Acta Astronautica 115 (2015) 277–285

Contents lists available at ScienceDirect

Acta Astronautica

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

Statistical Drake–Seager Equation for exoplanet and SETI searches ☆

Claudio Maccone [a,b,c,*]

[a] International Academy of Astronautics (IAA), Via Martorelli 43, Torino (Turin) 10155, Italy
[b] SETI Permanent Committee of the IAA, Via Martorelli 43, Torino (Turin) 10155, Italy
[c] IASF-INAF Associate, Milan, Italy

ARTICLE INFO

Article history:
Received 14 November 2014
Received in revised form
29 April 2015
Accepted 1 May 2015
Available online 22 May 2015

Keywords:
Statistical Drake equation
Statistical Seager Equation
Lognormal probability densities

ABSTRACT

In 2013, MIT astrophysicist Sara Seager introduced what is now called the Seager Equation (Refs. [20,21]): it expresses the number N of exoplanets with detectable signs of life as the product of six factors: Ns=the number of stars observed, fQ=the fraction of stars that are quiet, fHZ=the fraction of stars with rocky planets in the Habitable Zone, fO=the fraction of those planets that can be observed, fL=the fraction that have life, fS=the fraction on which life produces a detectable signature gas. This we call the "classical Seager equation".

Now suppose that each input of that equation is a positive random variable, rather than a sheer positive number. As such, each input random variable has a positive mean value and a positive variance that we assume to be numerically known by scientists. This we call the "Statistical Seager Equation". Taking the logs of both sides of the Statistical Seager Equation, the latter is converted into an equation of the type log(N)=SUM of independent random variables.

Let us now consider the possibility that, in the future, the number of physical inputs considered by Seager when she proposed her equation will actually increase, since scientists will know more and more details about the astrophysics of exoplanets. In the limit for an infinite number of inputs, i.e. an infinite number of independent input random variables, the Central Limit Theorem (CLT) of Statistics applies to the Statistical Seager Equation. Thus, the probability density function (pdf) of the output random variable log(N) will approach a Gaussian (normal) distribution in the limit, whatever the distribution of the input random variables might possibly be. But if log(N) approaches the normal distribution, then N approaches the lognormal distribution, whose mean value is the sum of the input mean values and whose variance is the sum of the input variances.

This is just what this author realized back in 2008 when he transformed the Classical Drake Equation into the Statistical Drake Equation (Refs. [10,11]). This discovery led to much more related work in the following years (Refs. [12–19]).

In this paper we study the lognormal properties of the Statistical Seager Equation relating them to the present and future knowledge for exoplanets searches from both the ground and space.

☆ This paper was presented during the 65th IAC in Toronto.

* Corresponding author at: SETI Permanent Committee of the IAA, Via Martorelli 43, Torino (Turin) 10155, Italy.
E-mail addresses: clmaccon@libero.it, claudio.maccone@iaamail.org
URL: http://www.maccone.com/

http://dx.doi.org/10.1016/j.actaastro.2015.05.005
0094-5765/

1. Introduction

As we stated in the Abstract, the Seager equation is mathematically equivalent to the Drake equation well known in SETI, the Search for ExtraTerrestrial Intelligence. Actually, all equations simply made up by an output equal to the multiplication of some independent inputs are equivalent to the Drake equation. For instance, the Dole equation that applies to the number of habitable planets for man in the Galaxy, was studied by this author in Ref. [13] and in Chapter 3 of Ref. [15] exactly in the same mathematical way the Drake equation was studied earlier by him in Refs. [10,11], and the Seager equation is studied in this paper. More prosaically, all these equations simply are the Law of Compound Probability that everyone learns about in every course on elementary probability theory. Sara Seager herself modestly called her equation "an extended Drake equation". Therefore, we prefer to describe the following important transition from the classical equation to the statistical one with the language of SETI, and so we now introduce first the classical, and later the Statistical Drake equation.

2. The classical Drake equation (1961)

The Drake equation is a now famous result (see Ref. [1] for the Wikipedia summary) in the fields of SETI (the Search for ExtraTerrestial Intelligence, 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 Extra-Terrestrial 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:

$$N = Ns \cdot fp \cdot ne \cdot fl \cdot fi \cdot fc \cdot fL \tag{1}$$

where

1) Ns 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 2015, this means that they must be aware of the Maxwell equations governing radio waves, as well as of computers and radio astronomy (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 equation (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!

3. Transition from the classical to the Statistical Drake equation

We claim the classical Drake equation (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 input variable appearing on the right-hand side of* (1).

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

4. 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 1977–1979.

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

$$\begin{cases} D_1 = Ns \\ D_2 = fp \\ D_3 = ne \\ D_4 = fl \\ D_5 = f \\ D_6 = fc \\ D_7 = fL \end{cases} \tag{2}$$

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…

5. 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 equation (3) and change it into a sum:

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

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

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

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

$$N = e^{Y}. \tag{6}$$

We are now ready to take Step 3.

6. 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 2.

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

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

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 equation (3).

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: keep in mind that, according to (5), once we selected a particular type of probability density function (pdf) for the last seven D_i 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 an important 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], pages 130–131) allowing us to compute the pdf of a certain new random variable Y that is a known function $Y = g(X)$ of 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

$$Y = g(X) \tag{8}$$

can be calculated according to the following 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

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

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 ith 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. 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 much simpler and much more practical uniform distribution instead, as shown in Section 7.

7. 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

$$f_{\text{uniform\_}D_i} = \frac{1}{b_i - a_i} \quad \text{with} \quad 0 \leq a_i \leq x \leq b_i \tag{10}$$

that follows at once from the normalization condition

$$\int_{a_i}^{b_i} f_{\text{uniform\_}D_i}(x)\, dx = 1\,. \tag{11}$$

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

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

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

$$\langle \text{uniform\_}D_i \rangle = \frac{a_i + b_i}{2}. \tag{12}$$

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

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

The second moment of the uniform distribution is thus

$$\left\langle \text{uniform\_}{D_i}^2 \right\rangle = \frac{{a_i}^2 + a_i b_i + {b_i}^2}{3}. \tag{13}$$

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

$$\sigma^2_{\text{uniform\_}D_i} = \left\langle \text{uniform\_}{D_i}^2 \right\rangle - \langle \text{uniform\_}D_i \rangle^2$$
$$= \frac{{a_i}^2 + a_i b_i + {b_i}^2}{3} - \frac{(a_i + b_i)^2}{4} = \frac{(b_i - a_i)^2}{12}. \tag{14}$$

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

$$\sigma_{\text{uniform\_}D_i} = \frac{b_i - a_i}{2\sqrt{3}}. \tag{15}$$

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)

$$\begin{cases} \langle \text{uniform\_}D_i \rangle = \frac{a_i + b_i}{2} \\ \sigma_{\text{uniform\_}D_i} = \frac{b_i - a_i}{2\sqrt{3}}. \end{cases} \tag{16}$$

Upon inverting this trivial linear system, one finds

$$\begin{cases} a_i = \langle \text{uniform\_}D_i \rangle - \sqrt{3}\,\sigma_{\text{uniform\_}D_i} \\ b_i = \langle \text{uniform\_}D_i \rangle + \sqrt{3}\,\sigma_{\text{uniform\_}D_i}. \end{cases} \tag{17}$$

This is of paramount importance for our application the Statistical Drake equation inasmuch 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 Eq. (17), respectively.*

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

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

8. Step 5: computing the logs of the 7 uniformly distributed Drake random variables D_i

Intuitively speaking, the natural log of a uniformly distributed random variable *may not* be 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

$$Y_i = \ln(D_i) \quad i = 1, \ldots, 7 \tag{18}$$

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

$$y = g(x) = \ln(x). \tag{19}$$

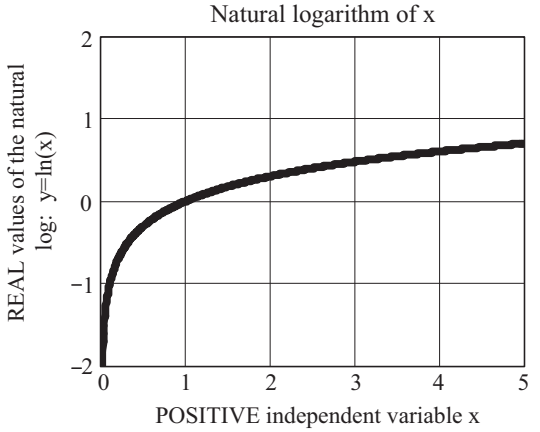

Fig. 1. The simple function $y = \ln(x)$.

That, upon inversion, yields the *single* root

$$x_1(y) = x(y) = e^y. \tag{20}$$

On the other hand, differentiating (19) one gets

$$g'(x) = \frac{1}{x} \quad \text{and} \quad g'(x_1(y)) = \frac{1}{x_1(y)} = \frac{1}{e^y} \tag{21}$$

where (20) was already used in the last step. By virtue of the uniform probability density function (10) and of (21), the general transformation law (9) finally yields

$$f_Y(y) = \sum_i \frac{f_X(x_i(y))}{|g'(x_i(y))|} = \frac{1}{b_i - a_i} \cdot \frac{1}{\left|\frac{1}{e^y}\right|} = \frac{e^y}{b_i - a_i}. \tag{22}$$

In other words, the requested pdf of Y_i is

$$f_Y(y) = \frac{e^y}{b_i - a_i} \quad i = 1, \ldots, 7 \quad \text{with} \quad \ln(a_i) \le y \le \ln(b_i). \tag{23}$$

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 $\ln(a_i) \le y \le \ln(b_i)$, as for every pdf, and it is easy to see that its normalization condition is fulfilled:

$$\int_{\ln(a_i)}^{\ln(b_i)} f_Y(y)dy = \int_{\ln(a_i)}^{\ln(b_i)} \frac{e^y}{b_i - a_i} dy = \frac{e^{\ln(b_i)} - e^{\ln(a_i)}}{b_i - a_i} = 1. \tag{24}$$

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 $\langle Y_i \rangle$ is given by

$$\langle Y_i \rangle = \int_{\ln(a_i)}^{\ln(b_i)} y \cdot f_Y(y)dy = \int_{\ln(a_i)}^{\ln(b_i)} \frac{y \cdot e^y}{b_i - a_i} dy = \frac{b_i\left[\ln(b_i) - 1\right] - a_i\left[\ln(a_i) - 1\right]}{b_i - a_i}. \tag{25}$$

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

$$\langle Y_i \rangle = \langle \ln(D_i) \rangle = \frac{b_i\left[\ln(b_i) - 1\right] - a_i\left[\ln(a_i) - 1\right]}{b_i - a_i}. \tag{26}$$

In order to find the variance also, we must first compute the mean value of the square of Y_i, that is

$$\left\langle Y_i^2 \right\rangle = \int_{\ln(a_i)}^{\ln(b_i)} y^2 \cdot f_Y(y)dy = \int_{\ln(a_i)}^{\ln(b_i)} \frac{y^2 \cdot e^y}{b_i - a_i} dy = \frac{b_i\left[\ln^2(b_i) - 2\ln(b_i) + 2\right] - a_i\left[\ln^2(a_i) - 2\ln(a_i) + 2\right]}{b_i - a_i}. \tag{27}$$

The *variance of* $Y_i = \ln(D_i)$ is now given by (27) minus the square of (26), that, after a few reductions, yield:

$$\sigma_{Y_i}^2 = \sigma_{\ln(D_i)}^2 = 1 - \frac{a_i b_i \left[\ln(b_i) - \ln(a_i)\right]^2}{(b_i - a_i)^2}. \tag{28}$$

Whence the corresponding standard deviation

$$\sigma_{Y_i} = \sigma_{\ln(D_i)} = \sqrt{1 - \frac{a_i b_i \left[\ln(b_i) - \ln(a_i)\right]^2}{(b_i - a_i)^2}}. \tag{29}$$

Let us now turn to 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 $j = \sqrt{-1}$)

$$\Phi_{Y_i}(\zeta) = \int_{-\infty}^{\infty} e^{j\zeta y} f_{Y_i}(y)\ dy. \tag{30}$$

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 9. 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 equation (3) and so we are now led to find the characteristic function of the random variable Y_i, i.e.

$$\begin{aligned}\Phi_{Y_i}(\zeta) &= \int_{-\infty}^{\infty} e^{j\zeta y} f_{Y_i}(y)\ dy = \int_{\ln(a_i)}^{\ln(b_i)} e^{j\zeta y} \frac{e^y}{b_i - a_i}\ dy \\ &= \frac{1}{b_i - a_i} \int_{\ln(a_i)}^{\ln(b_i)} e^{(1+j\zeta)y}\ dy = \frac{1}{b_i - a_i} \cdot \frac{1}{1 + j\zeta} \left[e^{(1+j\zeta)y}\right]_{n(a_i)}^{\ln(b_i)} \\ &= \frac{e^{(1+j\zeta)\ln(b_i)} - e^{(1+j\zeta)\ln(a_i)}}{(b_i - a_i)(1 + j\zeta)} = \frac{b_i^{1+j\zeta} - a^{1+j\zeta}}{(b_i - a_i)(1 + j\zeta)}.\end{aligned} \tag{31}$$

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

$$\Phi_{Y_i}(\zeta) = \frac{b_i^{1+j\zeta} - a^{1+j\zeta}}{(b_i - a_i)(1 + j\zeta)}. \tag{32}$$

9. 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

$$f_N(y;\mu,\sigma)=\frac{1}{y}\cdot\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(\ln(y)-\mu)^2}{2\,\sigma^2}}\quad(y\geq 0,\sigma\geq 0)\tag{33}$$

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 (CLT) 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 equation (3). The idea is to apply the CLT to the sum of random variables given by (4) and (5) *whatever their probability distributions can possibly be*. In other words, the CLT applied to the Statistical Drake equation (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

$$\begin{cases} Y=\sum_{i=1}^{7}Y_i \\ \langle Y\rangle=\sum_{i=1}^{7}\langle Y_i\rangle \\ \sigma_Y^2=\sum_{i=1}^{7}\sigma_{Y_i}^2. \end{cases}\tag{34}$$

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

10. 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 (6), tends to a normal random variable, and, as a consequence, it follows that the base random variable, like N in (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), that is, what is the probability density function of the lognormal distribution? To find it, set

$$y=g(x)=e^x.\tag{35}$$

This, upon inversion, yields the *single* root

$$x_1(y)=x(y)=\ln(y).\tag{36}$$

On the other hand, differentiating (35) one gets

$$g'(x)=e^x\quad\text{and}\quad g'(x_1(y))=e^{\ln(y)}=y.\tag{37}$$

where (21) was already used in the last step. The general transformation law (9) finally yields

$$f_N(y)=\sum_i\frac{f_X(x_i(y))}{|g'(x_i(y))|}=\frac{1}{|y|}f_Y(\ln(y)).\tag{38}$$

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. (33) is found, and the derivation of the lognormal distribution from the normal distribution is proved.

Table 1 summarizes all the properties of the lognormal distribution, whose demonstrations may be found in statistical textbooks (for instance see refs. [7–9]). The last two lines in Table 1, however, are about our own discovery of Eqs. (26) and (28) yielding μ and σ^2 of the lognormal distribution of the output of the statistical Drake (and Seager) equation when all inputs are supposed to be uniformly distributed and both the mean value and standard deviation of each input is assigned. Remember that this mean value and standard deviation may be immediately converted into the uniform distribution lower and upper values thanks to the two Eq. (17).

11. 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 becomes known to scientists. This capability to make room for more future factors in the statistical Drake equation we call "Data Enrichment Principle", and we regard it as the key to more profound future results in the fields of Astrobiology ad SETI.

Table 1
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	$f_N(n;\mu,\sigma)=\frac{1}{n}\cdot\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(\ln(n)-\mu)^2}{2\,\sigma^2}}\quad(n\geq 0,\sigma\geq 0)$
Mean value	$\langle N\rangle = e^{\mu}\,e^{\frac{\sigma^2}{2}}$
Variance	$\sigma_N^2 = e^{2\mu}\,e^{\sigma^2}\left(e^{\sigma^2}-1\right)$
Standard deviation	$\sigma_N = e^{\mu}\,e^{\frac{\sigma^2}{2}}\sqrt{e^{\sigma^2}-1}$
All the moments, i.e. kth moment	$\left\langle N^k\right\rangle = e^{k\mu}\,e^{k^2\frac{\sigma^2}{2}}$
Mode (=abscissa of the lognormal peak)	$n_{\mathrm{mod}\ e}=n_{peak}=e^{\mu}\,e^{-\ \sigma^2}$
Value of the mode peak	$f_N(n_{\mathrm{mode}})=\frac{1}{\sqrt{2\pi}\,\sigma}\cdot e^{-\mu}\cdot e^{\frac{\sigma^2}{2}}$
Median (=fifty–fifty probability value for N)	$\mathrm{median}=e^{\mu}$
Skewness	$\frac{K_3}{(K_2)^{\frac{3}{2}}}=\sqrt{e^{\sigma^2}-1}\left(e^{\sigma^2}+2\right)$
Kurtosis	$\frac{K_4}{(K_2)^2}=e^{4\,\sigma^2}+2\,e^{3\,\sigma^2}+3\,e^{2\,\sigma^2}-6$
Expression of μ in terms of the lower (a_i) and upper (b_i) limits of the Drake *uniform* input random variables D_i	$\mu=\sum_{i=1}^{7}\langle Y_i\rangle=\sum_{i=1}^{7}\left(\frac{b_i\left[\ln(b_i)-1\right]-a_i\left[\ln(a_i)-1\right]}{b_i-a_i}\right).$
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	$\sigma^2=\sum_{i=1}^{7}\sigma_{Y_i}^2=\sum_{i=1}^{7}\left(1-\frac{a_i\,b_i\left[\ln(b_i)-\ln(a_i)\right]^2}{(b_i-a_i)^2}\right).$

12. The Statistical Seager Equation

We now come to consider the Statistical Seager Equation. As described in the Abstract of this paper, in 2013, MIT astrophysicist Sara Seager introduced what is now called the Seager equation (see Refs. [20,21]): it expresses the number N of exoplanets with detectable signs of life as the product of six factors:

1) Ns = the number of stars observed,
2) fQ = the fraction of stars that are quiet,
3) fHZ = the fraction of stars with rocky planets in the Habitable Zone,
4) fO = the fraction of those planets that can be observed,
5) fL = the fraction that have life,
6) fS = the fraction on which life produces a detectable signature gas.

That is

$$N = Ns\cdot fQ\cdot fHZ\cdot fO\cdot fL\cdot fS \tag{39}$$

This we call the "classical Seager equation".

Mathematically speaking, Eq. (39) is exactly the same thing as Eq. (1): only the words change, and, of course, so does its scientific meaning.

Mathematically, one may thus immediately apply to (39) the full string of mathematical theorems that we described in all Eqs. (2)–(32) of this paper.

The first such step is clearly the transformation of the classical Seager equation (39) into the Statistical Seager Equation (looking the same, mathematically), having all numeric inputs replaced by positive random variables. No more comments are necessary.

The second step is asking the two questions:

1) What is the probability distribution of each of the six input positive random variables in (39)?
2) And what is the probability distribution of the resulting output N?

Our answers to these two questions are:

1) No analytical solution exists as long as the number of Input random variables is finite. Only numeric calculations may be done, but this requires writing a numeric code, that this author could not and would not write down because is he a mathematical physicist and not a computer programmer.
2) However, if we let the number of input random variables approach infinity (i.e. if we consider a high number of input random variables, like five to ten or even more ("how many" might be discussed later) then the solution of both the statistical Drake equation (1) and the Statistical Seager Equation (3) is immediate. In both cases the probability distribution of the output random variable N is a Lognormal Distribution:

a) Its real parameter μ is given by the sum of the mean values of all input random variables, and
b) Its positive parameter σ^2 is given by the sum of the variances of all input random variables, whatever their probability distribution might possibly be.

This is the result of applying the Central Limit Theorem of Statistics to both the Drake and Seager equations.

13. The extremely important particular case when the input random variables are uniform

Now we turn to the particular case of our theory when all the input random variables are uniform random variables. This case we regard as "extremely important" for all practical applications of both the Drake and the Seager statistical equations inasmuch as it is the only case when analytic formulae do exist expressing the two lognormal

parameters μ and σ directly in terms of the lower and upper limits of all the uniform input random variables.

In fact, define by

1) a_i the real and positive number representing the LOWER LIMIT of the range of the ith uniform input random variable.
2) b_i the real and positive number representing the UPPER LIMIT of the range of the ith uniform input random variable.
3) Clearly, it is assumed $b_i > a_i$.

Then

4) The mean value of the ith uniform input random variable is (obviously) given by (12), that is

$$\langle \text{Uniform\_}D_i \rangle = \frac{a_i + b_i}{2} \tag{40}$$

5) The standard deviation of the ith uniform input random variable is (less obviously) given by (15), that is

$$\sigma_{\text{Uniform\_}D_i} = \frac{b_i - a_i}{2\sqrt{3}}. \tag{41}$$

Above all, one has

6) The probability density function (pdf) of the output random variable N is the lognormal pdf (33), that is:

$$f_N(n) = \frac{1}{n} \cdot \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(\ln(n)-\mu)^2}{2\,\sigma^2}} \quad (n \geq 0) \tag{42}$$

7) The real and positive parameter μ appearing in the lognormal pdf (33) is expressed directly in terms of all the known a_i and b_i by Eq. (26), that is

$$\mu = \sum_{i=1}^{\text{number\_of\_inputs}} \left(\frac{b_i \,[\ln(b_i) - 1] - a_i \,[\ln(a_i) - 1]}{b_i - a_i} \right). \tag{43}$$

8) The real and positive parameter σ^2 appearing in the lognormal pdf (33) is expressed directly in terms of all the known a_i and b_i by Eq. (28), that is

$$\sigma^2 = \sum_{i=1}^{\text{number\_of\_inputs}} \left(1 - \frac{a_i\, b_i \,[\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2} \right). \tag{44}$$

9) In addition to providing the pdf of N explicitly by virtue of the three Eqs. (42)–(44), this case of the all-uniform input random variables also is "realistic" in that the uniform probability distribution is "the most uncertain one" in the sense of Shannon's Information Theory. This is intuitively obvious ("when you do not know where to go, you look around and all directions are equal to each other, i.e. your probability distribution in the azimuth is uniform between 0 and 2 π). But it may also be rigorously proven by applying the method of the Lagrange multipliers to the definition of Shannon's Entropy, as shown in the Appendix to this paper.
10) All the statistical properties of the lognormal output random variable N may be found analytically and are listed in Table 1.

This completes our summary of the discoveries that this author made back in 2008 about the Statistical Drake Equation and so necessarily also about the 2013 Statistical Seager Equation also. Especially useful is the simple case when all input variables are Uniformly distributed: then, the lognormal pdf of N is given by Eqs. (42)–(44).

14. Conclusion

The conclusion of this paper is that this author would respectfully advise Professor Sara Seager and her MIT Team to make use of Eqs. (42)–(44) in order to find the lognormal pdf of the number N of exoplanets with detectable signs of life, having previously estimated all the a_i and b_i numerically.

This research work could be part of the Phase A or B studies for the NASA planned TESS space mission, described at the web sites ⟨http://tess.gsfc.nasa.gov/⟩ and ⟨http://en.wikipedia.org/wiki/Transiting_Exoplanet_Survey_Satellite⟩).

Acknowledgents

The author wishes to acknowledge the cooperation and support of Professor Sara Seager and her Team about the two scientific meetings they had in 2014:

1) At the Istituto Nazionale di Astrofisica (INAF) in Milan, Italy, when they met on April 1st, 2014, and
2) At MIT, when they met on May 5th, 2014.

Appendix

Proof. of Shannon's 1948 Theorem stating that the Uniform distribution is the "most uncertain" one over any Finite range of values.

As it is well known, the Shannon entropy of any probability density function $p(x)$ is given by the integral

$$\text{Shannon\_Entropy\_of\_}p(x) = -\int_{-\infty}^{\infty} p(x) \log p(x)\; dx. \tag{45}$$

In modern textbooks this is also called Shannon differential entropy.

Now consider the case when a probability density function $p(x)$ is limited to a finite interval $a \leq x \leq b$. This is obviously the case with any physical positive random variable, such as the number N of extraterrestrial communicating civilizations in the Galaxy. *We now wish to prove that for any such finite random variable the maximum entropy distribution is the UNIFORM distribution over* $a \leq x \leq b$.

Shannon did not bother to prove this simple theorem in his 1948 papers since he probably regarded it as just too trivial. But we prefer to point out this theorem since, in the language of the statistical Drake equation, it sounds like: "Since we don't know what the probability distribution of any one of the Drake random variables D_i is, it is safer to assume that each of them has the maximum possible

entropy over $a \le x \le b$, i.e., that D_i is UNIFORMLY distributed there".

The proof of this theorem is as follows:

1) Start by assuming $a_i \le x \le b_i$.
2) Then form the linear combination of the entropy integral plus the normalization condition for D_i

$$\delta \int_{a_i}^{b_i} \left[-p(x)\log p(x) + \lambda p(x) \right] \, dx = 0 \tag{46}$$

where λ is a Lagrange multiplier.

Performing the variation, i.e. differentiating with respect to $p(x)$, one finds

$$-\log p(x) - 1 + \lambda = 0 \tag{47}$$

that is

$$p(x) = e^{\lambda - 1}. \tag{48}$$

Applying the normalization condition (constraint) to the last expression for $p(x)$ yields

$$1 = \int_{a_i}^{b_i} p(x) \, dx = \int_{a_i}^{b_i} e^{\lambda - 1} \, dx = e^{\lambda - 1} \int_{a_i}^{b_i} dx = e^{\lambda - 1}(b_i - a_i) \tag{49}$$

that is

$$e^{\lambda - 1} = \frac{1}{b_i - a_i} \tag{50}$$

and finally, from (48) and (50)

$$p(x) = \frac{1}{b_i - a_i} \quad \text{with} \quad a_i \le x \le b_i. \tag{51}$$

showing that the maximum-entropy probability distribution over any FINITE interval $a_i \le x \le b_i$ is just the UNIFORM distribution.

References

[1] ⟨http://en.wikipedia.org/wiki/Drake_equation⟩.
[2] ⟨http://en.wikipedia.org/wiki/Search_for_Extra-Terrestrial_Intelligence⟩.
[3] ⟨http://en.wikipedia.org/wiki/Astrobiology⟩.
[4] ⟨http://en.wikipedia.org/wiki/Frank_Drake⟩.
[5] A. Papoulis, S.U. Pillai, Probability, Random Variables and Stochastic Processes, fourth ed. Tata McGraw-Hill, New Delhi, ISBN 0-07-048658-1, 2002.
[6] ⟨http://en.wikipedia.org/wiki/Gamma_distribution⟩.
[7] ⟨http://en.wikipedia.org/wiki/Central_limit_theorem⟩.
[8] ⟨http://en.wikipedia.org/wiki/Cumulant⟩.
[9] ⟨http://en.wikipedia.org/wiki/Median⟩.
[10] C. Maccone. (2008, The Statistical Drake Equation, Paper #IAC-08-A4.1.4 presented on 1st October, 2008, at the 59th International Astronautical Congress (IAC), Glasgow, Scotland, UK, 29 September–3 October, 2008.
[11] C. Maccone, The Statistical Drake Equation, Acta Astronaut. 67 (2010) (2010) 1366–1383.
[12] C. Maccone, The Statistical Fermi Paradox, J. Br. Interplanet. Soc. 63 (2010) (2010) 222–239.
[13] C. Maccone, SETI and SEH (statistical equation for habitables), Acta Astronaut. 68 (2011) (2011) 63–75.
[14] C. Maccone, A mathematical model for evolution and SETI, Orig. Life Evolut. Biosph. (OLEB) 41 (2011) 609–619. Available online 3rd December, 2011.
[15] C. Maccone, 2012, Mathematical SETI, a 724-pages book published by Praxis–Springer in the fall of 2012. ISBN-10: 3642274366; ISBN-13: 978-3642274367, edition: 2012.
[16] C. Maccone, SETI, evolution and human history merged into a mathematical model Available online since April 23, 2013, Int. J. Astrobiol. 12 (3) (2013) 218–245.
[17] C. Maccone, Evolution and history in a new "Mathematical SETI" model, Acta Astronaut. 93 (2014) 317–344. Available online since 13 August 2013.
[18] C. Maccone, SETI as a part of big history, Acta Astronaut. 101 (2014) 67–80.
[19] C. Maccone, Evolution and mass extinctions as lognormal stochastic processes, Int. J. Astrobiol. 13 (4) (2014) 290–309.
[20] ⟨http://en.wikipedia.org/wiki/Sara_Seager#Seager_equation⟩.
[21] P. Gilster, Astrobiology: Enter the Seager Equation, Centauri Dreams, 11 September 2013, ⟨http://www.centauri-dreams.org/?p=28976⟩.