8-2019

Crowdsourced Sampling of a Composite Random Variable: Analysis, Simulation, and Experimental Test

Mark P. Silverman
Trinity College, mark.silverman@trincoll.edu

Scientific Research Publishing

Open Journal of Statistics, 2019, 9, 494-529
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Crowdsourced Sampling of a Composite Random Variable: Analysis, Simulation, and Experimental Test

M. P. Silverman

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

How to cite this paper: Silverman, M.P. (2019) Crowdsourced Sampling of a Composite Random Variable: Analysis, Simulation, and Experimental Test. *Open Journal of Statistics*, **9**, 494-529.
https://doi.org/10.4236/ojs.2019.94034

Received: July 12, 2019
Accepted: August 12, 2019
Published: August 15, 2019

Abstract

A composite random variable is a product (or sum of products) of statistically distributed quantities. Such a variable can represent the solution to a multi-factor quantitative problem submitted to a large, diverse, independent, anonymous group of non-expert respondents (the "crowd"). The objective of this research is to examine the statistical distribution of solutions from a large crowd to a quantitative problem involving image analysis and object counting. Theoretical analysis by the author, covering a range of conditions and types of factor variables, predicts that composite random variables are distributed log-normally to an excellent approximation. If the factors in a problem are themselves distributed log-normally, then their product is rigorously log-normal. A crowdsourcing experiment devised by the author and implemented with the assistance of a BBC (British Broadcasting Corporation) television show, yielded a sample of approximately 2000 responses consistent with a log-normal distribution. The sample mean was within ~12% of the true count. However, a Monte Carlo simulation (MCS) of the experiment, employing either normal or log-normal random variables as factors to model the processes by which a crowd of 1 million might arrive at their estimates, resulted in a visually perfect log-normal distribution with a mean response within ~5% of the true count. The results of this research suggest that a well-modeled MCS, by simulating a sample of responses from a large, rational, and incentivized crowd, can provide a more accurate solution to a quantitative problem than might be attainable by direct sampling of a smaller crowd or an uninformed crowd, irrespective of size, that guesses randomly.

Keywords

Crowdsourcing, Computer Modeling of Crowds, Monte Carlo Simulation, Large-Scale Sampling, Log-Normal Random Variable, Log-Normal

Distribution

1. Introduction: Estimation of an Unknown Composite Quantity by Large-Scale Sampling

The global reach of telecommunications media, including radio, television, and in particular the social media sites of the internet, make possible an ease and scale of statistical sampling hitherto inconceivable. Through use of these media, almost any question can, at least in principle, be posed to a large, anonymous, diverse, independent population of respondents, referred to in both technical and non-technical literature as the "crowd" [1]. This paper reports a comprehensive 1) analytical investigation, 2) Monte-Carlo simulation, and 3) experimental test of the distribution of a composite random variable (RV) representing a crowdsourced response to a question calling for a numerical answer. A composite RV is a product of two or more factor RVs. In the following sections it is shown that:

1) the most useful characteristic of a crowdsourced sample is its distribution function and not just a single statistic,

2) under conditions to be specified, a product of RVs is distributed log-normally to an excellent approximation, irrespective of the type or number or correlation of factor RVs,

3) computer simulation methods can model the response of a hypothetical rational crowd orders of magnitude larger than what actually might be practically attainable.

1.1. Background

To the author's knowledge, the first quantitative experiment in what today would be considered crowdsourcing was published by the English polymath and statistical innovator Sir Francis Galton in 1907 [2] [3]. Galton collected all the estimates of the weight of a dressed ox (*i.e.* the carcass weight) submitted by contestants at the annual West of England Fat Stock and Poultry Exhibition. To his surprise, he found that the sample median of 1207 pounds differed from the measured weight of 1198 pounds by a mere +0.8% and that the sample mean of 1197 pounds differed by an even smaller fractional error of −0.08%. The sample size was reported to be about 800. There was no mention of the sample distribution.

The idea underlying crowdsourcing—a term introduced in 2006—is that a large group of non-experts can collectively arrive at a more accurate estimate of some physical quantity or at a better decision regarding some policy, strategy, or treatment than a small group of experts [4]. This idea is a hypothesis to be examined experimentally, not a mathematical theorem, like Condorcet's jury theorem [5], subject to rigorous proof. Central among crowdsourcing issues in-

vestigated recently are questions regarding methods of sampling, quality control, bias elimination, and effectiveness [6] [7] [8] [9].

This paper addresses a different aspect of crowdsourcing closer in nature to the kind of experiment first performed by Galton. Questions whose responses can be represented numerically are especially suitable for statistical analysis. In this regard, the most useful statistical information to obtain from a crowdsourced sample is its distribution—*i.e.* the probability function for a discrete random variable (RV), or probability density function (PDF) for a continuous RV, or cumulative distribution function (CDF) for either kind of RV. For simplicity of discussion, the designation PDF will apply here to both discrete and continuous RVs. The importance of knowing the PDF or CDF of a distribution is that one can calculate from it, either theoretically or numerically, the exact population moments, which, depending on the size of an actual sample, can be significantly different from the sample moments. The population moments are estimates of the statistics that would result from a hypothetical infinitely large population of independent respondents. A virtually infinite sample size is what the internet and mass media have the potential to provide; it is also what computer-based Monte Carlo simulation (MCS) methods are already able to provide.

Throughout the past two decades, the author has conducted an array of experiments with students in his physics courses to investigate the validity of the crowdsourcing hypothesis [10]. In particular, tests were designed to examine whether groups of non-experts excelled over specialists in exercises relating to estimation, prediction, and deduction. Because sample sizes were relatively small (below 100), histograms of responses showed significant fluctuations, and the results did not appear to be accounted for by a universally applicable distribution. However, a larger-scale experiment (discussed in Section 4) to test crowdsourced sampling, implemented with the collaboration of a BBC One television show, yielded preliminary results that strongly suggested a log-normal distribution of estimates [10]. The present paper is the outcome of a more general and thorough analysis to extract information contained in a crowdsourced sample.

This paper reports a comprehensive study of the distribution of responses to a class of questions that calls for estimation of a composite random variable. A composite RV is formed by the product of two or more basis RVs. (The term "composite" is adopted from the designation of a "composite number" [11] as an integer expressible as the product of two or more integers, in contrast to a prime number.) This type of question is widely applicable to problems involving mathematics, statistics, physical sciences, engineering, bio-medical sciences, forensics, business and finance, military science, political science, archaeology, and other fields dependent upon quantitative reasoning.

An archetypical example of this class might be a question like the following: How many objects are contained within some partially disclosed geometric region? There are countless contexts in which such a question might arise and for which turning to a crowd for the answer may be good strategy. For example,

high-energy physicists may enlist a crowd to count events recorded in a complex bubble-chamber image; astronomers may enlist a crowd to search a deep-space image for some extraordinary astrophysical event or object; intelligence services may enlist a crowd to search reconnaissance images for locations or objects of military interest, archaeologists may enlist a crowd to search satellite images for structures associated with some cultural sites, and so on [12] [13] [14] [15] [16].

The specific problem examined in this paper is mathematically simple, but statistically informative: How many identical opaque objects are contained within a certain 3-dimensional volume of space seen only as a 2-dimensional image? The problem involves image analysis and object counting. A reasonable procedure to answer that question might entail the following: 1) Depending on the shape of the region, multiply together the appropriate geometric factors to obtain the volume, and then 2) multiply that volume by the numerical density, *i.e.* the number of objects in a unit volume. However, *none* of the needed numbers is known; all are representable by random variables whose realizations (*i.e.* estimates) by respondents in the crowd would be different. The sought-for RV would, in general, be a product (or sum of products) of 3 RVs relating to geometry and 1 RV characterizing the numerical density—or in all a product (or sum of products) of 4 RVs. The analyst is then faced with three general questions:

1) How are the basis RVs distributed?

2) What will be the distribution of the composite RV?

3) Which statistic of the composite RV should be taken to represent the physical value of the sought-for quantity?

By examining this archetypical question a) theoretically, b) computationally by Monte Carlo simulation, and c) experimentally, this paper addresses the preceding three questions.

1.2. Organization

The remainder of this paper is organized in the following way:

Section 2 investigates analytically the distribution of a composite random variable comprising independent basis RVs. Of particular interest are the cases in which the basis is either normally or log-normally distributed.

Section 3 investigates numerically by MCS the distributions of a composite variable comprising basis RVs whose distributions differ widely in shape parameters (skewness, kurtosis) for fixed location and scale parameters (mean, variance).

Section 4 reports 1) an experiment, implemented with the collaboration of a British national television show, to employ crowdsourcing as a means to estimate the number of opaque objects in a transparent receptacle, and 2) the use of MCS to predict the statistical results for a hypothetical much larger crowd incentivized to estimate rationally rather than guess randomly.

Section 5 concludes the paper with a summary of principal findings.

For the reader's convenience, the statistical abbreviations used in the paper

are listed below in alphabetical order.

BBC = British Broadcasting Corporation
CDF = cumulative distribution function
CF = characteristic function
CLT = central limit theorem
MCS = Monte Carlo simulation(s)
MGF = moment generating function
PDF = probability density function
RNG = random number generator
RV = random variable

2. Distribution of a Composite Random Variable

2.1. General Case

Consider a random variable Z defined by the product

$$Z = \prod_{i=1}^{N} X_i(\mu_i, \sigma_i) \tag{1}$$

where each basis variable $X_i(\mu_i, \sigma_i)$ in Equation (1) is characterized by its mean μ_i and standard deviation σ_i. At this point, the symbol X represents an arbitrary RV, and the parameters (μ_i, σ_i) for defining X_i were chosen to simplify the notation and analysis in sections to follow. Conventional statistical labeling of specific RVs that are relevant to this paper may include parameters different from the mean and standard deviation, as summarized in Table 1. The symbol $H(x)$ employed in Table 1 is the Heaviside function, also known as the step function, which we define here as

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{2}$$

(There are different definitions of $H(x)$ depending on the value assigned to $H(0)$ [17].) A statistical convention followed in this paper is to represent a random variable by an upper case letter, e.g. X, and a variate (*i.e.* sample or realization of the random variable) by a corresponding lower case letter, e.g. x.

Table 1. Representation of relevant random variables.

Distribution of RV X	Symbolic Representation	Significance of Parameters	PDF
normal or Gaussian	$N(\mu,\sigma^2)$	μ = mean of X σ = standard deviation of X	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
log-normal	$\Lambda(m,s^2)$	m = mean of $Y=\ln(X)$ s = standard deviation of Y	$\frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$
uniform	$U(a,b)$	a = lower boundary b = upper boundary	$\frac{1}{b-a}\left[H(x-a)-H(x-b)\right]$
Laplace	$La(\mu,\beta)$	μ = location parameter β = scale parameter	$\frac{1}{2\beta}\exp\left(-\frac{\lvert x-\mu\rvert}{\beta}\right)$

The natural logarithm of Z, which is a more convenient RV to work with, takes the form

$$Y = \ln(Z) = \sum_{i=1}^{N} \ln(X_i). \tag{3}$$

Reciprocally, one can write

$$Z = \exp(Y). \tag{4}$$

The strategy of the analysis in this section is to calculate the moment-generating function (MGF) of Y defined by the expectation operation

$$g_Y(t) \equiv \langle \exp(Yt) \rangle = \int e^{yt} p_Y(y) dy \tag{5}$$

in which $p_Y(y)$ is the PDF of Y, and t is a dummy variable the differentiation of which generates the statistical moments $k = 0,1,2,\cdots$ in the following way:

$$\langle Y^k \rangle = \left[d^k g_Y(t) / dt^k \right]_{t=0}. \tag{6}$$

If the MGF of a random variable does not exist, one can always use the characteristic function (CF) defined by

$$h_Y(t) \equiv \langle \exp(iYt) \rangle = \int e^{iyt} p_Y(y) dy \tag{7}$$

where Equation (7) is recognized as the Fourier transform of $p_Y(y)$ [18]. Each random variable is uniquely characterized by its MGF (if it exists) and CF [19]. By identifying the MGF or CF of Y, it may then be possible to determine the distribution of the sought-for composite variable Z.

Substitution of Equation (3) into Equation (5) leads to

$$g_Y(t) = \left\langle \exp\left(t \sum_{i=1}^{N} \ln(X_i) \right) \right\rangle = \left\langle \prod_{i=1}^{N} X_i^t \right\rangle = \prod_{i=1}^{N} \langle X_i^t \rangle \tag{8}$$

in which the last step—expectation of product equals product of expectations—is justified if the basis RVs are independent, as assumed to be the case in this section. This point will be revisited in Section 4.

From the form of Equation (1), a further condition of the analysis is that the basis RVs have well-defined means and variances. This is the same requirement as for the Central Limit Theorem (CLT) (see [19], 193-195). Re-express each $\boldsymbol{X_i}$ by the identity

$$X_i = \mu_i \left(1 + \frac{X_i - \mu_i}{\mu_i} \right) \equiv \mu_i (1 + \beta_i), \tag{9}$$

which defines the variable β_i, and substitute Equation (9) into Equation (8) to obtain

$$g_Y(t) = \prod_{i=1}^{N} \mu_i^t \left\langle (1 + \beta_i)^t \right\rangle. \tag{10}$$

If the basis variables X_i are to describe reasoned estimates rather than unrestricted random guesses, then it can be assumed that representative values of β_i are less than 1—*i.e.* that the expectations $\left\langle (X_i - \mu_i)^k \right\rangle$ are small compared

to μ_i^k for integer $k \geq 1$.

Expansion of the binomial factor in Equation (10) to order $O\left(\beta_i^3\right)$, followed by insertion of the expectation values

$$\begin{aligned} &\left\langle \beta_i \right\rangle = 0 \\ &\left\langle \beta_i^2 \right\rangle = \left(\sigma_i/\mu_i\right)^2 \quad \text{where } \sigma_i^2 = \left\langle \left(X-\mu_i\right)^2 \right\rangle \\ &\left\langle \beta_i^3 \right\rangle = \left(\lambda_i/\mu_i\right)^3 \quad \text{where } \lambda_i^3 = \left\langle \left(X-\mu_i\right)^3 \right\rangle \end{aligned} \tag{11}$$

leads to the approximate MGF

$$g_Y(t) \approx \exp\left(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2 + \frac{1}{6}\lambda_Y^3 t^3\right) \tag{12}$$

where

$$\mu_Y = \sum_{i=1}^{N}\left(\ln\left(\mu_i\right) - \frac{1}{2}\left(\sigma_i/\mu_i\right)^2 + \frac{1}{3}\left(\lambda_i/\mu_i\right)^3\right) \tag{13}$$

$$\sigma_Y^2 = \sum_{i=1}^{N}\left(\left(\sigma_i/\mu_i\right)^2 - \left(\lambda_i/\mu_i\right)^3\right) \tag{14}$$

$$\lambda_Y^3 = \sum_{i=1}^{N}\left(\lambda_i/\mu_i\right)^3 \tag{15}$$

respectively define the mean, variance, and skewness parameter λ_Y of Y. Under the conditions assumed in the foregoing analysis, MGF (12) shows that the distributions of Y, and therefore also Z, are not symmetric about the mean.

The author has been unable to find any source that identifies MGF (12) with a named distribution. However, upon neglect of skewness, Equation (12) takes the form

$$g_Y(t) \approx \exp\left(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right) \tag{16}$$

of the MGF of a normal RV [20]. By definition, if Y, as defined by Equation (3), is a normal RV denoted by $N\left(\mu_Y, \sigma_Y^2\right)$, then Z is a log-normal RV denoted by $\Lambda\left(\mu_Y, \sigma_Y^2\right)$; see **Table 1**. Note that the parameters defining the log-normal RV are the mean and variance of the associated normal RV and *not* the mean and variance of the log-normal RV itself.

For comparison, **Figure 1** shows plots of the PDF of a normal and log-normal distribution, as well as the PDFs of a uniform and Laplace distribution (which will be used in Section 3), all of the same mean $\left(\mu_X = 5\right)$ and standard deviation $\left(\sigma_X = 1\right)$. The figure illustrates that the significance of the standard deviation as a measure of statistical uncertainty (*i.e.* the width of the PDF) can vary markedly for different distributions, as summarized quantitatively in **Table 2**, which records the cumulative probability

$$\Delta_X \equiv \int_{\mu_X-\sigma_X}^{\mu_X+\sigma_X} p_X(x)\,\mathrm{d}x \tag{17}$$

of a variable X.

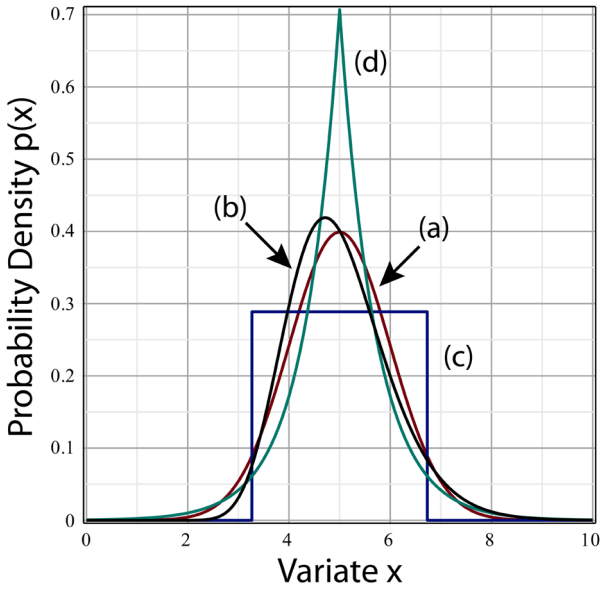

Figure 1. Graphical comparison of selected distributions of fixed mean $\mu = \langle x \rangle = 5$ and fixed standard deviation $\sigma = \sqrt{\langle x^2 \rangle - \mu^2} = 1$: (a) Gaussian (red), (b) log-normal (black), (c) uniform (blue), (d) Laplace (green).

Table 2. Comparative significance of 1 standard deviation uncertainty.

Distribution $X(a,b)$	**Mean** $\mu_X = 5$	**Variance** $\sigma_X^2 = 1$	**Probability** Δ_X $P(\lvert x-\mu_X\rvert \le \sigma_X)$
Normal $N(\mu_X, \sigma_X^2)$	μ_X	σ_X^2	$\operatorname{erf}\left(2^{-\frac{1}{2}}\right) \simeq 0.6827$
Log-Normal $\Lambda(m, s^2)$ $m = 1.5898$ $s = 0.1980$	$\exp\left(m + \frac{1}{2}s^2\right)$	$e^{2m}\left[e^{2s^2} - e^{s^2}\right]$	0.6940
Uniform $U(a,b)$ $a = 3.2679$ $b = 6.7321$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$	$\frac{1}{\sqrt{3}} \simeq 0.5774$
Laplace $La(\mu,\beta)$ $\mu = 5$ $\beta = 2^{-\frac{1}{2}} \simeq 0.7071$	μ	$2\beta^2$	$1 - e^{-\sqrt{2}} \simeq 0.7569$

Note that for variables *N*, *U*, and *La* the probability Δ_X that a sample falls within ± 1 standard deviation of the mean is a constant dependent on the type of distribution, but independent of the parameters of the distribution. For the log-normal variable, however, Δ_Λ has a complicated dependence on μ_Λ and σ_Λ

$$\Delta_\Lambda = \frac{1}{2}\left[\operatorname{erf}\left(\frac{\ln\left(\left(\mu_\Lambda + \sigma_\Lambda\right) + \frac{1}{4}\ln\left(\mu_\Lambda^2 + \sigma_\Lambda^2\right) - \ln\left(\mu_\Lambda\right)\right)}{\sqrt{2\ln\left(\mu_\Lambda^2 + \sigma_\Lambda\right) - 4\ln\left(\mu_\Lambda\right)}}\right)\right.$$

$$-\operatorname{erf}\left(\frac{\ln\left(\left(\mu_\Lambda-\sigma_\Lambda\right)+\frac{1}{4}\ln\left(\mu_\Lambda^2+\sigma_\Lambda^2\right)-\ln\left(\mu_\Lambda\right)\right)}{\sqrt{2\ln\left(\mu_\Lambda^2+\sigma_\Lambda\right)-4\ln\left(\mu_\Lambda\right)}}\right)\Bigg] \tag{18}$$

where the error function is defined by

$$\operatorname{erf}(x)\equiv\frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t\,. \tag{19}$$

The first column of Table 2 shows the values of the distribution parameters that lead to the fixed mean ($\mu_X=5$) and variance ($\sigma_X=1$) specified in the first row. The second and third columns of the table provide the theoretical relations connecting the parameters of each distribution to the mean and variance of the associated RVs.

In the analyses and experiments of this paper, it will be adequate to neglect the skewness of Y and adopt MGF (16), which identifies Y as a normal RV. In that case, it follows that Z takes the form

$$Z=\mathrm{e}^Y=\mathrm{e}^{\mu_Y+\sigma_Y W} \tag{20}$$

in which $W\equiv N(0,1)$ is a standard normal RV. The justification of Equation (20) is that an arbitrary normal RV $N\left(\mu,\sigma^2\right)$ can be written in the form [21]

$$N\left(\mu,\sigma^2\right)=\mu+\sigma W\,. \tag{21}$$

Equation (20) leads directly by integration to the expectation values of Z

$$\begin{aligned}\left\langle Z^k\right\rangle&=\left\langle \mathrm{e}^{k\mu_Y+k\sigma_Y W}\right\rangle=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(k\mu_Y+k\sigma_Y w\right)\mathrm{e}^{-\frac{w^2}{2}}\,\mathrm{d}w\\&=\exp\left(k\mu_Y+\frac{1}{2}k^2\sigma_Y^2\right)\end{aligned} \tag{22}$$

where the PDF of W is given in Table 1 by setting $\mu=0$ and $\sigma=1$ in the PDF of $N\left(\mu,\sigma^2\right)$.

From Equation (22), the mean and variance of the log-normal RV are then

$$\begin{aligned}\mu_Z&=\exp\left(\mu_Y+\frac{1}{2}\sigma_Y^2\right)\\\sigma_Z^2&=\exp\left(2\mu_Y\right)\left(\exp\left(2\sigma_Y^2\right)-\exp\left(\sigma_Y^2\right)\right)\end{aligned} \tag{23}$$

and the inverse relations, which will be needed later, can be shown to be

$$\begin{aligned}\mu_Y&=\ln\left(\mu_Z^2\Big/\sqrt{\mu_Z^2+\sigma_Z^2}\right)\\\sigma_Y^2&=\ln\left(\left(\mu_Z^2+\sigma_Z^2\right)\Big/\mu_Z^2\right)\end{aligned} \tag{24}$$

Although the RV Y is distributed symmetrically about its mean, the distribution of Z itself is skewed. From Equation (22) the third moment about the mean, to which skewness is proportional, can be shown to be

$$\left\langle\left(Z-\mu_Z\right)^3\right\rangle=\left\langle Z^3\right\rangle-3\left\langle Z^2\right\rangle\mu_Z+2\mu_Z^3=\mathrm{e}^{3\mu_Y}\left(\mathrm{e}^{\frac{9}{2}\sigma_Y^2}-\mathrm{e}^{\frac{5}{2}\sigma_Y^2}+\mathrm{e}^{\frac{3}{2}\sigma_Y^2}\right) \tag{25}$$

It is useful to note that Equation (20) provides an even more direct way than integration of the PDF at arriving at Equation (22) for the moments of Z since $\langle Z^k \rangle = \langle e^{Yk} \rangle$ takes the form of the MGF (16) of a normal RV, upon replacing the dummy variable t with the moment order k.

The seminal findings of this section may be summarized as follows:

1) *A random variable Z composed of the product of* 2 *or more factor RVs for which the ratio of standard deviation to mean is* <1 *is distributed log-normally to the extent that the skewness* (*and higher order moments*) *of* $\ln(Z)$ *can be neglected.*

2) *To find the parameters of the distribution of a log-normal RV Z, one first transforms the data* (*e.g. sample or simulation*) *by* $y_i = \ln(z_i)$ *to obtain the distribution of the associated normal RV Y which is symmetric about its mean.*

In concluding this section, a point of comparison is in order regarding the CLT for the sum of independent RVs and relation (20) for the product of independent RVs. In brief, the CLT holds that the sum (e.g. mean) of a sufficiently large number N of identically distributed, independent RVs $X_i(\mu_X, \sigma_X), i = 1, \cdots, N$ converges to a normal RV irrespective of the distribution of X, provided that the X_i have a well-defined mean and variance [22] [23]. In theory, the number N is infinitely large, but in practice it can be well below 10; see Ref [10], pp. 36-38. In contrast, the foregoing demonstration that a product of RVs is distributed approximately log-normally

$$Z = \prod_{i=1}^{N} X_i(\mu_i, \sigma_i) \to \Lambda\left(\sum_{i=1}^{N} \ln(\mu_i), \sum_{i=1}^{N} (\sigma_i/\mu_i)^2\right) \tag{26}$$

holds for any number of factors $N \geq 2$ under the previously specified conditions. Moreover, the individual independent factors $X_i(\mu_i, \sigma_i)$ need not have identical distribution parameters, nor even all be the same type of variable X. The parameters of Λ shown in Equation (26) are from Equations (13), (14), (15) with neglect of the skewness parameter and terms of order $(\sigma_i/\mu_i)^2$ in the mean. This reduction has been found satisfactory in accounting for the Monte Carlo simulations and experimental results discussed in later sections.

2.2. Special Case: Product of Normal RVs $X_i = N_i(\mu_i, \sigma_i^2)$

The log-normal distribution of a composite RV derived in the previous section is an approximate relation valid to the extent that certain conditions are fulfilled. In the special case where the factors X_i of the product (26) defining Z are normal RVs, an alternative expression for the PDF of $Y = \ln(Z)$ can be derived by means of the CF. This is an important case because the normal distribution satisfactorily describes measurements or estimates of many biomedical variables, physical variables, and variables relating to business management and finance, among others [24] [25] [26].

From Equation (8) for the MGF of Y and the definition (7) for the CF, one can write

$$h_Y(t)=\prod_{j=1}^{N}\left\langle X_j^{it}\right\rangle=\prod_{j=1}^{N}\int x_j^{it}p_{X_j}(x_j)\mathrm{d}x_j\equiv\int \mathrm{e}^{iyt}p_Y(y)\mathrm{d}y \tag{27}$$

for the CF of Y, where the summation index has been changed from i to j so as not to be confounded with the unit imaginary $i=\sqrt{-1}$. The inverse Fourier transform of Equation (27) then yields the PDF of Y

$$\begin{aligned}p_Y(y)&=(2\pi)^{-1}\int_{-\infty}^{\infty}\mathrm{e}^{-iyt}h_Y(t)\mathrm{d}t\\&=(2\pi)^{-1}\int_{-\infty}^{\infty}\mathrm{e}^{-iyt}\left[\prod_{j=1}^{N}\int x_j^{it}p_{X_j}(x_j)\mathrm{d}x_j\right]\mathrm{d}t\end{aligned} \tag{28}$$

in which the second equality of Equation (27) was substituted for $h_Y(t)$ in the first line of Equation (28). The PDF of Z is calculable from the PDF of Y by the following transformation (see Appendix 1):

$$p_Z(z)=z^{-1}p_Y(\ln(z)). \tag{29}$$

Substitution of relation (21) for each normal factor X_j into (27) leads to

$$h_Y(t)=\prod_{j=1}^{N}(2\pi)^{-\frac{1}{2}}\int_{-\mu_j/\sigma_j}^{\infty}(\mu_j+\sigma_j x)^{it}\mathrm{e}^{-x^2/2}\mathrm{d}x, \tag{30}$$

which can be re-expressed in the form

$$h_Y(t)=\prod_{j=1}^{N}(2\pi)^{-\frac{1}{2}}\mathrm{e}^{it\ln(\mu_j)}\int_{-\alpha_j^{-1}}^{\infty}\exp\left(it\ln(1+\alpha_j x)-x^2/2\right)\mathrm{d}x \tag{31}$$

where

$$\alpha_j=\sigma_j/\mu_j. \tag{32}$$

Equation (31) is an exact expression for the CF of Y, but, to the author's knowledge, cannot be integrated in closed form. However, for $\alpha_j<1$, expansion of the logarithm in a Taylor series to order α_j^2 results in the closed form expression

$$h_Y(t)=\prod_{j=1}^{N}\left[\frac{\mu_j^{it}\exp\left(-\frac{1}{2}\alpha_j^2t^2\Big/\left(1+i\alpha_j^2t\right)\right)}{\sqrt{1+i\alpha_j^2t}}\right]. \tag{33}$$

Substitution of CF (33) into Equations (28) and (29) provides a more accurate PDF of Y and Z than the PDF of log-normal (26).

If $\alpha_j^2<1$ for each factor X_j in Equation (27), then one can approximate $h_Y(t)$ in Equation (33) by

$$h_Y(t)\simeq\prod_{j=1}^{N}\mu_j^{it}\exp\left(-\frac{1}{2}\alpha_j^2t^2\right) \tag{34}$$

which, substituted into the integral in Equation (28), leads to the Gaussian distribution

$$Y=\ln\left(\prod_{i=1}^{N}N_i(\mu_i,\sigma_i^2)\right)\to N\left(\sum_{i=1}^{N}\ln(\mu_i),\sum_{i=1}^{N}(\sigma_i/\mu_i)^2\right) \tag{35}$$

for Y and the log-normal distribution (26) for Z.

As an example to illustrate the stages of the analysis, consider the composite RV

$$\begin{aligned} Z &= \prod_{i=1}^{4} X_i\left(\mu_i, \sigma_i\right) \\ &= N_1\left(1.0, (0.2)^2\right) N_2\left(4.0, (0.5)^2\right) N_3\left(6.0, (1.0)^2\right) N_4\left(10.0, (1.5)^2\right) \end{aligned} \tag{36}$$

and associated log-product $Y = \ln(Z)$. In **Figure 2** are plotted the real part $\mathrm{Re}(F)$ (red), imaginary part $\mathrm{Im}(F)$ (blue), and magnitude $|F|$ (dashed black) of the Fourier transform $F(t) = h_Y(t)$ given by Equation (33) as a function of t. Although t serves in the MGF as a dummy variable for computation of statistical moments by differentiation, in the CF t is equivalent to a spatial or temporal frequency [27] [28]. $|F|$ and $\mathrm{Re}(F)$ are seen to be symmetric, and $\mathrm{Im}(F)$ antisymmetric, about $t = 0$, extending over a range $\Delta t \approx 20$ from −10 to +10. **Figure 3** shows plots of $p_Y(y)$, Equation (28), as calculated by (1) numerical integration of the Fourier transform of the exact CF (31) (solid red), (2) the analytical approximation (33) to the CF (dashed blue), and (3) the PDF of the normal RV (35) (solid green). Profiles (1) and (2) are seen to be nearly indistinguishable, and both are well approximated by the Gaussian profile (3). **Figure 4** shows plots of $p_Z(z)$ as calculated by (1) numerical integration of the transformation (29) of the exact PDF of Y (solid red), and (2) the PDF of the approximate log-normal RV (26) (dashed blue). The exact and log-normal PDFs of Z closely match, apart from a slight forward shift of the peak of the log-normal profile.

2.3. Special Case: Product of Log-Normal RVs $X_i = \Lambda_i\left(m_i, s_i^2\right)$

The ubiquity of the normal distribution is primarily a consequence of the CLT, which is a limiting theorem for the sum of a large (in theory, infinite) number of random variables. Moreover, the distributed variable can take—or, as a matter of practicality, be thought to take—both positive and negative values, since the Gaussian PDF is normalized to unity only when integrated over the entire real axis. The log-normal distribution also occurs widely, particularly in reference to activities that involve counting, measuring, or observing the attributes of real physical things. Such activities underlie many kinds of problems for which crowdsourced solutions can be sought. The distributed variable then takes on only non-negative real values and is expected to be intrinsically skewed, since its least value cannot be below zero, whereas its upper limit is open.

Consider, therefore, a composite variable Z comprised of log-normal factors

$$Z = \prod_{i=1}^{N} \Lambda_i\left(m_i, s_i^2\right) \tag{37}$$

with PDF of the form (see [21], pp. 131-134)

$$p_Z\left(z \middle| m, s\right) = \frac{1}{\sqrt{2\pi} s z} \exp\left(-\left(\ln(z) - m\right)^2 \Big/ 2s^2\right). \tag{38}$$

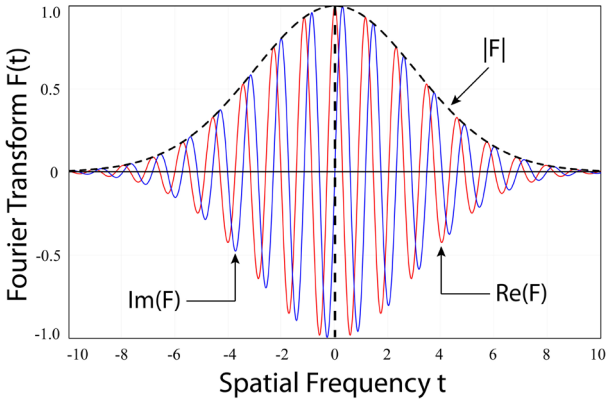

Figure 2. Fourier transform $F(t)$ of the characteristic function of $Y=\ln(Z)$, Equation (33), where $Z=\prod_{i=1}^{4}N_i(\mu_i,\sigma_i^2)$ is defined by parameters $\mu_i=\{1.0,4.0,6.0,10.0\}$ and $\sigma_i=\{0.2,0.5,1.0,1.5\}$: (a) real part (solid red), (b) imaginary part (solid blue), (c) magnitude (dashed black).

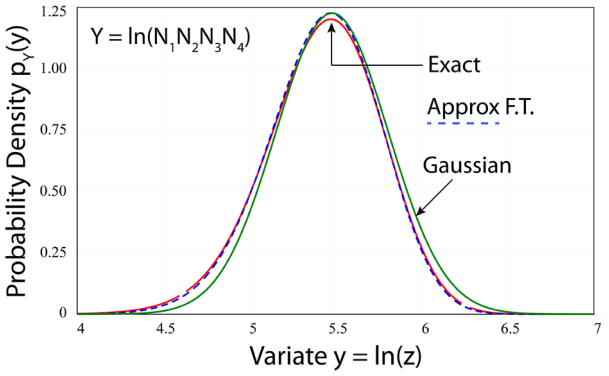

Figure 3. PDF of $Y=\ln Z$ defined in **Figure 2**, as calculated from the Fourier transform of the exact CF Equation (31) (solid red), the Fourier transform of the analytical approximation Equation (33) (dashed blue), and the Gaussian Equation (35) (solid green).

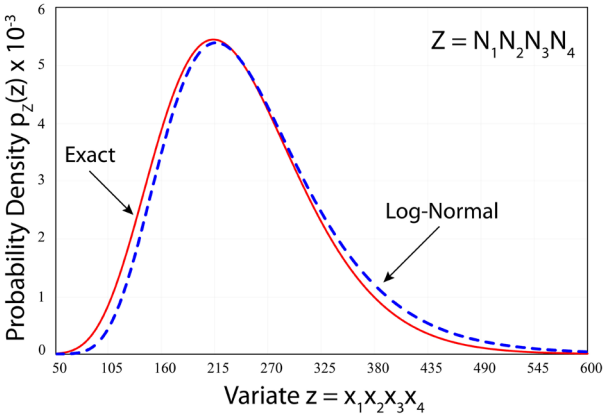

Figure 4. PDF Z defined in **Figure 2**, as calculated from the exact transformation relation (29) (solid red) and from the PDF of log-normal variable (26) (dashed blue).

It then readily follows from the inverse of Equation (29) (see Appendix 1) that the PDF of the variable $Y=\ln(Z)$ has the form

$$p_Y(y|m,s)=\frac{1}{\sqrt{2\pi}s}\exp\left(-(y-m)^2/2s^2\right) \tag{39}$$

which shows that Y is a Gaussian RV of mean m and variance s^2, *i.e.* $Y=N(m,s^2)$.

Thus, taking the log of Equation (37) leads to the chain of relations

$$Y=\ln(Z)=\sum_{i=1}^{N}\ln\left(\Lambda_i\left(m_i,s_i^2\right)\right)=\sum_{i=1}^{N}N_i\left(m_i,s_i^2\right)=N\left(\sum_{i=1}^{N}m_i,\sum_{i=1}^{N}s_i^2\right) \tag{40}$$

from which it follows that Z, itself, is a log-normal RV

$$Z=\Lambda\left(m,s^2\right) \tag{41}$$

with

$$\begin{aligned} m&=\sum_{i=1}^{N}m_i\\ s^2&=\sum_{i=1}^{N}s_i^2 \end{aligned} \tag{42}$$

Stated formally: *The product of log-normal RVs is a log-normal RV with parameters given by* Equation (42). Note that the preceding result, Equation (41), is exact; no approximations regarding either the number of factor RVs or the relative magnitudes of parameters m_i and s_i have been made.

From Equation (23) the mean and variance of Z, defined by Equation (37), is then

$$\begin{aligned} \mu_Z&=\exp\left(\sum_{i=1}^{N}m_i+\frac{1}{2}\sum_{i=1}^{N}s_i^2\right)\\ \sigma_Z^2&=\exp\left(2\sum_{i=1}^{N}m_i\right)\left(\exp\left(2\sum_{i=1}^{N}s_i^2\right)-\exp\left(\sum_{i=1}^{N}s_i^2\right)\right) \end{aligned} \tag{43}$$

3. Monte-Carlo Simulations of a Composite Random Variable

In this section the distribution of responses to the kind of archetypical problem posed at the end of Section 1.1 is examined numerically by means of Monte-Carlo simulations (MCS) employing four basic types of two-parameter RVs $X_i(\mu_i,\sigma_i)$: 1) normal, 2) uniform, 3) Laplace, and 4) log-normal. The means μ_i and standard deviations σ_i of the factor RVs are respectively those of the arguments of the four RVs in Equation (36):

$$\begin{aligned} (\mu_1,\sigma_1)&=(1.0,0.2)\\ (\mu_2,\sigma_2)&=(4.0,0.5)\\ (\mu_3,\sigma_3)&=(6.0,1.0)\\ (\mu_4,\sigma_4)&=(10.0,1.5) \end{aligned} \tag{44}$$

The four types of RVs differ markedly, however, in skewness and kurtosis, which characterize the shape of the PDF, as shown in **Figure 1**. Consider X_1 to

represent the numerical density of objects in a receptacle, and the variables X_2, X_3, X_4 to characterize the 3-dimensional receptacle geometry. The physical quantity for which an estimate is sought is then represented by the variable $Z = X_1 X_2 X_3 X_4$. If Z is satisfactorily described by a log-normal RV, then $Y = \ln(X_1 X_2 X_3 X_4)$ should be well-approximated by a Gaussian RV.

Each of the four simulations of the composite variable Z reported in the subsections to follow comprises $n = 10^6$ independent samples from a random number generator (RNG) corresponding to one of the four basis RVs listed above. The simulated variates $\{x_{i,j}\}$ $(i = 1,2,3,4; j = 1,\cdots,n)$ are partitioned into uniform bins of width $\Delta x = 0.1$; the resulting variates $\{y_j\}$, $\{z_j\}$ are partitioned into uniform bins of width $\Delta y = 0.1$, $\Delta z = 10.0$ (if $X = N, U, La$) or 15.0 (if $X = \Lambda$). To get a sense of scale, note that the product of the four means in Equation (44) is 240 and that $\ln(240) \approx 5.48$. It is to be expected, therefore, that, neglecting skewness, the histogram of Z should be centered at a point near 240, whereas the symmetric histogram of Y should be centered at close to 5.48, which lies between the centers of histograms X_2 and X_3.

Superposed on each of the generated histograms in the figures to follow will be the relevant theoretical PDF (solid red): 1) PDF of the corresponding RNG for the basis variables $\{X_i\}$, 2) log-normal PDF (26) (if $X = N, La, U$) or (41) (if $X = \Lambda$) for *Z*, and 3) normal PDF (35) (if $X = N, La, U$) or (40) (if $X = \Lambda$) for *Y*. The analysis of Section 2.1 leads to an important prediction concerning the four Monte Carlo simulations:

- *Each simulation, although generated with a different type of basis variable X, should lead within statistical uncertainties to identical histograms for Z and Y.*

The preceding prediction follows from the fact that the means and variances of Z and Y depend only on the means and variances (44) of the basis variables X_i, and *not* on the type of RV symbolized by *X*.

From the ungrouped variates of each MCS

$$z_j = x_{1,j} x_{2,j} x_{3,j} x_{4,j} \tag{45}$$

$$y_j = \ln(x_{1,j} x_{2,j} x_{3,j} x_{4,j}), \tag{46}$$

one can calculate the sample mean and sample variance of Z by two different approaches, both employing relations deduced from the method of maximum likelihood (ML) [29]. The first approach is to calculate the sample mean (m_Z) and sample variance (s_Z^2) directly from the set of variates (45)

SAMPLE: Z

$$m_Z = \frac{1}{n}\sum_{j=1}^{n} z_j$$
$$s_Z^2 = \frac{1}{n}\sum_{j=1}^{n}(z_j - m_Z)^2 \tag{47}$$

The second approach is to calculate the sample mean (m_Y) and sample variance (s_Y^2) from the set of Gaussian variates (46)

$$\text{SAMPLE: } Y \qquad \begin{aligned} m_Y &= \frac{1}{n}\sum_{j=1}^{n} y_j \\ s_Y^2 &= \frac{1}{n}\sum_{j=1}^{n}\left(y_j - m_Y\right)^2 \end{aligned} \tag{48}$$

and use relations (48) to deduce the sample mean $\left(M_Z\right)$ and sample variance $\left(S_Z^2\right)$ as follows from Equation (23)

$$\text{SAMPLE: } Z(Y) \qquad \begin{aligned} M_Z &= \exp\left(m_Y + \frac{1}{2}s_Y^2\right) \\ S_Z^2 &= \exp\left(2m_Y\right)\left(\exp\left(2s_Y^2\right) - \exp\left(s_Y^2\right)\right) \end{aligned} \tag{49}$$

Agreement of statistics (47) and (49) would be indicative that the variates of Z were distributed log-normally.

Comparison of sample statistics with theory for each of the simulations to follow are summarized in Table 3.

3.1. Normal Basis $X = N$

The normal distribution is defined by its mean and variance (see Table 1). The basis variables of the simulation are therefore $N_i\left(\mu_i, \sigma_i^2\right)$, $i = 1,2,3,4$, as shown in Equation (36) with parameters as defined in list (44). For purposes of comparing histogram shapes, it is noted that the skewness and kurtosis of a normally distributed RV are respectively

$$Sk_X^{(N)} \equiv \left\langle \left(\left(X - \mu_X\right)/\sigma_X \right)^3 \right\rangle = 0 \tag{50}$$

$$K_X^{(N)} \equiv \left\langle \left(\left(X - \mu_X\right)/\sigma_X \right)^4 \right\rangle = 3\,. \tag{51}$$

Skewness (50) is a measure of symmetry of the PDF with respect to the mean. Kurtosis (51) is a measure of the shape of the tails of the PDF. A distribution

Table 3. Statistics of Monte Carlo simulations of $Z = X_1X_2X_3X_4$.

Basis Parameters $X_i\left(\mu_i,\sigma_i\right)\left(i=1,2,3,4\right)$	$\left(\mu_i,\sigma_i\right)=\left(1,0.2\right),\left(4,0.5\right),\left(6,1.0\right),\left(10,1.5\right)$				
	Sample (n = 1,000,000)			**Theory** $Z=\Lambda\left(\mu_Y,\sigma_Y^2\right)$	
Basis Variables X_i	**Sample Z**	**Sample Y**	**Sample $Z(Y)$**	Y	Z
Normal	$m_Z = 240.01$ $s_Z = 79.54$	$m_Y = 5.43$ $s_Y = 0.34$	$M_Z = 240.48$ $S_Z = 83.91$	$\mu_Y = 5.48$ $\sigma_Y = 0.33$	$M_Z = 253.05$ $\Sigma_Z = 84.58$
Uniform	$m_Z = 240.02$ $s_Z = 79.53$	$m_Y = 5.43$ $s_Y = 0.33$	$M_Z = 240.23$ $S_Z = 82.11$	—	—
Laplace	$m_Z = 240.01$ $s_Z = 79.48$	$m_Y = 5.42$ $s_Y = 0.38$	$M_Z = 243.01$ $S_Z = 96.58$	—	—
Log-Normal	$m_Z = 239.97$ $s_Z = 79.50$	$m_Y = 5.43$ $s_Y = 0.32$	$M_Z = 239.97$ $S_Z = 79.49$	$\mu_Y = 5.43$ $\sigma_Y = 0.32$	$M_Z = 240.00$ $\Sigma_Z = 79.60$

with "fat tails" (leptokurtic) has a higher probability than normal of extreme events, in contrast to a distribution with "thin tails" (platykurtic) for which the probability of extreme events is lower than normal.

Figure 5 shows a panoramic plot of the histograms of X_1 (green), X_2, X_3, X_4 (gray), and $Y=\ln\left(X_1X_2X_3X_4\right)$ (blue), where $Z=X_1X_2X_3X_4$. As expected, all the histograms in the figure appear to be Gaussian, and the histogram of Y lies between the histograms of X_2 and X_3.

Panels A and B of Figure 6 respectively show in greater detail the histograms of Z and Y, bordered by the profiles of the corresponding log-normal and normal PDFs. In panel A, the right tail of the histogram is marginally less skewed than predicted by the log-normal model. In panel B, the left tail of the histogram is marginally more skewed than the symmetric profile of the Gaussian PDF. Nevertheless, in both panels, the theoretical profiles satisfactorily match the peak and overall shape of the histograms.

3.2. Uniform Basis $X = U$

A uniform RV $X\left(\mu,\sigma\right)=U\left(a,b\right)$ is symbolized by its upper and lower boundaries $\left(b>a\right)$. From Table 2 it follows that the mean and standard deviation of X are related to the boundary parameters by

$$\begin{aligned} a &= \mu-\sqrt{3}\sigma \\ b &= \mu+\sqrt{3}\sigma \end{aligned} \tag{52}$$

The basis RVs $X_i\left(\mu_i,\sigma_i\right), i=1,2,3,4$ of the simulation, which have the same means and variances as the basis RVs of Section 3.1, are then respectively

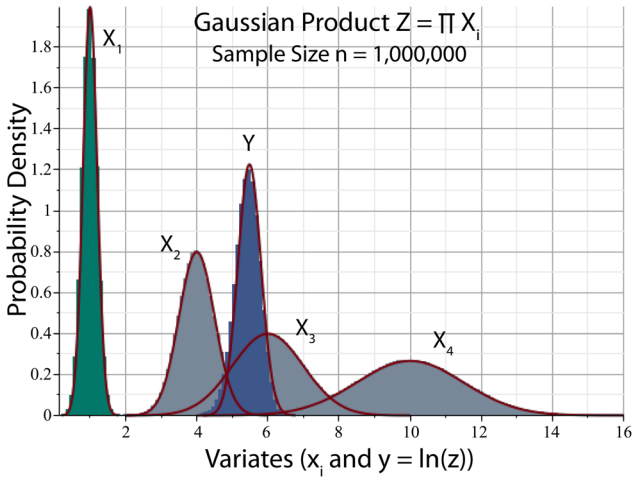

Figure 5. Monte-Carlo simulated histograms of normal variables $X_i\left(\mu_i,\sigma_i\right)=N_i\left(\mu_i,\sigma_i^2\right)$ with means μ_i and standard deviations σ_i listed in (44), and $Y=\ln\left(\prod_{i=1}^{4}X_i\right)$ (blue). X_1 (green) represents number density; X_2, X_3, X_4 (gray) represent geometric dimensions. The sample size is $n=10^6$. Each histogram is enveloped by its associated Gaussian PDF (red).

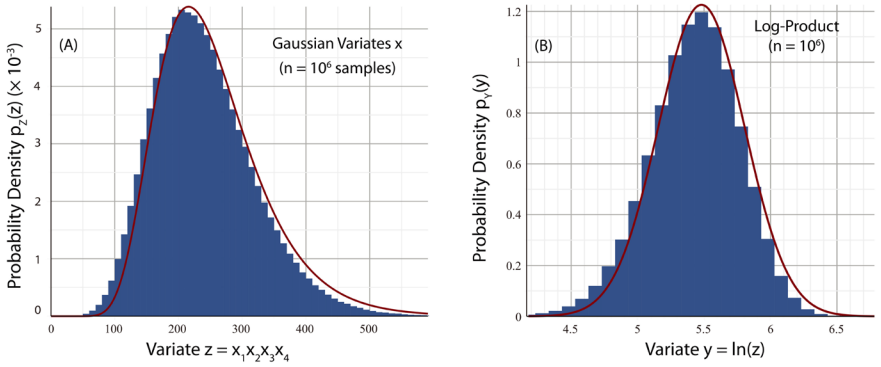

Figure 6. Panel A: Histogram of Gaussian product Z of **Figure 5** enveloped by PDF of log-normal variable (26) with values (44). Panel B: Histogram of $Y=\ln(Z)$ of **Figure 5** enveloped by PDF of Gaussian variable (35).

$$\begin{aligned} X_1 &= U_1(0.6536, 1.3464) \\ X_2 &= U_2(3.1340, 4.8660) \\ X_3 &= U_3(4.2679, 7.7321) \\ X_4 &= U_4(7.4019, 12.5981) \end{aligned} \tag{53}$$

The skewness and kurtosis of a uniformly distributed RV are

$$Sk_X^{(U)} = 0 \tag{54}$$

$$K_X^{(U)} = 9/5 = 1.8\,. \tag{55}$$

Figure 7 shows a panoramic plot of the histograms X_i, which have tails that drop vertically in comparison to the Gaussian histograms of **Figure 5**. Equation (55) establishes that a uniform RV is platykurtic, as is apparent from **Figure 1**. Nevertheless, the histogram of $Y=\ln(X_1X_2X_3X_4)$ is again well represented by a Gaussian PDF, which indicates that $Z=X_1X_2X_3X_4$ should be reasonably well described by a log-normal RV, as shown in greater detail in **Figure 8**.

3.3. Laplace Basis *X* = *La*

A Laplace RV $X(\mu,\sigma)=La(\mu,\beta)$ is symbolized by a location parameter μ corresponding to the mean of X and a scale parameter β related to the standard deviation of X by

$$\beta = 2^{-\frac{1}{2}}\sigma \tag{56}$$

(see **Table 2**). The four basis variables of the simulation, which have the same means and variances as the basis RVs of Section 3.1, are then respectively

$$\begin{aligned} X_1 &= La_1(1.0, 0.1414) \\ X_2 &= La_2(4.0, 0.3536) \\ X_3 &= La_3(6.0, 0.7071) \\ X_4 &= La_4(10.0, 1.0607) \end{aligned} \tag{57}$$

The skewness and kurtosis of a Laplace distributed RV are

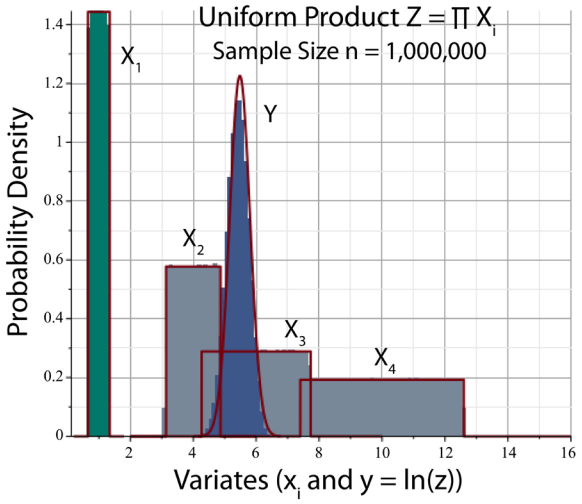

Figure 7. Monte-Carlo simulated histograms of uniform variables $X_i(\mu_i,\sigma_i)=U_i(a_i,b_i)$ with means μ_i and standard deviations σ_i listed in (44), and $Y=\ln\left(\prod_{i=1}^{4}X_i\right)$. Histograms X_i are enveloped by their associated uniform PDFs (red). Histogram Y is enveloped by the Gaussian PDF of **Figure 5**. Sample size, symbolic notation, and color coding are the same as in **Figure 5**.

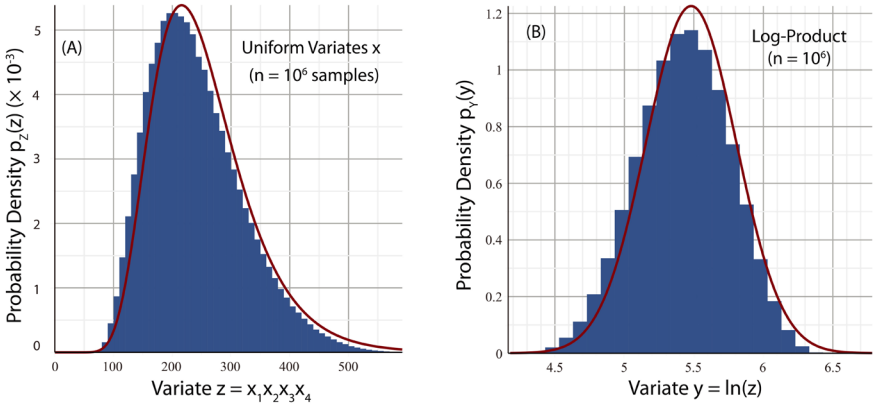

Figure 8. Panel A: Histogram of uniform product Z of **Figure 7** enveloped by PDF of log-normal variable (26) with values (44). Panel B: Histogram of $Y=\ln(Z)$ of **Figure 7** enveloped by PDF of Gaussian variable (35).

$$Sk_X^{(La)}=0 \tag{58}$$

$$K_X^{(La)}=6\,. \tag{59}$$

Figure 9 shows a panoramic plot of the histograms X_i, which have sharp cusps and fat tails in comparison to the Gaussian histograms of **Figure 5**. Equation (59) establishes quantitatively that a Laplace RV is leptokurtic, as is apparent from **Figure 1**. Nevertheless, the histogram of $Y=\ln(X_1X_2X_3X_4)$ is again well represented by a Gaussian PDF, which indicates that $Z=X_1X_2X_3X_4$ should again be a log-normal variable to good approximation, as shown in greater detail in **Figure 10**.

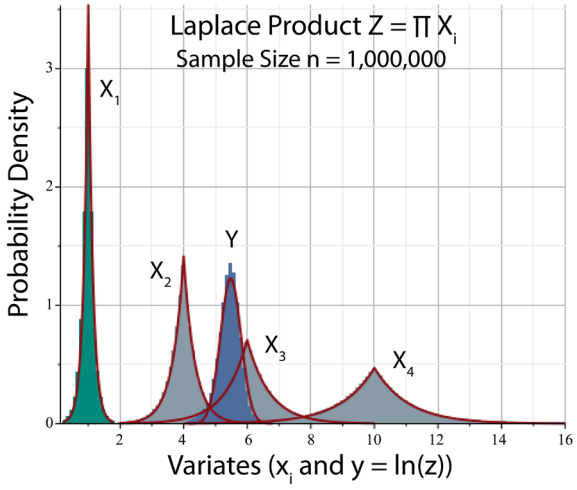

Figure 9. Monte-Carlo simulated histograms of Laplace variables $X_i(\mu_i,\sigma_i)=La_i(\mu_i,\beta_i)$ with means μ_i and standard deviations σ_i listed in (44), and $Y=\ln\left(\prod_{i=1}^{4}X_i\right)$. Histograms X_i are enveloped by their associated uniform PDFs (red). Histogram Y is enveloped by the Gaussian PDF of **Figure 5**. Sample size, symbolic notation, and color coding are the same as in **Figure 5**.

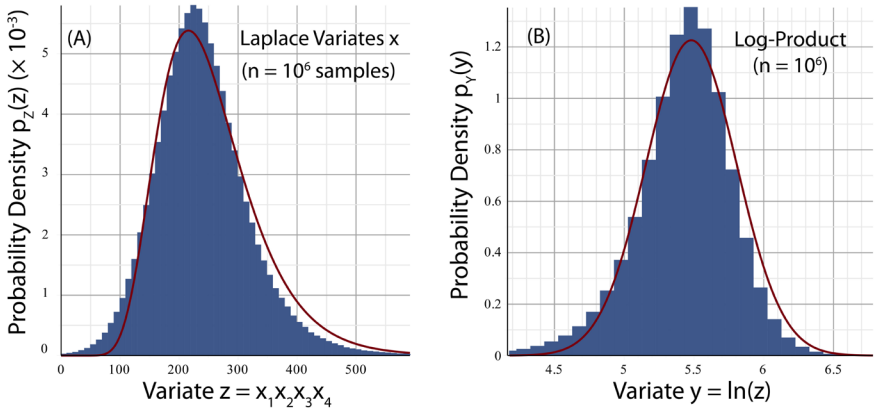

Figure 10. Panel A: Histogram of Laplace product Z of **Figure 9** enveloped by PDF of log-normal variable (26) with values (44). Panel B: Histogram of $Y=\ln(Z)$ of **Figure 9** enveloped by PDF of Gaussian variable (35).

3.4. Log-Normal Basis $X = \Lambda$

A log-normal RV $X(\mu,\sigma)=\Lambda(m,s^2)$ is symbolized by the mean and variance of the normal variable $Y=N(m,s^2)=\ln(X)$. From Equation (24), re-expressed below for convenience,

$$\begin{aligned} m &= \ln\left(\mu^2/\sqrt{\mu^2+\sigma^2}\right) \\ s^2 &= \ln\left(\left(\mu^2+\sigma^2\right)/\mu^2\right) \end{aligned} \tag{60}$$

it follows that the four log-normal basis variables with properties (44) are respectively

$$\begin{aligned} X_1 &= \Lambda_1(-0.0196, 0.1980) \\ X_2 &= \Lambda_2(1.3785, 0.1245) \\ X_3 &= \Lambda_3(1.7781, 0.1655) \\ X_4 &= \Lambda_4(2.2915, 0.1492) \end{aligned} \tag{61}$$

The skewness and kurtosis of a log-normal RV

$$Sk_X^{(\Lambda)} = \left(\exp\left(s^2\right)+2\right)\sqrt{\exp\left(s^2\right)-1} \tag{62}$$

$$K_X^{(\Lambda)} = \exp\left(4s^2\right)+2\exp\left(3s^2\right)+3\exp\left(2s^2\right)-3 \tag{63}$$

are not constants, but depend on the scale parameter s. Skewness (62) is greater than 0 for all values of $s>0$; kurtosis (63) is greater than 3 for all values of $s>0$.

Figure 11 shows a panoramic plot of the log-normal histograms X_i, which skew to the right in comparison to the symmetric shapes of the Gaussian basis histograms of Figure 5. The histograms of $Y=\ln\left(X_1X_2X_3X_4\right)$ and $Z=X_1X_2X_3X_4$ are seen to be precisely normal and log-normal, respectively, as predicted in Section 2.3 and shown in detail in Figure 12.

3.5. Commentary

The set of variates (45) comprise the response of a crowd to a problem for which the sought-for solution is a composite random variable Z. The information, or so-called "wisdom of the crowd" [1], lies in the distribution of Z from which the full population statistics can be determined. In comparing the MCS histograms

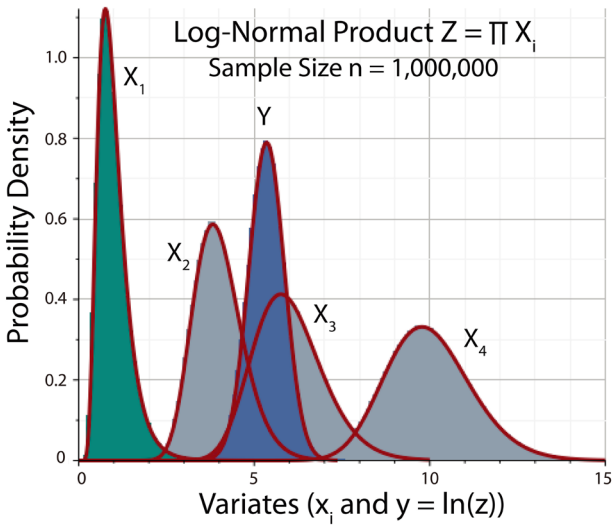

Figure 11. Monte-Carlo simulated histograms of log-normal variables $X_i(\mu_i,\sigma_i)=\Lambda_i\left(m_i,s_i^2\right)$ with means μ_i and standard deviations σ_i listed in (44), and $Y=\ln\left(\prod_{i=1}^{4}X_i\right)$. Histograms X_i are enveloped by their associated log-normal PDFs (red) (41). Histogram Y is enveloped by the Gaussian PDF (40). Sample size, symbolic notation, and color coding are the same as in Figure 5.

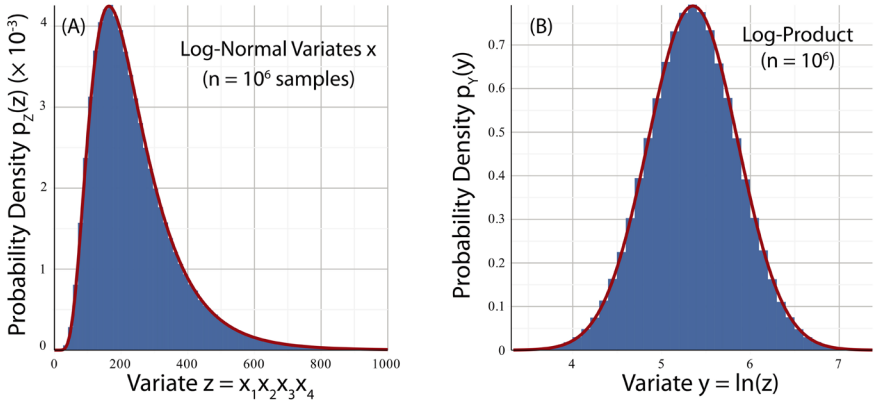

Figure 12. Panel A: Histogram of log-normal product Z of **Figure 11** enveloped by PDF of log-normal variable (41). Panel B: Histogram of $Y = \ln(Z)$ of **Figure 11** enveloped by PDF of Gaussian variable (40).

of Y and Z to the profiles of their respective PDFs, one should bear in mind that in general there is no underlying fundamental theory of crowd response. The log-normal model is not a fundamental theory such as one encounters in physics, and therefore the MCS histograms in Section 3 were not subjected to a chi-square goodness-of-fit test, as is often done in physics to compare experiment and theory.

The validity of the analytical model developed in this paper lies in how well it enables the analyst to predict an unknown quantity represented by the sampled variable Z, and not necessarily in how closely the complete distribution of the sample (*i.e.* histogram of Z) is matched by a log-normal distribution. However, if there is reason to believe that the basis variables X_i comprising the composite variable Z are distributed log-normally, then Z itself should be rigorously log-normal, and a goodness-of-fit test may then be appropriate. This point will be illuminated further in Section 4, which reports a crowdsourcing experiment and MCS to estimate the number of identical objects in a receptacle.

The preceding comments notwithstanding, **Figures 5-12** illustrate how well the predicted log-normal distribution fits the histograms of Z generated by basis variables of widely differing distribution shapes, as distinguished by their skewness and kurtosis. Simulations using normal or log-normal basis variables yielded the visually closest matches to the log-normal model. In the case of a log-normal basis, theory predicted, and MCS sustained, an exact log-normal distribution of Z.

4. Test of Crowdsourced Estimation

In a collaborative effort with the BBC The One Show (nearly exactly 100 years after Galton's pioneering statistical experiment), the author was able to obtain, using the wide reach of national television, a crowdsourced sample sufficiently large to test the log-normal hypothesis, namely, that under appropriate conditions composite random variables are distributed log-normally. Two kinds of

experiments were performed entailing crowdsourced estimates of 1) the weight of a tangible local object, and 2) the quantity of a remotely viewed object. (See Ref. [10] for a popular account.) Experiments of these kinds were conducted by the author in various physics classes during the past two decades, but no single sample was large enough to permit reliable inference of the statistical distribution. Pooling of results from different sample populations was not feasible since the conditions of the experiments were not all identical.

4.1. The Coin-Estimation Experiment

The experiment analyzed in detail here is of the second kind. Viewers of The One Show were shown on their televisions a transparent tumbler filled with opaque £1 coins. The tumbler rested on a table adjacent to two ordinary cylindrical glasses of water to provide clues to scale. No explicit dimensions of any objects were given. The challenge posed to viewers (*i.e.* the crowd) was to estimate the number of coins in the vessel.

The experimental estimates $z_k^{(\exp)}$, $k=1,\cdots,n_0$, were transmitted to the show by email, and the author subsequently received the full set of $n_0=1706$ anonymous responses, which ranged from a low of 42 to a high of 43,200.[1] The mean and median of the estimates were respectively $\overline{Z}^{(\exp)}=982$, $\tilde{Z}^{(\exp)}=695$. The true count was $N_c=1111$. If the mean is taken as the measure of crowd response—a standard statistical practice—the fractional error of the crowd was

$$\Delta N_c^{(\exp)}=\left(\overline{Z}^{(\exp)}-N_c\right)\Big/N_c=-11.6\%. \tag{64}$$

Although result (64) is not bad, it calls into question—at least to the author—how Galton's crowd of just 800 members (less than half the BBC sample size) could guess the weight of an ox to within a fractional error of less than 0.1%. One explanation might be that the participants at the fair comprised a crowd of experts familiar with livestock. The respondents to The One Show apparently had no special expertise in the estimation of quantity.

Figure 13 shows a scatter diagram of the estimates as a function of sample number, *i.e.* the order in which the estimates were received. Estimates in the approximate range between 0 and 1000 form a dense band; estimates from about 2000 to 10,000 resemble a foam of points the density of which falls off with increasing ordinate. The blue histogram labeled "Experiment" in **Figure 14** shows the distribution of estimates partitioned over $K=24$ bins of equal width ranging from 0 to 4000. Points that extended beyond 4000 are not shown, since the main body of the histogram would then be severely compressed. Superposed on the histogram of experimental results is the profile (dashed blue) of the corresponding log-normal PDF with sample parameters obtained by application of the method of maximum likelihood (ML) to a Gaussian $Y^{(\exp)}$ [30],

[1]Actually, the maximum value submitted was 25 million, which was about 15% of the entire BBC One network annual budget in the form of £1 coins in a small glass tumbler. The submission was rejected on the grounds that it was so preposterous as to be intended to undermine the experiment.

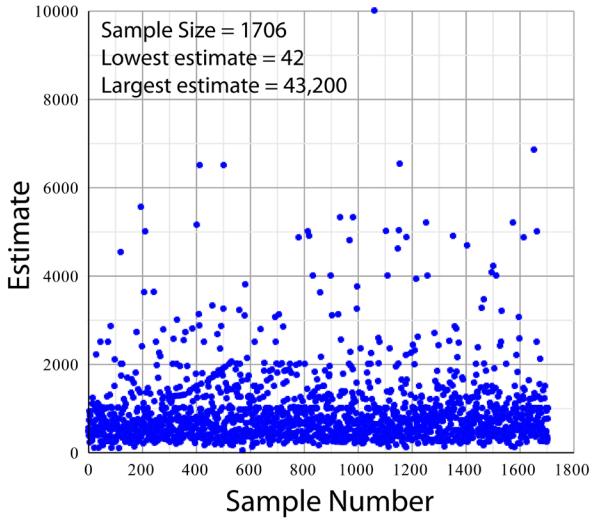

Figure 13. Estimates, in order of receipt, of the number of £1 coins in a tumbler displayed on the BBC One Show in 2007. The true count was 1111 coins; the sample size was 1706. Statistics of the experiment are given in **Table 4**.

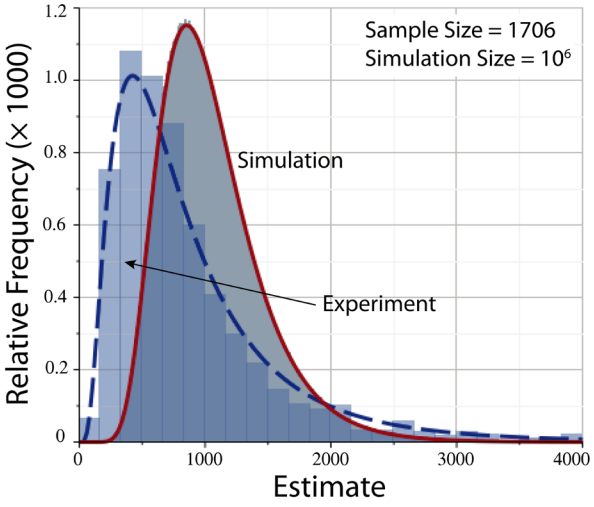

Figure 14. Comparison of the histogram (blue) of 1706 crowdsourced estimates with the histogram (gray) of 10^6 Monte Carlo simulated responses employing log-normal basis variables for coin density and tumbler geometry. The crowdsourced mean estimate was 982; the MCS mean was 1057; the true count was 1111. Relevant statistics are given in **Table 4**. Enveloping the histograms are the profiles of the log-normal PDFs for the sample (dashed blue) and simulation (solid red).

$$m_0 = \frac{1}{n_0}\sum_{k=1}^{n_0} y_k^{(\exp)} = 6.565 \tag{65}$$

$$s_0 = \sqrt{\frac{1}{n_0}\sum_{k=1}^{n_0}\left(y_k^{(\exp)} - m_0\right)^2} = 0.719 \tag{66}$$

where the variates $y_k^{(\exp)}$ are defined by

$$y_k^{(\exp)} = \ln\left(z_k^{(\exp)}\right). \tag{67}$$

Parameters m_0 and s_0 in Equation (65) are respectively the mean and standard deviation of $Y^{(\exp)}$. The gray histogram with red border in **Figure 14** will be discussed in Section 4.2.

Despite the caution about goodness-of-fit tests in Section 3.5, it is noteworthy that the fit of the log-normal PDF with parameters (65) to the histogram of experimental estimates actually does exceed the 5% acceptance threshold of a chi-square test for $\nu = 21$ degrees of freedom: $\chi^2_{21} = 6.4\%$. The number ν of degrees of freedom is given by

$$\nu = K - 1 - p \tag{68}$$

where $K = 24$ is the number of distribution categories (bins), $p = 2$ is the number of parameters (m_0, s_0) determined from the data, and the numeral 1 refers to the fact that the histogram is normalized to unit area, in which case knowledge of the values of $K-1$ bins determines the value of the remaining bin.

4.2. Monte Carlo Simulation of the Coin Estimation Experiment

Passing a goodness-of-fit test does not necessarily prove that a hypothesized theory is correct. Rather, it signifies that the theory should not be rejected on the basis of the tested data. The statistical significance of the experiment described in Section 4.1 is that the distribution of estimates of the number of coins (a composite RV) is *consistent* with a log-normal distribution for the given sample. Nevertheless, the implication of this result is of far-reaching practical importance:

If it is indeed the case that the estimates from a crowd of given size are distributed log-normally, then one should be able to simulate the estimates of a much larger crowd by constructing the appropriate basis variables that form the factors of the sought-for composite variable.

In other words, the analyst may be able to avoid sampling an impractically large crowd, yet still obtain reliable statistical information by a Monte Carlo simulation (MCS). In this section the responses from a hypothetical crowd of 1 million were simulated by applying the underlying reasoning and mathematical procedure described in Section 2.

Responses from a large crowd to a question that calls for a quantitative answer will presumably include some random guesses as well as reasoned estimates. As the author has emphasized elsewhere [10], a seminal principle to increasing the proportion of reliable estimates in crowdsourcing is to provide participants with a *personal incentive* to respond thoughtfully. Broadly speaking, there are two types of incentives. The first is to reward all respondents in some way for participating. For example, the author has used this method to provide extra credit toward the final course grade of all students in the class who executed certain tasks designed to measure the randomization of shuffled playing cards [31]. Another example of this reward structure is the internet-based Amazon Mechanical Turk which, according to Amazon, leverages "the skills of distributed Workers on a pay-per-task model" [32]. The second kind of incentive, which has

also been applied by the author in his physics classes as well as by The One Show in the experiment to estimate weight, is to reward only the respondent(s) whose estimate(s) comes closest to the true (or best) answer to the problem, once the answer becomes known. In this second approach, the members of the crowd are effectively in a competition where skill matters—unlike the case of a lottery where success depends primarily on probability and luck.

Let us assume, then, that members of the hypothetical crowd represented by the MCS are incentivized to deduce the number of coins as described in Section 2. A likely approach entails multiplying the numerical density of coins by the geometrical dimensions of the volume of the receptacle. The televised image of the tumbler showed it to have the shape of an inverted truncated right circular cone, or frustum, such as illustrated in **Figure 15**. The number Z of coins in the tumbler could then be calculated from the expression [33]

$$Z = (\pi/3)\left(R_1^2 + R_1 R_2 + R_2^2\right) HC \tag{69}$$

in which R_1 is the lower radius, R_2 is the upper radius, H is the height, and C is the numerical density of the coins. Because the upper and lower radii, height, and numerical density of coins are quantities unknown to the crowd, they must be treated as random variables. The author, himself, did not know the true numerical values, but, judging from the same image presented to the viewers, assigned random variables with the following estimated means and standard deviations (in units of cm)

$$\begin{aligned} R_1 &= X(3, 0.7) \\ R_2 &= X(5, 1.0) \\ H &= X(20, 2.0) \\ C &= X(1, 0.2) \end{aligned} \tag{70}$$

Monte Carlo simulations were then implemented for both normal variables $X = N$ and log-normal variables $X = \Lambda$.

Figure 16 shows a panoramic plot of the distributions of variables $X_i, i = 1, 2, 3, 4$, for both normal (dashed) and log-normal (solid) bases. Although the former (normal) are symmetric about the mean and the latter (log-normal) exhibit skewness, the difference in visual appearance of the two PDF profiles for each variable is relatively insignificant for the parameters shown in relations (70).

The gray histogram marked "Simulation" in **Figure 14** shows the outcome of a MCS comprising $n_s = 10^6$ samples from log-normal random number generators with parameters given by relations (70). The profile (solid red) of the histogram is the PDF of the log-normal variable $\Lambda(m_s, s_s)$ with Gaussian parameters

$$m_s = \frac{1}{n_s}\sum_{k=1}^{n_s} y_k^{(\mathrm{sim})} = 6.892 \tag{71}$$

$$s_s = \sqrt{\frac{1}{n_s}\sum_{k=1}^{n_s}\left(y_k^{(\mathrm{sim})} - m_s\right)^2} = 0.378 \tag{72}$$

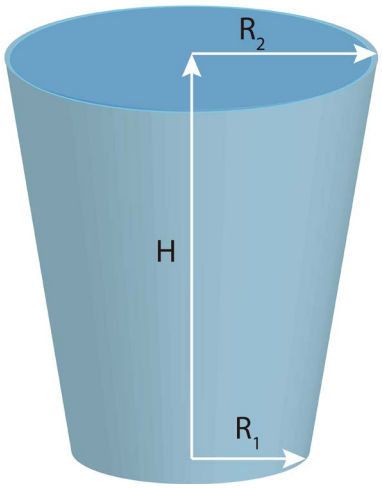

Figure 15. Geometry of the tumbler is a truncated right circular cone or frustum with dimensions given by independent random variables for height H, lower radius R_1 and upper radius R_2.

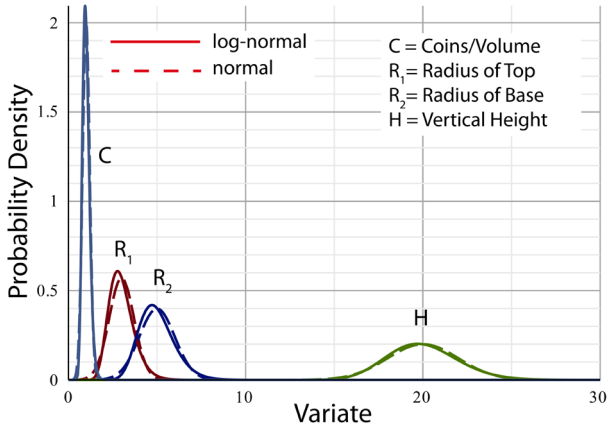

Figure 16. Distributions of the numerical density (C) and geometrical attributes (R_1, R_2, H) represented by normal (dashed) or log-normal (solid) random variables, used in the Monte Carlo simulations of **Figure 14**. The sample size was 10^6.

where variates $z_k^{(\mathrm{sim})}$, $k=1,\cdots,n_s$, are the simulated values of Z in Equation (69) and

$$y_k^{(\mathrm{sim})}=\ln\left(z_k^{(\mathrm{sim})}\right). \tag{73}$$

For the log-normal basis and sample size of 1 million, the match of theory and simulation in **Figure 14** is visually perfect at the scale shown. The predicted number of coins, given by both the theoretical expectation $\langle Z\rangle=\int_0^\infty zp_Z\left(z\middle|m_s,s_s\right)\mathrm{d}z$ and sample mean $\overline{Z}$, Equation (47), is 1057, which represents a fractional error

$$\Delta N_c^{(\mathrm{sim})}=\left(\overline{Z}^{(\mathrm{sim})}-N_c\right)\Big/N_c=-4.86\% \tag{74}$$

as summarized in **Table 4**. Thus, the MCS estimate was considerably closer to the true value $N_c=1111$ than the mean estimate of 982 by the crowd.

Table 4. Crowdsourced estimate of number of £1 coins in a tumbler.

	Simulation Variables				
Basis RVs $X(\mu_i,\sigma_i)$	$R_1(3,0.7)$	$R_2(5,1.0)$	$H(20,2.0)$		$C(1,0.2)$
Composite RVs	$Z=(\pi/3)(R_1^2+R_1R_2+R_2^2)HC$ $Y=\ln(Z)$				
	Mean $\bar{Z}$ **or** $\langle Z\rangle$	**S.E.** S_Z **or** Σ_Z	**Mean** $\bar{Y}$ **or** $\langle Y\rangle$	**S.D.** s_Y **or** σ_Y	**S.E.** S_Y **or** Σ_Y
Sample n **= 1706**	982	38.56	6.57	0.72	0.017
Simulation Log-Normal n **= 1,000,000**	1057	0.42	6.89	0.38	0.00038
Theoretical Expectations	1057	0.41	6.89	0.38	0.00038
Simulation Normal n **= 1,000,000**	1057	0.40	6.89	0.39	0.00039
Theoretical Expectations	1061	0.43	6.89	0.39	0.00039
True Count	**1111**				

Fractional Error: Experiment (n = 1706) −11.61%; Simulation (n = 1,000,000) −4.86%; Theory (n = 1,000,000) −4.50%.

The histogram obtained from the MCS with normal basis variables is nearly identical to that in Figure 14, and therefore not shown. The match with the corresponding theoretical log-normal PDF is marginally less close, but the higher mean $\bar{Z}=1061$ is marginally closer to N_c, yielding a fractional error of −4.50%. Since the standard error of the mean (*i.e.* the standard deviation divided by the square root of sample size) is 3.8×10^{-4}, the difference of means (1061 − 1057 = 4) is statistically significant in principle. In practical terms, however, the Monte Carlo simulation with either the log-normal or normal basis variables yielded effectively equivalent predictions. Since the individual estimates received from the respondents consisted solely of a single number of coins, it was not possible to conclude which of the two sets of basis variables more accurately described the crowd.

The most significant statistical outcome, however, is that the MCS predicted the number of coins in the tumbler much more closely than did the actual crowd. Results of the experiment and simulations are summarized in detail in Table 4. Theoretical means and sample means are distinguished respectively by expectation brackets like $\langle Z\rangle$ and overbars like $\bar{Z}$. Theoretical standard deviations (SD) and standard errors (SE) are symbolized by Greek letters (lower case and upper case sigma, respectively); sample SD and SE are symbolized by Roman letters (lower case and upper case s, respectively).

4.3. Commentary on the Experiment and Simulations

The coin estimation study raises several issues worth clarifying if the investigation is to provide a useful general methodology for seeking solutions to other quantitative problems by crowdsourcing.

1) Although sample size matters, the reason that the MCS did much better than the BBC crowd in estimating the number of coins in the tumbler was *not* primarily due to sample size. The populations sampled by crowdsourcing and by MCS were different not only in size but principally in their effective information content. This was seen by running the MCS with the same parameters (44) as before, but for a sample size comparable to that of the coin experiment, *i.e.* ~2000. The result was a 24-bin histogram that produced a sample mean of ~1048 and a shape that effectively overlapped the MCS histogram of **Figure 14**. The distinction between the two populations is that the BBC crowd contained a subpopulation of uninformed individuals who guessed randomly, whereas the random choices of the MCS were more tightly constrained by the variances assigned to the basis variables. In effect, the MCS population comprised a more rational crowd who used the visual cues better and made better use of a rudimentary knowledge of geometry.

2) Although the MCS of Section 4.2 estimated the number of coins by calculating the volume of a conical frustum, it is unlikely that respondents to The One Show arrived at their estimates in precisely the same way. Quite possibly, very few of the members of the crowd would have known what a frustum is or how to calculate its volume. It is not this geometrical detail that is important in determining the *distribution* of estimates, but only the act of estimating a volume and multiplying it by a numerical density. The crowd could have treated the glass tumbler simply as a rectangular solid. The independent variations of height, length, and width assumed by different respondents would have again generated estimates distributed log-normally to an excellent approximation, as demonstrated in Section 3. The fact that the sample mean of the crowd was reasonably accurate indicates that most respondents probably applied some kind of valid reasoning to obtain their answers. How closely the MCS estimate matches the true value of a composite variable depends on how well the analyst can model the statistical uncertainties in the factors upon which the sought-for variable depends.

3) It is especially noteworthy that the MCS estimates Z, defined in Equation (69), resulted in a virtually perfect log-normal distribution, as shown by **Figure 14**. This outcome suggests that the validity of the log-normal hypothesis of composite variables applies *beyond* what was explicitly demonstrated in the analysis of Section 2. In contrast to a composite RV like (26) which is formed by products of independent basis RVs, the products forming the variable Z in Equation (69) are *not* all independent. In particular, the product R_1R_2 is correlated with both R_1^2 and R_2^2. In the case of two correlated variables—call them U and V—one cannot assume, as was done in the last step of Equation (8), that the

expectation operation factors; in other words, $\langle UV\rangle \neq \langle U\rangle\langle V\rangle$.

One widely used measure of the degree of correlation between two random variables U, V is provided by the Pearson correlation coefficient $\rho_{U,V}$ defined by [34]

$$\rho_{U,V} \equiv \frac{\operatorname{cov}(U,V)}{\sigma_U \sigma_V} = \frac{\langle (U-\mu_U)(V-\mu_V)\rangle}{\sqrt{\langle (U-\mu_U)^2\rangle \langle (V-\mu_V)^2\rangle}}. \tag{75}$$

$\rho_{U,V}$ can range between −1 and +1. At the upper limit +1, V varies in the same direction and in perfect linearity with U; at the lower limit −1, V varies in the opposite direction in perfect linearity with U. If two random variables are independent, then $\rho_{U,V} = 0$, but the converse is not true; $\rho_{U,V} = 0$ does not prove that U and V are independent. Various interpretations have been given to $\rho_{U,V}$ [35] [36]. Perhaps the most useful quantitative interpretation is this [37]: The square of the correlation coefficient is equal to the fraction of the variance of variable V that is accounted for by a linear relationship with variable U. Other, more general, methods of testing for *non*linear dependence of two random variables are also known [38] [39].

To estimate the degree of correlation of terms in Equation (69) for the volume of the tumbler the Pearson correlation coefficient was used. Substitution of

$$\begin{aligned} U &\equiv R_1^2 \\ V &\equiv R_1 R_2 \end{aligned} \tag{76}$$

into Equation (75), where the radii R_1 and R_2 are given in Equation (70), resulted in correlation coefficients

$$\begin{aligned} \rho_{U,V}^{(N)} &= 0.741 \\ \rho_{U,V}^{(\Lambda)} &= 0.740 \end{aligned} \tag{77}$$

for normal (N) and log-normal (Λ) radius variables, respectively. The analysis is given in Appendix 2.

The author is unaware of any closed-form expression for the PDF or CDF of a sum of correlated or uncorrelated log-normal RVs, although it is known that the resulting RV is not rigorously log-normal [40]. Various approaches exist to approximating the sum of log-normal RVs under special circumstances (such as independent identically distributed terms), or to achieve accuracy in selected parts of the distribution profile (e.g. the tails), or to match the lowest moments (e.g. mean and variance) of an empirical distribution [40] [41] [42] [43]. No single analytical method appears to provide a satisfactory approximation for all conditions.

Nevertheless, the Monte Carlo simulations executed in the present study of crowdsourcing have shown by computational and graphical means that composite random variables are distributed log-normally to an excellent approximation for large sample size and log-normal basis RVs of low variance $(\sigma_i/\mu_i < 1)$, even if the composite variable comprises correlated terms.

5. Conclusions

This paper examined analytically, numerically, and experimentally the distribution of crowdsourced estimates of the solution to a problem seeking the number of objects in a partially revealed three-dimensional volume. Experimentally, the mean response of the crowd, which comprised approximately 2000 viewers of a BBC television show, was within ~12% of the true count. More significantly, the distribution of viewer responses was satisfactorily accounted for by a log-normal distribution.

Theoretical analyses of the product of independent random variables of low standard deviation-to-mean ratios showed that the product was distributed log-normally to an excellent approximation irrespective of the number of factors and their individual distributions. Monte Carlo tests of the theory were made with normal, uniform, Laplace, and log-normal factor variables, all of the same mean and variance, but differing widely in the shape statistics skewness and kurtosis. For independent factors of the log-normal type, the product was rigorously (not approximately) log-normal.

Monte Carlo simulations of the coin estimation experiment, employing basis variables of either the normal or log-normal type and a sample size of 1 million, resulted in mean estimates that were within ~5% of the true count. Particularly noteworthy is the fact that the sought-for composite variable comprised terms that were not independent, but linearly correlated. Nevertheless, the histogram of the product variable was, to all visual appearances, rigorously log-normal.

Telecommunications media and the internet have the potential to make possible large-scale crowdsourcing of problems like the archetype investigated here, which involved image analysis and object counting. However, the robustness of the log-normal distribution as a kind of universal distribution of composite random variables suggests that crowdsourcing can likewise be accomplished accurately by computer simulations of sufficiently large sample size, provided the underlying statistical model accurately accounts for the uncertainties of the factor variables.

Acknowledgements

The author thanks reporter Alexandra Freeman of the BBC The One Show for initiating contact regarding the planning of crowdsourcing experiments and providing the author with the resulting data files. The author also thanks Trinity College for partial support through the research fund associated with the George A. Jarvis Chair of Physics.

Conflicts of Interest

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

References

[1] Surowiecki, J. (2005) The Wisdom of the Crowds. Anchor, New York.

[2] Galton, F. (1907) Vox Populi (Voice of the People). *Nature*, **75**, 450-451. https://doi.org/10.1038/075450a0

[3] Galton, F. (1907) The Ballot Box. *Nature*, **75**, 509. https://doi.org/10.1038/075509e0

[4] Wazny, K. (2017) "Crowdsourcing" Ten Years in: A Review. *Journal of Global Health*, **7**, Article ID: 020602. https://doi.org/10.7189/jogh.07.020601 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5735781

[5] Boland, P.J. (1989) Majority Systems and the Condorcet Jury Theorem. *The Statistician*, **38**, 181-189. https://doi.org/10.2307/2348873

[6] Jin, Y., Carman, M., Zhu, Y. and Xiang, Y. (2018) A Technical Survey on Statistical Modeling and Design Methods for Crowdsourcing Quality Control. arXiv: 1812.02736v1.

[7] Baba, Y. and Kashima, H. (2013) Statistical Quality Estimation for General Crowdsourcing Tasks. In: Ghani, R., Senator, T.E., Bradley, P., Parekh, R. and He, J., Eds., *Proceedings of the* 19*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, New York, 8-9. https://doi.org/10.1145/2487575.2487600

[8] Guazzini, A., Vilone, D., Donati, C., Nardi, A. and Levnajic, Z. (2015) Modeling Crowdsourcing as Collective Problem Solving. *Nature Scientific Reports*, **5**, Article ID: 16557. https://doi.org/10.1038/srep16557

[9] Vaughan, J.W. (2018) Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research*, **18**, 1-46.

[10] Silverman, M.P. (2014) A Certain Uncertainty: Nature's Random Ways. Cambridge University Press, Cambridge, 457-514. https://doi.org/10.1017/CBO9781139507370

[11] Wikipedia (2019) Composite Number. https://en.m.wikipedia.org/wiki/Composite_number

[12] Feder, T. (2016) Crowdsourcing Platform Gets Results. *Physics Today*, **69**, 25-27. https://doi.org/10.1063/PT.3.3047

[13] Vera, A. and Salge, T.O. (2017) Crowdsourcing and Policing: Opportunities for Research and Practice. *European Police Science and Research Bulletin*, **16**, 143-154.

[14] Stinson, L. (2017) Want to Be a Space Archaeologist? Here's Your Chance. Wired. https://www.wired.com/2017/01/want-space-archaeologist-heres-chance/

[15] Pellerin, C. (2016) DoD Crowdsourcing Effort Produces Innovative Operational Approaches. DOD News. https://dod.defense.gov/News/Article/Article/1035881/dod-crowdsourcing-effort-produces-innovative-operational-approaches/

[16] Kelly, M.L. (2018) Intelligence Community Looking at Crowdsourcing for Predicting Geopolitical Events. National Public Radio. https://www.npr.org/2018/01/26/581142439/intelligence-community-looking-at-crowdsourcing-for-predicting-geopolitical-even

[17] Arfken, G. and Weber, H. (2005) Mathematical Methods for Physicists. Elsevier, Burlington, MA, 93.

[18] Hogg, R.V., McKean, J.W. and Craig, A.T. (2005) Introduction to Mathematical Statistics. Pearson/Prentice Hall, Upper Saddle River, NJ, 64.

[19] Kendall, M.G. and Stuart, A. (1963) The Advanced Theory of Statistics: Distribution Theory. Hafner, New York, 60-62.

[20] Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) Introduction to the Theory of Statistics. 3rd Edition, McGraw-Hill, New York, 540-541.

[21] Forbes, C., Evans, M., Hastings, N. and Peacock, B. (2011) Statistical Distributions. 4th Edition, Wiley, New York, 143-146. https://doi.org/10.1002/9780470627242

[22] Hald, A. (1952) Statistical Theory with Engineering Applications. Wiley, New York, 188-195.

[23] Jeffreys, H. (1961) Theory of Probability. 3rd Edition, Oxford University Press, London, 95-103.

[24] Altman, D.G. (1991) Practical Statistics for Medical Research. Chapman & Hall, New York, 132-146.

[25] Weaver, W. (1963) Lady Luck: The Theory of Probability. Anchor Books, Garden City, New York, 255-264.

[26] Boot, J.C.G. and Cox, E.B. (1974) Statistical Analysis for Managerial Decisions. McGraw-Hill, New York, 163-164.

[27] Bendat, J.S. and Piersol, A.G. (1966) Measurement and Analysis of Random Data. Wiley, New York, 41-43.

[28] Wilson, R.G. (1995) Fourier Series and Optical Transform Techniques in Contemporary Optics. Wiley, New York, 281-303.

[29] Kempthorne, O. and Folks, L. (1971) Probability, Statistics, and Data Analysis. Iowa State University Press, Ames, IA, 242-291.

[30] Wikipedia. Normal Distribution. https://en.wikipedia.org/wiki/Normal_distribution

[31] Silverman, M.P. (2019) Progressive Randomization of a Deck of Playing Cards: Experimental Tests and Statistical Analysis of the Riffle Shuffle. *Open Journal of Statistics*, **9**, 268-298. https://doi.org/10.4236/ojs.2019.92020

[32] Amazon Mechanical Turk. https://www.mturk.com/

[33] Wolfram MathWorld (2019) Conical Frustum. https://mathworld.wolfram.com/ConicalFrustum.html

[34] Kendall, M.G. and Stuart, A. (1961) The Advanced Theory of Statistics, Vol 2: Inference and Relationship. Charles Griffin & Company, London, 287-299.

[35] Chou, Y. (1969) Statistical Analysis with Business and Economic Applications. Holt, Rinehart, and Winston, New York, 614-623.

[36] Wikipedia (2019) Pearson Correlation Coefficient. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[37] Hoel, P.G. (1947) Introduction to Mathematical Statistics. Wiley, New York, 83-84.

[38] Bradley, J.V. (1968) Distribution-Free Statistical Tests. Prentice-Hall, Englewood Cliffs, NJ, 283-310.

[39] Wang, Y., Li, Y., Cao, H., Xiong, M., Shugart, Y. and Jin, L. (2015) Efficient Test for Nonlinear Dependence of Two Continuous Variables. *BMC Bioinformatics*, **16**, 260. https://doi.org/10.1186/s12859-015-0697-7

[40] Dufresne, D. (2004) The Log-Normal Approximation in Financial and Other Computations. *Advances in Applied Probability*, **34**, 747-773. https://doi.org/10.1017/S0001867800013094

[41] Fenton, L.F. (1960) The Sum of Lognormal Probability Distributions in Scatter Transmission Systems. *IRE Transactions on Communications Systems*, **8**, 57-67. https://doi.org/10.1109/TCOM.1960.1097606

[42] Wu, J., Mehta, N.B. and Zhang, J. (2005) Flexible Lognormal Sum Approximation

Method. 2005 *IEEE Global Telecommunications Conference*, St. Louis, MO, 28 November-2 December 2005, 3413-3417.

[43] Cobb, B.R., Rumi, R. and Salmeron, A. (2012) Approximating the Distribution of a Sum of Log-Normal Random Variables. *The Sixth European Workshop on Probabilistic Graphical Models*, Granada, Spain, 19-21 September 2012, 67-74. https://www.academia.edu/14441554/Approximating_the_Distribution_of_a_Sum_of_Log-normal_Random_Variables

Appendix 1

Probability Density Function of $Z = \exp(Y)$

Consider random variables Y and Z related by

$$Z = \exp(Y). \tag{78}$$

The cumulative probability function (CPF) of Z is defined by the relation

$$F_Z(z) = \Pr(Z \le z) = \int_{z_0}^{z} p_Z(z')\mathrm{d}z' \tag{79}$$

where z_0 is some constant reference point. The probability density function (PDF) of Z can be calculated from the CPF by differentiation (see Ref [20], pp. 60-62)

$$p_Z(z) = \mathrm{d}F_Z(z)/\mathrm{d}z. \tag{80}$$

Substitution of Equation (78) into (79) leads to the chain of deductions

$$F_Z(z) = \Pr\left(\mathrm{e}^Y \le z\right) = \Pr\left(Y \le \ln(z)\right) = \int_{-\infty}^{\ln(z)} p_Y(y)\mathrm{d}y. \tag{81}$$

Substitution of Equation (81) into (80) leads by the Leibniz integral formula (see Ref [17], p 590) to

$$p_Z(z) = z^{-1} p_Y\left(\ln(z)\right). \tag{82}$$

Appendix 2

Calculation of the Correlation Coefficient of Variables X^2 and XY

Consider the two composite variables

$$\begin{aligned} Z_1(\mu_1,\sigma_1) &= X(m_1,s_1)^2 \\ Z_2(\mu_2,\sigma_2) &= X(m_1,s_1)Y(m_2,s_2) \end{aligned} \tag{83}$$

where $X(m_1,s_1)$ and $Y(m_2,s_2)$ are independent RVs with respective means (m_i) and standard deviations (s_i), $i = 1,2$. The correlation coefficient defined by Equation (75) then takes the form

$$\rho_{Z_1Z_2} = \frac{\langle X^3Y\rangle - \langle X^2\rangle\langle X\rangle\langle Y\rangle}{\sqrt{\left(\langle X^4\rangle - \langle X^2\rangle^2\right)\left(\langle X^2\rangle\langle Y^2\rangle - \langle X\rangle^2\langle Y\rangle^2\right)}} \tag{84}$$

Equation (84) will be evaluated for the two basis distributions of Section 4.

Case 1: X and Y are normal RVs

Substitution of the variables

$$\begin{aligned} X(m_1,s_1) &= N\left(m_1,s_1^2\right) \\ Y(m_2,s_2) &= N\left(m_2,s_2^2\right) \end{aligned} \tag{85}$$

into Equations (83) and (84) leads to expectation values

$$\begin{aligned} \mu_1 &= m_1^2 + s_1^2 \quad \sigma_1^2 = 4m_1^2s_1^2 + 2s_1^4 \\ \mu_2 &= m_1m_2 \quad\quad \sigma_2^2 = m_1^2s_2^2 + m_2^2s_1^2 + s_1^2s_2^2 \end{aligned} \tag{86}$$

and the correlation coefficient

$$\rho_{Z_1Z_2}^{(N)} = \frac{\sqrt{2}m_1 m_2 s_1}{\sqrt{\left(2m_1^2 + s_1^2\right)\left(m_1^2 s_2^2 + m_2^2 s_1^2 + s_1^2 s_2^2\right)}} \tag{87}$$

Case 2: X and Y are log-normal RVs

Substitution of the variables (with parameters related by Equation (60))

$$\begin{aligned} X\left(m_1, s_1\right) &= \Lambda\left(a_1, b_1^2\right) \\ Y\left(m_2, s_2\right) &= \Lambda\left(a_2, b_2^2\right) \end{aligned} \tag{88}$$

into Equations (83) and (84) leads to expectation values

$$\begin{aligned} \mu_1 &= \exp\left(2a_1^2 + 2b_1^2\right) \\ \sigma_1^2 &= \exp\left(4a_1\right)\left(\exp\left(8b_1^2\right) - \exp\left(4b_1^2\right)\right) \end{aligned} \tag{89}$$

$$\begin{aligned} \mu_2 &= \exp\left(a_1 + a_2 + \frac{1}{2}b_1^2 + \frac{1}{2}b_2^2\right) \\ \sigma_2^2 &= \exp\left(2a_1 + 2a_2\right)\left(\exp\left(2b_1^2 + 2b_2^2\right) - \exp\left(b_1^2 + b_2^2\right)\right) \end{aligned} \tag{90}$$

and the correlation coefficient

$$\rho_{Z_1Z_2}^{(\Lambda)} = \frac{\exp\left(2b_1^2\right) - 1}{\sqrt{\exp\left(5b_1^2 + b_2^2\right) - \exp\left(4b_1^2\right) - \exp\left(b_1^2 + b_2^2\right) + 1}}. \tag{91}$$

With regard to the random variables representing the geometry of the tumbler in Section 4, application of the foregoing relations leads to

$$\begin{aligned} Z_1 \equiv R_1^2 &= N_1\left(3, 0.7^2\right)^2 \\ Z_2 \equiv R_1 R_2 &= N_1\left(3, 0.7^2\right) N_2\left(5, 1.0^2\right) \end{aligned} \tag{92}$$

and

$$\rho_{Z_1Z_2}^{(N)} = 0.741 \tag{93}$$

for Case 1, and to

$$\begin{aligned} Z_1 \equiv R_1^2 &= \Lambda_1\left(1.0721, 0.2302^2\right)^2 \\ Z_2 \equiv R_1 R_2 &= \Lambda_1\left(1.0721, 0.2302^2\right) \Lambda_2\left(1.5898, 0.1980^2\right) \end{aligned} \tag{94}$$

and

$$\rho_{Z_1Z_2}^{(\Lambda)} = 0.740 \tag{95}$$

for Case 2.

The correlation coefficients are virtually the same for the normal and log-normal bases, as one might have anticipated from the close match of the individual distribution functions displayed in Figure 16.