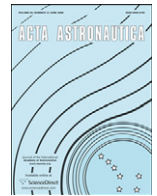




ELSEVIER

Contents lists available at ScienceDirect

Acta Astronautica

journal homepage: www.elsevier.com/locate/actaastro

The Statistical Drake Equation

Claudio Maccone*

Technical Director of the International Academy of Astronautics (IAA) and Co-Chair, SETI Permanent Study Group of the IAA

ARTICLE INFO

Article history:

Received 22 March 2010

Accepted 3 May 2010

Keywords:

Drake Equation

Statistics

SETI

ABSTRACT

We provide the statistical generalization of the Drake equation.

From a simple product of seven positive numbers, the Drake equation is now turned into the product of seven positive random variables. We call this “the Statistical Drake Equation”. The mathematical consequences of this transformation are then derived. The proof of our results is based on the Central Limit Theorem (CLT) of Statistics. In loose terms, the CLT states that the sum of any number of independent random variables, each of which may be ARBITRARILY distributed, approaches a Gaussian (i.e. normal) random variable. This is called the Lyapunov Form of the CLT, or the Lindeberg Form of the CLT, depending on the mathematical constraints assumed on the third moments of the various probability distributions. In conclusion, we show that:

- (1) The new random variable N , yielding the number of communicating civilizations in the Galaxy, follows the LOGNORMAL distribution. Then, as a consequence, the mean value of this lognormal distribution is the ordinary N in the Drake equation. The standard deviation, mode, and all the moments of this lognormal N are also found.
- (2) The seven factors in the ordinary Drake equation now become seven positive random variables. The probability distribution of each random variable may be ARBITRARY. The CLT in the so-called Lyapunov or Lindeberg forms (that both do not assume the factors to be identically distributed) allows for that. In other words, the CLT “translates” into our statistical Drake equation by allowing an arbitrary probability distribution for each factor. This is both physically realistic and practically very useful, of course.
- (3) An application of our statistical Drake equation then follows. The (average) DISTANCE between any two neighboring and communicating civilizations in the Galaxy may be shown to be inversely proportional to the cubic root of N . Then, in our approach, this distance becomes a new random variable. We derive the relevant probability density function, apparently previously unknown and dubbed “Maccone distribution” by Paul Davies.
- (4) DATA ENRICHMENT PRINCIPLE. It should be noticed that ANY positive number of random variables in the Statistical Drake Equation is compatible with the CLT. So, our generalization allows for many more factors to be added in the future as long as more refined scientific knowledge about each factor will be known to the scientists. This capability to make room for more future factors in the statistical Drake equation, we call the “Data Enrichment Principle,” and we regard it as the key to more profound future results in the fields of Astrobiology and SETI.

Finally, a practical example is given of how our statistical Drake equation works numerically. We work out in detail the case, where each of the seven random variables is uniformly distributed around its own mean value and has a given standard deviation.

* Mailing address: Via Martorelli 43, 10155 Torino (Turin), Italy.

E-mail address: maccon@libero.it

URL: <http://www.maccone.com/>

For instance, the number of stars in the Galaxy is assumed to be uniformly distributed around (say) 350 billions with a standard deviation of (say) 1 billion. Then, the resulting lognormal distribution of N is computed numerically by virtue of a MathCad file that the author has written. This shows that the mean value of the lognormal random variable N is actually of the same order as the classical N given by the ordinary Drake equation, as one might expect from a good statistical generalization.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

The Drake equation is now a famous result (see Ref. [1] for the Wikipedia summary) in the fields of the Search for ExtraTerrestrial Intelligence (SETI, see Ref. [2]) and Astrobiology (see Ref. [3]). Devised in 1961, the Drake equation was the first scientific attempt to estimate the number N of ExtraTerrestrial civilizations in the Galaxy, with which we might come in contact. Frank D. Drake (see Ref. [4]) proposed it as the product of seven factors:

$$N = N_s \cdot fp \cdot ne \cdot fl \cdot fi \cdot fc \cdot fl \quad (1)$$

where

- (1) N_s is the estimated number of stars in our Galaxy.
- (2) fp is the fraction (=percentage) of such stars that have planets.
- (3) ne is the number “Earth-type” such planets around the given star; in other words, ne is number of planets, in a given stellar system, on which the chemical conditions exist for life to begin its course: they are “ready for life”.
- (4) fl is fraction (=percentage) of such “ready for life” planets on which life actually starts and grows up (but not yet to the “intelligence” level).
- (5) fi is the fraction (=percentage) of such “planets with life forms” that actually evolve until some form of “intelligent civilization” emerges (like the first, historic human civilizations on Earth).
- (6) fc is the fraction (=percentage) of such “planets with civilizations”, where the civilizations evolve to the point of being able to communicate across the interstellar distances with other (at least) similarly evolved civilizations. As far as we know in 2008, this means that they must be aware of the Maxwell equations governing radio waves, as well as of computers and radioastronomy (at least).
- (7) fl is the fraction of galactic civilizations alive at the time when we, poor humans, attempt to pick up their radio signals (that they throw out into space just as we have done since 1900, when Marconi started the transatlantic transmissions). In other words, fl is the number of civilizations now transmitting and receiving, and this implies an estimate of “how long will a technological civilization live?” that nobody can make at the moment. Also, are they going to destroy themselves in a nuclear war, and thus live only a few decades of technological civilization? Or are they slowly becoming wiser, reject war, speak a single language (like English today), and merge into a single “nation”, thus living in peace for ages? Or will robots

take over one day making “flesh animals” disappear forever (the so-called “post-biological universe”)?

No one knows...

But let us go back to the Drake Eq. (1).

In the fifty years of its existence, a number of suggestions have been put forward about the different numeric values of its seven factors. Of course, every different set of these seven input numbers yields a different value for N , and we can endlessly play that way. But we claim that these are like... children plays!

We claim the classical Drake Eq. (1), as we shall call it from now on to distinguish it from our statistical Drake equation to be introduced in the coming sections, well, the classical Drake equation is scientifically inadequate in one regard at least: it just handles sheer numbers and does not associate an error bar to each of its seven factors. **At the very least, we want to associate an error bar to each D_i .**

Well, we have thus reached STEP ONE in our improvement of the classical Drake equation: replace each sheer number by a **probability distribution!**

The reader is now asked to look at the flow chart in the next page as a guide to this paper, please.

2. Step 1: Letting each factor become a random variable

In this paper, we adopt the notations of the great book “Probability, Random Variables and Stochastic Processes” by Athanasios Papoulis (1921–2002), now re-published as Papoulis-Pillai, Ref. [5]. The advantage of this notation is that it makes a neat distinction between probabilistic (or statistical: it is the same thing here) variables, always denoted by **capitals**, from non-probabilistic (or “deterministic”) variables, always denoted by lower-case letters. Adopting the Papoulis notation also is a tribute to him by this author, who was a Fulbright Grantee in the United States with him at the Polytechnic Institute (now Polytechnic University) of New York in the years 1977–79.

We thus introduce seven new (positive) random variables D_i (“D” from “Drake”) defined as

$$\begin{cases} D_1 = N_s \\ D_2 = fp \\ D_3 = ne \\ D_4 = fl \\ D_5 = fi \\ D_6 = fc \\ D_7 = fl \end{cases} \quad (2)$$

so that our **STATISTICAL Drake equation** may be simply rewritten as

$$N = \prod_{i=1}^7 D_i. \tag{3}$$

Of course, N now becomes a (positive) random variable too, having its own (positive) mean value and standard deviation. Just as each of the D_i has its own (positive) mean value and standard deviation...

... the natural question then arises: how are the seven mean values on the right related to the mean value on the left?

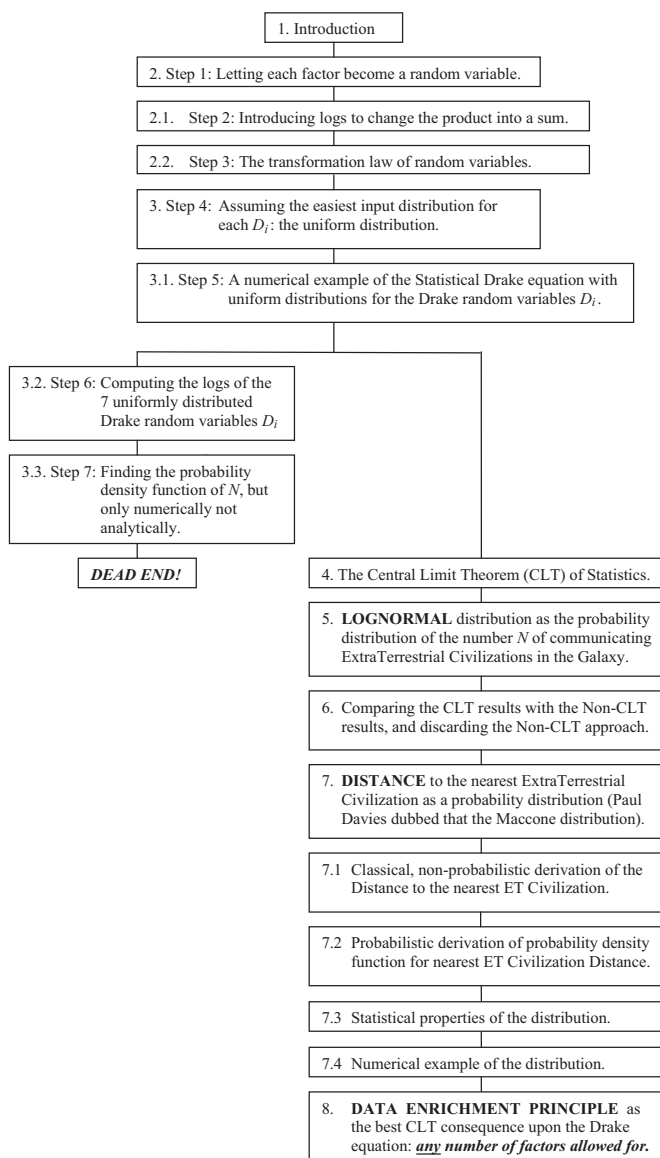
... and how are the seven standard deviations on the right related to the standard deviation on the left?

Just take the next step, STEP TWO.

2.1. Step 2: Introducing logs to change the product into a sum

Products of random variables are not easy to handle in probability theory. It is actually much easier to handle sums of random variables, rather than products, because:

- (1) The probability density of the sum of two or more independent random variables is the convolution of the relevant probability densities (worry not about the equations, right now).
- (2) The Fourier transform of the convolution simply is the product of the Fourier transforms (again, worry not about the equations, at this point).



So, let us take the natural logs of both sides of the Statistical Drake Eq. (3) and change it into a sum:

$$\ln(N) = \ln\left(\prod_{i=1}^7 D_i\right) = \sum_{i=1}^7 \ln(D_i). \quad (4)$$

It is now convenient to introduce eight new (positive) random variables defined as follows:

$$\begin{cases} Y = \ln(N) \\ Y_i = \ln(D_i) \quad i = 1, \dots, 7. \end{cases} \quad (5)$$

Upon inversion, the first equation of Eq. (5) yields the important equation, that will be used in the sequel

$$N = e^Y. \quad (6)$$

We are now ready to take STEP THREE.

2.2. Step 3: The transformation law of random variables

So far we did not mention at all the problem: “which probability distribution shall we attach to each of the seven (positive) random variables D_i ?”

It is not easy to answer this question because we do not have the least scientific clue to what probability distributions fit at best to each of the seven points listed in Section 1.

Yet, at least one trivial error must be avoided: claiming that each of those seven random variables must have a Gaussian (i.e. normal) distribution. In fact, the Gaussian distribution, having the well-known bell-shaped probability density function

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (\sigma \geq 0) \quad (7)$$

has its independent variable x ranging between $-\infty$ and ∞ and so it can apply to a **real** random variable X only, and never to **positive** random variables like those in the statistical Drake Eq. (3). Period.

Searching again for probability density functions that represent positive random variables, an obvious choice would be the gamma distributions (see, for instance, Ref. [6]). However, we discarded this choice too because of a different reason: please keep in mind that, according to Eq. (5), once we selected a particular type of probability density function (pdf) for the last seven of Eq. (5), then we must compute the (new and different) pdf of the logs of such random variables. And the pdf of these logs certainly is not gamma-type any more.

It is high time now to remind the reader of a certain theorem that is proved in probability courses, but, unfortunately, does not seem to have a specific name. It is the **transformation law** (so we shall call it, see, for instance, Ref. [5]) allowing us to compute the pdf of a certain new random variable Y that is a known function $Y=g(X)$ of an another random variable X having a known pdf. In other words, if the pdf $f_X(x)$ of a certain random variable X is known, then the pdf $f_Y(y)$ of the new random variable Y , related to X by the functional relationship

$$Y = g(X) \quad (8)$$

can be calculated according to this rule:

- (1) First, invert the corresponding non-probabilistic equation $y=g(x)$ and denote by $x_i(y)$ the various real roots resulting from this inversion.
- (2) Second, take notice whether these real roots may be either finitely- or infinitely many, according to the nature of the function $y=g(x)$.
- (3) Third, the probability density function of Y is then given by the (finite or infinite) sum

$$f_Y(y) = \sum_i \frac{f_X(x_i(y))}{|g'(x_i(y))|} \quad (9)$$

where the summation extends to all roots $x_i(y)$ and $|g'(x_i(y))|$ is the absolute value of the first derivative of $g(x)$, where the i -th root $x_i(y)$ has been replaced instead of x .

Since we must use this transformation law to transfer from the D_i to the $Y_i = \ln(D_i)$, it is clear that we need to start from a D_i pdf that is as simple as possible. The gamma pdf is not responding to this need because the analytic expression of the transformed pdf is very complicated (or, at least, it looked so to this author in the first instance). Also, the gamma distribution has two free parameters in it, and this “complicates” its application to the various meanings of the Drake equation. In conclusion, we discarded the gamma distributions and confined ourselves to the simpler uniform distribution instead, as shown in the next section.

3. Step 4: Assuming the easiest input distribution for each D_i : the uniform distribution

Let us now suppose that each of the seven D_i is distributed UNIFORMLY in the interval ranging from the lower limit $a_i \geq 0$ to the upper limit $b_i \geq a_i$.

This is the same as saying that the probability density function of each of the seven Drake random variables D_i has the equation

$$f_{\text{uniform}_D_i}(x) = \frac{1}{b_i - a_i} \quad \text{with} \quad 0 \leq a_i \leq x \leq b_i \quad (10)$$

as it follows at once from the normalization condition

$$\int_{a_i}^{b_i} f_{\text{uniform}_D_i}(x) dx = 1. \quad (11)$$

Let us now consider the mean value of such uniform D_i defined by

$$\begin{aligned} \langle \text{uniform}_D_i \rangle &= \int_{a_i}^{b_i} x f_{\text{uniform}_D_i}(x) dx = \frac{1}{b_i - a_i} \int_{a_i}^{b_i} x dx \\ &= \frac{1}{b_i - a_i} \left[\frac{x^2}{2} \right]_{a_i}^{b_i} = \frac{b_i^2 - a_i^2}{2(b_i - a_i)} = \frac{a_i + b_i}{2}. \end{aligned}$$

By words (as it is intuitively obvious): the **mean value of the uniform distribution** simply is the mean of the lower plus upper limit of the variable range

$$\langle \text{uniform}_D_i \rangle = \frac{a_i + b_i}{2}. \quad (12)$$

In order to find the variance of the uniform distribution, we first need finding the second moment

$$\begin{aligned} \langle \text{uniform_}D_i^2 \rangle &= \int_{a_i}^{b_i} x^2 f_{\text{uniform_}D_i}(x) dx \\ &= \frac{1}{b_i - a_i} \int_{a_i}^{b_i} x^2 dx = \frac{1}{b_i - a_i} \left[\frac{x^3}{3} \right]_{a_i}^{b_i} = \frac{b_i^3 - a_i^3}{3(b_i - a_i)} \\ &= \frac{(b_i - a_i)(a_i^2 + a_i b_i + b_i^2)}{3(b_i - a_i)} = \frac{a_i^2 + a_i b_i + b_i^2}{3}. \end{aligned}$$

The second moment of the uniform distribution is thus

$$\langle \text{uniform_}D_i^2 \rangle = \frac{a_i^2 + a_i b_i + b_i^2}{3}. \quad (13)$$

From Eqs. (12) and (13), we may now derive the variance of the uniform distribution

$$\begin{aligned} \sigma_{\text{uniform_}D_i}^2 &= \langle \text{uniform_}D_i^2 \rangle - \langle \text{uniform_}D_i \rangle^2 \\ &= \frac{a_i^2 + a_i b_i + b_i^2}{3} - \frac{(a_i + b_i)^2}{4} = \frac{(b_i - a_i)^2}{12}. \end{aligned} \quad (14)$$

Upon taking the square root of both sides of Eq. (14), we finally obtain the **standard deviation of the uniform distribution**:

$$\sigma_{\text{uniform_}D_i} = \frac{b_i - a_i}{2\sqrt{3}}. \quad (15)$$

We now wish to perform a calculation that is mathematically trivial, but rather unexpected from the intuitive point of view, and very important for our applications to the statistical Drake equation. Just consider the two simultaneous Eqs. (12) and (15)

$$\begin{cases} \langle \text{uniform_}D_i \rangle = \frac{a_i + b_i}{2} \\ \sigma_{\text{uniform_}D_i} = \frac{b_i - a_i}{2\sqrt{3}}. \end{cases} \quad (16)$$

Upon inverting this trivial linear system, one finds

$$\begin{cases} a_i = \langle \text{uniform_}D_i \rangle - \sqrt{3} \sigma_{\text{uniform_}D_i} \\ b_i = \langle \text{uniform_}D_i \rangle + \sqrt{3} \sigma_{\text{uniform_}D_i}. \end{cases} \quad (17)$$

This is of paramount importance for our application the Statistical Drake equation in as much as it shows that:

if one (scientifically) assigns the mean value and standard deviation of a certain Drake random variable D_i , then the lower and upper limits of the relevant uniform distribution are given by the two Eqs. (17), respectively.

In other words, there is a factor of $\sqrt{3} = 1.732$ included in the two Eqs. (17) that is not obvious at all to human intuition, and must indeed be taken into account.

The application of this result to the Statistical Drake equation is discussed in the next section.

3.1. Step 5: A numerical example of the statistical Drake equation with uniform distributions for the Drake random variables D_i

The first variable N_s in the classical Drake Eq. (1) is the number of stars in our Galaxy. Nobody knows how many they are exactly (!). Only **statistical** estimates can be made by astronomers, and they oscillate (say) around a mean

value of 350 billions (if this value is indeed correct!). This being the situation, we assume that our uniformly distributed random variable N_s has a mean value of 350 billions minus or plus a standard deviation of (say) one billion (we do not care whether this number is scientifically the best estimate as of August 2008: we just want to set up a numerical example of our Statistical Drake equation). In other words, we now assume that one has:

$$\begin{cases} \langle \text{uniform_}D_1 \rangle = 350 \times 10^9 \\ \sigma_{\text{uniform_}D_1} = 1 \times 10^9. \end{cases} \quad (18)$$

Therefore, according to Eq. (17), the lower and upper limit of our uniform distribution for the random variable $N_s = D_1$ are, respectively

$$\begin{cases} a_{N_s} = \langle \text{uniform_}D_1 \rangle - \sqrt{3} \sigma_{\text{uniform_}D_1} = 348.3 \times 10^9 \\ b_{N_s} = \langle \text{uniform_}D_1 \rangle + \sqrt{3} \sigma_{\text{uniform_}D_1} = 351.7 \times 10^9. \end{cases} \quad (19)$$

Similarly, we proceed for all the other six random variables in the Statistical Drake Eq. (3).

For instance, we assume that the fraction of stars that have planets is 50%, i.e. 50/100, and this will be the mean value of the random variable $fp = D_2$. We also assume that the relevant standard deviation will be 10%, i.e. that $\sigma_{fp} = 10/100$. Therefore, the relevant lower and upper limits for the uniform distribution of $fp = D_2$ turn out to be

$$\begin{cases} a_{fp} = \langle \text{uniform_}D_2 \rangle - \sqrt{3} \sigma_{\text{uniform_}D_2} = 0.327 \\ b_{fp} = \langle \text{uniform_}D_2 \rangle + \sqrt{3} \sigma_{\text{uniform_}D_2} = 0.673. \end{cases} \quad (20)$$

The next Drake random variable is the number ne of “Earth-type” planets in a given star system. Taking example from the Solar System, since only the Earth is truly “Earth-type”, the mean value of ne is clearly 1, but the standard deviation is not zero if we assume that Mars also may be regarded as Earth-type. Since there are thus two Earth-type planets in the Solar System, we must assume a standard deviation of $1/\sqrt{3} = 0.577$ to compensate the $\sqrt{3}$ appearing in Eq. (17) in order to finally yield two “Earth-type” planets (Earth and Mars) for the upper limit of the random variable ne . In other words, we assume that

$$\begin{cases} a_{ne} = \langle \text{uniform_}D_3 \rangle - \sqrt{3} \sigma_{\text{uniform_}D_3} = 0 \\ b_{ne} = \langle \text{uniform_}D_3 \rangle + \sqrt{3} \sigma_{\text{uniform_}D_3} = 2. \end{cases} \quad (21)$$

The next four Drake random variables have even more “arbitrarily” assumed values that we simply assume for the sake of making up a numerical example of our Statistical Drake equation with uniform entry distributions. So, **we really make no assumption about the astronomy, or the biology, or the sociology of the Drake equation: we just care about its mathematics.**

All our assumed entries are given in Table 1.

Please notice that, had we assumed all the standard deviations to equal **zero** in Table 1, then our Statistical Drake Eq. (3) would have obviously reduced to the classical Drake Eq. (1), and the resulting number of civilizations in the Galaxy would have turned out to be 3500:

$$N = 3500. \quad (22)$$

Table 1

Input values (i.e. mean values and standard deviations) for the seven Drake uniform random variables D_i . The first column on the left lists the seven input sheer numbers that also become the mean values (middle column). Finally, the last column on the right lists the seven input standard deviations. The bottom line is the classical Drake Eq. (1).

| | | |
|------------------------------------------------------------------------|------------------|--------------------------------------|
| $N_s := 350 \cdot 10^9$ | $\mu N_s := N_s$ | $\sigma N_s := 1 \cdot 10^9$ |
| $f_p := \frac{50}{100}$ | $\mu f_p := f_p$ | $\sigma f_p := \frac{10}{100}$ |
| $n_e := 1$ | $\mu n_e := n_e$ | $\sigma n_e := \frac{1}{\sqrt{3}}$ |
| $f_l := \frac{50}{100}$ | $\mu f_l := f_l$ | $\sigma f_l := \frac{10}{100}$ |
| $f_i := \frac{20}{100}$ | $\mu f_i := f_i$ | $\sigma f_i := \frac{10}{100}$ |
| $f_c := \frac{20}{100}$ | $\mu f_c := f_c$ | $\sigma f_c := \frac{10}{100}$ |
| $f_L := \frac{1000}{10^{10}}$ | $\mu f_L := f_L$ | $\sigma f_L := \frac{1000}{10^{10}}$ |
| $N := N_s \cdot f_p \cdot n_e \cdot f_l \cdot f_i \cdot f_c \cdot f_L$ | | $N = 3500$ |

This is an important **deterministic** number that we will use in the sequel of this paper for comparison with our **statistical** results on the mean value of N , i.e. $\langle N \rangle$. This will be explained in Sections 3.3 and 5.

3.2. Step 6: Computing the logs of the seven uniform distributed Drake random variables D_i

Intuitively speaking, the natural log of a uniformly distributed random variable **may not** be an another uniformly distributed random variable! This is obvious from the trivial diagram of $y = \ln(x)$ shown in Fig. 1.

So, if we have a uniformly distributed random variable D_i with lower limit a_i and upper limit b_i , the random variable

$$Y_i = \ln(D_i) \quad i = 1, \dots, 7 \quad (23)$$

must have its range limited in between the lower limit $\ln(a_i)$ and the upper limit $\ln(b_i)$. In other words, these are the lower and upper limits of the relevant probability density function $f_{Y_i}(y)$. But what is the actual analytic expression of such a pdf? To find it, we must resort to the general transformation law for random variables, defined by Eq. (9). Here, we obviously have

$$y = g(x) = \ln(x). \quad (24)$$

That, upon inversion, yields the **single** root

$$x_1(y) = x(y) = e^y. \quad (25)$$

On the other hand, differentiating Eq. (24) one gets

$$g'(x) = \frac{1}{x} \quad \text{and} \quad g'(x_1(y)) = \frac{1}{x_1(y)} = \frac{1}{e^y} \quad (26)$$

where Eq. (25) was already used in the last step. By virtue of the uniform probability density function (10) and of

Eq. (26), the general transformation law (9) finally yields

$$f_Y(y) = \sum_i \frac{f_X(x_i(y))}{|g'(x_i(y))|} = \frac{1}{b_i - a_i} \frac{1}{\left| \frac{1}{e^y} \right|} = \frac{e^y}{b_i - a_i}. \quad (27)$$

In other words, the requested pdf of Y_i is

$$f_{Y_i}(y) = \frac{e^y}{b_i - a_i} \quad i = 1, \dots, 7 \quad \ln(a_i) \leq y \leq \ln(b_i). \quad (28)$$

These are the probability density functions of the natural logs of all the uniformly distributed Drake random variables D_i .

This is indeed a positive function of y over the interval $\ln(a_i) \leq y \leq \ln(b_i)$, as for every pdf, and it is easy to see that its normalization condition is fulfilled

$$\int_{\ln(a_i)}^{\ln(b_i)} f_{Y_i}(y) dy = \int_{\ln(a_i)}^{\ln(b_i)} \frac{e^y}{b_i - a_i} dy = \frac{e^{\ln(b_i)} - e^{\ln(a_i)}}{b_i - a_i} = 1. \quad (29)$$

Next, we want to find the mean value and standard deviation of Y_i , since these play a crucial role for future developments. **The mean value $\langle Y_i \rangle$ of the random variables $Y_i = \ln(D_i)$ is given by**

$$\begin{aligned} \langle Y_i \rangle &= \int_{\ln(a_i)}^{\ln(b_i)} y f_{Y_i}(y) dy = \int_{\ln(a_i)}^{\ln(b_i)} \frac{y e^y}{b_i - a_i} dy \\ &= \frac{b_i[\ln(b_i) - 1] - a_i[\ln(a_i) - 1]}{b_i - a_i}. \end{aligned} \quad (30)$$

This is thus the mean value of the natural log of all the uniformly distributed Drake random variables D_i

$$\langle Y_i \rangle = \langle \ln(D_i) \rangle = \frac{b_i[\ln(b_i) - 1] - a_i[\ln(a_i) - 1]}{b_i - a_i}. \quad (31)$$

In order to find the variance also, we must first compute the mean value of the square of Y_i , i.e.

$$\begin{aligned} \langle Y_i^2 \rangle &= \int_{\ln(a_i)}^{\ln(b_i)} y^2 f_{Y_i}(y) dy = \int_{\ln(a_i)}^{\ln(b_i)} \frac{y^2 e^y}{b_i - a_i} dy \\ &= \frac{b_i[\ln^2(b_i) - 2\ln(b_i) + 2] - a_i[\ln^2(a_i) - 2\ln(a_i) + 2]}{b_i - a_i}. \end{aligned} \quad (32)$$

The **variance of $Y_i = \ln(D_i)$** is now given by Eq. (32) minus the square of Eq. (31), that, after a few reductions, yield:

$$\sigma_{Y_i}^2 = \sigma_{\ln(D_i)}^2 = 1 - \frac{a_i b_i [\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2}. \quad (33)$$

Whence the corresponding standard deviation

$$\sigma_{Y_i} = \sigma_{\ln(D_i)} = \sqrt{1 - \frac{a_i b_i [\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2}}. \quad (34)$$

Let us now turn to an another topic: the use of Fourier transforms, that, in probability theory, are called “characteristic functions”. Following again the notations of Papoulis (Ref. [5]) we call “characteristic function,” $\Phi_{Y_i}(\zeta)$, of an assigned probability distribution Y_i , the Fourier transform of the relevant probability density function, that is (with $j = \sqrt{-1}$)

$$\Phi_{Y_i}(\zeta) = \int_{-\infty}^{\infty} e^{j \zeta y} f_{Y_i}(y) dy. \quad (35)$$

The use of characteristic functions simplifies things greatly. For instance, the calculation of all moments of a

known pdf becomes trivial if the relevant characteristic function is known, and greatly simplified also are the proofs of important theorems of statistics, like the Central Limit Theorem that we will use in Section 4. Another important result is that the characteristic function of the sum of a finite number of independent random variables is simply given by the product of the corresponding characteristic functions. This is just the case we are facing in the Statistical Drake Eq. (4), and so we are now led to find the characteristic function of the random variable Y_i , i.e.

$$\begin{aligned} \Phi_{Y_i}(\zeta) &= \int_{-\infty}^{\infty} e^{j\zeta y} f_{Y_i}(y) dy = \int_{\ln(a_i)}^{\ln(b_i)} e^{j\zeta y} \frac{e^y}{b_i - a_i} dy \\ &= \frac{1}{b_i - a_i} \int_{\ln(a_i)}^{\ln(b_i)} e^{(1+j\zeta)y} dy = \frac{1}{b_i - a_i} (1/1+j\zeta) \left[e^{(1+j\zeta)y} \right]_{\ln(a_i)}^{\ln(b_i)} \\ &= \frac{e^{(1+j\zeta)\ln(b_i)} - e^{(1+j\zeta)\ln(a_i)}}{(b_i - a_i)(1+j\zeta)} = \frac{b_i^{1+j\zeta} - a_i^{1+j\zeta}}{(b_i - a_i)(1+j\zeta)}. \end{aligned} \quad (36)$$

Thus, the characteristic function of the natural log of the Drake uniform random variable D_i is given by

$$\Phi_{Y_i}(\zeta) = \frac{b_i^{1+j\zeta} - a_i^{1+j\zeta}}{(b_i - a_i)(1+j\zeta)}. \quad (37)$$

3.3. Step 7: Finding the probability density function of N , but only numerically, not analytically

Having found the characteristic functions $\Phi_{Y_i}(\zeta)$ of the logs of the seven input random variables D_i , we can now immediately find the characteristic function of the random variable $Y = \ln(N)$ defined by Eq. (5). In fact, by virtue of Eq. (4), of the well-known Fourier transform property stating that “the Fourier transform of a convolution is the product of the Fourier transforms,” and of Eq. (37), it immediately follows that $\Phi_Y(\zeta)$ equals the product of the seven $\Phi_{Y_i}(\zeta)$

$$\Phi_Y(\zeta) = \prod_{i=1}^7 \Phi_{Y_i}(\zeta) = \prod_{i=1}^7 \frac{b_i^{1+j\zeta} - a_i^{1+j\zeta}}{(b_i - a_i)(1+j\zeta)}. \quad (38)$$

The next step is to **invert** this Fourier transform in order to get the probability density function of the random variable $Y = \ln(N)$. In other words, we must compute the following inverse Fourier transform. (Fig. 1)

$$f_Y(y) = (1/2\pi) \int_{-\infty}^{\infty} e^{-j\zeta y} \Phi_Y(\zeta) d\zeta$$

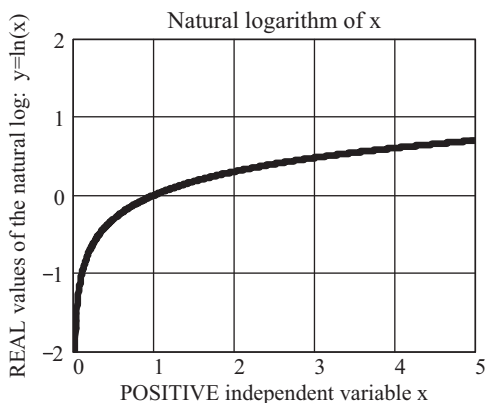


Fig. 1. The simple function $y = \ln(x)$.

$$\begin{aligned} &= (1/2\pi) \int_{-\infty}^{\infty} e^{-j\zeta y} \left[\prod_{i=1}^7 \Phi_{Y_i}(\zeta) \right] d\zeta \\ &= (1/2\pi) \int_{-\infty}^{\infty} e^{-j\zeta y} \left[\prod_{i=1}^7 \frac{b_i^{1+j\zeta} - a_i^{1+j\zeta}}{(b_i - a_i)(1+j\zeta)} \right] d\zeta. \end{aligned} \quad (39)$$

This author regrets that he was unable to compute the last integral **analytically**. He had to compute it **numerically** for the particular values of the 14 a_i and b_i that follow from Table 1 and Eq. 17. The result was the probability density function for $Y = \ln(N)$ plotted in the following Fig. 2.

We are now just one more step from finding the probability density of N , the number of ExtraTerrestrial Civilizations in the Galaxy predicted by our Statistical Drake Eq. (3). The point here is to transfer from the probability density function of Y to that of N , knowing that $Y = \ln(N)$, or alternatively, that $N = \exp(Y)$, as stated by Eq. (6). We must thus resort to the transformation law of random variables Eq. (9) by setting

$$y = g(x) = e^x. \quad (40)$$

This, upon inversion, yields the **single root**

$$x_1(y) = x(y) = \ln(y). \quad (41)$$

On the other hand, differentiating Eq. (40) one gets

$$g'(x) = e^x \quad \text{and} \quad g'(x_1(y)) = e^{\ln(y)} = y \quad (42)$$

where Eq. (41) was already used in the last step. The general transformation law (9) finally yields

$$f_N(y) = \sum_i \frac{f_X(x_i(y))}{|g'(x_i(y))|} = \frac{1}{|y|} f_Y(\ln(y)). \quad (43)$$

This probability density function $f_N(y)$ was computed numerically by using Eq. (43) and the numeric curve given by Eq. (39), and the result is shown in Fig. 3.

We now want to compute the mean value $\langle N \rangle$ of the probability density Eq. (43). Clearly, it is given by

$$\langle N \rangle = \int_0^{\infty} y f_N(y) dy. \quad (44)$$

This integral too was computed numerically, and the result was a **perfect match** with $N = 3500$ of Eq. (22), i.e.

$$\langle N \rangle = 3499.99880177509 + 0.000000124914686j. \quad (45)$$

Note that this result was computed numerically in the complex domain because of the Fourier transforms, and

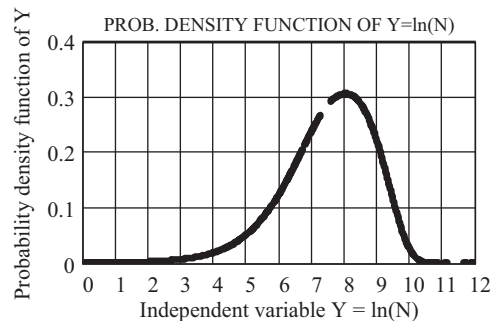


Fig. 2. Probability density function of $Y = \ln(N)$ computed numerically by virtue of the integral (39). The two “funny gaps” in curve are due to the numeric limitations in the MathCad numeric solver that the author used for this numeric computation.

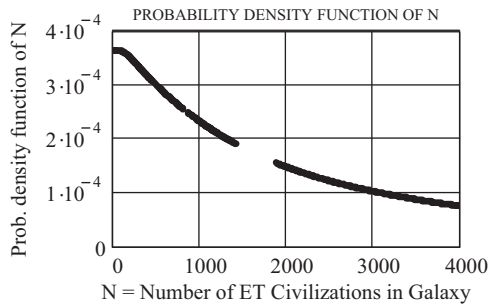


Fig. 3. The *numeric* (and not analytic) probability density function curve $f_N(y)$ of the number N of ExtraTerrestrial Civilizations in the Galaxy according to the Statistical Drake Eq. (3). We see that curve peak (i.e. the mode) is very close to low values of N , but the tail on the right is long, meaning that the resulting mean value $\langle N \rangle$ is of the order of thousands.

that the real part is virtually 3500 (as expected), while the imaginary part is virtually zero because of the rounding errors. So, this result is excellent, and proves that the theory presented so far is mathematically correct.

Finally, we want to consider the standard deviation. This also had to be computed numerically, resulting in

$$\sigma_N = 3953.42910143389 + 0.000000032800058i. \quad (46)$$

This standard deviation, higher than the mean value, implies that N might range in between 0 and 7453.

This completes our study of the probability density function of N if the seven uniform Drake input random variable D_i have the mean values and standard deviations listed in Table 1.

We conclude that, unfortunately, even under the simplifying assumptions that the D_i be uniformly distributed, it is impossible to solve the full problem analytically, since all calculations beyond Eq. (38) had to be performed numerically.

This is no good.

Shall we thus loose faith, and declare “impossible” the task of finding an analytic expression for the probability density function $f_N(y)$?

Rather surprisingly, the answer is “no”, and there is indeed a way out of this dead-end, as we shall see in the next section.

4. The central limit theorem (CLT) of statistics

Indeed there is a good, approximating analytical expression for $f_N(y)$, and this is the following **lognormal probability density function**

$$f_N(y, \mu, \sigma) = \frac{1}{y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\ln y) - \mu}{2\sigma^2}} \quad (y \geq 0). \quad (47)$$

To understand why, we must resort to what is perhaps the most beautiful theorem of Statistics: the Central Limit Theorem (abbreviated CLT). Historically, the CLT was in fact proven first in 1901 by the Russian mathematician Alexandr Lyapunov (1857–1918), and later (1920) by the Finnish mathematician Jarl Waldemar Lindeberg (1876–1932) under weaker conditions. These conditions are certainly fulfilled in the context of the Drake equation because of the “reality” of the astronomy, biology and sociology involved

with it, and we are not going to discuss this point any further here. A good, synthetic description of the Central Limit Theorem of Statistics is found at the Wikipedia site (Ref. [7]) to which the reader is referred for more details, such as the equations for the Lyapunov and the Lindeberg conditions, making the theorem “rigorously” valid.

Put in loose terms, the CLT states that, if one has a sum of random variables even NOT identically distributed, this sum tends to a normal distribution when the number of terms making up the sum tends to infinity. Also, the normal distribution mean value is the sum of the mean values of the addend random variables, and the normal distribution variance is the sum of the variances of the addend random variables.

Let us now write down the equations of the CLT in the form needed to apply it to our Statistical Drake Eq. (3). The idea is to apply the CLT to the sum of random variables given by Eqs. (4) and (5) **whatever their probability distributions can possibly be**. In other words, the CLT applied to the Statistical Drake Eq. (3) leads immediately to the following three equations:

- (1) The sum of the (arbitrarily distributed) independent random variables Y_i makes up the new random variable Y .
- (2) The sum of their mean values makes up the new mean value of Y .
- (3) The sum of their variances makes up the new variance of Y .

In equations

$$\begin{cases} Y = \sum_{i=1}^7 Y_i \\ \langle Y \rangle = \sum_{i=1}^7 \langle Y_i \rangle \\ \sigma_Y^2 = \sum_{i=1}^7 \sigma_{Y_i}^2. \end{cases} \quad (48)$$

This completes our synthetic description of the CLT for **sums** of random variables.

5. The lognormal distribution is the distribution of the number N of extraterrestrial civilizations in the Galaxy

The CLT may of course be extended to products of random variables upon taking the logs of both sides, just as we did in Eq. (3). It then follows that the exponent random variable, like Y in Eq. (6), tends to a normal random variable, and, as a consequence, it follows that the base random variable, like N in Eq. (6), tends to a lognormal random variable.

To understand this fact better in mathematical terms consider again of the transformation law (9) of random variables. The question is: what is the probability density function of the random variable N in Eq. (6), i.e. what is the probability density function of the lognormal distribution? To find it, set

$$y = g(x) = e^x. \quad (49)$$

This, upon inversion, yields the **single** root

$$x_1(y) = x(y) = \ln(y). \tag{50}$$

On the other hand, differentiating (49) one gets

$$g'(x) = e^x \text{ and } g'(x_1(y)) = e^{\ln(y)} = y \tag{51}$$

where Eq. (50) was already used in the last step. The general transformation law Eq. (9) finally yields

$$f_N(y) = \sum_i \frac{f_X(x_i(y))}{|g'(x_i(y))|} = \frac{1}{|y|} f_Y(\ln(y)). \tag{52}$$

Therefore, replacing the probability density on the right by virtue of the well-known normal (or Gaussian) distribution given by Eq. (7), the lognormal distribution of Eq. (47) is found, and the derivation of the lognormal distribution from the normal distribution is proved.

In view of future calculations, it is also useful to point out the so-called ‘‘Gaussian integral,’’ i.e.

$$\int_{-\infty}^{\infty} e^{-Ax^2} e^{Bx} dx = \sqrt{\frac{\pi}{A}} e^{\frac{B^2}{4A}}, \quad A > 0, \quad B = \text{real}. \tag{53}$$

This follows immediately from the normalization condition of the Gaussian Eq. (7), i.e.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1, \tag{54}$$

just upon expanding the square at the exponent and making the two replacements (we skip all steps)

$$\begin{cases} A = \frac{1}{2\sigma^2} > 0, \\ B = \frac{\mu}{\sigma^2} = \text{real}. \end{cases} \tag{55}$$

In the sequel of this paper, we shall denote the independent variable of the lognormal distribution (47) by a lower-case letter n to remind the reader that corresponding random variable N is the positive integer number of ExtraTerrestrial Civilizations in the Galaxy. In other words, n will be treated as a **positive real** number in all calculations to follow, because it is a ‘‘large’’ number (i.e. a continuous variable) compared to the only civilization that we know of, i.e. ourselves. In conclusion, *from now on the lognormal probability density function of N will be written as*

$$f_N(n) = \frac{1}{n} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-(\ln(n)-\mu)^2/(2\sigma^2)} \quad (n \geq 0). \tag{56}$$

Having so said, we now turn to the statistical properties of the lognormal distribution (56), i.e. to the statistical properties that describe the number N of ExtraTerrestrial Civilizations in the Galaxy.

Our first goal is to prove an equation yielding all the moments of the lognormal distribution (56), i.e. for every non-negative integer $k=0, 1, 2, \dots$ one has

$$\langle N^k \rangle = e^{k\mu} e^{k^2 \frac{\sigma^2}{2}}. \tag{57}$$

The relevant proof starts with the definition of the k -th moment

$$\begin{aligned} \langle N^k \rangle &= \int_0^{\infty} n^k f_N(n) dn \\ &= \int_0^{\infty} n^k \frac{1}{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(\ln(n)-\mu)^2/(2\sigma^2)} dn. \end{aligned}$$

One then transforms the above integral by virtue of the substitution

$$\ln[n] = z. \tag{58}$$

The new integral in z is then seen to reduce to the Gaussian integral (53) (we skip all steps here) and Eq. (57) follows

$$= e^{k\mu} e^{k^2 \frac{\sigma^2}{2}}.$$

Upon setting $k=0$ into Eq. (57), the normalization condition for $f_N(n)$ follows

$$\int_0^{\infty} f_N(n) dn = 1. \tag{59}$$

Upon setting $k=1$ into Eq. (57), the important **mean value of the random variable N** is found

$$\langle N \rangle = e^{\mu} e^{\frac{\sigma^2}{2}}. \tag{60}$$

Upon setting $k=2$ into Eq. (57), the mean value of the square of the random variable N is found

$$\langle N^2 \rangle = e^{2\mu} e^{2\sigma^2}. \tag{61}$$

The **variance of N** now follows from the last two formulae

$$\sigma_N^2 = e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1). \tag{62}$$

The square root of this is the important **standard deviation formula for the N random variable**

$$\sigma_N = e^{\mu} e^{\sigma^2/2} \sqrt{e^{\sigma^2} - 1}. \tag{63}$$

The third moment is obtained upon setting $k=3$ into Eq. (57)

$$\langle N^3 \rangle = e^{3\mu} e^{\frac{9\sigma^2}{2}}. \tag{64}$$

Finally, upon setting $k=4$, the fourth moment of N is found

$$\langle N^4 \rangle = e^{4\mu} e^{8\sigma^2}. \tag{65}$$

Our next goal is to find the cumulants of N . In principle, we could compute all the cumulants K_i from the generic i -th moment μ'_i by virtue of the recursion formula (see Ref. [8])

$$K_i = \mu'_i - \sum_{k=1}^{i-1} \binom{i-1}{k-1} K_k \mu'_{i-k}. \tag{66}$$

In practice, however, here we shall confine ourselves to the computation of the first four cumulants only because they only are required to find the skewness and kurtosis of the distribution. Then, the first four cumulants in terms of the first four moments read

$$\begin{cases} K_1 = \mu'_1 \\ K_2 = \mu'_2 - K_1^2 \\ K_3 = \mu'_3 - 3K_1 K_2 - K_1^3 \\ K_4 = \mu'_4 - 4K_1 K_3 - 3K_2^2 - 6K_2 K_1^2 - K_1^4. \end{cases} \tag{67}$$

These equations yield, respectively

$$K_1 = e^{\mu} e^{\sigma^2/2}. \tag{68}$$

$$K_2 = e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1). \tag{69}$$

$$K_3 = e^{3\mu} e^{\frac{9\sigma^2}{2}}. \tag{70}$$

$$K_4 = e^{4\mu+2\sigma^2} (e^{\sigma^2} - 1)^3 (e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6). \quad (71)$$

From these we derive the skewness

$$\frac{K_3}{(K_4)^{3/2}} = (e^{\sigma^2} + 2) \sqrt{\frac{e^{-6\mu} e^{-3\sigma^2}}{(e^{\sigma^2} - 1)^5 (e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6)^3}}, \quad (72)$$

and the kurtosis

$$\frac{K_4}{(K_2)^2} = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6. \quad (73)$$

Finally, we want to find the mode of the lognormal probability density function, i.e. the abscissa of its peak. To do so, we must first compute the derivative of the probability density function $f_N(n)$ of Eq. (56), and then set it equal to zero. This derivative is actually the derivative of the ratio of two functions of n , as it plainly appears from Eq. (56). Thus, let us set for a moment

$$E(n) = \frac{(\ln(n) - \mu)^2}{2\sigma^2} \quad (74)$$

where “E” stands for “exponent”. Upon differentiating this, one gets

$$E'(n) = \frac{1}{2\sigma^2} 2(\ln(n) - \mu) \frac{1}{n}. \quad (75)$$

But the lognormal probability density function (56), by virtue of Eq. (74), now reads

$$f_N(n) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{e^{-E(n)}}{n}. \quad (76)$$

So that its derivative is

$$\frac{df_N(n)}{dn} = \frac{1}{\sqrt{2\pi}\sigma} \frac{-e^{-E(n)} E'(n) n - 1 e^{-E(n)}}{n^2} \quad (77)$$

Setting this derivative equal to zero means setting

$$E'(n)n + 1 = 0. \quad (78)$$

That is, upon replacing Eq. (75),

$$\frac{1}{\sigma^2} (\ln(n) - \mu) + 1 = 0. \quad (79)$$

Rearranging, this becomes

$$\ln(n) - \mu + \sigma^2 = 0 \quad (80)$$

and finally

$$n_{\text{mode}} \equiv n_{\text{peak}} = e^{\mu} e^{-\sigma^2}. \quad (81)$$

This is the most likely number of ExtraTerrestrial Civilizations in the Galaxy.

How likely? To find the value of the probability density function $f_N(n)$ corresponding to this value of the mode, we must obviously substitute Eq. (81) into Eq. (56). After a few rearrangements, one then gets

$$f_N(n_{\text{mode}}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\mu} e^{\sigma^2/2}. \quad (82)$$

This is “how likely” the most likely number of ExtraTerrestrial Civilizations in the Galaxy is, i.e. it is the peak height in the lognormal probability density function $f_N(n)$.

Next to the mode, the median m (Ref. [9]) is one more statistical number used to characterize any probability distribution. It is defined as the independent variable abscissa m such that a realization of the random variable will take up a value lower than m with 50% probability or a value higher than m with 50% probability again. In other words, the median m splits up our probability density in exactly two equally probable parts. Since the probability of occurrence of the random event equals the area under its density curve (i.e. the definite integral under its density curve), then the median m (of the lognormal distribution, in this case) is defined as the integral upper limit m :

$$\int_0^m f_N(n) dn = \int_0^m \frac{1}{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\ln(n)-\mu)^2}{2\sigma^2}} dn = \frac{1}{2}. \quad (83)$$

In order to find m , we may **not** differentiate Eq. (83) with respect to m , since the “precise” factor $\frac{1}{2}$ on the right would then disappear into a zero. On the contrary, we may try to perform the obvious substitution

$$z^2 = \frac{(\ln(n)-\mu)^2}{2\sigma^2} \quad z \geq 0. \quad (84)$$

into the integral (83) to reduce it to the following integral defining the error function $erf(z)$

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz. \quad (85)$$

Then, after a few reductions that we skip for the sake of brevity, the full Eq. (83) is turned into

$$\frac{1}{2} + erf\left(\frac{\ln(m)-\mu}{\sqrt{2}\sigma}\right) = \frac{1}{2} \quad (86)$$

i.e.

$$erf\left(\frac{\ln(m)-\mu}{\sqrt{2}\sigma}\right) = 0. \quad (87)$$

Since from the definition Eq. (85) one obviously has $erf(0)=0$, Eq. (87) becomes

$$\frac{\ln(m)-\mu}{\sqrt{2}\sigma} = 0 \quad (88)$$

whence finally

$$\text{median} = m = e^{\mu}. \quad (89)$$

This is the median of the lognormal distribution of N. In other words, this is the number of ExtraTerrestrial civilizations in the Galaxy such that, with 50% probability the actual value of N will be lower than this median, and with 50% probability it will be higher.

In conclusion, we feel useful to summarize all the equations that we derived about the random variable N in the following Table 2.

We want to complete this section about the lognormal probability density function (56) by finding out its **numeric values** for the inputs to the Statistical Drake Eq. (3) listed in Table 1.

According to the CLT, the **mean value** μ to be inserted into the lognormal density Eq. (56) is given (according to the second Eq. (48)) by the sum of all the mean

Table 2

Summary of the properties of the lognormal distribution that applies to the random variable N =number of ET communicating civilizations in the Galaxy.

| Random variable | N =number of communicating ET civilizations in Galaxy |
|----------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Probability distribution | Lognormal |
| Probability density function | $f_N(n) = \frac{1}{n} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\ln n - \mu)^2}{2\sigma^2}} \quad (n \geq 0)$ |
| Mean value | $\langle N \rangle = e^\mu e^{\sigma^2/2}$ |
| Variance | $\sigma_N^2 = e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1)$ |
| Standard deviation | $\sigma_N = e^\mu e^{\frac{\sigma^2}{2}} \sqrt{e^{\sigma^2} - 1}$ |
| All the moments, i.e. k -th moment | $\langle N^k \rangle = e^{k\mu} e^{\frac{k^2}{2}\sigma^2}$ |
| Mode (=abscissa of the lognormal peak) | $n_{\text{mode}} \equiv n_{\text{peak}} = e^{\mu - \sigma^2}$ |
| Value of the mode peak | $f_N(n_{\text{mode}}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\mu} e^{\sigma^2/2}$ |
| Median (=fifty-fifty probability value for N) | Median = $m = e^\mu$ |
| Skewness | $\frac{K_3}{(K_4)^{3/2}} = (e^{\sigma^2} + 2) \sqrt{\frac{e^{-6\mu} e^{-3\sigma^2}}{(e^{\sigma^2} - 1)^5 (e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6)^3}}$ |
| Kurtosis | $\frac{K_4}{(K_2)^2} = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6$ |
| Expression of μ in terms of the lower (a_i) and upper (b_i) limits of the Drake uniform input random variables D_i | $\mu = \sum_{i=1}^7 \langle Y_i \rangle = \sum_{i=1}^7 \frac{b_i[\ln(b_i)-1] - a_i[\ln(a_i)-1]}{b_i - a_i}$ |
| Expression of σ^2 in terms of the lower (a_i) and upper (b_i) limits of the Drake uniform input random variables D_i | $\sigma^2 = \sum_{i=1}^7 \sigma_{Y_i}^2 = \sum_{i=1}^7 \left(1 - \frac{a_i b_i [\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2} \right)$ |

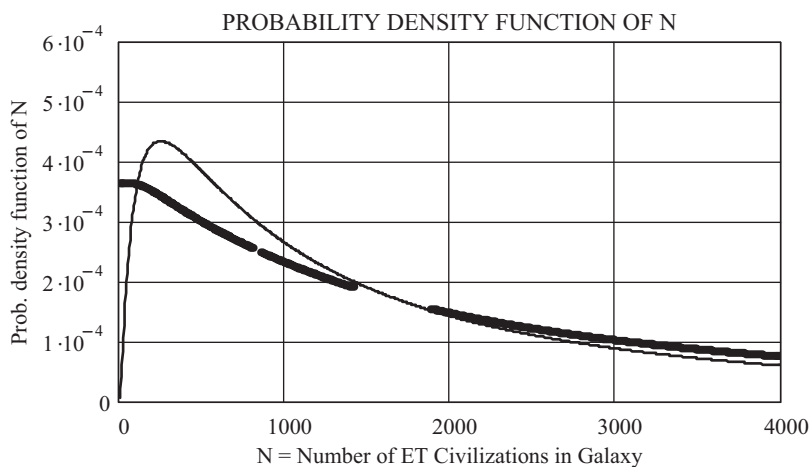


Fig. 4. Comparing the two probability density functions of the random variable N found: (1) at the end of Section 3.3. in a purely numeric way and without resorting to the CLT at all (thick curve) and (2) analytically by using the CLT and the relevant lognormal approximation (thin curve).

values $\langle Y_i \rangle$, that is, by virtue of Eq. (31), by

$$\mu = \sum_{i=1}^7 \langle Y_i \rangle = \sum_{i=1}^7 \frac{b_i[\ln(b_i)-1] - a_i[\ln(a_i)-1]}{b_i - a_i}. \quad (90)$$

Upon replacing the 14 a_i and b_i listed in Table 1 into Eq. (90), the following **numeric mean value** μ is found

$$\mu \approx 7.462176. \quad (91)$$

Similarly, to get the numeric variance σ^2 one must resort to the last of Eq. (48) and to Eq. (33):

$$\sigma^2 = \sum_{i=1}^7 \sigma_{Y_i}^2 = \sum_{i=1}^7 \left(1 - \frac{a_i b_i [\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2} \right) \quad (92)$$

yielding the following **numeric variance** σ^2 to be inserted into the lognormal pdf Eq. (56)

$$\sigma^2 \approx 1.938725 \quad (93)$$

whence the numeric standard deviation σ

$$\sigma \approx 1.392381. \quad (94)$$

Upon replacing these two numeric values Eqs. (91) and (94) into the lognormal pdf Eq. (56), the latter is perfectly determined. It is plotted in Fig. 4 as the thin curve.

In other words, Fig. 4 shows the lognormal distribution for the number N of ExtraTerrestrial Civilizations in the Galaxy derived from the Central Limit Theorem as applied to the Drake equation (with the input data listed in Table 1).

We now like to point out the most important statistical properties of this lognormal pdf:

- (1) **Mean Value of N.** This is given by Eq. (60) with μ and σ given by Eqs. (91) and (94), respectively:

$$\langle N \rangle = e^\mu e^{\sigma^2/2} \approx 4589.559. \tag{95}$$

In other words, there are 4590 ET Civilizations in the Galaxy according to the Central Limit Theorem of Statistics with the inputs of Table 1. This number 4590 is HIGHER than the 3500 foreseen by the classical Drake equation working with sheer numbers only, rather than with probability distributions. Thus, Eq. (95) IS GOOD FOR NEWS FOR SETI, since it shows that the expected number of ETs is HIGHER with an adequate statistical treatment than just with the too simple Drake sheer numbers of Eq. (1).

- (2) **Variance of N.** The variance of the lognormal distribution is given by Eq. (62) and turns out to be a huge number

$$\sigma_N^2 = e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1) \approx 125328623. \tag{96}$$

- (3) **Standard deviation of N.** The standard deviation of the lognormal distribution is given by Eq. (63) and turns out to be

$$\sigma_N = e^\mu e^{\frac{\sigma^2}{2}} \sqrt{e^{\sigma^2} - 1} = 11195. \tag{97}$$

Again, this is GOOD NEWS FOR SETI. In fact, such a high standard deviation means that N may range from very low values (zero, theoretically, and one since humanity exists) up to tens of thousands ($4590 + 11,195 = 15,785$ is Eq. (95)+Eq. (97)).

- (4) **Mode of N:** the mode (= peak abscissa) of the lognormal distribution of N is given by Eq. (81), and has a surprisingly low numeric value

$$n_{\text{mode}} \equiv n_{\text{peak}} = e^\mu e^{-\sigma^2} \approx 250. \tag{98}$$

This is well shown in Fig. 4: the mode peak is very pronounced and close to the origin, but the right tail is

long, and this means that the mean value of the distribution is much higher than the mode: $4590 \gg 250$.

- (5) **Median of N:** the median (= fifty-fifty abscissa, splitting the pdf in two exactly equi-probable parts) of the lognormal distribution of N is given by Eq. (89), and has the numeric value

$$n_{\text{median}} \equiv e^\mu \approx 1740. \tag{99}$$

In words, assuming the input values listed in Table 1, we have exactly a 50% probability that the actual value of N is lower than 1740, and 50% that it is higher than 1740.

6. Comparing the CLT results with the non-CLT results

The time is now ripe to compare the CLT-based results about the lognormal distribution of N , just described in Section 5, against the Non-CLT-based results obtained numerically in Section 3.3.

To do so in a simple, visual way, let us plot on the same diagram two curves (see Fig. 4):

- (1) The numeric curves appearing in Fig. 2 and obtained after laborious Fourier transform calculations in the complex domain, and
- (2) The lognormal distribution (56) with numeric μ and σ given by Eqs. (91) and (94), respectively.

We see in Fig. 4 that the two curves are virtually coincident for values of N larger than 1500. This is a consequence of the law of large numbers, of which the CLT is just one of the many facets.

Similarly, it happens for natural log of N , i.e. the random variable Y of Eq. (5), that is plotted in Fig. 5 both in its normal curve version (thin curve) and in its numeric version, obtained via Fourier transforms and already shown in Fig. 2.

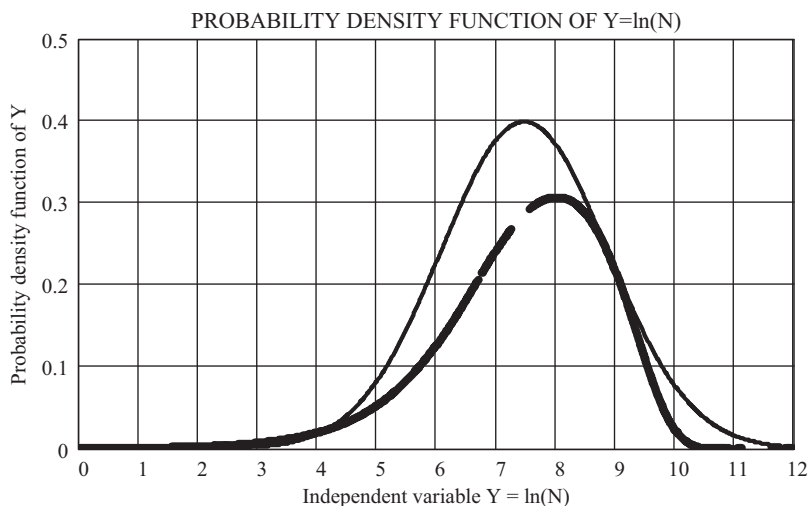


Fig. 5. Comparing the two probability density functions of the random variable $Y = \ln(N)$ found: (1) at the end of Section 3.3. in a purely numeric way and without resorting to the CLT at all (thick curve) and (2) analytically by using the CLT and the relevant normal (Gaussian) approximation (thin Gaussian curve).

The conclusion is simple: from now on we shall discard forever the numeric calculations and we will stick only to the equations derived by virtue of the CLT, i.e. to the lognormal Eq. (56) and its consequences.

7. Distance of the nearest extraterrestrial civilization as a probability distribution

As an application of the Statistical Drake Equation developed in the previous sections of this paper, we now want to consider the problem of estimating the distance of the ExtraTerrestrial Civilization nearest to us in the Galaxy. In all Astrobiology textbooks (see for instance, Ref. [10]) and in several web sites, the solution to this problem is reported with only slight differences in the mathematical proofs among the various authors. In the first of the coming two sections (Section 7.1), we derive the expression for this “ET_Distance” (as we like to denote it) in the classical, non-probabilistic way: in other words, this is the classical, deterministic derivation. In the second Section 7.2, we provide the probabilistic derivation, arising from our Statistical Drake Equation, of the corresponding probability density function $f_{ET_distance}(r)$: here, r is the distance between us and the nearest ET civilization assumed as the independent variable of its own probability density function. The ensuing sections provide more mathematical details about this $f_{ET_distance}(r)$ such as its mean value, variance, standard deviation, all central moments, mode, median, cumulants, skewness and kurtosis.

7.1. Classical, non-probabilistic derivation of the distance of the nearest ET civilization

Consider the Galactic Disk and assume that:

- (1) The diameter of the Galaxy is (about) 100,000 light years, (abbreviated ly) i.e. its radius, R_{Galaxy} , is about 50,000 ly.
- (2) The thickness of the Galactic Disk at half-way from its center, h_{Galaxy} , is about 16,000 ly. Then,
- (3) the volume of the Galaxy may be approximated as the volume of the corresponding cylinder, i.e.

$$V_{Galaxy} = \pi R_{Galaxy}^2 h_{Galaxy}. \tag{100}$$

- (4) Now consider the sphere around us having a radius r . The volume of such as sphere is

$$V_{Our_Sphere} = \frac{4}{3} \pi \left(\frac{ET_Distance}{2} \right)^3. \tag{101}$$

In the last equation, we had to divide the distance “ET_Distance” between ourselves and the nearest ET Civilization by 2, because we are now going to make the unwarranted assumption that all ET Civilizations are equally spaced from each other in the Galaxy! This is a crazy assumption, clearly, and should be replaced by more scientifically grounded assumptions as soon as we know more about our Galactic Neighbourhood. At the moment, however, this is the best guess that we can make, and so we shall take it for granted, although we are aware that this is weak point in the reasoning.

Having thus assumed that ET Civilizations are UNIFORMLY SPACED IN THE GALAXY, we can write down this proportion

$$\frac{V_{Galaxy}}{N} = \frac{V_{Our_Sphere}}{1}. \tag{102}$$

That is, upon replacing both Eq. (100) and Eq. (101) into Eq. (102)

$$\frac{\pi R_{Galaxy}^2 h_{Galaxy}}{N} = \frac{\frac{4}{3} \pi \left(\frac{ET_Distance}{2} \right)^3}{1}. \tag{103}$$

The only unknown in the last equation is ET_Distance, and so we may solve for it, thus getting the

(AVERAGE) DISTANCE BETWEEN ANY PAIR OF NEIGHBORING CIVILIZATIONS IN THE GALAXY

$$ET_Distance = \frac{\sqrt[3]{6R_{Galaxy}^2 h_{Galaxy}}}{\sqrt[3]{N}} = \frac{C}{\sqrt[3]{N}} \tag{104}$$

where the positive constant C is defined by

$$C = \sqrt[3]{6R_{Galaxy}^2 h_{Galaxy}} \approx 28,845 \text{ light years}. \tag{105}$$

Eqs. (104) and (105) are the starting point for our first application of the Statistical Drake equation, that we discuss in detail in the coming sections of this paper.

7.2. Probabilistic derivation of the probability density function for ET_Distance

The probability density function (pdf) yielding the distance of the ET Civilization nearest to us in the Galaxy and presented in this section, was discovered by this author on September 5th, 2007. He did not disclose it to other scientists until the SETI meeting run by the famous mathematical physicist and popular science author, Paul Davies, at the “Beyond” Center of the University of Arizona at Phoenix, on February 5–8, 2008. This meeting was also attended by SETI Institute experts Jill Tarter, Seth Shostak, Doug Vakoch, Tom Pierson and others. During the author’s talk, Paul Davies suggested to call “the Maccone distribution” the new probability density function that yields the ET_Distance and is derived in this section.

Let us go back to Eq. (104). Since N is now a random variable (obeying the lognormal distribution), it follows that the ET_Distance must be a random variable as well. Hence, it must have some unknown probability density function that we denote by

$$f_{ET_Distance}(r) \tag{106}$$

where r is the new independent variable of such a probability distribution (it is denoted by r to remind the reader that it expresses the three-dimensional radial distance separating us from the nearest ET civilization in a full spherical symmetry of the space around us).

The question then is: what is the unknown probability distribution (106) of the ET_Distance?

We can answer this question upon making the two formal substitutions

$$\begin{cases} N \rightarrow x \\ ET_distance \rightarrow y \end{cases} \tag{107}$$

into the transformation law (8) for random variables. As a consequence, Eq. (104) takes form

$$y = g(x) = \frac{C}{\sqrt[3]{x}} = Cx^{-1/3}. \quad (108)$$

In order to find the unknown probability density $f_{ET_Distance}(r)$, we now apply the rule of Eq. (9) to Eq. (108). First, notice that Eq. (108), when inverted to yield the various roots $x_i(y)$, yields a **single** real root only

$$x_1(y) = \frac{C^3}{y^3}. \quad (109)$$

Then, the summation in Eq. (9) reduces to one term only. Second, differentiating Eq. (108) one finds

$$g'(x) = -\frac{C}{3} x^{-4/3}. \quad (110)$$

Thus, the relevant absolute value reads

$$|g'(x)| = \left| -\frac{C}{3} x^{-4/3} \right| = \frac{C}{3} x^{-4/3}. \quad (111)$$

Upon replacing Eq. (111) into Eq. (9), we then find

$$|g'(x_1)| = \frac{C}{3} x^{-4/3} = \frac{C}{3} \left[\frac{C^3}{y^3} \right]^{-4/3} = \frac{C}{3} \left[\frac{C}{y} \right]^{-4} = \frac{y^4}{3C^3}. \quad (112)$$

This is the denominator of Eq. (9). The numerator simply is the lognormal probability density function (56) where the old independent variable x must now be re-written in terms of the new independent variable y by virtue of Eq. (109). By doing so, we finally arrive at the new probability density function $f_Y(y)$

$$f_Y(y) = \frac{3C^3}{y^4} \cdot \frac{1}{C^3} \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{\left(\ln\left[\frac{C^3}{y^3}\right] - \mu\right)^2}{2\sigma^2}}.$$

Rearranging and replacing y by r , the final form is

$$f_{ET_Distance}(r) = \frac{3}{r} \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{\left(\ln\left[\frac{C^3}{r^3}\right] - \mu\right)^2}{2\sigma^2}}. \quad (113)$$

Now, just replace C in Eq. (113) by virtue of Eq. (105). Then:

We have discovered the probability density function yielding the probability of finding the nearest ExtraTerrestrial Civilization in the Galaxy in the spherical shell between the distances r and $r+dr$ from Earth:

$$f_{ET_Distance}(r) = \frac{3}{r} \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{\left(\ln\left[\frac{6R_{Galaxy}^2}{r^3} h_{Galaxy}\right] - \mu\right)^2}{2\sigma^2}} \quad (114)$$

holding for $r \geq 0$.

7.3. Statistical properties of this distribution

We now want to study this probability distribution in detail. Our next questions are:

- (1) What is its mean value?
- (2) What are its variance and standard deviation?
- (3) What are its moments to any higher order?
- (4) What are its cumulants?

- (5) What are its skewness and kurtosis?
- (6) What are the coordinates of its peak, i.e. the mode (peak abscissa) and its ordinate?
- (7) What is its median?

The first three points in the list are all covered by the following theorem: all the moments of Eq. (113) are given by (here k is the generic and non-negative integer exponent, i.e. $k=0, 1, 2, 3, \dots \geq 0$)

$$\begin{aligned} \langle ET_Distance^k \rangle &= \int_0^\infty r^k f_{ET_Distance}(r) dr \\ &= \int_0^\infty r^k \frac{3}{r} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(\ln\left[\frac{C^3}{r^3}\right] - \mu\right)^2}{2\sigma^2}} dr \\ &= C^k e^{-k\frac{\mu}{3}} \frac{\sigma^2}{18}. \end{aligned} \quad (115)$$

To prove this result, one first transforms the above integral by virtue of the substitution

$$\ln\left[\frac{C^3}{r^3}\right] = z. \quad (116)$$

Then, the new integral in z is then seen to reduce to the known Gaussian integral (53) and, after several reductions that we skip for the sake of brevity, Eq. (115) follows from Eq. (53). In other words, we have proven that

$$\langle ET_Distance^k \rangle = C^k e^{-k\frac{\mu}{3}} \frac{\sigma^2}{18}. \quad (117)$$

Upon setting $k=0$ into Eq. (117), the normalization condition for $f_{ET_Distance}(r)$ follows

$$\int_0^\infty f_{ET_Distance}(r) dr = 1. \quad (118)$$

Upon setting $k=1$ into Eq. (117), the important **mean value of the random variable ET_Distance** is found.

$$\langle ET_Distance \rangle = C e^{-\frac{\mu}{3}} \frac{\sigma^2}{18}. \quad (119)$$

Upon setting $k=2$ into Eq. (117), the mean value of the square of the random variable ET_Distance is found

$$\langle ET_Distance^2 \rangle = C^2 e^{-\frac{2\mu}{3}} \frac{\sigma^2}{9}. \quad (120)$$

The variance of ET_Distance now follows from the last two formulae with a few reductions:

$$\begin{aligned} \sigma_{ET_Distance}^2 &= \langle ET_Distance^2 \rangle - \langle ET_Distance \rangle^2 \\ &= C^2 e^{-\frac{2\mu}{3}} \frac{\sigma^2}{9} (e^{\frac{\sigma^2}{9}} - 1). \end{aligned} \quad (121)$$

So, the variance of ET_Distance is

$$\sigma_{ET_Distance}^2 = C^2 e^{-\frac{2\mu}{3}} \frac{\sigma^2}{9} (e^{\frac{\sigma^2}{9}} - 1). \quad (122)$$

The square root of this is the important **standard deviation of the ET_Distance random variable**

$$\sigma_{ET_Distance} = C e^{-\frac{\mu}{3}} \frac{\sigma}{3} \sqrt{e^{\frac{\sigma^2}{9}} - 1}. \quad (123)$$

The third moment is obtained upon setting $k=3$ into Eq. (117)

$$\langle ET_Distance^3 \rangle = C^3 e^{-\mu} \frac{\sigma^2}{27}. \quad (124)$$

Finally, upon setting $k=4$ into Eq. (117), the fourth moment of ET_Distance is found

$$\langle ET_Distance^4 \rangle = C^4 e^{-\frac{4\mu}{3}} \frac{\sigma^2}{81}. \quad (125)$$

Our next goal is to find the cumulants of the ET_Distance. In principle, we could compute all the cumulants K_i from the generic i -th moment μ_i by virtue

of the recursion formula (see Ref. [8])

$$K_i = \mu'_i - \sum_{k=1}^{i-1} \binom{i-1}{k-1} K_k \mu'_{i-k}. \quad (126)$$

In practice, however, here we shall confine ourselves to the computation of the first four cumulants, because they only are required to find the skewness and kurtosis of the distribution (113). Then, the first four cumulants in terms of the first four moments read

$$\begin{cases} K_1 = \mu'_1 \\ K_2 = \mu'_2 - K_1^2 \\ K_3 = \mu'_3 - 3K_1K_2 - K_1^3 \\ K_4 = \mu'_4 - 4K_1K_3 - 3K_2^2 - 6K_2K_1^2 - K_1^4. \end{cases} \quad (127)$$

These equations yield, respectively:

$$K_1 = C e^{-\mu/3} e^{\sigma^2/18}. \quad (128)$$

$$K_2 = C^2 e^{-2\mu/3} e^{\sigma^2/9} (e^{\sigma^2/9} - 1). \quad (129)$$

$$K_3 = C^3 e^{-\mu} (e^{\sigma^2/2} - 3e^{5\sigma^2/18} + 2e^{\sigma^2/6}). \quad (130)$$

$$\begin{aligned} K_4 &= \\ &= C^4 e^{4\mu/3} (e^{8\sigma^2/9} - 4e^{5\sigma^2/9} - 3e^{4\sigma^2/9} + 12e^{\sigma^2/3} - 6e^{2\sigma^2/9}). \end{aligned} \quad (131)$$

From these, we derive the skewness

$$\frac{K_3}{(K_4)^{3/2}} = \frac{e^{-\mu} (e^{\sigma^2/2} - 3e^{5\sigma^2/18} + 2e^{\sigma^2/6})}{C^3 (e^{8\sigma^2/9} - 4e^{5\sigma^2/9} - 3e^{4\sigma^2/9} + 12e^{\sigma^2/3} - 6e^{2\sigma^2/9})^{3/2}}. \quad (132)$$

and the kurtosis

$$\frac{K_4}{(K_2)^2} = e^{4\sigma^2/9} + 2e^{\sigma^2/3} + 3e^{2\sigma^2/9} - 6. \quad (133)$$

Next we want to find the mode of this distribution, i.e. the abscissa of its peak. To do so, we must first compute the derivative of the probability density function $f_{ET_Distance}(r)$ of Eq. (113), and then set it equal to zero. This derivative is actually the derivative of the ratio of two functions of r , as it plainly appears from Eq. (113). Thus, let us set for a moment

$$E(r) = \frac{(\ln \left[\frac{C^3}{r^3} \right] - \mu)^2}{2\sigma^2}. \quad (134)$$

where “E” stands for “exponent”. Upon differentiating, one gets

$$\begin{aligned} E'(r) &= \frac{1}{2\sigma^2} 2 \left(\ln \left[\frac{C^3}{r^3} \right] - \mu \right) \frac{1}{C^3} C^3 (-3)r^{-4} \\ &= \frac{1}{\sigma^2} \left(\ln \left[\frac{C^3}{r^3} \right] - \mu \right) (-3) \frac{1}{r}. \end{aligned} \quad (135)$$

But the probability density function (113) now reads

$$f_{ET_Distance}(r) = \frac{3}{\sqrt{2\pi}\sigma} \cdot \frac{e^{-E(r)}}{r}. \quad (136)$$

So that its derivative is

$$\frac{df_{ET_Distance}(r)}{dr} = \frac{3}{\sqrt{2\pi}\sigma} \frac{-e^{-E(r)} E'(r) r - 1 e^{-E(r)}}{r^2}$$

$$= \frac{3}{\sqrt{2\pi}\sigma} \frac{-e^{-E(r)} [E'(r)r + 1]}{r^2}. \quad (137)$$

Setting this derivative equal to zero means setting

$$E'(r)r + 1 = 0. \quad (138)$$

That is, upon replacing Eq. (135) into Eq. (138), we get

$$\frac{1}{\sigma^2} \left(\ln \left[\frac{C^3}{r^3} \right] - \mu \right) (-3) \frac{1}{r} r + 1 = 0. \quad (139)$$

Rearranging, this becomes

$$-3 \left(\ln \left[\frac{C^3}{r^3} \right] - \mu \right) + \sigma^2 = 0 \quad (140)$$

that is

$$-3 \ln \left[\frac{C^3}{r^3} \right] + 3\mu + \sigma^2 = 0 \quad (141)$$

whence

$$\ln \left[\frac{C}{r} \right] = \frac{\mu}{3} + \frac{\sigma^2}{9} \quad (142)$$

and finally

$$r_{\text{mode}} \equiv r_{\text{peak}} = C e^{-\frac{\mu}{3}} e^{-\frac{\sigma^2}{9}}. \quad (143)$$

This is the most likely ET_Distance from Earth.

How likely?

To find the value of the probability density function $f_{ET_Distance}(r)$ corresponding to this value of the mode, we must obviously replace Eq. (143) into Eq. (113). After a few rearrangements, which we skip for the sake of brevity, one gets

$$\begin{aligned} \text{Peak Value of } f_{ET_Distance}(r) &\equiv f_{ET_Distance}(r_{\text{mode}}) \\ &= \frac{3}{C\sqrt{2\pi}\sigma} \cdot \frac{\mu}{e^3} \cdot \frac{\sigma^2}{e^{18}}. \end{aligned} \quad (144)$$

This is the peak height in the pdf $f_{ET_Distance}(r)$.

Next to the mode, the median m (Ref. [9]) is one more statistical number used to characterize any probability distribution. It is defined as an independent variable abscissa m such that a realization of the random variable will take up a value lower than m with 50% probability or a value higher than m with 50% probability again. In other words, the median m splits up our probability density in exactly two equally probable parts. Since the probability of occurrence of the random event equals the area under its density curve (i.e. the definite integral under its density curve), then the median m (of the Maccone distribution, Eq. (113)) is defined as the integral upper limit m

$$\int_0^m f_{ET_Distance}(r) dr = \frac{1}{2}. \quad (145)$$

Upon replacing Eq. (113), this becomes

$$\int_0^m \frac{3}{r} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\ln \left[\frac{C^3}{r^3} \right] - \mu\right)^2 / (2\sigma^2)} dr = \frac{1}{2}. \quad (146)$$

In order to find m , we may **not** differentiate Eq. (146) with respect to m , since the “precise” factor $\frac{1}{2}$ on the right would then disappear into a zero. On the contrary, we

may try to perform the obvious substitution

$$z^2 = \frac{\left(\ln\left[\frac{C^3}{r^3}\right] - \mu\right)^2}{2\sigma^2} \quad z \geq 0. \tag{147}$$

into the integral (146) to reduce it to the integral (85) defining the error function erf(z). Then, after a few reductions that we leave to the reader as an exercise, the full Eq. (145), defining the median, is turned into the corresponding equation involving the error function erf(x) as defined by Eq. (85)

$$\frac{1}{2} + \operatorname{erf}\left(\frac{\ln\left[\frac{C^3}{m^3}\right] - \mu}{\sqrt{2}\sigma}\right) = \frac{1}{2} \tag{148}$$

i.e.

$$\operatorname{erf}\left(\frac{\ln\left[\frac{C^3}{m^3}\right] - \mu}{\sqrt{2}\sigma}\right) = 0. \tag{149}$$

Since from the definition Eq. (85) one obviously has erf(0)=0, Eq. (149) yields

$$\frac{\ln\left[\frac{C^3}{m^3}\right] - \mu}{\sqrt{2}\sigma} = 0 \tag{150}$$

whence finally

$$\text{median} = m = C e^{-\mu/3}. \tag{151}$$

This is the median of the Maccone distribution of ET_distance. In other words, this is the distance from the Sun such that, with 50% probability the actual value of

ET_distance will be smaller than this median, and with 50% probability it will be higher.

In conclusion, we feel useful to summarize all the equations that we derived about the random variable ET_distance in the following Table 3.

7.4. Numerical example of the ET_Distance distribution

In this section, we provide a numerical example of the analytic calculations carried on so far.

Consider the Drake Equation input values reported in Table 1. Then, the graph of the corresponding probability density function of the nearest ET_Distance, f_ET_Distance(r), is shown in Fig. 6.

From Fig. 6, we see that **the probability of finding ExtraTerrestrials is practically zero up to a distance of about 500 light years from Earth.** Then, it starts increasing with the increasing distance from Earth, and reaches its maximum at

$$r_{\text{mode}} \equiv r_{\text{peak}} = C e^{-\mu/3} e^{-\sigma^2/9} \approx 1933 \text{ light years}. \tag{152}$$

This is the MOST LIKELY VALUE of the distance at which we can expect to find the nearest ExtraTerrestrial civilization.

It is **not**, however, the mean value of the probability distribution (113) for f_ET_Distance(r). In fact, the probability density Eq. (113) has an infinite tail on the right, as clearly shown in Fig. 6, and hence its mean value must be higher

Table 3

Summary of the properties of the probability distribution that applies to the random variable ET_Distance yielding the (average) distance between any two neighboring communicating civilizations in the Galaxy.

| | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Random variable | ET_Distance between any two neighboring ET civilizations in Galaxy assuming they are UNIFORMLY distributed throughout the whole Galaxy volume. |
| Probability distribution | Unnamed (Paul Davies suggested "Maccone distribution") |
| Probability density function | $f_{\text{ET_Distance}}(r) = \frac{3}{r} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(\ln\left[\frac{6R_{\text{Galaxy}}^2 h_{\text{Galaxy}}}{r^3}\right] - \mu\right)^2}{2\sigma^2}}$ |
| (Defining the positive numeric constant C) | $C = \sqrt[3]{6R_{\text{Galaxy}}^2 h_{\text{Galaxy}}} \approx 28,845 \text{ light years}$ |
| Mean value | $\langle \text{ET_Distance} \rangle = C e^{-\mu/3} e^{\sigma^2/18}$ |
| Variance | $\sigma_{\text{ET_Distance}}^2 = C^2 e^{-2\mu/3} e^{\frac{\sigma^2}{9}} (e^{\frac{\sigma^2}{9}} - 1)$ |
| Standard deviation | $\sigma_{\text{ET_Distance}} = C e^{-\mu/3} e^{\frac{\sigma^2}{18}} \sqrt{e^{\frac{\sigma^2}{9}} - 1}$ |
| All the moments, i.e. k-th moment | $\langle \text{ET_Distance}^k \rangle = C^k e^{-k\mu/3} e^{k^2 \frac{\sigma^2}{18}}$ |
| Mode (=abscissa of the probability density function peak) | $r_{\text{mode}} \equiv r_{\text{peak}} = C e^{-\mu/3} e^{-\frac{\sigma^2}{9}}$ |
| Value of the mode peak | $\text{Peak value of } f_{\text{ET_Distance}}(r) \equiv f_{\text{ET_Distance}}(r_{\text{mode}}) = \frac{3}{C\sqrt{2\pi}\sigma} \cdot e^{\mu/3} \cdot e^{\frac{\sigma^2}{18}}$ |
| Median (=fifty-fifty probability value for ET_Distance) | Median = m = C e^{-μ/3} |
| Skewness | $\frac{K_3}{(K_4)^{3/2}} = \frac{e^{-\mu} (e^{\frac{\sigma^2}{9}} - 3e^{\frac{5\sigma^2}{18}} + 2e^{\frac{\sigma^2}{6}})}{C^3 (e^{\frac{8\sigma^2}{9}} - 4e^{\frac{5\sigma^2}{9}} - 3e^{\frac{\sigma^2}{9}} + 12e^{\frac{\sigma^2}{3}} - 6e^{\frac{2\sigma^2}{3}})^{3/2}}$ |
| Kurtosis | $\frac{K_4}{(K_2)^2} = \frac{e^{\frac{4\sigma^2}{9}} + 2e^{\frac{\sigma^2}{3}} + 3e^{\frac{2\sigma^2}{9}} - 6}{6}$ |
| Expression of μ in terms of the lower (a _i) and upper (b _i) limits of the Drake uniform input random variables D _i | $\mu = \sum_{i=1}^7 \langle Y_i \rangle = \sum_{i=1}^7 \frac{b_i[\ln(b_i) - 1] - a_i[\ln(a_i) - 1]}{b_i - a_i}$ |
| Expression of σ ² in terms of the lower (a _i) and upper (b _i) limits of the Drake uniform input random variables D _i | $\sigma^2 = \sum_{i=1}^7 \sigma_{Y_i}^2 = \sum_{i=1}^7 \left(1 - \frac{a_i b_i [\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2}\right)$ |

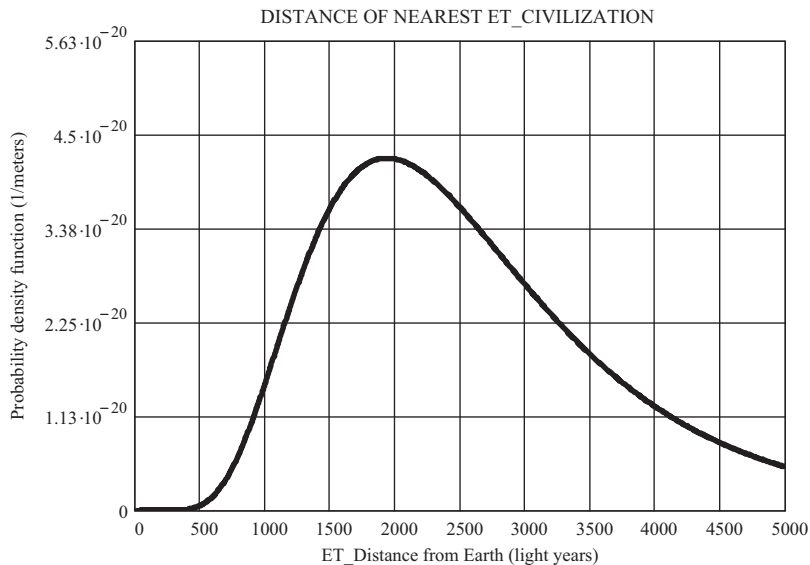


Fig. 6. This is the probability of finding the nearest ExtraTerrestrial Civilization at the distance r from Earth (in light years) if the values assumed in the Drake Equation are those shown in Table 1. The relevant probability density function $f_{ET_Distance}(r)$ is given by Eq. (113). Its mode (peak abscissa) equals 1933 light years, but its mean value is higher since curve has a long tail on the right: the mean value equals in fact 2670 light years. Finally, the standard deviation equals 1309 light years: *THIS IS GOOD NEWS FOR SETI, inasmuch as the nearest ET Civilization might lie at just 1 sigma=2670-1309=1361 light years from us.*

than its peak value. As given by Eq. (119), its mean value is

$$r_{mean_value} = Ce^{-\mu/3} e^{\sigma^2/18} \approx 2670 \text{ light years.} \quad (153)$$

This is the MEAN (value of the) DISTANCE at which we can expect to find ExtraTerrestrials.

After having found the above two distances (1933 and 2670 light years, respectively), the next natural question that arises is: “what is the range, forth and back around the mean value of the distance, within which we can expect to find ExtraTerrestrials with “the highest hopes?”. The answer to this question is given by the notion of standard deviation, that we already found to be given by Eq. (123)

$$\sigma_{ET_Distance} = Ce^{-\frac{\mu}{3}} e^{\frac{\sigma^2}{18}} \sqrt{e^{\frac{\sigma^2}{9}} - 1} \approx 1309 \text{ light years.} \quad (154)$$

More precisely, this is the so-called 1-sigma (distance) level. Probability theory then shows that the nearest ExtraTerrestrial civilization is expected to be located within this range, i.e. within the two distances of $(2670-1309)=1361$ light years and $(2670+1309)=3979$ light years, with probability given by the integral of $f_{ET_Distance}(r)$ taken in between these two lower and upper limits, i.e.

$$\int_{1361 \text{ light years}}^{3979 \text{ light years}} f_{ET_Distance}(r) dr \approx 0.75 = 75\%. \quad (155)$$

In plain words: with 75% probability, the nearest ExtraTerrestrial civilization is located in between the distances of 1361 and 3979 light years from us, having assumed the input values to the Drake Equation given by Table 1. If we change those input values, then all the numbers change again.

8. The “DATA ENRICHMENT PRINCIPLE” as the best CLT consequence upon the statistical Drake equation (any number of factors allowed)

As a fitting climax to all the statistical equations developed so far, let us now state our “DATA ENRICHMENT PRINCIPLE”. It simply states that **“The Higher the Number of Factors in the Statistical Drake equation, The Better”**.

Put in this simple way, it simply looks like a new way of saying that the CLT lets the random variable Y approach the normal distribution when the number of terms in the sum (4) approaches infinity. And this is the case, indeed. However, **our “Data Enrichment Principle” has more profound methodological consequences** that we cannot explain now, but hope to describe more precisely in one or more coming papers.

9. Conclusions

We have sought to extend the classical Drake equation to let it encompass Statistics and Probability.

This approach appears to pave the way to future, more profound investigations intended not only to associate “error bars” to each factor in the Drake equation, but especially to increase the number of factors themselves. In fact, this seems to be the only way to incorporate into the Drake equation more and more new scientific information as soon as it becomes available. In the long run, the Statistical Drake equation might just become a huge computer code, growing up in size and especially in the depth of the scientific information it contained. It would thus be the humanity’s first “Encyclopaedia Galactica”.

Unfortunately, to extend the Drake equation to Statistics, it was necessary to use a mathematical apparatus that is more sophisticated than just the simple product of seven numbers.

The first IAC presentation of the Statistical Drake Equation was made by the author on October 1st, 2008, at the 59th International Astronautical Congress held in Glasgow, Scotland, UK (Ref. [11]).

When this author had the honour and privilege to present his results at the SETI Institute on April 11th, 2008, in front of an audience also including Professor Frank Drake, he felt he had to add these words: “My apologies, Frank, for disrupting the beautiful simplicity of your equation”.

Acknowledgements

The author is grateful to Drs. Jill Tarter, Paul Davies, Seth Shostak, Doug Vakoch, Tom Pierson, Carol Oliver, Paul Shuch and Kathryn Denning for attending his first presentation ever about these topics at the “Beyond” Center of the University of Arizona at Phoenix on February 8th, 2008. He also would like to thank Dr. Dan Werthimer

and his School of SETI young experts for keeping alive the interplay between experimental and theoretical SETI. But the greatest “thanks” goes of course to the Teacher to all of us: Professor Frank Donald Drake, whose equation opened a new way of thinking about the past and the future of Humans in the Galaxy.

References

- [1] <http://en.wikipedia.org/wiki/Drake_equation>.
- [2] <<http://en.wikipedia.org/wiki/SETI>>.
- [3] <<http://en.wikipedia.org/wiki/Astrobiology>>.
- [4] <http://en.wikipedia.org/wiki/Frank_Drake>.
- [5] Athanasios Papoulis, S. Unnikrishna Pillai, in: *Probability, Random Variables and Stochastic Processes*, fourth edition, Tata McGraw-Hill, New Delhi, 2002 ISBN 0-07-048658-1.
- [6] <http://en.wikipedia.org/wiki/Gamma_distribution>.
- [7] <http://en.wikipedia.org/wiki/Central_limit_theorem>.
- [8] <<http://en.wikipedia.org/wiki/Cumulants>>.
- [9] <<http://en.wikipedia.org/wiki/Median>>.
- [10] Jeffrey Bennett, Seth Shostak, in: *Life in the Universe*, second edition, Pearson-Addison-Wesley, San Francisco, 2007 ISBN 0-8053-4753-4. See in particular page 404.
- [11] Claudio Maccone, The Statistical Drake Equation, Paper presented on October 1, 2008 at the 59th International Astronautical Congress (IAC) held in Glasgow, Scotland, UK, September 29–October 3, 2008. Paper #IAC-08-A4.1.4.