# Defense Intelligence Reference Document

[redacted]

*Acquisition Threat Support*

11 March 2010

ICOD: 1 December 2009

# An Introduction to the Statistical Drake Equation

# An Introduction to the Statistical Drake Equation

**Prepared by:**

(b)(3):10 USC 424

**Defense Intelligence Agency**

**Author:**

(b)(6)

**Administrative Note**



This product is one in a series of advanced technology reports produced in FY 2009 under the Defense Intelligence Agency, (b)(3):10 USC 424 Advanced Aerospace Weapon System Applications (AAWSA) Program. Comments or questions pertaining to this document should be addressed to (b)(3):10 USC 424;(b)(6), AAWSA Program Manager, Defense Intelligence Agency, ATTN: (b)(3):10 USC 424 Bldg 6000, Washington, DC 20340-5100.

# Contents

# An Introduction to the Statistical Drake Equation

## 1. Introduction

**SETI (an acronym for "Search for Extraterrestrial Intelligence") is a relatively new branch of scientific research, having begun only in 1959. Its goal is to ascertain whether alien civilizations exist in the universe, how far from us they exist, and possibly how much more advanced than us they may be.**

**As of 2009, the only physical tools we know that could help us get in touch with aliens are the electromagnetic waves an alien civilization could emit and we could detect. This forces us to use the largest radiotelescopes on Earth for SETI research, because the higher our collecting area of electromagnetic radiation is, the higher our sensitivity is (that is, the farther in space we can probe). Yet, even by using the largest radiotelescopes on Earth (the 310-meter dish at Arecibo, for instance), we cannot search for aliens beyond, say, a few hundred light years away. This is a very, very small amount of space around us within our galaxy, the Milky Way, that is about 100,000 light years in diameter. Thus, current SETI can cover only a very tiny fraction of the galaxy, and it is not surprising that in the past 50 years of SETI searches, NO extraterrestrial civilization was discovered. Quite simply, we did not get far enough!**

**This demands the construction of much more powerful and radically new radiotelescopes. Rather than big and heavy metal dishes, whose mechanical problems hamper SETI research too much, we are now turning to "software radiotelescopes," where a large number of small dishes (ATA = Allen Telescope Array, and ALMA = Atacama Large Millimeter/submillimeter Array) or even just of simple dipoles (LOFAR = Low Frequency Array) using state-of-the-art electronics and very-high-speed computing can outperform the classical radiotelescopes in many regards. The final dream in this field is the SKA (= Square Kilometer Array), currently being designed and expected to be completed around 2020.**

## 2. The Key Question: How Far are They ?

But still, the key question remains: how far are they?

Or, more correctly, how far do we expect the NEAREST extraterrestrial civilization to be from the Solar System in the galaxy?

This question was first faced in a scientific manner back in 1961 by the same scientist who also was the first experimental SETI radio astronomer ever: the American, Frank Donald Drake (born 1930). He first considered the shape and size of the galaxy where we are living: the Milky Way. This is a spiral galaxy measuring some 100,000 light years in diameter and some 16,000 light years in thickness of the Galactic Disk at half-way from its center. That is:

The diameter of the galaxy is (about) 100,000 light years, (abbreviated ly) i.e., its radius, $R_{Galaxy}$, is about 50,000 ly.

The thickness of the Galactic Disk at half-way from its center, $h_{Galaxy}$, is about 16,000 ly.

The volume of the galaxy may then be approximated as the volume of the corresponding cylinder, i.e.

$$V_{Galaxy} = \pi R_{Galaxy}^2 h \,. \tag{1}$$

Now consider the sphere around us having a radius *r*. The volume of such a sphere is

$$V_{Our\_Sphere} = \frac{4}{3}\pi\left(\frac{\text{ET\_Distance}}{2}\right)^3 \tag{2}$$

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 Neighborhood. 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 a weak point in the reasoning.

Furthermore, let us denote by *N* the total number of civilizations now living in the galaxy, including ourselves. Of course, this number *N* is unknown. We only know that $N \geq 1$ since one civilization does at least exist!

Having thus assumed that ET civilizations are UNIFORMLY SPACED IN THE GALAXY, we can then write down the proportion:

$$\frac{V_{Galaxy}}{N} = \frac{V_{Our\_Sphere}}{1} \,. \tag{3}$$

That is, upon replacing both (1) and (2) into (3):

$$\frac{\pi R_{Galaxy}^2 h}{N} = \frac{\frac{4}{3}\pi\left(\frac{\text{ET\_Distance}}{2}\right)^3}{1} \,. \tag{4}$$

The last equation contains two unknowns: *N* and ET_Distance, and so we don't know which one it is better to solve for.

However, we may suppose that, by resorting to the (rather uncertain) knowledge that we have about the Evolution of the galaxy through the last 10 billion years or so, we might somehow compute an approximate value for *N*.

Then, we may solve (4) for ET_Distance thus obtaining the (AVERAGE) DISTANCE BETWEEN ANY PAIR OF NEIGHBORING CIVILIZATIONS IN THE GALAXY (DISTANCE LAW)

$$\text{ET\_Distance}(N) = \frac{\sqrt[3]{6\,R_{Galaxy}^2\,h}}{\sqrt[3]{N}} = \frac{C}{\sqrt[3]{N}} \tag{5}$$

where the positive constant *C* is defined by

$$C = \sqrt[3]{6\,R_{Galaxy}^2\,h_{Galaxy}} \approx 28845 \text{ light years}. \tag{6}$$

Equations (5) and (6) are the starting point to understand the origin of the Drake equation that we discuss in detail in Section 3 of this paper.

Let us just complete this section by pointing out three different numerical cases of the distance law (5):

- We know that we exist, so *N* may not be smaller than 1, i.e., $N \geq 1$. Suppose then that we are alone in the galaxy, i.e., that *N*=**1**. Then the distance law (5) yields as distance to the nearest civilization from us just the constant C, i.e., 28,845 light years. This is about the distance in between ourselves and the center of the galaxy (i.e. the Galactic Bulge). Thus, this result seems to suggest that, if we do not find any extraterrestrial civilization around us in these outskirts of the galaxy where we live, we should look around the Galactic Center first. And this is indeed what is happening, i.e., many SETI searches are actually pointing the antennas towards the Galactic Center, looking for beacons (see, for instance ref. [**1**]).

- Suppose next that *N*=**1000**, i.e. there are about a thousand extraterrestrial communicating civilizations in the whole galaxy right now. Then the distance law (5) yields an average distance of 2,885 light years. This is a distance that most radiotelescopes in Earth may not reach for SETI searches right now: hence the need to build larger radiotelescopes, like ALMA, LOFAR and the SKA.

- Suppose finally that *N*=**1000000**, i.e., there are a million communicating civilizations now in the galaxy. Then the distance law (5) yields an average distance of 288 light years. This is within the (upper) range of distances that our current radiotelescopes may reach for SETI searches, and that justifies all SETI searches that have been done so far in the first fifty years of SETI (1960-2010).

In conclusion, interpolating the above three special cases of *N*, we may say that the distance law (5) yields the following key diagram of the average ET distance vs. the assumed number of communicating civilizations, *N*, in the galaxy right now (Figure 1):

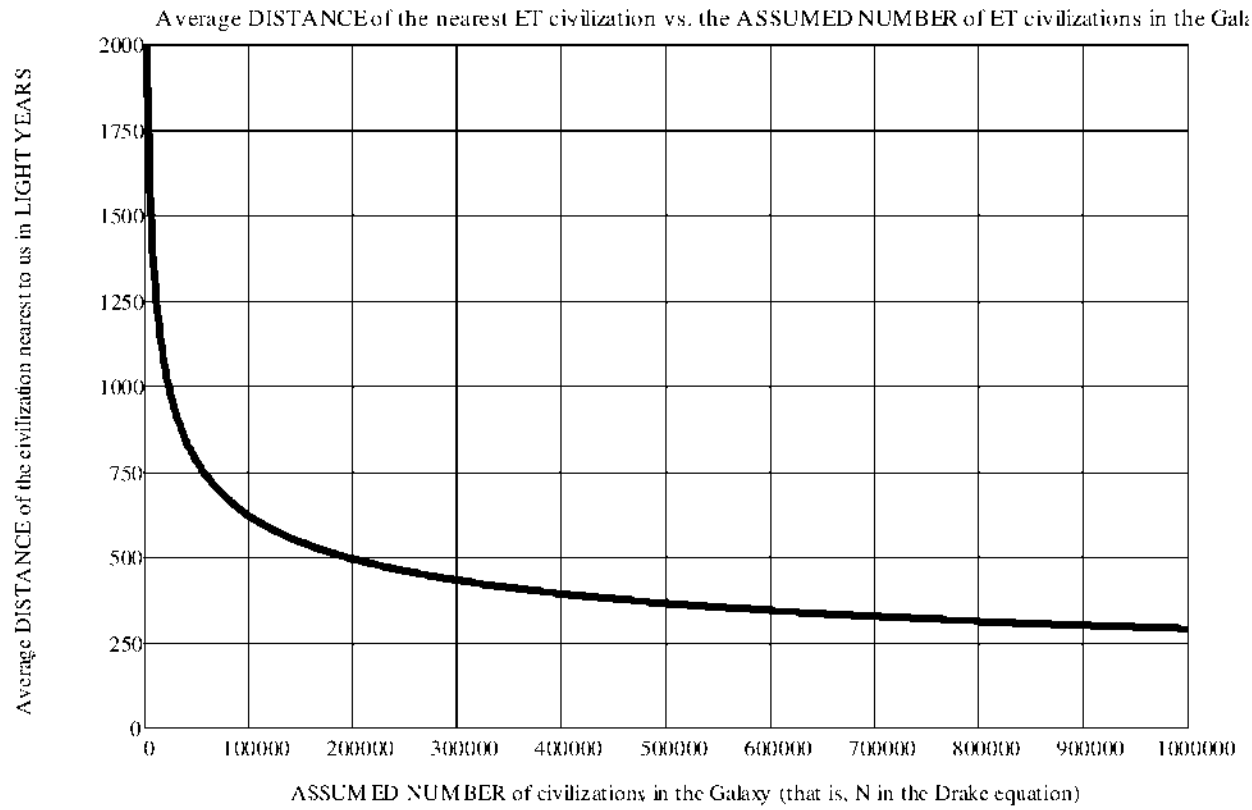



**Figure 1. DISTANCE LAW; i.e., the Average Distance (plot along the vertical axis in light years) Versus the NUMBER of Communicating Civilizations ASSUMED to Exist in the Galaxy Right Now**

## 3. Computing *N* By Virtue of the Drake Equation (1961)

In the previous section, the problem of finding how close the nearest ET civilization may be was "solved" by reducing it to the computation of *N*, the total number of extraterrestrial civilizations now existing in this galaxy. In this section the famous Drake equation is described, that was proposed back in 1961 by Frank Donald Drake (born 1930) to estimate the numerical value of *N*. We believe that no better introductory description of the Drake equations exists other than the one given by Carl Sagan in his 1983 book "Cosmos" (ref. [2]), in its turn based on the famous TV series "Cosmos." So, in this paragraph we report Carl Sagan's description of the Drake equation unabridged.

"But is there anyone out there to talk to? With a third or a half a trillion stars in our Milky Way galaxy alone, could ours be the only one accompanied by an inhabited planet? How much more likely it is that technical civilizations are a cosmic commonplace, that the galaxy is pulsing and humming with advanced societies, and, therefore, that the nearest such culture is not so very far away – perhaps transmitting from antennas established on a planet of a naked-eye star just next door. Perhaps when we look up at the sky at night, near one of those faint pinpoints of light is a world on which someone quite different from us is then glancing idly at a star *we* call the Sun and entertaining, for just a moment, an outrageous speculation.

It is very hard to be sure. There may be several impediments to the evolution of a technical civilization. Planets may be rarer than we think. Perhaps the origin of life is not so easy as our laboratory experiments suggest. Perhaps the evolution of advanced life forms is improbable. Or it may be that complex life forms evolve more readily, but intelligence and technical societies require an unlikely set of coincidences – just as the evolution of the human species depended on the demise of the dinosaurs and the ice-age recession of the forests in whose trees our ancestors screeched and dimly wondered. Or perhaps civilizations arise repeatedly, inexorably, on innumerable planets in the Milky Way, but are generally unstable; so all but a tiny fraction are unable to survive their technology and succumb to greed and ignorance, pollution and nuclear war.

It is possible to explore this great issue further and make a crude estimate of N, the number of advanced civilizations in the galaxy. We define an advanced civilization as one capable of radio astronomy. This is, of course, a parochial if essential definition. There may be countless worlds on which the inhabitants are accomplished linguists or superb poets but indifferent radio astronomers. We will not hear from them. N can be written as the product or multiplication of a number of factors, each a kind of filter, every one of which must be sizable for there to be a large number of civilizations:

- *Ns*, the number of stars in the Milky Way galaxy.
- *fp*, the fraction of stars that have planetary systems.
- *ne*, the number of planets in a given system that are ecologically suitable for life.
- *fl*, the fraction of otherwise suitable planets on which life actually arises.
- *fi*, the fraction of inhabited planets on which an intelligent form of life evolves.
- *fc*, the fraction of planets inhabited by intelligent beings on which a communicative technical civilization develops.
- *fL*, the fraction of planetary lifetime graced by a technical civilization.

Written out, the equation reads

$$N = Ns \cdot fp \cdot ne \cdot fl \cdot fi \cdot fc \cdot fL \qquad (7)$$

All of the *f*'s are fractions, having values between 0 and 1; they will pare down the large value of *Ns*.

To derive *N* we must estimate each of these quantities. We know a fair amount about the early factors in the equation, the number of stars and planetary systems. We know very little about the later factors, concerning the evolution of intelligence or the lifetime of technical societies. In these cases our estimates will be little better than guesses. I invite you, if you disagree with my estimates below, make your own choices and see what implications your alternative suggestions have for the number of advanced civilizations in the galaxy. One of the great virtues of this equation, due to Frank Drake of Cornell, is that it involves subjects ranging from stellar and planetary astronomy to organic chemistry, evolutionary biology, history, politics and abnormal psychology. Much of the Cosmos is in the span of the Drake equation.

We know *Ns*, the number of stars in the Milky Way galaxy, fairly well, by careful counts of stars in a small but representative region of the sky. It is a few hundred billion; some recent estimates place it at $4 \times 10^{11}$. Very few of these stars are of the massive short-lived variety that squander their reserves of thermonuclear fuel. The great majority have lifetimes of billions or more years in which they are shining stably, providing a suitable energy source for the energy and evolution of life on nearby planets.

There is evidence that planets are a frequent accompaniment of star formation: in the satellite systems of Jupiter, Saturn and Uranus, which are like miniature solar systems; in theories of the origin of the planets; in studies of double stars; in observations of accretion disks around stars; and is some preliminary investigations of gravitational perturbations of nearby stars.[1] Many, perhaps even most, stars may have planets. We take the fraction of stars that have planets, *fp*, as roughly equal to 1/3. Then the total number of planetary systems in the galaxy would be $Ns\ fp \sim 1.3 \times 10^{11}$ (the symbol ~ means "approximately equal to"). If each system were to have about ten planets, as ours does, the total number of worlds in the galaxy would be more than a trillion, a vast arena for the cosmic drama.

In our own solar system there are several bodies that may be suitable for life of some sort: the Earth certainly, and perhaps Mars, Titan and Jupiter. Once life originates, it tends to be very adaptable and tenacious. There must be many different environments suitable for life in a given planetary system. But conservatively we choose $ne=2$. Then the number of planets in the galaxy suitable for life becomes $Ns\ fp\ ne \sim 3 \times 10^{11}$.

Experiments show that under the most common cosmic conditions the molecular basis of life is readily made, the building blocks of molecules able to make copies of themselves. We are now on less certain grounds; there may, for example, be impediments in the evolution of the genetic code, although I think this is unlikely over billions of years of primeval chemistry. We choose $fl \sim 1/3$, implying a total number of planets in the Milky Way on which life has arisen at least once as $Ns\ fp\ ne\ fl \sim 1 \times 10^{11}$, a hundred billion inhabited worlds. That in itself is a remarkable conclusion. But we are not yet finished.

The choices of *fi* and *fc* are more difficult. On the one hand, many individually unlikely steps had to occur in biological evolution and human history for our present intelligence and technology to develop. On the other hand, there must be quite different pathways to an advanced civilization of specified capabilities. Considering the apparent difficulty in the evolution of large organisms, represented by the Cambrian explosion, let us choose $fi \times fc = 1/100$, meaning that only 1 per cent of planets on which life arises actually produce a technical civilization. This estimate represents some middle ground among the varying scientific options. Some think that the equivalent of the step from the emergence of trilobites to the domestication of fire goes like a shot in all planetary systems; others think that, even given ten or fifteen billion years, the evolution of a technical civilization is unlikely. This is not a subject on which we can do much experimentation as long as our investigations are limited to a single planet. Multiplying

[1] Carl Sagan was writings these lines back in the 1970's, when no extrasolar planets had been discovered yet. The first such discovery occurred in 1995, when Michel Mayor and Didier Queloz, working at the "Observatoire de Haute Provence" in France, discovered the first extrasolar planet orbiting the nearby star 51 Peg. This first extrasolar planet was hence named 51 Peg B. Many more extrasolar planets were discovered around nearby stars ever since. As of April 2009, 347 extrasolar planets (exoplanets) are listed in the Extrasolar Planets Encyclopaedia.

these factors together, we find *Ns fp ne fl fi fc* ~ 1 x $10^9$, a billion planets on which technical civilizations have arisen at least once. But that is very different from saying that there are a billion planets on which technical civilizations now exist. For this we must also estimate *fL*.

What percentage of the lifetime of a planet is marked by a technical civilization? The Earth has harbored a technical civilization characterized by radio astronomy for only a few decades out of a lifetime of a few billion years. So far, then, for our planet *fL* is less than $1/10^8$, a millionth of a percent. And it is hardly out of the question that we might destroy ourselves tomorrow. Suppose this were a typical case, and the destruction so complete that no other technical civilization – of the human or any other species – were able to emerge in the five or so billion years remaining before the Sun dies. Then *Ns fp ne fl fi fc fL* ~ 10, and, at a given time there would be only a tiny smattering, a handful, a pitiful few technical civilizations in the galaxy, the steady state number maintained as emerging societies replace those recently self-immolated. The number *N* might be even as small as 1 if civilizations tend to destroy themselves soon after reaching a technological phase; there might be no one for us to talk with but ourselves. And that we do but poorly. Civilizations would take billions of years of tortuous evolution, and then snuff themselves out in an instant of unforgivable neglect.

But consider the alternative, the prospect that at least some civilizations learn to live with technology; that the contradictions posed by the vagaries of past brain evolution are consciously resolved and do not lead to self destruction; or that, even if major disturbances occur, they are reveres in the subsequent billions of years of biological evolution. Such societies might live to a prosperous old age, their lifetimes measured perhaps on geological or stellar evolutionary time scales. If 1 percent of civilizations can survive technological adolescence, take the proper fork at this critical historical branch point and achieve maturity, then *fL* ~ 1/100, *N* ~ $10^7$, and the number of extant civilizations in the galaxy is in the millions. Thus, for all our concern about the possible unreliability of our estimates of the early factors in the Drake equation, which involve astronomy, organic chemistry and evolutionary biology, the principal uncertainty comes to economics and politics and what, on Earth, we call human nature. It seems fairly clear that if self-destruction is not the overwhelmingly preponderant fate of galactic civilizations, then the sky is softly humming with messages from the stars.

These estimates are stirring. They suggest that the receipt of a message from space is, even before we decode it, a profoundly hopeful sign. It means that someone has learned to live with high technology; that it is possible to survive technological adolescence. This alone, quite apart from the contents of the message, provides a powerful justification for the search for other civilizations.

## 4. The Drake Equation is Over-Simplified

In the nearly fifty years (1961-2009) elapsed since Frank Drake proposed his equation, a number of scientists and writers tried to find out which numerical values of its seven independent variables are more realistic in agreement with our present-day knowledge. Thus there is a considerable amount of literature about the Drake equation nowadays, and, as one can easily imagine, the results obtained by the various authors largely differ from one another. In other words, the value of *N*, that various authors obtained by different assumptions about the astronomy, the biology and the sociology implied by the Drake equation, may range from a few tens (in the pessimist's view) to some

million or even billions in the optimist's opinion. A lot of uncertainty is thus affecting our knowledge of *N* as of 2010. In all cases, however, the final result about *N* has always been a sheer number, i.e., a positive integer number ranging from 1 to millions or billions. This is precisely the aspect of the Drake equation that this author regarded as "too simplistic" and improved mathematically in his paper #IAC-08-A4.1.4, entitled "The Statistical Drake Equation" and presented on October 1st, 2008, at the 59th International Astronautical Congress (IAC) held in Glasgow, Scotland, UK, September 29th thru October 3rd, 2008. That paper is attached herewith as Appendix B. Newcomers to SETI and to the Drake equation, however, may find that paper too difficult to be understood mathematically at a first reading. Thus, I shall now explain the content of that paper "by speaking easily." I thank the reader for his or her attention.

## 5. The Statistical Drake Equation

We start by an example.

Consider the first independent variable in the Drake equation (7), i.e., *Ns*, the number of stars in the Milky Way galaxy. Astronomers tell us that *approximately* there should be about 350 millions stars in the galaxy. Of course, nobody has counted (or even seen in the photographic plates) *all* the stars in the galaxy! There are too many practical difficulties preventing us from doing so: just to name one, the dust clouds that don't allow us to see even the Galactic Bulge (i.e. the central region of the galaxy) in the visible light (although we may "see it" at radio frequencies like the famous neutral hydrogen line at 1420 MHz). So, it doesn't make any sense to say that $Ns = 350 \times 10^6$, or, say (even worse) that the number of stars in the galaxy is (say) 354,233,321, or similar fanciful exact integer numbers. That is just silly and non-scientific. Much more scientific, on the contrary, is to say that the number of stars in the galaxy is 350 million plus or minus, say, 50 millions (or whatever values the astronomers may regard as more appropriate, since this is just an example to let the reader understand the difficulty).

Thus, it makes sense to REPLACE each of the seven independent variables in the Drake equation (7) by a MEAN VALUE (350 millions, in the above example) PLUS OR MINUS A CERTAIN STANDARD DEVIATION (50 millions, in the above example).

By doing so, we have made a great step ahead: we have abandoned the too-simplistic equation (7) and replaced it by something more sophisticated and scientifically more serious: the STATISTICAL Drake equation. In other words, we have transformed the classical and simplistic Drake equation (7) into an advanced statistical tool for the investigation of a host of facts hardly known to us in detail. In other words still:

- We replace each independent variable in (7) by a RANDOM VARIABLE, labeled $D_i$ (from Drake).
- We assume that the MEAN VALUE of each $D_i$ is the same numerical value previously attributed to the corresponding independent variable in (7).
- But now we also ADD A STANDARD DEVIATION $\sigma_{D_i}$ on each side of the mean value, that is provided by the knowledge gathered by scientists in each discipline encompassed by each $D_i$.

Having so done, the next question is:

How can we find out the PROBABILITY DISTRIBUTION for each $D_i$?

For instance, shall that be a Gaussian, or what?

This is a difficult question, for nobody knows, for instance, the probability distribution of the number of stars in the galaxy, not to mention the probability distribution of the other six variables in the Drake equation (7).

There is a brilliant way to get around this difficulty, though.

We start by excluding the Gaussian because each variable in the Drake equation is a POSITIVE (or, more precisely, a non-negative) random variable, while the Gaussian applies to REAL random variables only. So, the Gaussian is out. Then, one might consider the large class of well-studied and positive probability densities called "the gamma distributions," but it is then unclear why one should adopt the gamma distributions and not any other. The solution to this apparent conundrum comes from Shannon's Information Theory and a theorem that he proved in 1948: "The probability distribution having maximum entropy (= uncertainty) over any FINITE range of real values is the UNIFORM distribution over that range," This is proven in Appendix A of the present document.

So, at this point, we assume that each of the seven $D_i$ in (7) is a UNIFORM random variable, whose mean value and standard deviation is known by the scientists working in the respective field (let it be astronomy, or biology, or sociology). Notice that, for such a uniform distribution, the knowledge of the mean value $\mu_{D_i}$ and of the standard deviation $\sigma_{D_i}$ automatically determines the RANGE of that random variable in between its lower (called $a_i$) and upper (called $b_i$) limits: in fact these limits are given by the equations

$$\begin{cases} a_i = \mu_{D_i} - \sqrt{3}\,\sigma_{D_i} \\ b_i = \mu_{D_i} + \sqrt{3}\,\sigma_{D_i} \end{cases} \tag{8}$$

(the "surprising" factor $\sqrt{3}$ in the above equations comes from the definitions of mean value and standard deviation: please see equations (12), (15) and (17) in Appendix B for the relevant proof). So the uniform distribution of each random variable $D_i$ is perfectly determined by its mean value and standard deviation, and so are all its other properties.

The next problem is the following:

OK, since we now know everything about each uniformly distributed $D_i$, what is the probability distribution of *N*, given that *N* is the product (7) of all the $D_i$?

In other words, not only do we want to find the analytical expression of the probability density function of *N*, but we also want to relate its mean value $\mu_N$ to all mean values $\mu_{D_i}$ of the $D_i$, and its standard deviation $\sigma_N$ to all standard deviations $\sigma_{D_i}$ of the $D_i$.

This is a difficult problem.

It occupied the author's mind for no less than about ten years (1997-2007).

It is actually an ANALYTICALLY UNSOLVABLE problem, in that, to the best of this author's knowledge, it is IMPOSSIBLE to find an analytic expression for any FINITE PRODUCT of uniform random variables $D_i$. This result is proven in Sections 2 thru 3.3 of Appendix B (unfortunately!).

## 6. Solving the Statistical Drake Equation By Virtue of the Central Limit Theorem (CLT) of Statistics

The solution to the problem of finding the analytical expression for the probability density function of *N* in the statistical Drake equation was found by this author in September 2007. The key steps are the following:

- Take the natural logs of both sides of the statistical Drake equation (7). This changes the product into a sum.
- The mean values and standard deviations of the logs of the random variables $D_i$ may all be expressed analytically in terms of the mean values and standard deviations of the $D_i$.
- Recall the Central Limit Theorem (CLT) of statistics, stating that (loosely speaking) if you have a SUM of independent random variables, each of which is ARBITRARILY DISTRIBUTED (hence, also including uniformly distributed), then, when the number of terms in the sum increases indefinitely (i.e. for a sum of random variables infinitely long)... the SUM RANDOM VARIABLE TENDS TO A GAUSSIAN.
- Thus, the natural log of *N* tends to a Gaussian.
- Thus, *N* tends to the LOGNORMAL DISTRIBUTION.
- The mean value and standard deviations of this lognormal distribution of *N* may all be expressed analytically in terms of the mean values and standard deviations of the logs of the $D_i$ already found previously.

This result is fundamental.

All the relevant equations are summarized in the following Table 1. This table is actually the same as Table 2 of the author's original paper IAC-08-A4.1.4, entitled "The Statistical Drake Equation" and presented by him at the International Astronautical Congress (IAC) held in Glasgow, UK, on October 1st, 2008. This original paper is reproduced in Appendix B.

To sum up, not only is it found that *N* approaches the completely known lognormal distribution for an INFINITY of factors in the statistical Drake equation (7), but the way is paved to further applications by removing the condition that the number of terms in the product (7) must be FINITE.

This possibility of ADDING ANY NUMBER OF FACTORS IN THE DRAKE EQUATION (7) was not envisaged, of course, by Frank Drake back in 1961, when "summarizing" the evolution of life in the galaxy in SEVEN simple STEPS. But today, the number of factors in the Drake equation should already be increased: for instance, there is no mention in the original Drake equation of the possibility that asteroidal impacts might destroy the life on Earth at any time, and this is because the demise of the dinosaurs at the K/T impact had not been yet understood by scientists in 1961, and was so only in 1980!

In practice, the number of factors should INCREASE as much as necessary in order to get better and better estimates of *N* as long as our scientific knowledge increases. This is called the "Data Enrichment Principle" and believe should be the next important goal in the study of the statistical Drake equation.

Finally, a numerical example explaining how the statistical Drake equation works in the practice will be given in the next section.

**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 | $f_N(n) = \frac{1}{n} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\ln(n)-\mu)^2}{2\sigma^2}} \quad (n \geq 0)$ |
| Mean value | $(N) = e^{\mu} e^{\frac{\sigma^2}{2}}$ |
| 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. *k*-th moment | $\left\langle N^k \right\rangle = e^{k\mu} e^{k^2 \cdot \frac{\sigma^2}{2}}$ |
| Mode (= abscissa of the lognormal peak) | $n_{\text{mode}} \equiv n_{\text{peak}} = e^{\mu} e^{-\sigma^2}$ |
| Value of the Mode Peak | $f_N(n_{\text{mode}}) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\mu} \cdot e^{\frac{\sigma^2}{2}}$ |
| Median (= fifty-fifty probability value for *N*) | $\text{median} = m = e^{\mu}$ |
| Skewness | $\frac{K_3}{(K_4)_2^3} = \left(e^{\sigma^2} + 2\right) \sqrt{\frac{e^{-6\mu} e^{-3\sigma^2}}{\left(e^{\sigma^2} - 1\right)^5 \left(e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6\right)^3}}$ |
| Kurtosis | $\frac{K_4}{(K_2)^2} = e^{4\sigma^2} + 2\,e^{3\sigma^2} + 3\,e^{2\sigma^2} - 6$ |
| Expression of $\mu$ 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 $\sigma^2$ in terms of the lower ($a_i$) and upper ($b_i$) limits of the Drake **uniform** input random variables $D_i$ | $\sigma^2 = \sum_{i-1}^{7} \sigma_{Y_i}^2 = \sum_{i-1}^{7} 1 - \frac{a_i b_i [\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2}$ |

# 7. An Example Explaining the Statistical Drake Equation

To understand how things work in practice for the statistical Drake equation, please consider the following table 2. It is made up of three columns:

- 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 equation (7). We see that, for this particular set of seven inputs, the classical Drake equation (i.e. the product of the seven numbers) yields a total of 3500 communicating extraterrestrial civilizations existing in the galaxy right now.

| | | |
|---|---|---|
| $Ns := 350 \cdot 10^9$ | $\mu Ns := Ns$ | $\sigma Ns := 1 \cdot 10^9$ |
| $fp := \frac{50}{100}$ | $\mu fp := fp$ | $\sigma fp := \frac{10}{100}$ |
| $ne := 1$ | $\mu ne := ne$ | $\sigma ne := \frac{1}{\sqrt{3}}$ |
| $fl := \frac{50}{100}$ | $\mu fl := fl$ | $\sigma fl := \frac{10}{100}$ |
| $fi := \frac{20}{100}$ | $\mu fi := fi$ | $\sigma fi := \frac{10}{100}$ |
| $fc := \frac{20}{100}$ | $\mu fc := fc$ | $\sigma fc := \frac{10}{100}$ |
| $fL := \frac{10000}{10^{10}}$ | $\mu fL := fL$ | $\sigma fL := \frac{1000}{10^{10}}$ |
| $N := Ns \cdot fp \cdot ne \cdot fl \cdot fi \cdot fc \cdot fL$ | $N = 3500$ | |

**Table 2. Input Values (i.e. mean values and standard deviations) for the Seven Drake Uniform Random Variables Di .** 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 equation (7).

The statistical Drake equation, however, provides a much more articulated answer than just the above sheer number $N = 3500$. In fact, a MathCad code written by this author and capable of performing all the numerical calculations required by the statistical Drake equation for a given set of seven input mean values plus seven input standard deviations, yields for $N$ the lognormal distribution (thin curve) plotted in Figure 2. We see immediately that the peak of this thin curve (i.e. the mode) falls at about $n_{\text{mode}} \equiv n_{\text{peak}} = e^{\mu} e^{-\sigma^2} \approx 250$ (this is equation (99) of Appendix B), while the median (fifty-fifty value splitting the lognormal density in two parts with equal undergoing areas) falls at about $n_{\text{median}} \equiv e^{\mu} \approx 1740$ . These seem to be smaller values than $N = 3500$ provided by the classical Drake equations, but it's a wrong impression due to a poor "intuitive" understanding of what statistics is! In fact, neither the mode nor the median are the "really important" values: the really important value for $N$ is the MEAN VALUE! Now if you look at the thin curve in Figure 2 below (i.e. the lognormal distribution arising from the Central Limit Theorem), you see that this curve has a LONG TAIL ON THE RIGHT! In other words, it does NOT immediately go down to nearly zero beyond the peak of the mode. Thus, when you actually compute the mean value, you should not be too

surprised to find out that it equals $\langle N \rangle = e^{\mu} e^{\frac{\sigma^2}{2}} \approx 4589.559$ ~ 4590 communicating civilizations now in the galaxy. This is the important number, and it is HIGHER than the 3500 provided by the classical Drake equation. Thus, in conclusion, THE STATISTICAL EXTENSION of the classical Drake equation INCREASES OUR HOPES to find an extraterrestrial civilization!

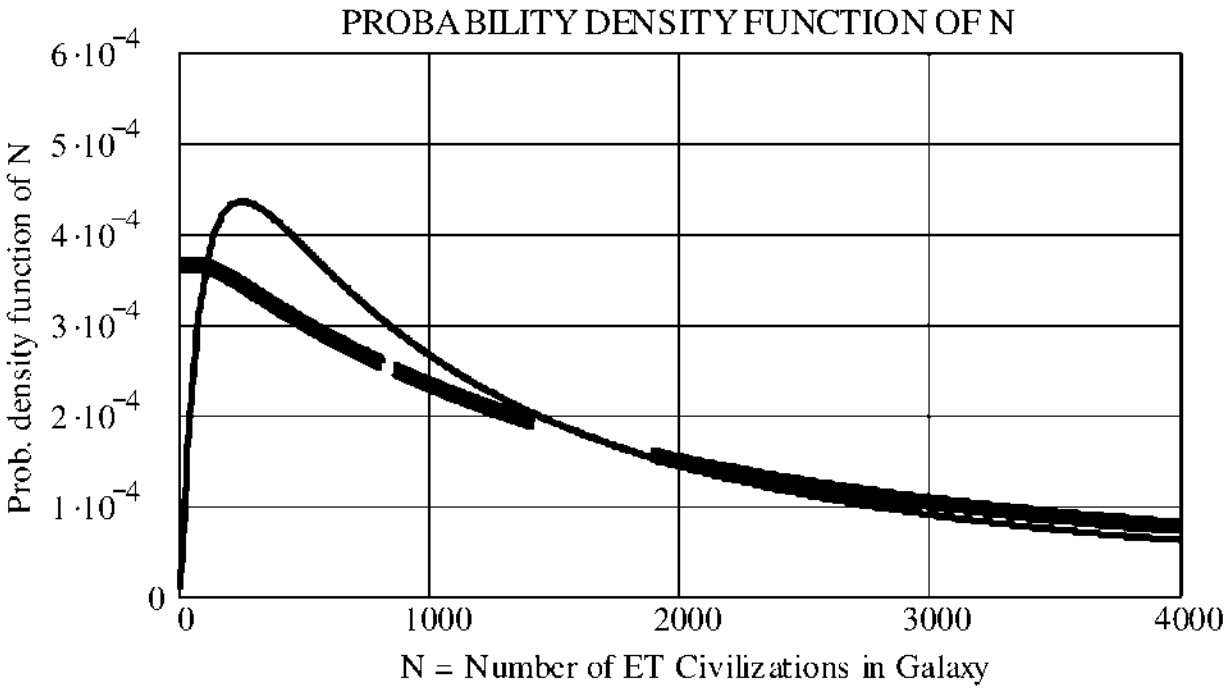



**Figure 2. Comparing the Two Probability Density Functions of the Random Variable N Found (1) Without Resorting to the CLT at All (thick curve) and (2) Using the CLT and the Relevant Lognormal Approximation (thin curve).**

Even more so our hopes are increased when we go on to consider the standard deviation associated with the mean value 4590. In fact, the standard deviation is given by equation (97) of Appendix B. This yields $\sigma_N = e^{\mu} e^{\frac{\sigma^2}{2}} \sqrt{e^{\sigma^2} - 1} = 11195$ and so the expected number of *N* may actually be even much higher than the 4590 provided by the mean value alone! The "upper limit of the one-sigma confidence interval" (as statisticians call it), i.e. the sum 4590+11195 = 15,785, yields a higher number still! (Note: the "lower limit of the one-sigma confidence interval is ZERO because the lognormal distribution is POSITIVE (or, more correctly, non-negative)). Finally, the reader should note that the thick curve depicted in Figure 2 is just the NUMERICAL solution of the statistical Drake equation for a FINITE number of 7 input factors. Figure 2 actually shows that this curve "is well interpolated" by the lognormal distribution (thin curve), i.e., by the neat analytical expression provided by the Central Limit Theorem for an INFINITE number of factors in the Drake equation. That is, in conclusion, Figure 2 visually shows that taking 7 factors or an infinity of factors "is almost the same thing" already for a value as small as 7.

# 8. Finding the Probability Distribution of the Et-Distance By Virtue of the Statistical Drake Equation

Having solved the statistical Drake equation by finding the lognormal distribution, we are now in a position to solve the ET-DISTANCE problem by resorting to statistics again, rather than just to the purely deterministic Distance Law (5), as we did in Section 2. This is "scientifically more serious" than just the purely deterministic Distance Law (5) inasmuch as the new statistical Distance Law will yield a PROBABILITY DENSITY for the Distance, with the relevant mean value and standard deviation. In other words, the Distance Law (5) itself becomes a random variable whose probability distribution, mean value and standard deviation must be computed by "replacing" into (5) the fact that *N* is now known to follow the lognormal distribution. This is mathematically described in detail in Section 7 of Appendix A.

The important new result is the PROBABILITY DENSITY FOR THE DISTANCE, the equation of which is

$$f_{\text{ET\_Distance}}(r) = \frac{3}{r} \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{\left(\ln\left[\frac{6\,R_{Galaxy}^2\,h_{Galaxy}}{r^3}\right]-\mu\right)^2}{2\sigma^2}} \tag{9}$$

holding for $r \geq 0$. This is equation (**114**) of Appendix B.

Starting from this equation, the MEAN VALUE OF THE random variable ET_DISTANCE is computed as

$$\langle \text{ET\_Distance} \rangle = C\,e^{-\frac{\mu}{3}}\,e^{\frac{\sigma^2}{18}} \tag{10}$$

which is equation (119) of Appendix B, and finally the ET_DISTANCE STANDARD DEVIATION

$$\sigma_{\text{ET\_Distance}} = C\,e^{-\frac{\mu}{3}}\,e^{\frac{\sigma^2}{18}}\sqrt{e^{\frac{\sigma^2}{9}}-1} \tag{11}$$

which is equation (123) of Appendix B. Of course, all other descriptive statistical quantities, such as moments, cumulants etc. can be computed upon starting from the probability density (9), and the result is Table two hereafter, that is Table 3 of Appendix B.

Finally, to complete this section, as well as this "introduction to the statistical Drake equation," the numerical values that equations (**10**) and (11) yield for the Input Table 1 are determined. They are, respectively:

$$r_{mean\ \ value} = C\,e^{-\frac{\mu}{3}}\,e^{\frac{\sigma^2}{18}} \approx 2{,}670 \text{ light years} \tag{12}$$

which is equation (153) of Appendix B, and

$$\sigma_{\text{ET\_Distance}} = C e^{-\frac{\mu}{3}} e^{\frac{\sigma^2}{18}} \sqrt{e^{\frac{\sigma^2}{9}} - 1} \approx 1{,}309 \text{ light years} \tag{13}$$

which is equation (154) of Appendix B.

**Table 2. 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 |
| Probability density function | $f_{\text{ET\_Distance}}(r) = \frac{3}{r} \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{\left(\ln\left[\frac{6 R_{Galaxy}^2 h_{Galaxy}}{r^3}\right] - \mu\right)^2}{2\sigma^2}}$ |
| Numerical constant C related to the Milky Way size | $C = \sqrt[3]{6\, R_{Galaxy}^2\, h_{Galaxy}} \approx 28{,}845 \text{ light years}$ |
| Mean value | $\langle \text{ET\_Distance} \rangle = C\, e^{-\frac{\mu}{3}}\, e^{\frac{\sigma^2}{18}}$ |
| Variance | $\sigma^2_{\text{ET Distance}} = C^2\, e^{-\frac{2}{3}\mu}\, e^{\frac{\sigma^2}{9}} \left( e^{\frac{\sigma^2}{9}} - 1 \right)$ |
| Standard deviation | $\sigma_{\text{ET Distance}} = C\, e^{-\frac{\mu}{3}}\, e^{\frac{\sigma^2}{18}} \sqrt{e^{\frac{\sigma^2}{9}} - 1}$ |
| All the moments, i.e. k-th moment | $\langle \text{ET\_Distance}^k \rangle = C^k\, e^{-\frac{k\mu}{3}}\, e^{\frac{k^2 \cdot \sigma^2}{18}}$ |
| Mode (= abscissa of the lognormal peak) | $r_{\text{mode}} \equiv r_{peak} = C\, e^{-\frac{\mu}{3}}\, e^{-\frac{\sigma^2}{9}}$ |
| Value of the Mode Peak | Peak Value of $f_{\text{ET\_Distance}}(r)$ = $\equiv f_{\text{ET Distance}}(r_{\text{mode}}) = \frac{3}{C\sqrt{2\pi}\,\sigma} \cdot e^{\frac{\mu}{3}} \cdot e^{\frac{\sigma^2}{18}}$ |
| Median (= fifty-fifty probability value for N) | $\text{median} = m = C e^{-\frac{\mu}{3}}$ |
| Skewness | $\frac{K_3}{(K_4)_2^3} = \frac{e^{-\mu}\left( e^{\frac{\sigma^2}{2}} - 3\, e^{\frac{5\sigma^2}{18}} + 2\, e^{\frac{\sigma^2}{6}} \right)}{C^3 \left( e^{\frac{8\sigma^2}{9}} - 4\, e^{\frac{5\sigma^2}{9}} - 3 e^{\frac{4\sigma^2}{9}} + 12\, e^{\frac{\sigma^2}{3}} - 6 e^{\frac{2\sigma^2}{9}} \right)^{\frac{3}{2}}}$ |
| Kurtosis | $\frac{K_4}{(K_2)^2} = e^{\frac{4\sigma^2}{9}} + 2\, e^{\frac{\sigma^2}{3}} + 3\, e^{\frac{2\sigma^2}{9}} - 6$ |
| Expression of $\mu$ in terms of the lower (ai) and upper (bi) limits of the Drake uniform input random variables Di | $\mu = \sum_{i=1}^{7} (Y_i) = \sum_{i=1}^{7} \frac{b_i[\ln(b_i) - 1] - a_i[\ln(a_i) - 1]}{b_i - a_i}$ |
| Expression of $\sigma^2$ in terms of the lower (ai) and upper (bi) limits of the Drake uniform input random variables Di | $\sigma^2 = \sum_{i=1}^{7} \sigma^2_{Y_i} = \sum_{i=1}^{7} 1 - \frac{a_i b_i [\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2}$ |

It is clarifying to draw the graph of the ET_Distance probability density (9):

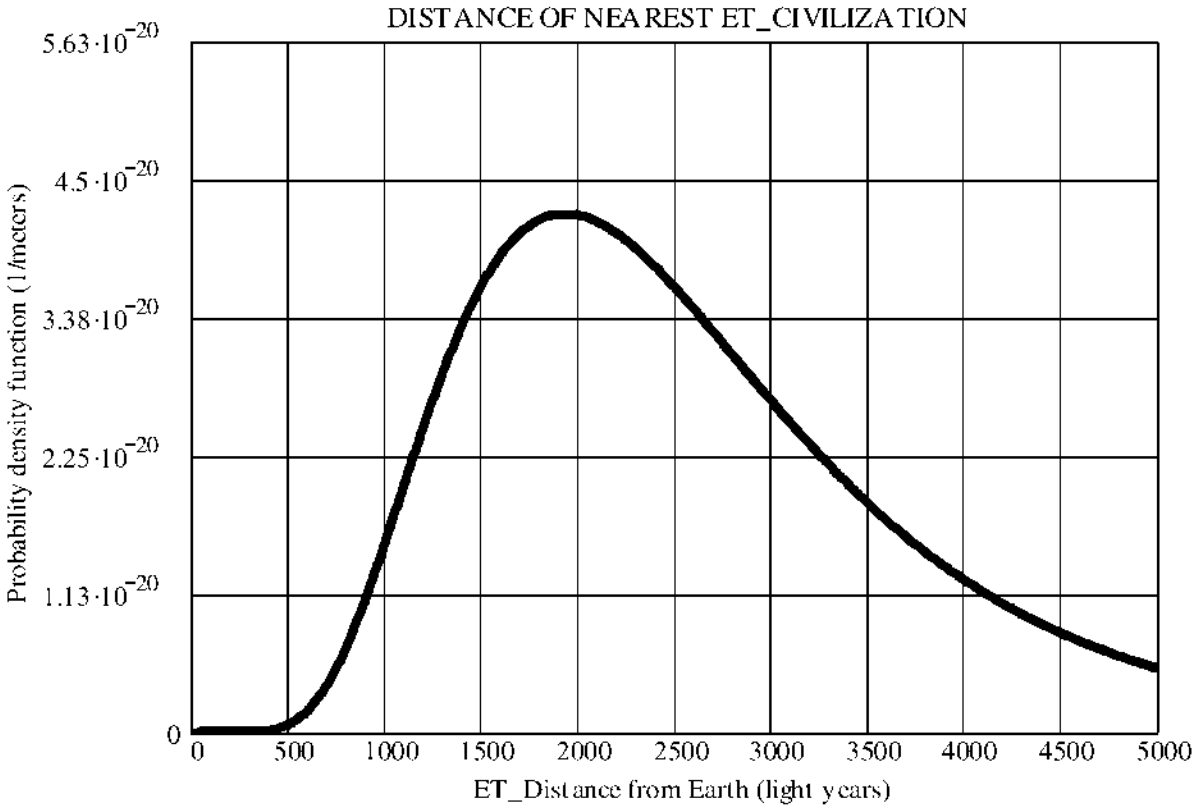



**Figure 3. The Probability of Finding the Nearest Extraterrestrial Civilization at the distance *r* From Earth (in light years) if the Values Assumed in the Drake Equation are Those Shown in Input Table 1.** The relevant probability density function $f_{\text{ET\_Distance}}(r)$ is given by equation (9). Its mode (peak abscissa) equals 1933 light years, but its mean value is higher since the 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 FDR SETI, inasmuch as the nearest ET galaxy civilization might lie at just* 1 sigma = 2670-1309 = 1361 light years *from us.*

From Figure 3 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

$$r_{\text{mode}} \equiv r_{peak} = C e^{-\frac{\mu}{3}} e^{-\frac{\sigma^2}{9}} \approx 1{,}933 \text{ light years}. \qquad (14)$$

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

It is not the mean value of the probability distribution (9) for $f_{\text{ET Distance}}(r)$. In fact, the probability density (9) has an infinite tail on the right, as clearly shown in Figure 3, and hence its mean value must be higher than its peak value. As given by (10) and (12), its mean value is $r_{mean\ value} = C e^{-\frac{\mu}{3}} e^{18} \approx 2670$ light years. 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, back and forth 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 (11) and (13),

$$\sigma_{\text{ET\_Distance}} = C\, e^{-\frac{\mu}{3}}\, e^{\frac{\sigma^2}{18}} \sqrt{e^{\frac{\sigma^2}{9}} - 1} \approx 1309 \text{ light years}.$$

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 $f_{\text{ET\_Distance}}(r)$ taken in between these two lower and upper limits, that is:

$$\int_{1361\text{lightyears}}^{3979\text{lightyears}} f_{\text{ET\_Distance}}(r)\, dr \approx 0.75 = 75\% \qquad (15)$$

In plain words: with 75 percent 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, of course.

## 9. 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.

## 10. 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 in size and especially in the depth of the scientific information it contains. It would thus be 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.

# Appendix A: Proof of Shannon's 1948 Theorem Stating That the Uniform Distribution is the "Most Uncertain" One Over a Finite Range of Values

Information Theory was initiated by Claude Shannon (1916-2001) in his well-known 1948 two papers:

Reprinted with corrections from *The Bell System Technical Journal*
Vol. 27, pp. 379–423, 623–656, July, October, 1948

A Mathematical Theory of Communication

By C. E. SHANNON

In this Appendix, we wish to draw attention to a couple of theorems that Shannon proves on pages 36 and 37 of his work, and read, respectively (note that Shannon omits the upper and lower limits of all integrals in the first theorem: they are minus infinity and plus infinity, respectively):

5. Let $p(x)$ be a one-dimensional distribution. The form of $p(x)$ giving a maximum entropy subject to the condition that the standard deviation of $x$ be fixed at $\sigma$ is Gaussian. To show this we must maximize

$$H(x) = -\int p(x) \log p(x)\, dx$$

with

$$\sigma^2 = \int p(x) x^2\, dx \quad \text{and} \quad 1 = \int p(x)\, dx$$

as constraints. This requires, by the calculus of variations, maximizing

$$\int \left( -p(x) \log p(x) + \lambda p(x) x^2 + \mu p(x) \right) dx.$$

The condition for this is

$$-1 - \log p(x) + \lambda x^2 + \mu = 0$$

and consequently (adjusting the constants to satisfy the constraints)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}.$$

and

7. If $x$ is limited to a half line ($p(x) = 0$ for $x < 0$) and the first moment of $x$ is fixed at $a$

$$a = \int_0^\infty p(x)\, x\, dx.$$

then the maximum entropy occurs when

$$p(x) = \frac{1}{a} e^{-x/a}$$

and is equal to $\log ea$.

Now, we wish to point out that there is a third possible case, other than the two given by Shannon. This is the case when the probability density function $p(x)$ is limited to a FINITE INTERVAL $a \le x \le b$. This is obviously the case with any physical POSITIVE random variable, such as a distance, or the number *N* of extraterrestrial communicating civilizations in the ,". And it is easy to prove that for any such finite random variable the maximum entropy distribution is the UNIFORM distribution over $a \le x \le b$. Shannon did not bother to prove this simple theorem in his 1948 papers since he probably regarded it as 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 $a_i \le x \le b_i$, i.e., that $D_i$ is UNIFORMLY distributed there.

The proof of this theorem is along the same lines as for the previous two cases discussed by Shannon:

We start by assuming that $a_i \le x \le b_i$.

We then form the linear combination of the entropy integral plus the normalization condition for $D_i$

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

where $\lambda$ is a Lagrange multiplier.

Performing the variation, one finds

$$-\log p(x) - 1 + \lambda = 0 \text{ that is: } p(x) = e^{\lambda - 1}.$$

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

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

that yields

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

and finally

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

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

## Appendix B: Original Text of the Author's Paper #IAC-08-A4.1.4 Titled the Statistical Drake Equation

**IAC-08-A4.1.4**

# THE STATISTICAL DRAKE EQUATION

***Claudio Maccone***

*Co-Vice Chair, SETI Permanent Study Group, International Academy of Astronautics*

*Address: Via Martorelli, 43 - Torino (Turin) 10155 - Italy*
URL: http://www.maccone.com/ - E-mail: clmaccon@libero.it

**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 found also.
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. 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.

## 1. INTRODUCTION

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 Astrohiology (see ref. [3]). Devised in 1960, 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:

$$N = Ns \cdot fp \cdot ne \cdot fl \cdot fi \cdot fc \cdot fL . \qquad (1)$$

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 heing 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 numher 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!

We claim the classical Drake equation (1), as we shall call it from now on to distinguish it from our statistical Drake equation to he 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 distribation!***

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's 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 1977-78-79.

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

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

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

$$N = \prod_{i=1}^{7} D_i \ . \qquad (3)$$

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?

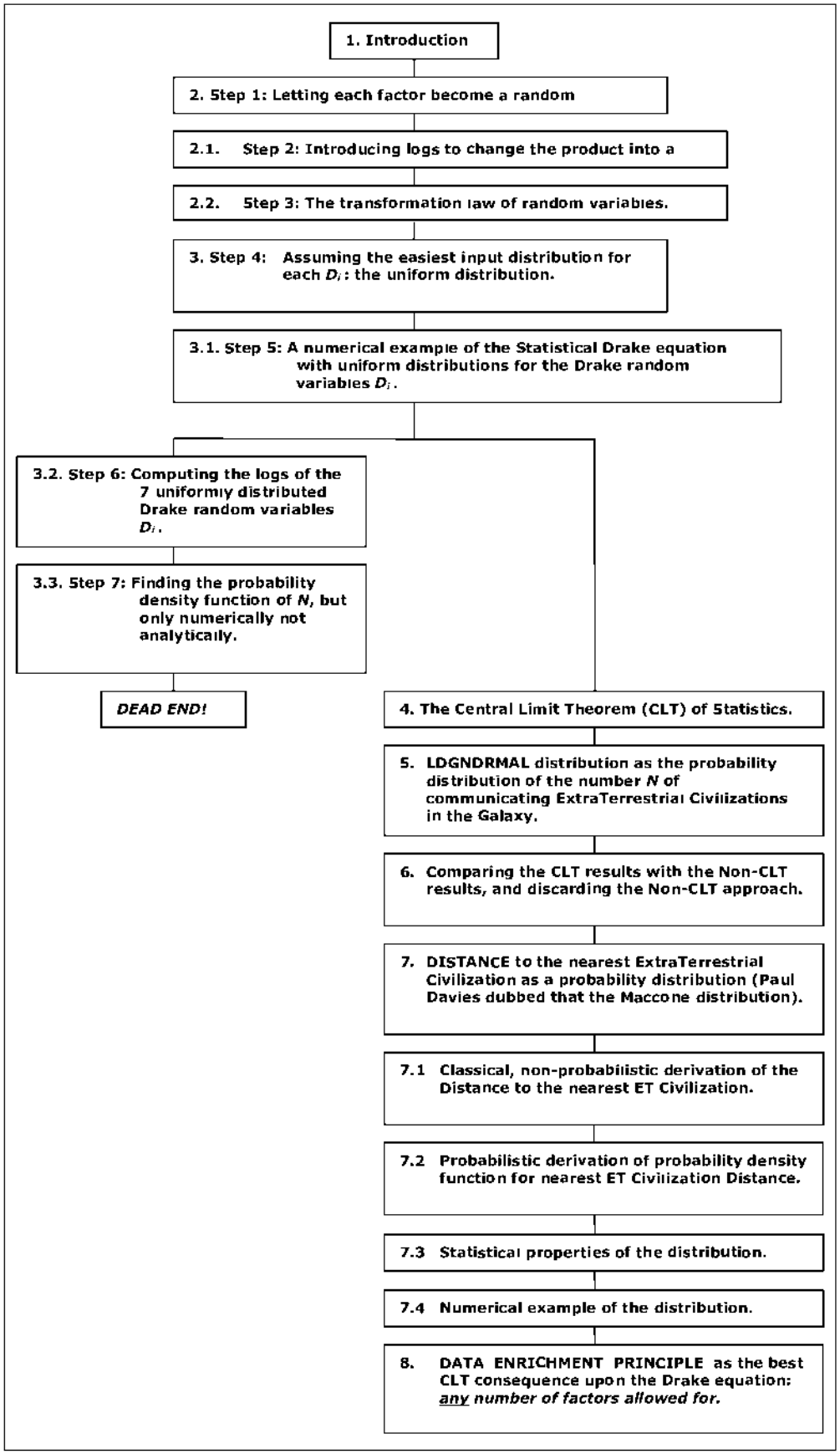
… and how are the seven standard deviations on the right related to the standard deviation on the left?

Just take the next step…

## 3. 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)


1. Introduction
2. Step 1: Letting each factor become a random
2.1. Step 2: Introducing logs to change the product into a
2.2. Step 3: The transformation law of random variables.
3. Step 4: Assuming the easiest input distribution for each $D_i$: the uniform distribution.
3.1. Step 5: A numerical example of the Statistical Drake equation with uniform distributions for the Drake random variables $D_i$.
3.2. Step 6: Computing the logs of the 7 uniformly distributed Drake random variables $D_i$.
3.3. Step 7: Finding the probability density function of N, but only numerically not analytically.
DEAD END!
4. The Central Limit Theorem (CLT) of Statistics.
5. LDGNDRMAL distribution as the probability distribution of the number N of communicating ExtraTerrestrial Civilizations in the Galaxy.
6. Comparing the CLT results with the Non-CLT results, and discarding the Non-CLT approach.
7. DISTANCE to the nearest ExtraTerrestrial Civilization as a probability distribution (Paul Davies dubbed that the Maccone distribution).
7.1 Classical, non-probabilistic derivation of the Distance to the nearest ET Civilization.
7.2 Probabilistic derivation of probability density function for nearest ET Civilization Distance.
7.3 Statistical properties of the distribution.
7.4 Numerical example of the distribution.
8. DATA ENRICHMENT PRINCIPLE as the best CLT consequence upon the Drake equation: any number of factors allowed for.

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

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

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

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

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

$$N = e^{Y}. \qquad (6)$$

We are now ready to take STEP THREE.

## 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

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

has its independent variable x ranging between $-\infty$ and $\infty$ and so it can apply to a ***real*** random variable $Y$ only, and never to ***positive*** random variables like those in the statistical Drake equation (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 (5), once we selected a particular type of probability density function (pdf) for the last seven of equations (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 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

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

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 the 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

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

where the summation extends to all roots $x_i(y)$ and $\left|g'(x_i(y))\right|$ 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 nest section.

## 4. 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

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

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

$$\langle \text{uniform\_D}_i \rangle = \int_{a_i}^{b_i} x\, f_{\text{uniform\_D}_i}(x)\,dx = \frac{1}{b_i - a_i}\int_{a_i}^{b_i} x\,dx$$

$$= \frac{1}{b_i - a_i}\left[\frac{x^2}{2}\right]_{a_i}^{b_i} = \frac{b_i^2 - a_i^2}{2(b_i - a_i)} = \frac{a_i + b_i}{2}.$$

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

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

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

$$\langle \text{uniform\_D}_i^{\,2} \rangle = \int_{a_i}^{b_i} x^2\, f_{\text{uniform\_D}_i}(x)\,dx$$

$$= \frac{1}{b_i - a_i}\int_{a_i}^{b_i} x^2\,dx = \frac{1}{b_i - a_i}\left[\frac{x^3}{3}\right]_{a_i}^{b_i} = \frac{b_i^3 - a_i^3}{3(b_i - a_i)}$$

$$= \frac{(b_i - a_i)(a_i^2 + a_i b_i + b_i^2)}{3(b_i - a_i)} = \frac{a_i^2 + a_i b_i + b_i^2}{3}.$$

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

$$\sigma^2_{\text{uniform\_D}_i} = \langle \text{uniform\_D}_i^{\,2} \rangle - \langle \text{uniform\_D}_i \rangle^2$$

$$= \frac{a_i^2 + a_i b_i + b_i^2}{3} - \frac{(a_i + b_i)^2}{4} = \frac{(b_i - a_i)^2}{12}. \quad (14)$$

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

$$\boxed{\sigma_{\text{uniform\_D}_i} = \frac{b_i - a_i}{2\sqrt{3}}}. \quad (15)$$

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

$$\begin{cases} \langle \text{uniform\_D}_i \rangle = \dfrac{a_i + b_i}{2} \\ \sigma_{\text{uniform\_D}_i} = \dfrac{b_i - a_i}{2\sqrt{3}}. \end{cases} \quad (16)$$

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 inasmuch as it shows that:

***if ane (scientifically) assigns the mean value and standard deviation of a certain Drake random variable $D_i$, then the lower and upper limits af the relevant uniform distributian are given by the two equations (17), respectively.***

In other words, there is a factor of $\sqrt{3}=1.732$ included in the two equations (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 *Ns* in the classical Drake equation (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 *Ns* has a mean value of 350 billions minus or plus a standard deviation of (say) one billion (we don't 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:

$$\begin{cases} \langle \text{uniform\_D}_1 \rangle = 350 \cdot 10^9 \\ \sigma_{\text{uniform\_D}_1} = 1 \cdot 10^9. \end{cases} \tag{18}$$

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

$$\begin{cases} a_{Ns} = \langle \text{uniform\_D}_1 \rangle - \sqrt{3}\,\sigma_{\text{uniform\_D}_1} = 348.3 \cdot 10^9 \\ b_{Ns} = \langle \text{uniform\_D}_1 \rangle + \sqrt{3}\,\sigma_{\text{uniform\_D}_1} = 351.7 \cdot 10^9 \end{cases} \tag{19}$$

Similarly we proceed for all the other six random variables in the Statistical Drake equation (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

$$\begin{cases} a_{fp} = \langle \text{uniform\_D}_2 \rangle - \sqrt{3}\,\sigma_{\text{uniform\_D}_2} = 0.327 \\ b_{fp} = \langle \text{uniform\_D}_2 \rangle + \sqrt{3}\,\sigma_{\text{uniform\_D}_2} = 0.673 \end{cases} \tag{20}$$

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 (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

$$\begin{cases} a_{ne} = \langle \text{uniform\_D}_3 \rangle - \sqrt{3}\,\sigma_{\text{uniform\_D}_3} = 0 \\ b_{ne} = \langle \text{uniform\_D}_3 \rangle + \sqrt{3}\,\sigma_{\text{uniform\_D}_3} = 2 \end{cases} \tag{21}$$

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 equation (3) would have obviously reduced to the classical Drake equation (1), and the resulting number of civilizations in the Galaxy would have turned out to be 3500:

$$\boxed{N = 3500}. \tag{22}$$

This is the 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.

| | | |
|---|---|---|
| $Ns := 350 \cdot 10^9$ | $\mu Ns := Ns$ | $\sigma Ns := 1 \cdot 10^9$ |
| $fp := \frac{50}{100}$ | $\mu fp := fp$ | $\sigma fp := \frac{10}{100}$ |
| $ne := 1$ | $\mu ne := ne$ | $\sigma ne := \frac{1}{\sqrt{3}}$ |
| $fl := \frac{50}{100}$ | $\mu fl := fl$ | $\sigma fl := \frac{10}{100}$ |
| $fi := \frac{20}{100}$ | $\mu fi := fi$ | $\sigma fi := \frac{10}{100}$ |
| $fc := \frac{20}{100}$ | $\mu fc := fc$ | $\sigma fc := \frac{10}{100}$ |
| $fL := \frac{10000}{10^{10}}$ | $\mu fL := fL$ | $\sigma fL := \frac{1000}{10^{10}}$ |
| $N := Ns \cdot fp \cdot ne \cdot fl \cdot fi \cdot fc \cdot fL$ | $N = 3500$ | |

**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 equation (1).

### 3.2 STEP 6: COMPUTING THE LOGS OF THE 7 UNIFORMY 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 below:

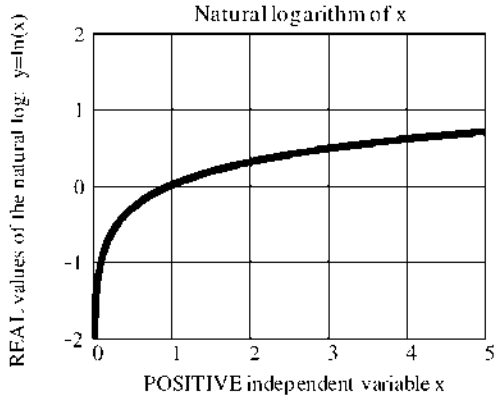



**Figure 1.** The simple function $y = \ln(x)$.

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

$$Y_i = \ln(D_i) \quad i = 1,...,7 \tag{23}$$

must have its range limited in between the lower limit $ln(a_i)$ and the upper limit $ln(b_i)$. In other words, this 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 equation (9). Here we obviously have

$$y = g(x) = \ln(x) \tag{24}$$

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

$$x_1(y) = x(y) = e^y . \tag{25}$$

On the other hand, differentiating (24) one gets

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

where (25) was already used in the last step. By virtue of the uniform probability density function (10) and of (26), the general transformation law (9) finally yields

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

In other words, the requested pdf of $Y_i$ is

$$\boxed{f_{Y_i}(y)=\frac{e^y}{b_i-a_i} \quad i=1,\ldots,7} \qquad \boxed{\ln(a_i)\le y\le \ln(b_i)} \qquad (28)$$

***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:

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

.

...(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 $\langle Y_i\rangle$ is given by

$$\langle Y_i\rangle=\int_{\ln(a_i)}^{\ln(b_i)} y\cdot f_{Y_i}(y)\,dy=\int_{\ln(a_i)}^{\ln(b_i)}\frac{y\cdot e^y}{b_i-a_i}\,dy$$

$$=\frac{b_i\left[\ln(b_i)-1\right]-a_i\left[\ln(a_i)-1\right]}{b_i-a_i}. \qquad (30)$$

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

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

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

$$\langle Y_i^2\rangle=\int_{\ln(a_i)}^{\ln(b_i)} y^2\cdot f_{Y_i}(y)\,dy=\int_{\ln(a_i)}^{\ln(b_i)}\frac{y^2\cdot e^y}{b_i-a_i}\,dy$$

$$=\frac{b_i\left[\ln^2(b_i)-2\ln(b_i)+2\right]-a_i\left[\ln^2(a_i)-2\ln(a_i)+2\right]}{b_i-a_i}.$$

...(32)

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

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

Whence the corresponding standard deviation

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

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 $j=\sqrt{-1}$ )

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

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

$$\Phi_{Y_i}(\zeta)=\int_{-\infty}^{\infty} e^{j\zeta y} f_{Y_i}(y)\,dy=\int_{\ln(a_i)}^{\ln(b_i)} e^{j\zeta y}\frac{e^y}{b_i-a_i}\,dy$$

$$= \frac{1}{b_i - a_i}\int_{\ln(a_i)}^{\ln(b_i)} e^{(1+j\zeta)y}\,dy = \frac{1}{b_i - a_i}\cdot\frac{1}{1+j\zeta}\left[e^{(1+j\zeta)y}\right]_{\ln(a_i)}^{\ln(b_i)}$$

$$= \frac{e^{(1+j\zeta)\ln(b_i)} - e^{(1+j\zeta)\ln(a_i)}}{(b_i - a_i)(1+j\zeta)} = \frac{b_i^{1+j\zeta} - a_i^{1+j\zeta}}{(b_i - a_i)(1+j\zeta)}. \qquad (36)$$

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

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

### 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 (5). In fact, by virtue of (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 (37), it immediately follows that $\Phi_Y(\zeta)$ equals the product of the seven $\Phi_{Y_i}(\zeta)$:

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

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

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

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-j\zeta y}\left[\prod_{i=1}^{7}\Phi_{Y_i}(\zeta)\right]d\zeta$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-j\zeta y}\left[\prod_{i=1}^{7}\frac{b_i^{1+j\zeta} - a_i^{1+j\zeta}}{(b_i - a_i)(1+j\zeta)}\right]d\zeta. \qquad (39)$$

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 equations 17. The result was the probability density function for $Y = \ln(N)$ plotted in the following Figure 2.

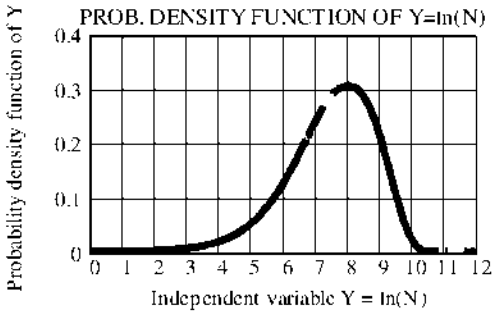



**Figure 2.** Probability density function of $Y = \ln(N)$ computed numerically by virtue of the integral (39). The two "funny gaps" in the curve are due to the numeric limitations in the MathCad numeric solver that the author used for this numeric computation.

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 equation (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 (6). We must thus resort to the transformation law of random variables (9) by setting

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

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

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

On the other hand, differentiating (40) one gets

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

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

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

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

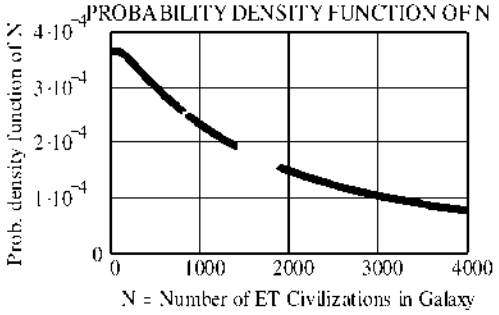



**Figure 3.** The ***numeric*** (and not analytic) probability density function curve $f_N(y)$ of the number $N$ of ExtraTerrestrial Civilizations in the Galaxy according to the Statistical Drake equation (3). We see that the curve peak (i.e. the mode) is very close to low values of $N$, but the tail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

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

$$\langle N \rangle = \int_0^\infty y\, f_N(y)\, dy\,. \qquad (44)$$

This integral too was computed numerically, and the result was a ***perfect match*** with $N$=3500 of (22), that is

$$\langle N \rangle = 3499.99880\ 177509 + 0.00000012\ 4914686i \quad (45)$$

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

$$\sigma_N = 3953.42910\ 143389 + 0.00000003\ 2800158i\ . \quad (46)$$

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 equation (38) had to be performed numerically.***

***This is no gaod.***

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.

## 5. 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***

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

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 (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:

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

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

## 6. THE LOGNORMAL DISTRIBTION 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 equation (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 equation (6), that is, what is the probability density function of the lognormal distribution? To find it, set

$$y = g(x) = e^x . \tag{49}$$

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

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

On the other hand, differentiating (49) one gets

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

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

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

Therefore, replacing the probability density on the right by virtue of the well-known normal (or Gaussian) distribution given by equation (7), the lognormal distribution of equation (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", that is:

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

This follows immediately from the normalization condition of the Gaussian (7), that is

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

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

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

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***

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

Having so said, we now turn to the statistical properties of the lognormal distribution (55), 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), that is, for every non-negative integer $k = 0, 1, 2, \ldots$ one has

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

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

$$\langle N^k \rangle = \int_0^\infty n^k \cdot f_N(n)\, dn$$

$$= \int_0^\infty n^k \cdot \frac{1}{n} \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{(\ln[n]-\mu)^2}{2\sigma^2}}\, dn$$

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 (57) follows

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

Upon setting $k = 0$ into (56), the normalization condition for $f_N(n)$ follows

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

Upon setting $k = 1$ into (56), the important ***mean value of the randam variable N*** is found

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

Upon setting $k = 2$ into (56), 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 af N*** now follows from the last two formulae:

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

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

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

The third moment is obtained upon setting $k = 3$ into (56)

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

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 $\mu_i'$ by virtue of the recursion formula (see ref. [8])

$$K_i = \mu_i' - \sum_{k=1}^{i-1} \binom{i-1}{k-1} K_k\, \mu_{n-k}'\,. \tag{66}$$

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:

$$\begin{cases} K_1 = \mu_1' \\ K_2 = \mu_2' - K_1^2 \\ K_3 = \mu_3' - 3K_1K_2 - K_1^3 \\ K_4 = \mu_4' - 4K_1K_3 - 3K_2^2 - 6K_2K_1^2 - K_1^4. \end{cases} \quad (67)$$

These equations yield, respectively:

$$K_1 = e^{\mu} e^{\frac{\sigma^2}{2}}. \quad (68)$$

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

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

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

From these we derive the skewness

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

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 equation (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 (57). 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

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

But the lognormal probability density function (56), by virtue of (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

$$\frac{df_{\text{ET Distance}}'(r)}{dr} = \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{-e^{E(n)}E'(n) \cdot n - 1 \cdot e^{E(n)}}{n^2}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{-e^{-E(n)}\left[E'(n) \cdot n + 1\right]}{n^2}. \quad (77)$$

Setting this derivative equal to zero means setting

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

That is, upon replacing (75),

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

Rearranging, this becomes

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

and finally

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

***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 replace (81) into (56). After a few rearrangements, one then gets

$$\boxed{f_N(n_{\text{mode}}) = \frac{1}{\sqrt{2\pi}\,\sigma}\cdot e^{-\mu}\cdot e^{\frac{\sigma^2}{2}}}. \qquad (82)$$

***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 $m$:

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

In order to find $m$, we may ***not*** differentiate (83) with respect to $m$, since the "precise" factor ½ on the right would then disappear into a zero. On the contrary, we may try to perform the obvious substitution

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

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

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

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

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

that is

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

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

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

whence finally

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

***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.

| 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 | $\langle N\rangle = e^{\mu}e^{\frac{\sigma^2}{2}}$ |
| 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. $k$-th moment | $\langle N^k \rangle = e^{k\mu} e^{k^2 \cdot \frac{\sigma^2}{2}}$ |
| Mode (= abscissa of the lognormal peak) | $n_{\text{mode}} \equiv n_{\text{peak}} = e^{\mu} e^{-\sigma^2}$ |
| Value of the Mode Peak | $f_N(n_{\text{mode}}) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{\mu} \cdot e^{\frac{\sigma^2}{2}}$ |
| Median (= fifty-fifty probability value for $N$) | $\text{median} = m = e^{\mu}$ |
| Skewness | $\frac{K_3}{(K_4)_2^3} = \left(e^{\sigma^2} + 2\right)\sqrt{\frac{e^{-6\mu} e^{-3\sigma^2}}{\left(e^{\sigma^2}-1\right)^6 \left(e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6\right)^3}}$ |
| Kurtosis | $\frac{K_4}{(K_2)^2} = e^{4\sigma^2} + 2\,e^{3\sigma^2} + 3\,e^{2\sigma^2} - 6$ |
| Expression of $\mu$ 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 $\sigma^2$ in terms of the lower ($a_i$) and upper ($b_i$) limits of the Drake **uniform** input random variables $D_i$ | $\sigma^2 = \sum_{i=1}^{7} \sigma_{Y_i}^2 = \sum_{i=1}^{7} 1 - \frac{a_i b_i [\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2}$ |

**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.

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 equation (3) listed in Table 1.

According to the CLT, the ***mean value*** $\mu$ to be inserted into the lognormal density (56) is given (according to the second equation (48)) by the sum of all the mean values $\langle Y_i \rangle$, that is, by virtue of (31), by:

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

Upon replacing the 14 $a_i$ and $b_i$ listed in Table 1 into (90), the following ***numeric mean value*** $\mu$ is found

$$\boxed{\mu \approx 7.462176} \quad (91)$$

Similarly, to get the numeric variance $\sigma^2$ one must resort to the last of equations (48) and to (33):

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

yielding the following ***numeric variance*** $\sigma^2$ to be inserted into the lognormal pdf (56)

$$\boxed{\sigma^2 \approx 1.938725} \quad (93)$$

whence the ***numeric standard deviation*** $\sigma$

$$\boxed{\sigma \approx 1.392381}. \quad (94)$$

Upon replacing these two numeric values (84) and (86) into the lognormal pdf (56), the latter is perfectly determined. It is plotted in Figure 4 hereafter as the thin curve.

In other words, ***Figure 4 shows the lognormal distribution for the number N of ExtraTerrestrial Civilizations in the Galaxy derived from the Central Limit Thearem as applied ta the Drake equatian (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 equation (60) with $\mu$ and $\sigma$ given by (91) and (94), respectively:

$$\boxed{\langle N \rangle = e^{\mu} e^{\frac{\sigma^2}{2}} \approx 4589.559} \quad (95)$$

***In other words, there are 4590 ET Civilizations in the Galaxy according 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 equation (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 (1).***

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

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

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

$$\boxed{\sigma_N = e^{\mu} e^{\frac{\sigma^2}{2}} \sqrt{e^{\sigma^2} - 1} = 11195} \quad (97)$$

***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+11195=15785 is (95)+(97)).**

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

$$\boxed{n_{\text{mode}} \equiv n_{\text{peak}} = e^{\mu} e^{-\sigma^2} \approx 250} \quad (98)$$

This is well shown in Figure 4: the mode peak is very pronounced and close to the origin, but the right tail is high, and this means that the mean value of the distribution is much higher than the mode: 4590>>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 (89), and has the numeric value:

$$\boxed{n_{\text{median}} \equiv e^{\mu} \approx 1740} \quad (99)$$

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.

## 7. 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:

1) The numeric curves appearing in Figure 2 and obtained after laborious Fourier transform calculations in the complex domain, and
2) The lognormal distribution (56) with numeric $\mu$ and $\sigma$ given by (91) and (94) respectively.

We see 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 (5), that is plotted in Figure 5 both in its normal curve version (thin curve) and in its numeric version, obtained via Fourier transforms and already shown in Figure 2.

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

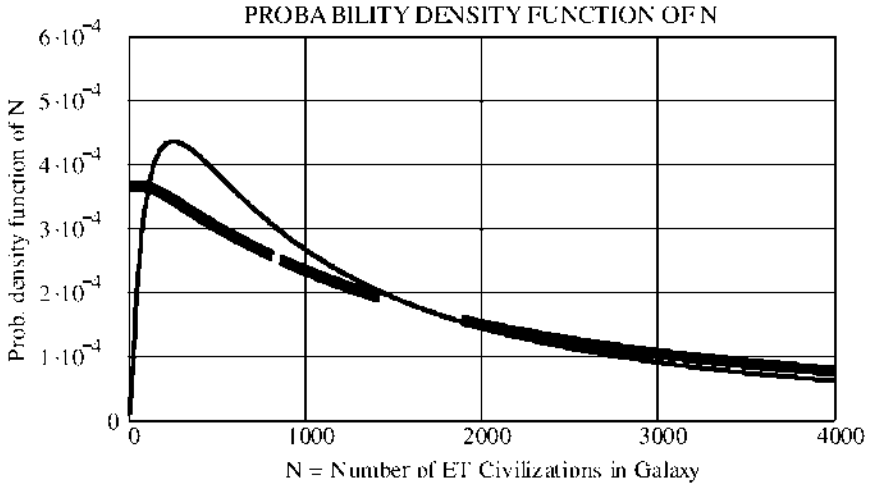



**Figure 4.** Comparing the two probability density functions of the random variable *N* found:
1) At the end of Section 3.3, in a purely numeric way and without resorting to the CLT at all (thick curve) and
2) Analytically by using the CLT and the relevant lognormal approximation (thin curve).

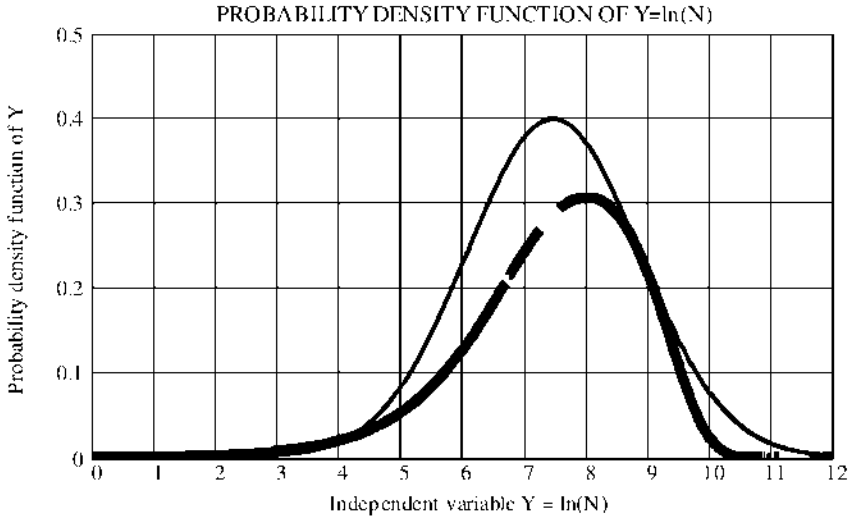



**Figure 5.** Comparing the two probability density functions of the random variable Y=ln(N) found:
1) At the end of Section 3.3, in a purely numeric way and without resorting to the CLT at all (thick curve) and
2) Analytically by using the CLT and the relevant normal (Gaussian) approximation (thin Gaussian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

## 8. 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 $f_{\text{ET\_Distance}}(r)$ : 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 $f_{\text{ET\_Distance}}(r)$ such as its mean value, variance, standard deviation, all central moments, mode, median, cumulants, skewness and kurtosis.

## 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.

$$V_{Galaxy} = \pi R_{Galaxy}^2 h \qquad (100)$$

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

$$V_{Our\_Sphere} = \frac{4}{3}\pi\left(\frac{\text{ET\_Distance}}{2}\right)^3 \qquad (101)$$

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 space 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:***

$$\frac{V_{Galaxy}}{N} = \frac{V_{Our\_Sphere}}{1}. \qquad (102)$$

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

$$\frac{\pi R_{Galaxy}^2 h}{N} = \frac{\frac{4}{3}\pi\left(\frac{\text{ET\_Distance}}{2}\right)^3}{1}. \qquad (103)$$

*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 NEIGHBOURING CIVILIZATIONS IN THE GALAXY***

$$\boxed{\text{ET\_Distance} = \frac{\sqrt[3]{6 R_{Galaxy}^2 h}}{\sqrt[3]{N}} = \frac{C}{\sqrt[3]{N}}} \qquad (104)$$

where the positive constant $C$ is defined by

$$\boxed{C = \sqrt[3]{6 R_{Galaxy}^2 h_{Galaxy}} \approx 28845 \text{ light years}}. \qquad (105)$$

Equations (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.

## 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 5-6-7-8, 2008. This meeting was also attended by SETI Institute experts Jill Tarter, Seth Shostak, Doug Vakoch, Tom Pierson and others. During this 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 equation (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

$$f_{\text{ET\_Distance}}(r) \tag{106}$$

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

$$\begin{cases} N \to x \\ \text{ET\_distance} \to y \end{cases} \tag{107}$$

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

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

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

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

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

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

Thus, the relevant absolute value reads

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

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

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

This is the denominator of (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 (109). By doing so, we finally arrive at the new probability density function $f_Y(y)$

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

Rearranging and replacing $y$ by $r$, the final form is:

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

Now, just replace $C$ in (113) by virtue of (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:***

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

holding for $r \geq 0$.

**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 (113) are given by (here $k$ is the generic and non-negative integer exponent, i.e. $k = 0,1,2,3,... \geq 0$)

$$\left\langle \text{ET\_Distance}^k \right\rangle = \int_0^\infty r^k \cdot f_{\text{ET\_Distance}}(r)\, dr$$

$$= \int_0^\infty r^k \cdot \frac{3}{r} \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{\left(\ln\left[\frac{C^3}{r^3}\right]-\mu\right)^2}{2\sigma^2}}\, dr$$

$$= C^k\, e^{-k\frac{\mu}{3}}\, e^{k^2\frac{\sigma^2}{18}}. \qquad (115)$$

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

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

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, (115) follows from (53). In other words, we have proven that

$$\boxed{\left\langle \text{ET\_Distance}^k \right\rangle = C^k\, e^{k\frac{\mu}{3}}\, e^{k^2\frac{\sigma^2}{18}}}. \qquad (117)$$

Upon setting $k = 0$ into (117), the normalization condition for $f_{\text{ET\_Distance}}(r)$ follows

$$\int_0^\infty f_{\text{ET\_Distance}}(r)\, dr = 1. \qquad (118)$$

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

$$\boxed{\left\langle \text{ET\_Distance} \right\rangle = C\, e^{\frac{\mu}{3}}\, e^{\frac{\sigma^2}{18}}}. \qquad (119)$$

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

$$\left\langle \text{ET\_Distance}^2 \right\rangle = C^2\, e^{-\frac{2}{3}\mu}\, e^{\frac{2}{9}\sigma^2}. \qquad (120)$$

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

$$\sigma^2_{\text{ET\_Distance}} = \left\langle \text{ET\_Distance}^2 \right\rangle - \left\langle \text{ET\_Distance} \right\rangle^2$$

$$= C^2\, e^{-\frac{2}{3}\mu}\, e^{\frac{\sigma^2}{9}} \left( e^{\frac{\sigma^2}{9}} - 1 \right). \qquad (121)$$

So, ***the variance of ET_Distance is***

$$\boxed{\sigma^2_{\text{ET\_Distance}} = C^2\, e^{-\frac{2}{3}\mu}\, e^{\frac{\sigma^2}{9}} \left( e^{\frac{\sigma^2}{9}} - 1 \right)}. \qquad (122)$$

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

$$\boxed{\sigma_{\text{ET\_Distance}} = C\, e^{-\frac{\mu}{3}}\, e^{\frac{\sigma^2}{18}} \sqrt{e^{\frac{\sigma^2}{9}} - 1}}. \qquad (123)$$

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

$$\left\langle \text{ET\_Distance}^3 \right\rangle = C^3\, e^{-\mu}\, e^{\frac{\sigma^2}{2}}. \qquad (124)$$

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

$$\left\langle \text{ET\_Distance}^4 \right\rangle = C^4\, e^{-\frac{4}{3}\mu}\, e^{\frac{8}{9}\sigma^2}. \qquad (125)$$

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 $\mu_i'$ by virtue of the recursion formula (see ref. [8])

$$K_i = \mu_i' - \sum_{k=1}^{i-1} \binom{i-1}{k-1} K_k \, \mu_{n-k}' . \qquad (126)$$

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:

$$\begin{cases} K_1 = \mu_1' \\ K_2 = \mu_2' - K_1^2 \\ K_3 = \mu_3' - 3K_1 K_2 - K_1^3 \\ K_4 = \mu_4' - 4K_1 K_3 - 3K_2^2 - 6K_2 K_1^2 - K_1^4 . \end{cases} \qquad (127)$$

These equations yield, respectively:

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

$$K_2 = C^2 e^{-\frac{2\mu}{3}} e^{\frac{\sigma^2}{9}} \left( e^{\frac{\sigma^2}{9}} - 1 \right) . \qquad (129)$$

$$K_3 = C^3 e^{-\mu} \left( e^{\frac{\sigma^2}{2}} - 3 e^{\frac{5\sigma^2}{18}} + 2 e^{\frac{\sigma^2}{6}} \right) . \qquad (130)$$

$$K_4 = \qquad (131)$$
$$= C^4 e^{\frac{4\mu}{3}} \left( e^{\frac{8\sigma^2}{9}} - 4 e^{\frac{5\sigma^2}{9}} - 3 e^{\frac{4\sigma^2}{9}} + 12 e^{\frac{\sigma^2}{3}} - 6 e^{\frac{2\sigma^2}{9}} \right) .$$

From these we derive the skewness

$$\frac{K_3}{(K_4)^{\frac{3}{2}}} =$$

$$= \frac{e^{\mu} \left( e^{\frac{\sigma^2}{2}} - 3 e^{\frac{5\sigma^2}{18}} + 2 e^{\frac{\sigma^2}{6}} \right)}{C^3 \left( e^{\frac{8\sigma^2}{9}} - 4 e^{\frac{5\sigma^2}{9}} - 3 e^{\frac{4\sigma^2}{9}} + 12 e^{\frac{\sigma^2}{3}} - 6 e^{\frac{2\sigma^2}{9}} \right)^{\frac{3}{2}}} . \qquad \ldots(132)$$

and the kurtosis

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

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 $f_{\text{ET\_Distance}}(r)$ of (113), and then set it equal to zero. This derivative is actually the derivative of the ratio of two functions of $r$, as its plainly appears from (113). Thus, let us set for a moment

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

where "E" stands for "exponent," Upon differentiating,
one gets

$$E'(r) = \frac{1}{2\sigma^2} \cdot 2 \left( \ln\left[ \frac{C^3}{r^3} \right] - \mu \right) \cdot \frac{1}{\frac{C^3}{r^3}} \cdot C^3 \cdot (-3) \cdot r^{-4}$$
$$= \frac{1}{\sigma^2} \cdot \left( \ln\left[ \frac{C^3}{r^3} \right] - \mu \right) \cdot (-3) \frac{1}{r} . \qquad (135)$$

But the probability density function (113) now reads

$$f_{\text{ET\_Distance}}(r) = \frac{3}{\sqrt{2\pi}\sigma} \cdot \frac{e^{-E(r)}}{r} \qquad (136)$$

So that its derivative is

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

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

Setting this derivative equal to zero means setting

$$E'(r)\cdot r + 1 = 0 \qquad (138)$$

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

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

Rearranging, this becomes

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

that is

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

whence

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

and finally

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

***This is the most likely* ET_Distance *from Earth.***

How likely ?

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

$$\boxed{\text{Peak Value of } f_{\text{ET\_Distance}}(r) \equiv f_{\text{ET\_Distance}}(r_{\text{mode}}) = \frac{3}{C\sqrt{2\pi}\sigma}\cdot e^{\frac{\mu}{3}}\cdot e^{\frac{\sigma^2}{18}}.}$$

...(144)

***This is the peak height in the pdf*** $f_{\text{ET\_Distance}}(r)$.

Next to the mode, the median $m$ (ref. [9]) is one more statistical number used to characterize any probability distribution. It is defined as 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 $m$:

$$\int_0^m f_{\text{ET Distance}}(r)dr = \frac{1}{2} \qquad (145)$$

Upon replacing (113), this becomes

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

In order to find $m$, we may ***not*** differentiate (146) with respect to $m$, since the "precise" factor ½ on the right would then disappear into a zero. On the contrary, we may try to perform the obvious substitution

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

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

| | |
|---|---|
| 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 | $f_{\text{ET\_Distance}}(r) = \frac{3}{r} \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{\left(\ln\left\lvert \frac{6\,R_{Galaxy}^2\,h_{Galaxy}}{r^3}\right\rvert - \mu\right)^2}{2\sigma^2}}$ |
| (Defining the positive numeric constant $C$) | $C = \sqrt[3]{6\,R_{Galaxy}^2\,h_{Galaxy}} \approx 28845$ light years |
| Mean value | $\langle \text{ET\_Distance} \rangle = C\,e^{-\frac{\mu}{3}}\,e^{\frac{\sigma^2}{18}}$ |
| Variance | $\sigma^2_{\text{ET Distance}} = C^2\,e^{-\frac{2}{3}\mu}\,e^{\frac{\sigma^2}{9}}\left(e^{\frac{\sigma^2}{9}} - 1\right)$ |
| Standard deviation | $\sigma_{\text{ET Distance}} = C\,e^{-\frac{\mu}{3}}\,e^{\frac{\sigma^2}{18}}\sqrt{e^{\frac{\sigma^2}{9}} - 1}$ |
| All the moments, i.e. $k$-th moment | $\langle \text{ET\_Distance}^k \rangle = C^k\,e^{\frac{k\mu}{3}}\,e^{\frac{k^2\sigma^2}{18}}$ |
| Mode (= abscissa of the probability density function peak) | $r_{\text{mode}} \equiv r_{peak} = C\,e^{-\frac{\mu}{3}}\,e^{-\frac{\sigma^2}{9}}$ |
| Value of the Mode Peak | Peak Value of $f_{\text{ET\_Distance}}(r) =$ $\equiv f_{\text{ET\_Distance}}(r_{\text{mode}}) = \frac{3}{C\sqrt{2\pi}\,\sigma} \cdot e^{\frac{\mu}{3}} \cdot e^{\frac{\sigma^2}{18}}$ |
| Median (= fifty-fifty probability value for ET_Distance) | $\text{median} = m = Ce^{\frac{\mu}{3}}$ |
| Skewness | $\frac{K_3}{(K_4)_2^3} = \frac{e^{-\mu}\left(e^{\frac{\sigma^2}{2}} - 3e^{\frac{5\sigma^2}{18}} + 2e^{\frac{\sigma^2}{6}}\right)}{C^3\left(e^{\frac{8\sigma^2}{9}} - 4e^{\frac{5\sigma^2}{9}} - 3e^{\frac{4\sigma^2}{9}} + 12e^{\frac{\sigma^2}{3}} - 6e^{\frac{2\sigma^2}{9}}\right)^{\frac{3}{2}}}$ |
| Kurtosis | $\frac{K_4}{(K_2)^2} = e^{\frac{4\sigma^2}{9}} + 2e^{\frac{\sigma^2}{3}} + 3e^{\frac{2\sigma^2}{9}} - 6$ |
| Expression of $\mu$ 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 $\sigma^2$ in terms of the lower ($a_i$) and upper ($b_i$) limits of the Drake **uniform** input random variables $D_i$ | $\sigma^2 = \sum_{i=1}^{7}\sigma^2_{Y_i} = \sum_{i=1}^{7} 1 - \frac{a_i b_i[\ln(b_i) - \ln(a_i)]^2}{(b_i - a_i)^2}$ |

**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.

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

that is

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

Since from the definition (147) one obviously has erf(0)=0, (149) yields

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

whence finally

$$\boxed{\text{median}=m=C\,e^{\frac{\mu}{3}}}. \qquad (151)$$

***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.

## 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 values reported in Table 1. Then, the graph of the corresponding probability density function of the nearest ET_Distance, $f_{\text{ET Distance}}(r)$, is shown in Figure 6.

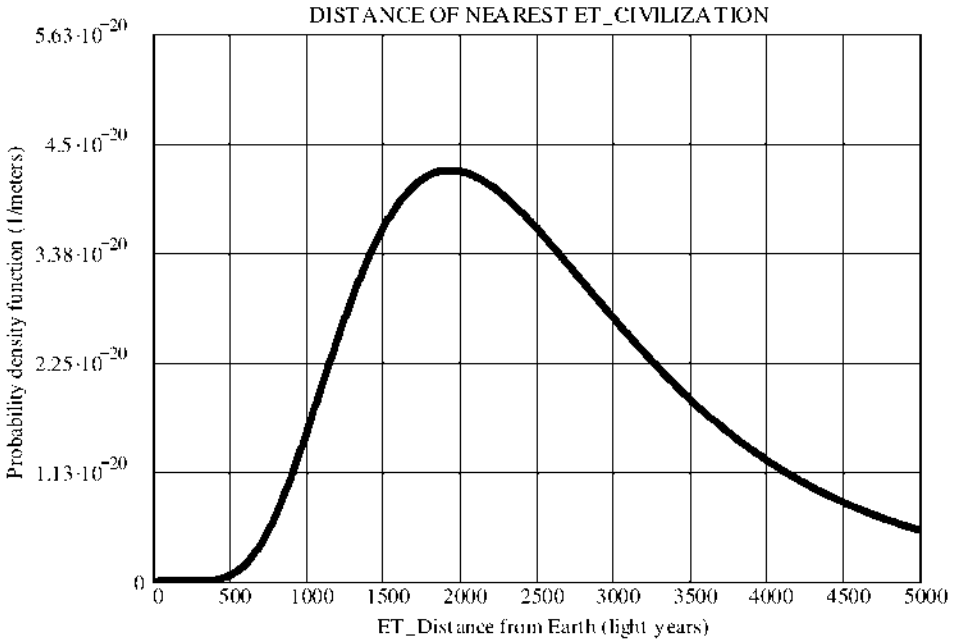



**Figure 6.** This is the probability of finding the nearest ExtraTerrestrial Civilization at the distance $r$ from Earth (in light years) if the values assumed in the Drake Equation are those shown in Table 1. The relevant probability density function $f_{\text{ET Distance}}(r)$ is given by equation (113). Its mode (peak abscissa) equals 1933 light years, but its mean value is higher since the curve has a high 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.***

From Figure 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

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

***This is the MOST LIKELY VALUE of the distance ut 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 (113) has an infinite tail on the right, as clearly shown in Figure 6, and hence its mean value must be higher than its peak value. As given by (119), its mean value is

$$r_{mean\_value} = C e^{\frac{\mu}{3}} e^{\frac{\sigma^2}{18}} \approx 2670 \text{ light years}. \quad (153)$$

***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 (123)

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

More precisely, this is the so called 1-sigma (distance) level. Probability theory then shows that the nearest ExtraTerrestrial civilization is expected to he 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, that is:

$$\int_{1361 \text{lightyears}}^{3979 \text{lightyears}} f_{ET\ Distance}(r)\, dr \approx 0.75 = 75\% \quad (155)$$

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.

## 9. 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 stutes thut "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.

## 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 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.

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 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 D. 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 and S. Unnikrishna Pillai, "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 and Seth Shostak, "Life in the Universe", Second Edition, Pearson - Addison-Wesley, San Francisco, 2007, ISBN 0-8053-4753-4. See in particular page 404.

## References

[1] Benford, Gregory, Jim and Dominic, "Cost Optimized Interstellar Beacons: SETI", arXiv.org web site (22 Oct. 2008).

[2] Carl Sagan, "Cosmos", Random House, New York, 1983. See in particular the pages 298-302.

[3] Bennet, Jeffrey, and Shostak, Seth, "Life in the Universe", second edition, Pearson – Addison Wesley, San Francisco, 2007. See in particular page 404.

[4] C. Maccone, "The Statistical Drake Equation", paper #IAC-08-A4.1.4 presented on October 1st, 2008, at the 59th International Astronautical Congress (IAC) held in Glasgow, Scotland, UK, September 29th thru October 3rd, 2008.