utilized in Section 2, to specify the parameters of the distribution uniquely.

5. Conclusions

Although measures to prevent and treat coronavirus have progressed significantly since the beginning of the pandemic, the disease is still prevalent throughout the world with more than 20 million current cases at the time of writing [45]. While COVID-related mortality is down in populations where vaccines are available, the disease can nevertheless lead to a wide range of organ damage [46] and post-infection disabilities [47] [48] in surviving patients. Despite the continued seriousness of the pandemic, the general fatigue of dealing with the disease coupled with a strong desire to return to normality has resulted in the loosening of public health measures nearly everywhere and at all levels of government from municipal to national. For example, the U.S. government has decided to end public health emergency measures for COVID in May 2023 [49], although the baseline of daily COVID hospitalizations has been over 25,000 [50], the daily average number of cases in the U.S. currently exceeds 30,000 [51], and many Americans have had multiple infections and are experiencing post-acute sequelae of SARS-CoV-2 infection.

Under such prevailing circumstances, where protective measures by public health agencies are either being phased out or else ignored by a disaffected public [52], individuals concerned for their health are left to their own means of protecting themselves and those whom they care about. Apart from being vaccinated, there are basically two actions over which an individual has control. The first is to try to diminish the risk of infection by wearing an appropriate face mask (e.g. N95, KN95) whenever exposed to groups of people in close proximity. The second is to diminish the risk of spreading infection once one has shown symptoms of COVID. The latter action requires self-isolation for an appropriate time period, the determination of which is the focal point of this paper.

It has been the objective of this paper to derive probability functions, using the sparse data from recent investigations, that enable medical practitioners, epidemiologists, and members of the public to estimate the risk of infectivity as a function of time following the onset of symptoms (or first positive antigen test). The analysis reported here, based on empirical consistency (Section 2) and on general statistical principles (Section 3) rather than on any specific dynamical model of pathogen transmission and growth, supports the propositions that the period of infectivity and the probability of infectivity follow lognormal distributions. As the coronavirus evolves new variants and new information becomes

available, the lognormal parameters of risk curves, like those in **Figures 1-3**, may need to be updated, but the mathematical form of the distribution is expected to remain unchanged.

As a final point worth commenting on, the quantitative measure of a probability is not in general something that either a lay person or even a medical professional may know how to think about. For example, what is one to make of the prediction from **Figure 2** that there is a 10% chance of being infectious after about 13 days of isolation? Is 10% a low or a significant probability? Actually, it can be both. From the perspective of the author, a nuclear physicist whose work entails consideration of the consequences of adverse events, a prudent policy is to examine probability *in appropriate context*.

What matters is not just the probability, but the potential consequence—*i.e.* the product of a probability and a measure (or personal sense) of the associated loss. This product is not likely to be the same for everyone or for different circumstances. For example, a carpenter might take a 25% chance of rain as a threshold of whether to engage in outdoor construction that day. Accordingly, a 10% chance of rain is a low probability. But COVID is not rain; it is, depending on demographics, age [53], health, and immunization status, a potentially lethal or debilitating disease to have or to spread. And a 10% chance of transmitting it to family, friends, and community may well be a risk not to be taken. (After all, if it were the case that 1 out of 10 commercial flights developed engine trouble after takeoff, would you travel by air?) From a practical, epidemiological perspective, the socially responsible action for an infected person to take is to remain in self-isolation until receiving 2 negative antigen tests 48 hours apart [54] irrespective of the number of days (beyond 5) following onset of symptoms.

Acknowledgements

Partial support of this project was provided by Trinity College through the research fund of the G. A. Jarvis Chair of Physics.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Berstein, A. (2021) Coronavirus: How Long Is a Person Contagious? Medical News Today. https://www.medicalnewstoday.com/articles/how-long-is-a-person-contagious-with-coronavirus

[2] Sun, L.H. and Achenbach, J. (2022) When You Have Covid, Here's How You Know You Are No Longer Contagious. *Washington Post*. https://www.washingtonpost.com/health/2022/08/01/covid-contagious-period-isolation

[3] Gillespie, C. (2022) How Long after Having COVID-19 Are You Contagious? Health. https://www.health.com/condition/infectious-diseases/coronavirus/how-long-after-coronavirus-are-you-contagious

[4] Ries, J. (2022) How Long before Someone with COVID-19 Isn't Contagious? Healthline. https://www.healthline.com/health-news/how-long-before-someone-with-covid-19-isnt-contagious

[5] Marshall, M. (2022) How Long Is Someone with COVID Contagious? CBS News Boston. https://www.cbsnews.com/boston/news/how-long-is-someone-with-covid-contagious

[6] CDC (2021) Scientific Brief: SARS-CoV-2 Transmission. https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/sars-cov-2-transmission.html#anchor_1619805184733

[7] CDC (2022) Ending Isolation and Precautions for People with COVID-19: Interim Guidance. https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html

[8] Johnston, C., Hughes, H., Lingard, S., Hailey, S. and Healy, B. (2022) Immunity and Infectivity in Covid-19. *The BMJ*, **378**, e061402. https://doi.org/10.1136/bmj-2020-061402

[9] Tsao, J., Kussman, A. and Segovia, N.A. (2022) Prevalence of Positive Rapid Antigen Tests after 7-Day Isolation Following SARS-Cov-2 Infection in College Athletes during Omicron Variant Predominance. *JAMA Network Open*, **5**, e2237149. https://doi.org/10.1001/jamanetworkopen.2022.37149

[10] Wu, Y., Kang, L. and Guo, Z. (2022) Incubation Period of COVID-19 Caused by Unique SARS-Cov-2 Strains: A Systematic Review and Meta-Analysis. *JAMA Network Open*, **5**, e2228008. https://doi.org/10.1001/jamanetworkopen.2022.28008

[11] Calabuig, J.M., Garcia-Raffi, L.M., Garcia-Valiente, A. and Sanchez-Perez, E.A. (2021) Kaplan-Meier Type Survival Curves for COVID-19: A Health Data Based Decision-Making Tool. *Frontiers in Public Health*, **9**, Article ID: 646863. https://doi.org/10.3389/fpubh.2021.646863

[12] Schlager, T. and Whillans, A.V. (2022) People Underestimate the Probability of Contracting the Coronavirus from Friends. *Humanities and Social Sciences Communications*, **9**, 59. https://doi.org/10.1057/s41599-022-01052-4

[13] Silverman, M.P. (2022) Exact Statistical Distribution and Correlation of Human Height and Weight: Analysis and Experimental Confirmation. *Open Journal of Statistics*, **12**, 743-787. https://doi.org/10.4236/ojs.2022.125044

[14] Silverman, M.P. and Lipscombe, T.C. (2022) Exact Statistical Distribution of the Body Mass Index (BMI): Analysis and Experimental Confirmation. *Open Journal of Statistics*, **12**, 324-356. https://doi.org/10.4236/ojs.2022.123022

[15] Limpert, E., Stahel, W.A. and Abbt, M. (2001) Log-Normal Distributions across the Sciences: Keys and Clues, *BioScience*, **51**, 341-352. https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

[16] O'Connor, P. and Kleyner, A. (2011) Practical Reliability Engineering. John Wiley & Sons, New York, 33-41. https://doi.org/10.1002/9781119961260

[17] Sartwell, P.E. (1950) The Distribution of Incubation Periods of Infectious Disease. *The American Journal of Hygiene*, **51**, 310-318. https://doi.org/10.1093/oxfordjournals.aje.a119397

[18] Lee, H., Kim, K., Choi, K., Hong, S., Son, H. and Ryu, S. (2020) Incubation Period of the Coronavirus Disease 2019 (COVID-19) in Busan, South Korea. *Journal of Infection and Chemotherapy*, **26**, 1011-1013. https://doi.org/10.1016/j.jiac.2020.06.018

[19] Karlin, S. (1966) A First Course in Stochastic Processes. Academic, New York, 189-201.

[20] Silverman, M.P. (2014) A Certain Uncertainty: Nature's Random Ways. Cambridge University Press, Cambridge, 495-506. https://doi.org/10.1017/CBO9781139507370

[21] Altman, D.G. (1999) Practical Statistics for Medical Research. Chapman & Hall/CRC, London, 368-371.

[22] van Kampen, J.J.A., van de Vijver, D.A.M.C., Fraaij, P.L.A., *et al.* (2021) Duration and Key Determinants of Infectious Virus Shedding in Hospitalized Patients with Coronavirus Disease-2019. *Nature Communications*, **12**, Article No. 267. https://doi.org/10.1038/s41467-020-20568-4

[23] Davies, M, Bramwell, L.R., Jeffery, N., Bunce, B., *et al.* (2022) Persistence of Clinically Relevant Levels of SARS-CoV2 Envelope Gene Subgenomic RNAs in Non-Immuno-Compromised Individuals. *International Journal of Infectious Diseases*, **116**, 418-425. https://doi.org/10.1016/j.ijid.2021.12.312

[24] Hafter, N. (2021) What's the Difference between a PCR and Antigen COVID-19 Test? UMass Chan Medical School, Worcester. https://www.umassmed.edu/news/news-archives/2021/11/whats-the-difference-between-a-pcr-and-antigen-covid-19-test

[25] Hakki, S., Zhou, J., Jonnerby, J., Singanayagam, A., *et al.* (2022) Onset and Window of SARS-Cov-2 Infectiousness and Temporal Correlation with Symptom Onset: A Prospective, Longitudinal, Community Cohort Study. *The Lancet Respiratory Medicine*, **10**, 1061-1073. https://doi.org/10.1016/S2213-2600(22)00226-0

[26] Imperial College London (2022, August 18) Real-World Study Details Average Duration of Infectiousness for COVID-19. ScienceDaily. https://www.sciencedaily.com/releases/2022/08/220818190625.htm

[27] Ke, R., Martinez, P.P., Smith, R.L., Gibson, L.L., *et al.* (2022) Daily Longitudinal Sampling of SARS-Cov-2 Infection Reveals Substantial Heterogeneity in Infectiousness. *Nature Microbiology*, **7**, 640-652. https://doi.org/10.1038/s41564-022-01105-z

[28] Rosenkrantz, R.D. (1989) Where Do We Stand on Maximum Entropy (1978). In: Rosenkrantz, R.D., Ed., *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, Springer, Berlin, 210-314. https://doi.org/10.1007/978-94-009-6581-2_10

[29] Jaynes, E.T. (1957) Information Theory and Statistical Mechanics. *Physical Review*, **106**, 620-630. https://doi.org/10.1103/PhysRev.106.620

[30] Jaynes, E.T. (1957) Information Theory and Statistical Mechanics, II. *Physical Review*, **108**, 171-190. https://doi.org/10.1103/PhysRev.108.171

[31] Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[32] Silverman, M.P. (2015) Cheating or Coincidence? Statistical Method Employing the Principle of Maximum Entropy for Judging Whether a Student Has Committed Plagiarism. *Open Journal of Statistics*, **5**, 143-157. https://doi.org/10.4236/ojs.2015.52018

[33] Sivia, D.S. and Skilling, J. (2006) Data Analysis: A Bayesian Tutorial. Oxford University Press, Oxford, 120-124.

[34] Silverman, M.P. (2012) Unpublished Notes on Thermal and Statistical Physics.

[35] Wikipedia Contributors (2023) Truncated Normal Distribution.

https://en.wikipedia.org/w/index.php?title=Truncated_normal_distribution&oldid=1142956876

[36] Taavitsainen, A. and Vanhanen, R. (2017) On the Maximum Entropy Distributions of Inherently Positive Nuclear Data. *Nuclear Instruments and Methods in Physics Research A*, **854**, 156-162. https://doi.org/10.1016/j.nima.2016.11.061

[37] Chou, Y. (1969) Statistical Analysis: With Business and Economic Applications. Holt, Rinehart, and Winston, New York, 242-244.

[38] Hogg, R.V., McKean, J.W. and Craig, A.T. (2005) Introduction to Mathematical Statistics. Pearson/Prentice Hall, Upper Saddle River, 220-225.

[39] Forbes, C., Evans, M., Hastings, N. and Peacock, B. (2011) Statistical Distributions. 4th Edition, Wiley, New York, 152-156. https://doi.org/10.1002/9780470627242

[40] Arfken, G.B. and Weber, H.J. (2005) Mathematical Methods for Physicists. 6th Edition, Elsevier, New York, 83-85, 669-670, 975.

[41] Silverman, M.P., Strange, W. and Lipscombe T.C. (2004) The Distribution of Composite Measurements: How to Be Certain of the Uncertainties in What We Measure. *American Journal of Physics*, **72**, 1068-1081. https://doi.org/10.1119/1.1738426

[42] Mood, A.M., Graybill, F.A. and Boes, D.C. (1963) Introduction to the Theory of Statistics. McGraw-Hill, New York, 181.

[43] Taylor, J.R. (1997) An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books, Sausalito, 60-62, 73-77, 146-148.

[44] Wikipedia Contributors (2023, March 19) Maximum Entropy Probability Distribution. In Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Maximum_entropy_probability_distribution&oldid=1145413438

[45] COVID-19 Coronavirus Pandemic (2023) Worldometer. https://www.worldometers.info/coronavirus

[46] Jain, U. (2020) Effect of COVID-19 on the Organs. *Cureus*, **12**, e9540. https://doi.org/10.7759/cureus.9540

[47] Mayo Clinic, COVID-19: Long-Term Effects (2023). https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/coronavirus-long-term-effects/art-20490351

[48] Sutherland, S. (2023) The Brain and Long Covid. *Scientific American*, **328**, 26-33.

[49] LaFraniere, S. and Welland, N. (2023) U.S. Plans to End Public Health Emergency for Covid in May. *New York Times*. https://www.nytimes.com/2023/01/30/us/politics/biden-covid-public-health-emergency.html

[50] Topol, E. (2023) Ground Truths (March 7): A Break from Covid Waves and a Breakthrough for Preventing Long Covid. https://erictopol.substack.com/p/a-break-from-covid-waves-and-a-breakthrough

[51] Allen, J., Almukhtar, S., Aufrichtig, A., Barnard, A., *et al.* (2023) Coronavirus in the U.S.: Latest Map and Case Count. *New York Times*. https://www.nytimes.com/interactive/2021/us/covid-cases.html

[52] Weber, L. and Achenbach, J. (2023) Covid Backlash Hobbles Public Health and Future Pandemic Response. *The Washington Post*. https://www.washingtonpost.com/health/2023/03/08/covid-public-health-backlash

[53] Harris, E., (2023) Most COVID-19 Deaths Worldwide Were among Older People. *JAMA*, **329**, 704. https://jamanetwork.com/journals/jama/fullarticle/2801723

https://doi.org/10.1001/jama.2023.1554

[54] CDC (2022), Isolation and Precautions for People with COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/your-health/isolation.html

Appendix—Glossary of Abbreviations

BMI	Body Mass Index
BMJ	British Medical Journal
CDC	Centers for Disease Control and Prevention
cdf	Cumulative Distribution Function
ccdf	Complementary Cumulative Distribution Function
cf	Characteristic Function
CLT	Central Limit Theorem
COVID	Corona Virus Disease
EPT	Error Propagation Theory
JAMA	Journal of the American Medical Society
MC	Medical Center
MS	Medical School
mgf	Moment Generating Function
PCR	Polymerase Chain Reaction
pdf	Probability Density Function
PME	Principle of Maximum Entropy
RV	Random Variable
SARS-Cov-2	Severe Acute Respiratory Syndrome Coronavirus 2
WHO	World Health Organization

Appendix—Glossary of Mathematical Symbols

Upper-case letters represent random variables (RV)	X, Y, Z, etc.
Lower-case letters represent variates (realizations)	x, y, z, etc.
Probability density function of continuous RV X	$p_X(x)$
Cumulative distribution function (cdf) of RV X	$F_X(x)$
Complementary cdf of RV X	$\tilde{F}_X(x)$
Symbolic notation of normal distribution of mean m and variance s^2	$N(m,s^2)$
Standard normal distribution	$N(0,1)$
Symbolic notation of lognormal distribution X for which $\ln X = N(m,s^2)$	$\Lambda(m,s^2)$
Symbolic notation of gamma distribution of shape parameter k and scale parameter θ	$\Gamma(k,\theta)$
Expectation value of a random variable X	$\langle X\rangle$
Statistical moment of order k	M_k
Mean of a random variable X	μ_X
Variance of a random variable X	σ_X^2
Skewness of a random variable X	Sk_X
Probability of infectivity at day t	p_t
Mean number of positive tests at day t	n_t
Binomial probability function of m positive tests, given mean number n_t in a cohort of size n_0	$P(m\|n_0,n_t)$

Multinomial probability function of set of positive tests $\{m_t\}$ given associated set of probabilities $\{p_t\}$ in a cohort of size n_0	$P\left(\{m_t\}\middle\vert n_0,\{p_t\}\right)$
Shannon entropy, given set of probabilities $\{p_t\}$	$H\left(p_1,\cdots,p_T\right)$
Dirac delta function (centered on x_0)	$\delta\left(x-x_0\right)$