10-2019

Extraction of Information from Crowdsourcing: Experimental Test Employing Bayesian, Maximum Likelihood, and Maximum Entropy Methods

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

Scientific Research Publishing

Open Journal of Statistics, 2019, 9, 571-600
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Extraction of Information from Crowdsourcing: Experimental Test Employing Bayesian, Maximum Likelihood, and Maximum Entropy Methods

M. P. Silverman

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

How to cite this paper: Silverman, M.P. (2019) Extraction of Information from Crowdsourcing: Experimental Test Employing Bayesian, Maximum Likelihood, and Maximum Entropy Methods. *Open Journal of Statistics*, **9**, 571-600.
https://doi.org/10.4236/ojs.2019.95038

Received: September 18, 2019
Accepted: October 21, 2019
Published: October 24, 2019

Abstract

A crowdsourcing experiment in which viewers (the "crowd") of a British Broadcasting Corporation (BBC) television show submitted estimates of the number of coins in a tumbler was shown in an antecedent paper (Part 1) to follow a log-normal distribution $\Lambda\left(m,s^2\right)$. The coin-estimation experiment is an archetype of a broad class of image analysis and object counting problems suitable for solution by crowdsourcing. The objective of the current paper (Part 2) is to determine the location and scale parameters $\left(m,s\right)$ of $\Lambda\left(m,s^2\right)$ by both Bayesian and maximum likelihood (ML) methods and to compare the results. One outcome of the analysis is the resolution, by means of Jeffreys' rule, of questions regarding the appropriate Bayesian prior. It is shown that Bayesian and ML analyses lead to the same expression for the location parameter, but different expressions for the scale parameter, which become identical in the limit of an infinite sample size. A second outcome of the analysis concerns use of the sample mean as the measure of information of the crowd in applications where the distribution of responses is not sought or known. In the coin-estimation experiment, the sample mean was found to differ widely from the mean number of coins calculated from $\Lambda\left(m,s^2\right)$. This discordance raises critical questions concerning whether, and under what conditions, the sample mean provides a reliable measure of the information of the crowd. This paper resolves that problem by use of the principle of maximum entropy (PME). The PME yields a set of equations for finding the most probable distribution consistent with given prior information and *only* that information. If there is no solution to the PME equations for a specified sample mean and sample variance, then the sample mean is an unreliable sta-

tistic, since no measure can be assigned to its uncertainty. Parts 1 and 2 together demonstrate that the information content of crowdsourcing resides in the distribution of responses (very often log-normal in form), which can be obtained empirically or by appropriate modeling.

Keywords

Crowdsourcing, Bayesian Priors, Maximum Likelihood, Principle of Maximum Entropy, Parameter Estimation, Log-Normal Distribution

1. Introduction

In a previous paper [1] to be designated Part 1, the author described a crowdsourcing experiment, implemented in collaboration with a British Broadcasting Corporation (BBC) television show, to solve a quantitative problem involving image analysis and object counting. The objective of the experiment was twofold: 1) to compare the true solution with the solution obtained by sampling the estimates submitted by a large number of participating BBC viewers (the "crowd"), and 2) to find the statistical distribution of the individual responses from the crowd.

The present paper, to be designated Part 2, extends the statistical analysis of crowdsourcing further. Whereas Part 1 was concerned primarily with the identity and universality of the distribution of crowd responses, Part 2 investigates the parameters by which this distribution is defined and discusses the procedure to be employed when the distribution of crowd responses is not known.

1.1. Estimation of Distribution Parameters

In contrast to impressions fostered by popularized accounts of crowdsourcing [2], whereby the "wisdom" of a crowd is represented by a single statistic such as the sample mean, the information provided by a crowdsourced sample is contained in the distribution of responses [1]. Knowledge of this distribution permits the analyst to calculate, theoretically or numerically, all desired statistics and their associated uncertainties and correlations. Moreover, the mathematical expression for the distribution, as given by the probability density function (PDF) or the cumulative distribution function (CDF), permits the analyst to deduce the population statistics of an arbitrarily large sample, which can differ significantly from the sample statistics of a practically attainable crowd.

Part 1 focused primarily on identifying, and demonstrating the universality of, the distribution of crowdsourced responses to a large class of quantitative problems. This class includes problems whose solutions are representable by a composite random variable (RV), *i.e.* a variable expressible as a product (or sum of products) of other random variables. Statistical analysis of the crowdsourced responses was shown to follow a log-normal distribution. More generally, theoretical analysis and Monte Carlo simulation (MCS) demonstrated that, for a

sufficiently large sample size, the distribution of any composite RV comprising factor variables of low relative uncertainty is log-normal to an excellent approximation. (Relative uncertainty is defined by the ratio of standard deviation to mean.) If the factor variables are themselves independent and log-normal, then the composite variable is rigorously (not approximately) log-normal itself.

The present article is concerned with estimation of the parameters that define the log-normal distribution. In general, statistical estimation can be classified into two methodologies: maximum likelihood and Bayesian [3]. Each method has its perceived advantages and disadvantages, which have been discussed at great length—and sometimes quite heatedly—in the statistical literature; see, for example, [4] [5] [6] [7]. As an atomic and nuclear physicist whose researches ordinarily involve probability and uncertainty [8], the author has used both methods, depending on the specific problem at hand.

The method of maximum likelihood (ML) is the simpler and easier to use; it was the method employed in Part 1 to extract information from both the crowdsourced sample and much larger MCS sample. The crux of the method, elaborated in the following section, is to compose from the data and known PDF a conditional probability density referred to as the likelihood function, and to solve for the parameters that maximize this function. The ML method is most successful when the likelihood function is unimodal and sharply peaked.

The Bayesian method is more complicated for several reasons. First, in general, it requires the analyst to assess the probability of the sought-for parameters prior to any experimental information about them. This prior probability function is referred to simply as "the prior". Much of the past debate over Bayesian methods centered on the alleged subjectivity of the prior. Subsequent research, rooted in mathematical group theory (*i.e.* theory of invariants) has established a rigorous procedure for finding an objective prior for most well-behaved PDFs; see Ref [4], pp. 378-396.

The second complication to the Bayesian method, as applied in the present case, is that the distributions relevant to crowdsourcing (normal and log-normal) are defined by two parameters: a location parameter m and a scale parameter s. The crux of the Bayesian method, as elaborated in the following section, is to integrate over the likelihood and prior so as to obtain a posterior probability function (more simply referred to as "the posterior") from which the statistics of the parameters can be calculated. However, for a two-parameter distribution there are two non-equivalent priors that may apply, depending on whether the analyst is interested in estimating only one or both of the parameters.

Despite the preceding complications, Bayesian methods afford a standardized procedure for incorporating new data by which to progressively update the posterior probability.

1.2. Organization

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

Section 2 derives the likelihood function and estimation relations for a RV described by a log-normal distribution. Section 3 elaborates on the question of dual priors and derives the corresponding posterior probability densities for a log-normal RV. Section 4 applies the ML and Bayesian methods of parameter estimation to the image analysis and object counting problem of Part 1. Section 5 examines the problem of parameter estimation when the distribution of responses by the crowd is not known, and addresses the question of reliability when two different statistical methods yield significantly different results. Section 6 concludes the paper with a summary of principal findings.

As a matter of statistical terminology, the samples of a random variable are referred to as variates. In keeping with standard statistical notation, a random variable will be denoted by an upper-case letter (e.g. Z), and its variates will be denoted by a corresponding lower-case letter (e.g. z).

2. Maximum Likelihood Estimate of Log-Normal Parameters

A random variable Z is log-normal, as symbolized by

$$Z = \Lambda\left(m, s^2\right), \tag{1}$$

if the variable Y, defined and symbolized by

$$Y = \ln\left(Z\right) = N\left(m, s^2\right), \tag{2}$$

is described by a normal (also called Gaussian) distribution. Reciprocally, one can express Z in the form

$$Z = \exp\left(Y\right). \tag{3}$$

The parameters m and s in Equations (1) and (2) are respectively the mean and standard deviation of the normal RV Y whose PDF takes the familiar form

$$p_Y\left(y|m,s\right) = \frac{1}{\sqrt{2\pi}s}\exp\left(-\left(y-m\right)^2/2s^2\right). \tag{4}$$

The PDF of the original log-normal variable Z, derived in Part 1 from Equation (2), is

$$p_Z\left(z|m,s\right) = \frac{1}{\sqrt{2\pi}sz}\exp\left(-\left(\ln\left(z\right)-m\right)^2/2s^2\right). \tag{5}$$

The q^{th} statistical moment of Z for $q = 0,1,2,\cdots$, derived in Part 1, is given by

$$\left\langle Z^q\right\rangle = \exp\left(qm + \frac{1}{2}q^2s^2\right), \tag{6}$$

from which the mean m_Z and variance s_Z^2 directly follow

$$m_Z = \left\langle Z\right\rangle = \exp\left(m + \frac{1}{2}s^2\right), \tag{7}$$

$$s_Z^2 = \left\langle Z^2\right\rangle - \left\langle Z\right\rangle^2 = \exp\left(2m\right)\left(\exp\left(2s^2\right) - \exp\left(s^2\right)\right). \tag{8}$$

Given the set of variates $\{z_k\}$, $k = 1,\cdots,n$, obtained, for example, as solutions to a problem by crowdsourcing or by MCS in which the sought-for quan-

tity Z is taken to be log-normal, the likelihood function $L\left(\{z_k\}|m,s\right)$ is defined to be

$$L\left(\{z_k\}|m,s\right)=\prod_{k=1}^{n}p_Z\left(z_k|m,s\right), \tag{9}$$

where the factors on the right side are evaluations of PDF (5). Equation (9) quantifies the conditional probability of the data, given the distribution parameters m, s.

Since the extremum of a function and of its logarithm occur at the same point, it is more convenient to find the maximum of the log-likelihood

$$\mathscr{L}\left(\{z_k\}|m,s\right)=\ln\left(\prod_{k=1}^{n}p_Z\left(z_k|m,s\right)\right)=\sum_{k=1}^{n}\ln\left(p_Z\left(z_k|m,s\right)\right), \tag{10}$$

which, upon substitution of Equation (5), takes the form

$$\mathscr{L}\left(\{z_k\}|m,s\right)=-n\ln(s)-\left[\sum_{k=1}^{n}\left(m-\ln\left(z_k\right)\right)^2\Big/2s^2\right]-\sum_{k=1}^{n}\ln\left(z_k\right)-\frac{n}{2}\ln(2\pi). \tag{11}$$

The last two terms of Equation (11) are independent of the parameters and could have been omitted. Solution of the maximization equations

$$\frac{\partial\mathscr{L}}{\partial m}=-s^{-2}\sum_{k=1}^{n}\left(m-\left(\ln z_k\right)\right)=0, \tag{12}$$

$$\frac{\partial\mathscr{L}}{\partial s}=-ns^{-1}+s^{-3}\sum_{k=1}^{n}\left(m-\ln\left(z_k\right)\right)^2=0, \tag{13}$$

leads to the ML parameters

$$\hat{m}=n^{-1}\sum_{k=1}^{n}y_k=n^{-1}\sum_{k=1}^{n}\ln\left(z_k\right), \tag{14}$$

$$\hat{s}^2=n^{-1}\sum_{k=1}^{n}\left(y_k-\hat{m}\right)^2=n^{-1}\sum_{k=1}^{n}\left(\ln\left(z_k\right)-\hat{m}\right)^2, \tag{15}$$

in which the ML solution $\hat{m}$ was substituted for the variable m in Equation (13).

It is to be noted for use later that (a) the first equality of Equation (14) is precisely the form of the sample mean of Y for a sample of size n, and (b) the first equality of Equation (15) differs from the *unbiased* sample variance of Y for which the normalizing factor of a sample of size n is $(n-1)^{-1}$, rather than n^{-1} [9]. For sufficiently large n, the distinction between the ML variance and unbiased sample variance is insignificant and will be disregarded in this paper[1].

The variance and correlation of the ML parameters are elements of a 2-dimensional correlation matrix C obtained from the Hessian matrix H (*i.e.* matrix of second derivatives) according to [8] [10]

$$C=-H^{-1}, \tag{16}$$

[1]The term "unbiased" means that the expectation value of the sample variance equals the theoretical population variance. This is not the case for the ML variance. A heuristic justification for the factor $(n-1)^{-1}$ is that there can be no variance for a sample of size 1, and thus the unbiased variance should become indeterminate.

in which

$$H=\begin{pmatrix}\dfrac{\partial^2\mathscr{L}}{\partial m^2} & \dfrac{\partial^2\mathscr{L}}{\partial m\partial s}\\ \dfrac{\partial^2\mathscr{L}}{\partial s\partial m} & \dfrac{\partial^2\mathscr{L}}{\partial s^2}\end{pmatrix}. \tag{17}$$

Upon differentiation of Equations (12) and (13) and use of Equations (14) and (15), the coherence matrix (16) reduces to

$$C=\begin{pmatrix}\sigma_m^2 & \rho_{ms}\\ \rho_{ms} & \sigma_s^2\end{pmatrix}=\begin{pmatrix}\hat{s}^2/n & 0\\ 0 & \hat{s}^2/2n\end{pmatrix}. \tag{18}$$

One sees, therefore, that the ML parameters $\hat{m}$ and $\hat{s}$ are uncorrelated and that the standard error (*i.e.* standard deviation of the mean) of each is inversely proportional to the square root of the sample size, as expected.

3. Bayesian Estimate of Log-Normal Parameters

3.1. Bayesian Posterior for a Two-Dimensional Parameter Space

Although past uses of Bayes' theorem for estimation and prediction were at times controversial, the theorem itself is a fundamental part of the principles of statistics. Succinctly expressed in terms of hypotheses (H) and data (D), Bayes' theorem takes a simple form

$$P\left(H\middle|D\right)=\frac{P\left(D\middle|H\right)P\left(H\right)}{P\left(D\right)}, \tag{19}$$

where $P\left(H\right)$ is the prior, $P\left(D\middle|H\right)$ is the likelihood, and $P\left(H\middle|D\right)$ is the posterior; the denominator $P\left(D\right)$ is a normalization constant to be calculated, when needed, by summing or integrating over the full range of the numerator.

Applying Equation (19) in detail to the set of log-normal variates of Section 2 leads to the posterior probability

$$p_{(2)}^{(n)}\left(m,s\middle|\{z_k\}\right)=\frac{p_{(2)}^{(n)}\left(\{z_k\}\middle|m,s\right)\pi_{(2)}\left(m,s\right)}{\iint p_{(2)}^{(n)}\left(\{z_k\}\middle|m,s\right)\pi_{(2)}\left(m,s\right)\mathrm{d}m\mathrm{d}s}, \tag{20}$$

in which $\pi_{(2)}\left(m,s\right)$ is the prior probability of parameters m, s and the likelihood function is given by Equations (9) and (5). The subscript (2) in Equation (20) signifies that the parameter space is 2-dimensional; the superscript (n) marks the total sample size with variates denoted individually by $k=1,\cdots,n$. The range of m extends from $-\infty$ to ∞; the range of s extends from 0 to ∞. These ranges hold throughout the entire paper and will, therefore, be omitted from display so that equations will appear less cluttered.

The denominator in Equation (20) is an integral over a log-normal PDF, which is difficult to perform as such. However, since the posterior in Equation (20) is a conditional probability density for the parameters and *not* for the variates, a major simplification can be achieved by applying Bayes' theorem to the associated normal variable Y, Equation (2). The expression for the posterior then

takes the form

$$p_{(2)}^{(n)}\left(m,s\left|\{y_k\}\right.\right)=\frac{p_{(2)}^{(n)}\left(\{y_k\}\left|m,s\right.\right)\pi_{(2)}\left(m,s\right)}{\iint p_{(2)}^{(n)}\left(\{y_k\}\left|m,s\right.\right)\pi_{(2)}\left(m,s\right)\mathrm{d}m\mathrm{d}s}, \tag{21}$$

where the likelihood function (the numerator of (21)) is now taken to be

$$L\left(\{y_k\}\left|m,s\right.\right)=\prod_{k=1}^{n}p_Y\left(y_k\left|m,s\right.\right), \tag{22}$$

instead of Equation (9), and the set of variates $\{y_k\}$ is obtained from Equation (2)

$$y_k=\ln\left(z_k\right), \tag{23}$$

for $k=1,\cdots,n$. The integral in the denominator of Equation (21) now involves the Gaussian density (4), instead of the log-normal density (5). Actually, had the calculation proceeded as originally formulated in Equation (20), a transformation of integration variable would have resulted in an expression equivalent to Equation (21). The product $\prod_{k=1}^{n}z_k$ in the denominator of the log-normal likelihood (see Equations (5) and (9)) would have canceled from both numerator and denominator of Equation (20) since it is not a function of the integration variables *m*, *s*. In view of the equivalence of posteriors (20) and (21), the same symbol $p_{(2)}^{(n)}$ is retained.

To evaluate the right side of Equation (21), one must have an appropriate expression for the prior $\pi\left(m,s\right)$. A general rule for determining the prior probability in a large class of estimation problems was developed by Jeffreys [11] based on the requirement that it be invariant under certain transformations of the parameters. Applied to the 2-dimensional parameter space of the normal distribution (4), Jeffreys' rule takes the form

$$\pi\left(m,s\right)\propto\sqrt{\det\left(M\left(m,s\right)\right)}, \tag{24}$$

in which $\det\left(M\left(m,s\right)\right)$ is the determinant of the Hessian matrix

$$M\left(m,s\right)=\begin{pmatrix}\partial_{mm'}^2 & \partial_{ms'}^2\\ \partial_{sm'}^2 & \partial_{ss'}^2\end{pmatrix}\int\sqrt{p_Y\left(y\left|m,s\right.\right)p_Y\left(y\left|m's'\right.\right)}\mathrm{d}y\Big|_{\substack{m'=m\\ s'=s}}, \tag{25}$$

where the differential operators $\left(\partial_{uv}^2\equiv\partial^2/\partial u\partial v\right)$ act on the integral to the right. Substitution of Equation (4) into Equation (25) results in the matrix

$$M\left(m,s\right)=\begin{pmatrix}\left(4s^2\right)^{-1} & 0\\ 0 & \left(2s^2\right)^{-1}\end{pmatrix}. \tag{26}$$

Evaluation of the determinant in Equation (24) then yields the prior

$$\pi_{(2)}\left(m,s\right)\propto s^{-2}. \tag{27}$$

Constant factors in Equation (26) are unimportant since they cancel from the expression (21) for the posterior, and one can replace the proportionality in (27) with an equality.

Substitution of prior (27) into Equation (21) leads to the posterior probability density

$$p_{(2)}^{(n)}\left(m,s\left|\{y_k\}\right.\right)=\frac{n^{(n+1)/2}S^n\exp\left(-\frac{n}{2s^2}\left[\left(m-\bar{Y}\right)^2+S^2\right]\right)}{\sqrt{\pi}2^{(n-1)/2}\Gamma\left(n/2\right)s^{n+2}},\tag{28}$$

in which

$$\bar{Y}=n^{-1}\sum_{k=1}^{n}y_k=n^{-1}\sum_{k=1}^{n}\ln\left(z_k\right),\tag{29}$$

$$\overline{Y^2}=n^{-1}\sum_{k=1}^{n}y_k^2=n^{-1}\sum_{k=1}^{n}\ln\left(z_k\right)^2,\tag{30}$$

$$S^2=\overline{Y^2}-\bar{Y}^2,\tag{31}$$

and

$$\Gamma\left(x\right)=\int_0^\infty u^{x-1}\mathrm{e}^{-u}\mathrm{d}u\tag{32}$$

is the gamma function.

Since the set of variates $\{y_k\}$ obtained by sampling the crowd enters Equation (28) in the form of two statistics, a sample mean (29) and sample variance (31), the posterior probability function will be designated $p_{(2)}^{(n)}\left(m,s\left|\bar{Y},S\right.\right)$ in the remainder of the article.

3.2. Confidence Intervals and Expectation Values

Plots of solutions to the equation $p_{(2)}^{(n)}\left(m,s\left|\bar{Y},S\right.\right)=c$ for different values of the conditional probability c and sample size n form contours analogous to equipotential lines in electrostatics. Viewed as a topographical map, the peak—or point of highest probability density—provides a graphical means of estimating the best set of parameters $\left(\tilde{m},\tilde{s}\right)$ to be inferred from the sample statistics $\bar{Y}$, S. This is the Bayesian counterpart to the ML procedure of maximizing the likelihood function analytically.

Figure 1 illustrates this point for sample sizes n = 10, 100, 1000, 4000, given hypothetical sample statistics $\bar{Y}=5$, $S=1$ chosen for the convenience of visual display. Values of m and s for contours of fixed c for sample size n = 10 vary widely along each contour. For sample size n = 1000, however, the contours for the same values c are tightly compressed and encompass a maximum at or near the point $\left(\tilde{m},\tilde{s}\right)=\left(5,1\right)$. For n = 4000, the single black contour c = 15 encircles the presumptive maximum point even more tightly. In the limit of arbitrarily large n, the uncertainty in location of the point of maximum probability density in parameter space will be arbitrarily small. (Note that $p_{(2)}^{(n)}$ is a probability density, not a probability, and can take values greater than unity).

Numerical estimates of the parameters m, s, are obtained from PDF (28) by calculating the expectation values

$$\bar{m}=\iint mp_{(2)}^{(n)}\left(m,s\left|\bar{Y},S\right.\right)\mathrm{d}s\mathrm{d}m=\int mp_{(2m)}^{(n)}\left(m\left|\bar{Y},S\right.\right)\mathrm{d}m=\bar{Y},\tag{33}$$

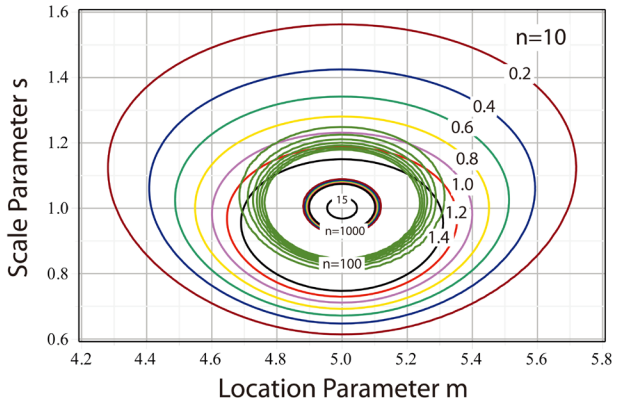

Figure 1. Contours of the posterior $p_{(2)}^{(10)}(m,s)=c$ for constants c = 0.2 (red), 0.4 (blue), 0.6 (green), 0.8 (yellow), 1.0 (violet), 1.2 (orange), 1.4 (black). Contours of $p_{(2)}^{(100)}(m,s)=c$ are shown in green for the same values of c. Contours of $p_{(2)}^{(1000)}(m,s)=c$ are shown as variants of the foregoing colors for the same values of c. The greater the sample size n, the more compressed the contours. The central black contour surrounding point (5.0, 1.0) is c = 15 for n = 4000.

$$\overline{s}=\iint sp_{(2)}^{(n)}\left(m,s\left|\overline{Y},S\right.\right)\mathrm{d}m\mathrm{d}s=\int sp_{(2s)}^{(n)}\left(m\left|\overline{Y},S\right.\right)\mathrm{d}s=\frac{\Gamma\left(\left(n-1\right)/2\right)}{\Gamma\left(n/2\right)}\sqrt{\frac{n}{2}}S\,, \quad (34)$$

where the second equality in Equations (33) and (34) defines the *marginal* probability densities for m and s respectively, as indicated by subscripts (2m) and (2s)

$$p_{(2m)}^{(n)}\left(m\left|\overline{Y},S\right.\right)=\int p_{(2)}^{(n)}\left(m,s\left|\overline{Y},S\right.\right)\mathrm{d}s=\frac{S^n\Gamma\left(\left(n+1\right)/2\right)}{\sqrt{\pi}\Gamma\left(n/2\right)\left[\left(m-\overline{Y}\right)^2+S^2\right]^{\left(\frac{n+1}{2}\right)}}\,, \quad (35)$$

$$p_{(2s)}^{(n)}\left(s\left|\overline{Y},S\right.\right)=\int p_{(2)}^{(n)}\left(m,s\left|\overline{Y},S\right.\right)\mathrm{d}m=\frac{n^{n/2}S^n\exp\left(-nS^2/2s^2\right)}{2^{\frac{n}{2}-1}\Gamma\left(n/2\right)s^{n+1}}\,. \quad (36)$$

From Equations (33), (29), and (14), it is seen that the Bayesian mean $\overline{m}$ is identical to the maximum likelihood $\hat{m}$. However, the two estimates of s given by Equations (34) and (15) differ, since expansion of Equation (15) yields

$$\hat{s}^2=n^{-1}\sum_{k=1}^{n}\left(\hat{m}-\ln\left(z_k\right)\right)^2=n^{-1}\left(\sum_{k=1}^{n}\ln\left(z_k\right)^2\right)-\hat{m}^2=\overline{Y^2}-\overline{Y}^2=S^2\,. \quad (37)$$

In the limit of an infinite sample size, the numerical coefficient of S in Equation (34) reduces to

$$\lim_{n\to\infty}\left(\frac{\Gamma\left(\left(n-1\right)/2\right)}{\Gamma\left(n/2\right)}\sqrt{\frac{n}{2}}\right)=1\,, \quad (38)$$

in which case the Bayesian and ML scale parameters become identical. For finite sample sizes, a series expansion of the coefficient yields the Bayesian scale parameter to 4th order in n^{-1},

$$\bar{s} = \left(1 + \frac{3}{4}n^{-1} + \frac{25}{32}n^{-2} + \frac{105}{128}n^{-3} + \frac{1659}{2048}n^{-4}\right)S \tag{39}$$

in comparison with the ML scale parameter $\hat{s} = S$.

Figure 2 and Figure 3 respectively show plots of the marginal densities (35) and (36) conditioned on sample statistics $\bar{Y} = 5$, $S = 1$ for different values of sample size n. As n increases, density (35) (solid curves) in Figure 2 peaks sharply about the location $\bar{m}$. (The dashed curves will be discussed in Section 3.3.) The functional form of $p_{(2)}^{(n)}\left(m\middle|\bar{Y},S\right)$, which corresponds to a Cauchy distribution [12] for n = 1, approaches the PDF of a Gaussian distribution as n increases. Similarly, density (36) in Figure 3 is highly skewed to the right for low n, but approaches the PDF of a Gaussian distribution narrowly centered on $\bar{s}$ as n increases. The functional form of (36) for arbitrary sample size can be cast into the PDF of a gamma distribution by the change of variable $\lambda = s^{-2}$.

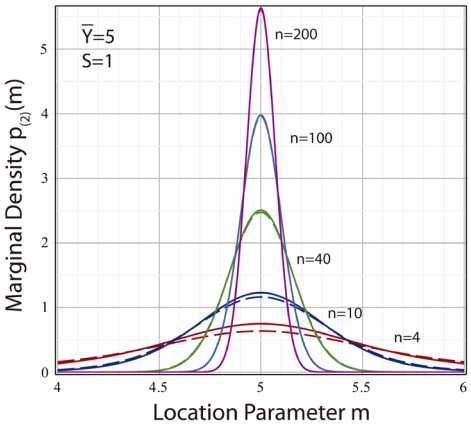

Figure 2. Marginal probability density of location parameter $p_{(2)}^{(n)}\left(m\middle|\bar{Y},S\right)$ (solid curves) and $p_{(1)}^{(n)}\left(m\middle|\bar{Y},S\right)$ (dashed curves) for sample sizes n = 4 (red), 10 (dark blue), 40 (green), 100 (light blue), 200 (violet). The marginal densities are conditioned on data $\bar{Y} = 5$, $S = 1$.

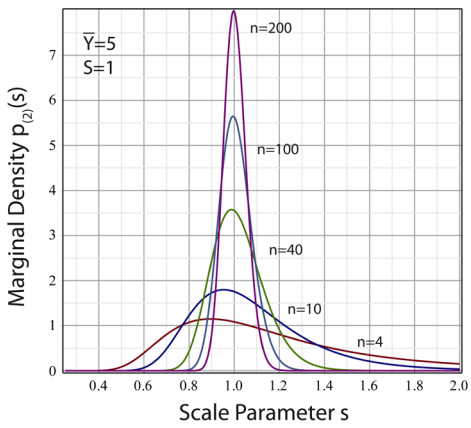

Figure 3. Marginal probability density of scale parameter $p_{(2)}^{(n)}\left(s\middle|\bar{Y},S\right)$ for the same sample sizes and color-coding as in Figure 3.

The uncertainties in Bayesian estimates, Δm^2 and Δs^2, which can be compared with the ML uncertainties derived from the correlation matrix (18), are obtained again as expectation values of the marginal density functions (35) and (36) as follows

$$\Delta m^2 \equiv \int (m-\bar{m})^2 p_{(2m)}^{(n)}\left(m\middle|\bar{Y},S\right)\mathrm{d}m = \frac{S^2}{n-2}, \tag{40}$$

$$\Delta s^2 \equiv \int (s-\bar{s})^2 p_{(2s)}^{(n)}\left(s\middle|\bar{Y},S\right)\mathrm{d}s = nS^2\left[\frac{1}{n-2}-\frac{\Gamma\left((n-1)/2\right)^2}{2\Gamma\left(n/2\right)^2}\right]. \tag{41}$$

In the limit of infinite sample size, Equations (40) and (41) respectively reduce to

$$\lim_{n\to\infty}\left(\Delta m^2\big|_{\text{Bayes}}\right) = \frac{S^2}{n} = \sigma_m^2\big|_{\text{ML}}, \tag{42}$$

$$\lim_{n\to\infty}\left(\Delta s^2\big|_{\text{Bayes}}\right) = \frac{S^2}{2n} = \sigma_s^2\big|_{\text{ML}}, \tag{43}$$

which again shows large-sample agreement between Bayesian and maximum likelihood statistics. For finite sample sizes, series expansion of the Bayesian uncertainty (41) to 4th order in n^{-1} yields

$$\Delta s^2 = \left(1+\frac{15}{4}n^{-1}+\frac{83}{8}n^{-2}+\frac{1605}{64}n^{-3}\right)\frac{S^2}{2n}. \tag{44}$$

As expected on the basis of the Central Limit Theorem [13], the PDFs of the marginal distributions (35) and (36) reduce to the following Gaussian forms for large n

$$\begin{aligned} p_{(2m)}^{(n\gg 1)}\left(m\middle|\bar{Y},S\right) &\to \frac{1}{\sqrt{2\pi\sigma_m^2}}\exp\left(-\left(m-\bar{Y}\right)^2/2\sigma_m^2\right) \\ &= \frac{1}{\sqrt{2\pi\left(S^2/n\right)}}\exp\left(-\left(m-\bar{Y}\right)^2/2\left(S^2/n\right)\right), \end{aligned} \tag{45}$$

$$\begin{aligned} p_{(2s)}^{(n\gg 1)}\left(s\middle|\bar{Y},S\right) &\to \frac{1}{\sqrt{2\pi\sigma_s^2}}\exp\left(-\left(s-S\right)^2/2\sigma_s^2\right) \\ &= \frac{1}{\sqrt{2\pi\left(S^2/2n\right)}}\exp\left(-\left(s-S\right)^2/\left(S^2/n\right)\right), \end{aligned} \tag{46}$$

signifying that $\bar{m}\sim N\left(\bar{Y},S^2/n\right)$ and $\bar{s}\sim N\left(S,S^2/2n\right)$ in the large-sample approximation.

A summary of the means and variances of the log-normal parameters obtained by both ML and Bayesian methods is given in **Table 1**.

3.3. Bayesian Posterior for a One-Dimensional Parameter Space

The log-normal distribution $\Lambda\left(m,s^2\right)$, as demonstrated in Part 1, describes the distribution of crowdsourced estimates $\{z_k\}$ of the solution Z to a quantitative problem involving products of random variables. Both parameters (m, s) are

Table 1. Comparison of maximum likelihood and Bayesian estimates of parameters (m, s).

Statistic	Maximum Likelihood		Bayesian Expectation Values		
	Symbol	Value	Symbol	Value	Limit $n\to\infty$
location m	$\hat{m}$	$\overline{Y}$	$\overline{m}=\langle m\rangle\big\vert_{(2)}^{(n)}$	$\overline{Y}$	
scale s	$\hat{s}$	$S\equiv\sqrt{\overline{Y^2}-\overline{Y}^2}$	$\overline{s}=\langle s\rangle\big\vert_{(2)}^{(n)}$	$\left(\frac{n}{2}\right)^{\frac{1}{2}}\frac{\Gamma\left((n-1)/2\right)}{\Gamma(n/2)}S$	S
$\mathrm{var}(m)$	$\hat{\sigma}_m^2$	S^2/n	$\Delta m^2=\left\langle\left(m-\overline{m}\right)^2\right\rangle\big\vert_{(2)}^{(n)}$	$\frac{S^2}{n-2}$	S^2/n
$\mathrm{var}(s)$	$\hat{\sigma}_s^2$	$S^2/2n$	$\Delta s^2=\left\langle\left(s-\overline{s}\right)^2\right\rangle\big\vert_{(2)}^{(n)}$	$\left[\frac{n}{n-2}-\frac{n\Gamma\left((n-1)/2\right)^2}{2\Gamma(n/2)^2}\right]S^2$	$S^2/2n$
Sample Statistics:			$\overline{Y}=\frac{1}{n}\sum_{k=1}^{n}\ln(z_k)$	$\overline{Y^2}=\frac{1}{n}\sum_{k=1}^{n}\ln(z_k)^2$	

needed to determine the population statistics of Z, as shown explicitly by Equation (6). It is to be recalled, however, that m and s are respectively the mean and standard deviation of a normal random variable $Y\equiv\ln(Z)=N\left(m,s^2\right)$. For the purposes of this paper and its antecedent, which is to extract information from sampling or simulating the responses of a crowd, Z is the quantity of interest, and Y is merely an intermediary for obtaining the parameters m and s.

Under other circumstances, however, an analyst may be interested in the normal variable Y, but desire only to know its mean value, *i.e.* the location parameter m and its distribution. In such a case, it may seem reasonable simply to follow the approach of Section 3.2—namely, to use the marginal probability density $p_{(2)}^{(n)}\left(m\middle|\overline{Y},S\right)$. Surprisingly, the matter of how to proceed in this case is controversial. Arguments against the preceding approach claim that it leads to "marginalization paradoxes" [14] [15], whereas counter-arguments point out that such paradoxes are specious and arise as a result of ambiguities in the use of language and reasoning [16].

According to critics of using $p_{(2)}^{(n)}\left(m\middle|\overline{Y},S\right)$, the correct Bayesian approach for estimating the posterior by which to calculate one parameter of a two-parameter distribution is to return to Jeffrey's rule, Equation (24), and determine the prior $\pi_{(1)}(m,s)$ for a one-dimensional parameter space. Implementing this instruction leads to a matrix with the single element M_{11} of matrix (26) whose substitution in Equation (24), then yields the prior

$$\pi_{(1)}(m,s)\propto s^{-1}. \tag{47}$$

Use of prior (47) in Equation (21) with subsequent integration over s as in Equation (35) results in the posterior

$$p_{(1)}^{(n)}\left(m\middle|\overline{Y},S\right)=\frac{S^{n-1}\Gamma(n/2)}{\sqrt{\pi}\Gamma\left((n-1)/2\right)\left[\left(m-\overline{Y}\right)^2+S^2\right]^{\frac{n}{2}}}, \tag{48}$$

where the subscript (1) explicitly denotes a prior for a one-dimensional parameter space. Comparison of Equations (48) and (35) shows that $p_{(1)}^{(n)} = p_{(2)}^{(n-1)}$.

The dashed curves in Figure 2 are plots of $p_{(1)}^{(n)}$ as a function of m for increasing values of n, conditioned on the same sample statistics as the plots of $p_{(2)}^{(n)}$. For sample sizes n greater than about 10, the two posterior probability densities are equivalent for all practical purposes.

4. Bayesian Analysis of the Coin Estimation Experiment

Part 1 reported a crowdsourcing experiment devised by the author and implemented with the collaboration of a BBC television show. In brief, the experiment involved a transparent tumbler in the shape of a conical frustum filled with £1 coins. Viewers saw the 3-dimensional tumbler as a 2-dimensional projection on their television screens. Viewers were asked to submit by email their estimates of the number of coins in the tumbler, which were subsequently transmitted to the author for analysis. The number of participants was $n = 1706$. Objectives of the experiment were 1) to determine the statistical distribution of the viewers' estimates and 2) to gauge how closely a statistical analysis of crowd responses matched the true count, which was $N_c = 1111$. The sample mean $\overline{Z}$, sample variance (biased $\hat{S}_Z^2$ or unbiased S_Z^2), and standard error $S_{\overline{Z}}$ of the responses from the BBC viewers were calculated to be

$$\overline{Z} = \frac{1}{n}\sum_{k=1}^{n} z_k = 982, \tag{49}$$

$$\hat{S}_Z^2 = \frac{1}{n}\sum_{k=1}^{n} z_k^2 = 1592.65^2 \quad \text{or} \quad S_Z^2 = \frac{1}{n-1}\sum_{k=1}^{n} z_k^2 = 1593.29^2, \tag{50}$$

$$S_{\overline{Z}} = \frac{S_Z}{\sqrt{n}} = 38.56, \tag{51}$$

where Z is the random variable representing the estimated number of coins submitted by a participant in the crowd.

The sample of estimates was satisfactorily accounted for by a log-normal distribution as shown by the histogram (gray bars) in Figure 4. Superposed on the histogram is the theoretical PDF (dark-red solid curve) of log-normal variable $\Lambda(\hat{m}, \hat{s})$ with ML parameters $\hat{m} = \overline{Y} \simeq 6.5651$, $\hat{s} = S \simeq 0.7186$ calculated from the sample mean $\overline{Y}$ and variance S^2 of the associated normal variable $Y = \ln(Z)$.

From the relations of the previous section as summarized in Table 1, the expectation value of the Bayesian location parameter $\overline{m}$ is seen to be identical to the ML parameter $\hat{m}$,

$$\overline{m} = \hat{m} = \overline{Y} = 6.5651, \tag{52}$$

and the Bayesian scale parameter $\overline{s}$, Equation (34), for a sample size n = 1706, is

$$\overline{s} = 1.0004399 \times S = 0.7189, \tag{53}$$

which differs from the ML parameter $\hat{s}$ only in the fourth decimal place. Thus,

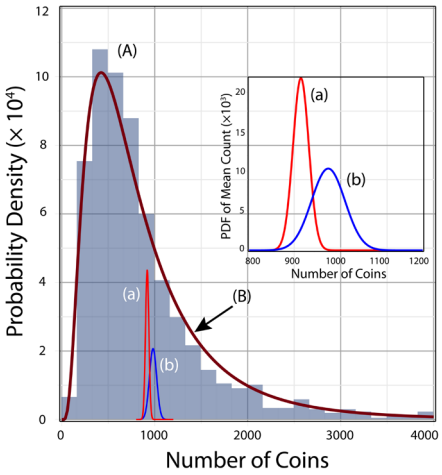

Figure 4. Histogram (gray bars) of estimates of the number of coins submitted by viewers of the BBC show bordered by log-normal PDF (dark red). Superposed are Gaussian distributions of (a) the mean of the log-normal distribution (bright red) and (b) the sample mean (blue) for a sample size $n = 1706$. Plots (a) and (b) are shown in greater detail in the insert.

for a sample size of nearly 2000, the ML and Bayesian analyses lead to statistically equivalent log-normal parameters.

Substitution of the Bayesian (or ML) parameters, Equations (52) and (53), into the log-normal expectation values (7) and (8) for the mean, variance, and standard error of Z results in an estimate of the number of coins significantly different from that of Equation (49)

$$\langle Z\rangle = \exp\left(\overline{m} + \frac{1}{2}\overline{s}^2\right) = 919, \tag{54}$$

$$\sigma_Z^2 = \exp(2\overline{m})\left(\exp\left(2\overline{s}^2\right) - \exp\left(\overline{s}^2\right)\right) = 756.26^2, \tag{55}$$

$$\sigma_{\langle Z\rangle} = \sigma_Z/\sqrt{n} = 18.31. \tag{56}$$

According to the Central Limit Theorem (CLT) [13], the distribution of the mean of a random variable with finite first and second moments approaches a Gaussian distribution in the limit of an effectively infinite sample size. **Figure 4** shows the Gaussian distributions, labeled (a) for the Bayesian-estimated mean $\langle Z\rangle$ and (b) for the sample mean $\overline{Z}$, superposed on the histogram as well as in greater detail in the insert. The difference in estimates of the two means in units of the standard error of the mean $\sigma_{\langle Z\rangle}$ is

$$\frac{\left|\overline{Z} - \langle Z\rangle\right|}{\sigma_{\langle Z\rangle}} \approx 3.5, \tag{57}$$

corresponding to a P-value(Ref [8], pp. 66-72):

$$\Pr\left(\left|\overline{Z} - \langle Z\rangle\right|/\sigma_{\langle Z\rangle} \geq 3.5\right) = 4.7\times 10^{-4}. \tag{58}$$

The low probability (58) signifies that it is very unlikely that the difference in the two means occurred as a matter of chance.

Since the sample mean has been the statistic routinely used in numerous crowdsourcing applications, the large discrepancy between Equations (49) and (54) raises questions crucial to the extraction and interpretation of crowdsourced information:

1) Is there some fundamental statistical principal that justifies use of the sample mean as a measure of the collective response of the crowd?

2) Why does the sample mean differ so markedly from the Bayesian (or ML) estimate of the mean number of coins?

3) Which estimate of Z—(a) the sample mean (49) obtained directly from the variates $\{z_k\}$ or (b) the population mean (54) of the log-normal distribution $\Lambda\left(\overline{m},\overline{s}^2\right)$—more accurately reflects the information contained in the collective response of the crowd?

These questions are resolved in Section 5.3 by first examining a third estimation procedure based on the principle of maximum entropy (PME).

5. Crowdsourcing and the Maximum Entropy Distribution

When the probability distribution of a random variable is known, the maximum likelihood or Bayesian methods can be used to estimate the parameters of that distribution, as was done in previous sections. However, in numerous applications of crowdsourcing—starting with the original experiment of Sir Francis Galton in 1907 [17] [18]—where the statistical distribution was not reported, the sample mean was taken to represent the crowd's collective response. This section examines whether, and under what conditions, such a choice can be justified.

Given incomplete statistical information of a random variable, there is a procedure for finding the most objective probability distribution—*i.e.* the distribution least biased by unwarranted assumptions—consistent with the known information. This is the distribution that maximizes entropy subject to the constraints of prior information. The so-called principle of maximum entropy (PME) has a vast literature [19] [20], since it is widely used throughout the physical sciences and engineering. It was employed initially to provide a foundation for equilibrium statistical mechanics [21] [22] and has subsequently been shown to be a general inferential method applicable to almost any problem involving probability and uncertainty [23]. For example, besides applications to physics, the author has used the PME as a means to ascertain whether students have cheated on assignments [24]. A brief summary, not intended to be rigorous in all details but merely to provide enough background for readers unfamiliar with the PME to understand its application here, is given in the following section.

5.1. Principle of Maximum Entropy (PME)

Suppose $p(z)$, $z=0,\cdots,\infty$, is the probability for outcome z of the random variable Z, which represents the possible estimates of the number of coins by the crowd. Given the discrete nature of the problem, z should be a non-negative in-

teger, but it is written as the argument of a function rather than as an index because, where summation is required, it will be treated as a continuous variable to be integrated. The practical justification for the continuum approximation is that it leads to useful closed-form expressions. The mathematical justification lies in the fact that the range is infinite, and the mean and variance of the system are assumed to be large compared to the unit interval. Thus, treatment of $p(z)$ as a continuous PDF is analogous to the well-known procedures for transforming a discrete distribution like the binomial or Poisson into a Gaussian.

The entropy H_0 of a system whose states (*i.e.* possible outcomes) z occur with probability $p(z)$ is given by [25]

$$H_0 = -\sum_{z=0}^{\infty} p(z)\ln\left(p(z)\right), \tag{59}$$

and corresponds to the quantity designated by Shannon as "information" in communication theory [26]. Although it may not be apparent, the right side of Equation (59) is equivalent, up to a universal constant factor (Boltzmann's constant), to the thermodynamic and statistical mechanical expressions for entropy of systems in thermodynamic equilibrium [27] [28].

Suppose further that all that is known of the system, in addition to the non-negative range of outcomes, are the first and second moments of Z, or equivalently the mean and variance. In other words, the prior information can be summarized as

$$1 = \sum_{z=0}^{\infty} p(z) = \int_0^{\infty} p(z)\mathrm{d}z, \tag{60}$$

$$\langle Z\rangle = \sum_{z=0}^{\infty} zp(z) = \int_0^{\infty} zp(z)\mathrm{d}z = \alpha_1, \tag{61}$$

$$\langle Z^2\rangle = \sum_{z=0}^{\infty} z^2 p(z) = \int_0^{\infty} z^2 p(z)\mathrm{d}z = \alpha_2, \tag{62}$$

in which Equation (60) is the completeness relation for $p(z)$ to be a probability (for discrete z) or probability density (for continuous z). Moments (61) and (62), respectively defined by the first equality and calculated by the second equality, take the known numerical values (α_1, α_2) given by the third equality. Then, according to the PME, the least-biased distribution $p(z)$ can be obtained by maximizing the functional

$$\begin{aligned} H = &-\sum_{z=0}^{\infty} p(z)\ln\left(p(z)\right) + \lambda_0\left[1 - \sum_{z=0}^{\infty} p(z)\right] \\ &+ \lambda_1\left[\alpha_1 - \sum_{z=0}^{\infty} zp(z)\right] + \lambda_2\left[\alpha_2 - \sum_{z=0}^{\infty} z^2 p(z)\right], \end{aligned} \tag{63}$$

with respect to each independent probability $p(z')$, $z' = 0, \cdots, \infty$, where the three factors λ_k, $k = 0,1,2$, are Lagrange multipliers.

Implementation of the maximization procedure

$$\partial H / \partial p(z') = 0, \tag{64}$$

by means of the orthonormality relation of independent probabilities

$$\partial p(z)/\partial p(z') = \delta_{zz'}, \tag{65}$$

in which $\delta_{zz'}$ is the Kronecker delta function [29], leads directly to the solution

$$p(z|\lambda_1,\lambda_2) = \frac{\exp\left(-\lambda_1 z - \lambda_2 z^2\right)}{Q(\lambda_1,\lambda_2)}, \tag{66}$$

where the multiplier λ_0 has been absorbed into the partition function

$$\begin{aligned} Q(\lambda_1,\lambda_2) &\equiv \int_0^\infty \exp\left(-\lambda_1 z - \lambda_2 z^2\right) \mathrm{d}z \\ &= \frac{1}{2}\sqrt{\frac{\pi}{\lambda_2}}\left(1+\operatorname{erf}\left(\frac{\lambda_1}{2\sqrt{\lambda_2}}\right)\right)\exp\left(\left(\lambda_1^2/4\lambda_2\right)\right) \end{aligned}, \tag{67}$$

to yield

$$p(z|\lambda_1,\lambda_2) = \frac{2\sqrt{\frac{\lambda_2}{\pi}}\exp\left(-\lambda_1 z - \lambda_2 z^2\right)\mathrm{e}^{-\lambda_1^2/4\lambda_2}}{\left(1+\operatorname{erf}\left(\lambda_1/2\sqrt{\lambda_2}\right)\right)}. \tag{68}$$

The error function $\operatorname{erf}(x)$ is defined by the integral

$$\operatorname{erf}(x) \equiv \frac{2}{\sqrt{\pi}}\int_0^x \exp\left(-t^2\right)\mathrm{d}t, \tag{69}$$

which yields limiting values $\operatorname{erf}(0)=0$, $\operatorname{erf}(\infty)=1$, and has odd symmetry $\operatorname{erf}(-x) = -\operatorname{erf}(x)$.

PDF (68) satisfies the completeness integral (60). From the definition of the partition function in Equation (67), it follows that the first two moments of the distribution can be calculated from the derivatives

$$\langle Z\rangle = -\partial \ln\left(Q(\lambda_1,\lambda_2)\right)/\partial\lambda_1, \tag{70}$$

$$\left\langle Z^2\right\rangle = -\partial \ln\left(Q(\lambda_1,\lambda_2)\right)/\partial\lambda_2, \tag{71}$$

which, when substituted into Equations (61) and (62), yield expressions for determining the Lagrange multipliers

$$\frac{\lambda_1}{2\lambda_2} + \frac{1}{\sqrt{\pi\lambda_2}}\frac{\mathrm{e}^{-\lambda_1^2/4\lambda_2}}{1-\operatorname{erf}\left(\lambda_1/2\sqrt{\lambda_2}\right)} = \alpha_1, \tag{72}$$

$$\frac{\lambda_1^2}{4\lambda_2^2} + \frac{1}{2\lambda_2} - \frac{\lambda_1}{2\lambda_2^{3/2}}\frac{\mathrm{e}^{-\lambda_1^2/4\lambda_2}}{\sqrt{\pi}\left(1-\operatorname{erf}\left(\lambda_1/2\sqrt{\lambda_2}\right)\right)} = \alpha_2. \tag{73}$$

At this point[2] the analysis is considerably facilitated by a change of variables from (λ_1,λ_2) to the variables (a,b) defined by

$$\lambda_1 = -a/b^2, \tag{74}$$

$$\lambda_2 = 1/2b^2, \tag{75}$$

[2]To calculate moments of a distribution from the partition function, differentiation must be with respect to the Lagrange multipliers (λ_1,λ_2) and not the transformed variables (a,b).

which, when substituted into Equation (68), result in the PDF

$$p(z|a,b)=\frac{\sqrt{\frac{2}{\pi}}\exp\left(-(z-a)^2\right)}{b\left(1+\operatorname{erf}\left(a/\sqrt{2}\,b\right)\right)}. \tag{76}$$

The form of PDF (76) gives the impression that a is a location parameter (mean) and b is a scale parameter (standard deviation). This is not strictly correct, as can be seen by substituting Equations (74) and (75) into Equations (72) and (73) to obtain

$$\langle Z\rangle = a+bq(a,b)=\alpha_1, \tag{77}$$

$$\langle Z^2\rangle = a^2+b^2+abq(a,b)=\alpha_2, \tag{78}$$

where

$$q(a,b)=\frac{\sqrt{2/\pi}\,\mathrm{e}^{-a^2/2b^2}}{1+\operatorname{erf}\left(a/\sqrt{2}b\right)}. \tag{79}$$

However, $\lim_{a/b\to\infty} q(a,b)\to 0$ in which case $a=\langle Z\rangle=\alpha_1$ and $b^2=\langle Z^2\rangle-\langle Z\rangle^2=\alpha_2-\alpha_1^2$. Thus, if the distribution is sharply defined, then, for all practical purposes, the error function in Equation (76) is equal to 1, and $p(z|a,b)$ becomes a Gaussian PDF extending over the full real axis with mean a > 0 and standard deviation b.

5.2. Maximum Likelihood Solution to the Maximum Entropy Equations

To solve the set of PME Equations (77)-(79) for a and b one must supply the values of α_1 and α_2, which constitute prior information, but which in practice must be estimated from the sample whose theoretical distribution is not part of the prior information. The optimal estimation procedure is to use the sample averages

$$\alpha_1 \approx \overline{Z}=n^{-1}\sum_{i=1}^{n} z_i, \tag{80}$$

$$\alpha_2 \approx \overline{Z^2}=n^{-1}\sum_{i=1}^{n} z_i^2, \tag{81}$$

again symbolized by overbars to distinguish them from theoretical expectation values symbolized by angular brackets. Justification of (80) and (81) derives from a general result of probability theory that maximizing the entropy subject to constraints (61) and (62) is equivalent to maximizing the likelihood function over the manifold of sampling distributions selected by maximum entropy. (See Ref. [23], pages 270-271). An explicit demonstration of this result as it pertains to the present problem is given in **Appendix 1**.

Equations (77)-(79) are highly nonlinear in the variables a and b. One way to solve the set of equations is graphically by plotting the variation of b as a function of a subject to each of the two constraints

$$a + bq(a,b) - \alpha_1 = 0, \tag{82}$$

and

$$a^2 + b^2 + abq(a,b) - \alpha_2 = 0, \tag{83}$$

and finding the common point $(\hat{a},\hat{b})$ of intersection. As an example that illuminates the present discussion, consider a hypothetical set of estimates with sample mean $\alpha_1 = 981$ and sample mean-square $\alpha_2 = 1156^2$. The mean α_1 was chosen to be very close to the mean coin estimate Equation (49) of the BBC viewers, but the variance $\alpha_2 - \alpha_1^2 = 612^2$ is significantly lower than the sample variance Equation (50). The left panel of Figure 5 shows implicit plots of Equations (82) and (83) with intersection at P yielding the solution $(\hat{a},\hat{b}) = (740, 780)$.

Figure 6 shows the variation with z of the corresponding maximum entropy PDF $p(z|\hat{a},\hat{b})$ defined by Equation (76). Comparison of Figure 6 with the log-normal PDF (solid red curve) in Figure 4 shows that the PME distribution for the illustrative data (α_1,α_2) fails to reproduce the observed distribution in at least two ways: 1) it does not tend toward 0 for estimates z in the vicinity of 0, and 2) it decreases toward 0 much faster than a heavy-tailed power law as z increases toward infinity.

The right panel of Figure 5 shows implicit plots of Equations (82) and (83) for values of α_1 and α_2 corresponding to the actual sample mean and sample mean-square of responses obtained in the coin-estimation experiment. The two curves do *not* intersect, and therefore there is *no* maximum-entropy solution—and no associated PDF—for this sample under the conditions specified by Equations (80) and (81).

So that the reader does not misinterpret these results, it is to be emphasized that the failure of the PME to yield a solution under some specified conditions is

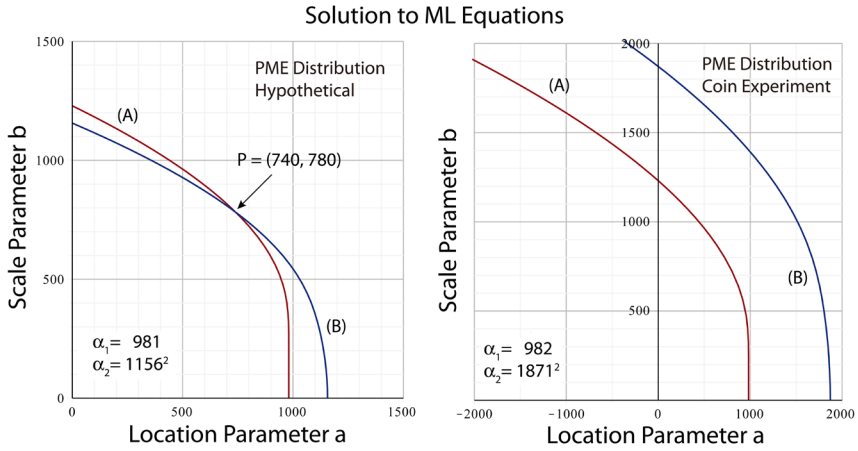

Figure 5. Implicit plots of the PME relations (A) Equation (82) (red) and (B) Equation (83) (blue), The point of intersection of the two curves in the left panel mark the solution $(\hat{a},\hat{b}) = (740, 780)$. The plots of the right panel, corresponding to sample statistics α_1 and α_2 of the crowdsourced coin experiment do not intersect, indicating there is insufficient information for a maximum-entropy solution.

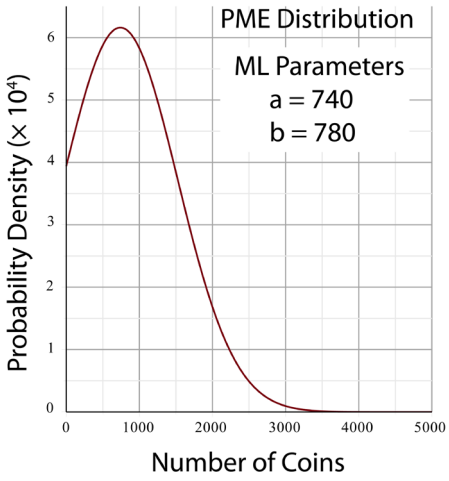

Figure 6. Plot of the maximum-entropy PDF, Equation (76), employing the parameters $(\hat{a},\hat{b})$ obtained in the left panel of **Figure 5**.

not a failure of the method. Rather, it is useful information signifying that the prior information was insufficient to provide a solution, and that additional or more consistent information is required. Thus, to persist in this approach to finding the mean response of the crowd in the absence of a known distribution, one might have to include in the prior information the sample mean-cube α_3 and the sample mean-quartic α_4 and so on, until a satisfactory solution was obtained. However, to construct a solution incrementally by including higher-order sample moments is a very unsatisfactory way to proceed, since the mathematics soon becomes impractically complicated. Moreover, from a conceptual perspective, the need for such an approach is entirely unnecessary because the actual distribution can be deduced or estimated from the crowdsourced sample.

Recall that the rationale for using the PME in the first place arose from ignorance of the distribution, and that under such circumstances the PME furnishes the least biased distribution by which to interpret the sample mean and variance. However, the distribution of a wide class of crowdsourced samples is knowable, if only the analyst were to extract it from the set of responses: it is the log-normal distribution [1]. Knowing this, one could then construct the best log-normal for the sample by finding only 2 parameters (m, s)—as was done in Part 1 and previous sections of Part 2—rather than having to solve a set of $q > 2$ nonlinear equations of constraint involving q sample moments.

It may be argued that the complexity of the analysis in Sections 5.1 and 5.2 could be avoided if one simply omitted from the prior information the requirement that $z \geq 0$. Permitting z to range over the entire real axis would then yield a PME distribution of pure Gaussian form

$$p(z|a,b) = \frac{1}{\sqrt{2\pi b^2}} \exp\left(-(z-a)^2/2b^2\right), \tag{84}$$

in which parameters a and b^2 are unambiguously the population mean and

variance. Thus, given the mean and variance as the only prior information, it follows from the PME that 1) the most objective distribution is Gaussian, and 2) the theoretical mean and variance can be estimated directly from the sample mean and sample variance. In other words, it may seem that reducing the prior information would lead unfailingly to a PME solution (*i.e.* Equation (84)) with easily obtainable parameters. However, although omission of known information may simplify the mathematics, it yields an *unreliable* solution, as discussed in the following section.

5.3. Answers to the Three Questions of Section 4

In regard to Question (1), consider a large set of crowdsourced responses to a problem for which the analyst receives just the sample mean and standard deviation, and not the full set of responses. Under these conditions, the resulting maximum entropy distribution is a normal distribution, Equation (84), and the use of maximum likelihood or Bayesian methods for estimating the mean of a normal distribution is *precisely* the sample mean, as expressed by Equation (49) for the coin estimation experiment. Thus, use of the sample mean to estimate the population mean when the actual distribution is unknown is justified by the PME.

Moreover, the reverse logic also applies. To use the sample mean and standard deviation as the statistics representing the crowd's collective answer to a problem is to assume implicitly that the responses received from the crowd were normally distributed. However, in the example of the coin-estimation experiment, that assumption is incorrect, as evidenced by the histogram of **Figure 4** which has the form of a log-normal, not a Gaussian, distribution. Furthermore, as demonstrated analytically and by MCS in [1], one can expect all crowdsourced estimates that involve products of random variables to be approximately or rigorously of log-normal form. The answer, therefore, to Question (2) at the end of Section 4 is now clear. One does not expect the means calculated from two different, nonequivalent distributions to be the same.

There remains Question (3): Which statistic better represents the information of the crowd—the sample mean of a falsely presumed Gaussian distribution or the expectation value calculated from the appropriate log-normal distribution? The answer to this question is somewhat subjective, since it depends on how one views the process of crowdsourcing and what one expects to learn from it.

One way of thinking might be the following. Recall that the idea underlying crowdsourcing is to pose a problem to a large number of diverse, independent-minded people, who collectively represent a wide range of proficiencies and experiences, and see what answers they provide. It is assumed that the crowd will include some members who know enough to address the problem rationally, some members who will guess randomly, and most of the rest whose responses fall somewhere in-between. Since the crowd is large and their responses anonymous, it is not possible to distinguish the experts from the random guessers, so one might just as well average all solutions with equal weighting, which is

what the sample mean does. The fact that the sample mean 982 [Equation (49)] of the coin-estimation experiment was closer to the true number $N_c = 1111$ than the estimate 919 [Equation (54)] based on the log-normal distribution might seem to support this viewpoint.

There is, however, a different way to think about the question—but first examine Figure 7, which shows the log-normal distribution of the coin estimates (plot A) and three Gaussian distributions (plots B, C, D).

On theoretical grounds alone, the log-normal plot A manifests the most important statistical properties to be expected of the responses of a crowd to a problem calling for a positive numerical answer. The PDF $p_Z^{(\Lambda)}(z|m,s)$ must be 0 for all $z \le 0$, since every viewer could see that the tumbler had at least 1 coin (and, in fact, many more coins than 1). The shape of the plot—main body of roughly Gaussian form coupled to a highly skewed right tail—graphically displays the distinction between informed respondents (main body) and random guessers (outliers under the heavy tail). Thus, without knowing which respondents submitted which estimates, the log-normal PDF appropriately weights each estimate depending on its value relative to the totality of estimates. If the most accurate estimate of N_c should actually differ significantly from the mean of $\Lambda(\bar{m},\bar{s}^2)$, that indicates that the crowd as a whole was not knowledgeable in regard to the posed problem.

A log-normal curve can be approximated by a normal curve, as carried out in detail in **Appendix 2**. The resulting Gaussian, which takes the form

$$p_Z^{(N)}(z|m,s) = \frac{1}{\sqrt{2\pi}\left(e^m s\right)} \exp\left(\left(z - e^m\right)^2 \Big/ 2\left(e^m s\right)^2\right), \tag{85}$$

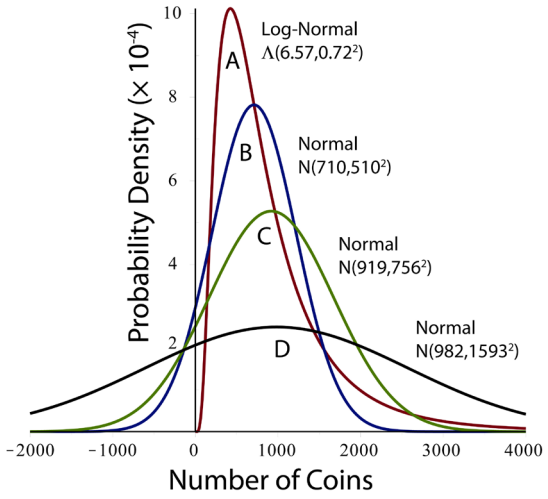

Figure 7. (a) PDF of log-normal $\Lambda(\bar{m},\bar{s}^2)$ (red) fit to the empirical distribution of coin estimates submitted by BBC viewers; (b) PDF of Gaussian approximation (blue) to $\Lambda(\bar{m},\bar{s}^2)$ as derived in **Appendix 2**; (c) PDF of a Gaussian with mean and variance of $\Lambda(\bar{m},\bar{s}^2)$ (green); (d) PDF of Gaussian with the sample mean and sample variance of the coin estimation experiment.

is shown as plot B in Figure 7. The log-normal parameters $\left(\bar{m},\bar{s}^2\right)=\left(6.57,0.72^2\right)$ of the coin-estimation experiment result in the Gaussian mean and variance $\left(M_{\mathrm{B}},S_{\mathrm{B}}^2\right)=\left(710,510^2\right)$, respectively. Plot B overlaps most of plot A although it lacks the heavy right tail, and a small fraction of the area under plot B falls in the unphysical region of negative estimates. The center $M_{\mathrm{B}}=\mathrm{e}^{\bar{m}}$ of this Gaussian is actually the median of the log-normal distribution $\Lambda\left(\bar{m},\bar{s}^2\right)$ comprising plot A.

A second, lesser accurate normal approximation to the log-normal plot A is obtained simply by substituting the log-normal mean and variance $\left(M_{\mathrm{C}},S_{\mathrm{C}}^2\right)=\left(919,756^2\right)$ of Equations (7) and (8) into a Gaussian PDF. The resulting distribution comprises plot C in Figure 7. The peak of plot C is located closer to N_c than the peak of plot B, but plot C is wider, overlaps plot A less, ascribes higher probability than plot B to the outliers in the heavy tail of plot A, and extends more significantly into the unphysical negative region.

The final Gaussian, plot D, is the distribution predicted by the PME with sample mean and sample variance $\left(M_{\mathrm{D}},S_{\mathrm{D}}^2\right)=\left(982,1593^2\right)$ of the coin-estimation experiment with neglect of the non-negativity of the range of outcomes. Although the peak is closest to N_c of the four plots, plot D has the greatest width (and therefore greatest uncertainty), overlaps the true distribution (plot A) the least, gives the greatest weight to the outliers of plot A, and extends furthest into the domain of unphysical negative estimates. By weighting each estimate the same, the sample mean (center of plot D) ignores the distinction between informed respondents and wild guessers that is a critical part of the structure of plot A. In view of the adverse features of plot D, one must ask whether the fact that the mean of plot D, rather than the mean of plot A, lies closer to N_c is in any way significant.

The answer is "No". Observe that the center of plot D can be displaced even further toward N_c simply by increasing the number of outliers with values greater than 3 or more times the value of N_c. *In short, a statistic that can be made more accurate by the inclusion of estimates that are increasingly wrong is not reliable*. Note that the effect of outliers on the theoretical mean of plot A is much weaker because (1) the exponential part of the log-normal PDF $p_Z^{(\Lambda)}\left(z\middle|m,s\right)$ is a function of $\ln\left(z\right)$ rather than z, and (2) the non-exponential part of $p_Z^{(\Lambda)}\left(z\middle|m,s\right)$ decreases inversely with increasing z.

Given the preceding observations regarding the plots of Figure 7—and the fact that a more informed application of the PME, which includes the correct range of outcomes, leads to *no* solution at all—it is clear that the mean of plot D, irrespective of its value, is an unreliable statistic. Thus, an alternative answer to Question (3) might go as follows. The most important information that can be extracted from a crowdsourced sample is its distribution (and not any individual statistic) because the distribution helps the analyst gauge the overall knowledge of the crowd and therefore the reliability of the sample. After all, there is no mathematical or statistical principle that guarantees that a crowdsourced answer to

a problem will necessarily be correct, even in the limit of an arbitrarily large crowd.

5.4. Quantitative Measure of Information Content

The entropy of a distribution is a measure of its information content. Because the word "information" has different meanings in different fields of science and engineering that employ statistical reasoning, this section uses "information" as it is interpreted in physics—*i.e.* as a measure of uncertainty. The greater the entropy of a *particular* distribution, the greater is the uncertainty (and the lower is the reliability) of its predictive capability. The word "particular" is italicized above for emphasis so as to avoid misconstruing the objective of the method of maximum entropy.

When all one knows about a statistical system is partial prior information such as the mean and variance, the PME provides an inferential method to find the most probable distribution consistent with that information and *only* that information. This is the distribution that is consistent with the prior information in the greatest number of ways—*i.e.* which maximizes the entropy of the system. On the other hand, if an analyst has to chose between two known distributions for purposes of prediction, the better choice is the distribution for which the number of possible outcomes *inconsistent* with the observed properties of the system is fewer—*i.e.* the distribution with lower entropy.

The two distributions of relevance in this analysis of crowdsourcing are the log-normal and normal distributions whose entropies, given by Equation (59), are respectively evaluated to be

$$H_0^{(\Lambda)} = -\int_0^\infty p^{(\Lambda)}\left(z|m,s\right)\ln\left(p^{(\Lambda)}\left(z|m,s\right)\right)\mathrm{d}z = \ln\left(\sqrt{2\pi\mathrm{e}}s\right)+m, \tag{86}$$

$$H_0^{(N)} = -\int_0^\infty p^{(N)}\left(z|a,b\right)\ln\left(p^{(N)}\left(z|a,b\right)\right)\mathrm{d}z = \ln\left(\sqrt{2\pi\mathrm{e}}b\right), \tag{87}$$

where the log-normal and Gaussian PDFs are respectively given by Equations (5) and (84). Substituting into Equations (86) and (87) the parameters obtained from the coin-estimation experiment (repeated below for convenience)

$$\text{Log-Normal}\quad \Lambda\left(m,s^2\right) \qquad (m,s) = (6.57, 0.72), \tag{88}$$

$$\text{Normal}\quad N\left(a_{\text{sample}}, b_{\text{sample}}^2\right) \qquad (a,b) = (982.17, 1593.65), \tag{89}$$

yields entropies

$$H_0^{(\Lambda)} = 7.65, \tag{90}$$

and

$$H_0^{(N)} = 8.79, \tag{91}$$

in units of nats (*i.e.* natural entropy units), since the natural logarithm is used in the definition of entropy in physics. (In communication theory, the logarithm to base 2 is usually employed, in which case entropy is expressed in bits, *i.e.* binary digits).

Although the numerical difference between relations (90) and (91), $H_0^{(N)} - H_0^{(\Lambda)} = 1.14$, may appear unremarkable, the micro-statistical implications are actually beyond imagining. The number Ω of possible samples of size n consistent with the known prior information of a distribution formed from a particular sample—what in physics would be termed the multiplicity or number of accessible microstates [30]—is given by an adaptation of the Boltzmann formula [31] [32]

$$\ln(\Omega) = nH_0 . \tag{92}$$

The greater the entropy, the greater is the number of possible outcomes of any draw from the distribution describing the sample. It then follows from Equation (92) that the relative uncertainty—*i.e.* ratio of microstates—of the two distributions parameterized by (88) and (89) describing the BBC crowd of size n = 1706 is

$$\frac{\Omega^{(N)}}{\Omega^{(\Lambda)}} = \exp\left(n\left(H^{(N)} - H^{\Lambda}\right)\right) \approx 4.5\times 10^{844} . \tag{93}$$

Numbers of the order of the ratio (93) rarely, if ever, occur even in physics on a cosmological scale. The import of (93) is that a vast number of Gaussian microstates—*i.e.* outcomes of the distribution (84) compatible with the prior information (89)—describe outcomes (e.g. negative numbers of coins) that are not compatible with the physical conditions of the experiment or the statistics of the crowd response as deducible from (88).

Section 5.3 and the foregoing analysis of Section 5.4 call for revisiting **Figure 4**, in which inserts (a) and (b) respectively show the distributions of the means (M_{A}, M_{D}) of plots A and D of **Figure 7**. Based on the central limit theorem (CLT), these means are distributed normally with variances smaller than the variances of plots A and D by the factor n. Although insert (b) (the sample mean) is a little wider than insert (a) (the log-normal mean), it nevertheless appears sharply localized around the sample mean M_{D}. Does this indicate that the sample mean is a reliable measure of the information content of the crowd in the coin-estimation experiment?

The answer again is "No". In brief, all that the CLT tells us in regard to the coin-estimation experiment is this: if the experiment is run a large number of times n, then the variation (standard deviation) of the mean result will be narrower than the variation for a single run in proportion to $n^{-1/2}$. This is perfectly valid as applied to insert (a) since it derives from a legitimate single-run distribution function(of log-normal form) illustrated by the histogram A or plot B in **Figure 4** or plot A in **Figure 7**. For the CLT to be valid the single-run distribution function must have finite first and second moments. However, it has been shown by use of the PME that, given the sample mean, sample variance, and appropriate non-negative range of the coin-estimation experiment as prior information, *no* compatible single-run distribution function exists. Thus, the distribution depicted by insert (b) is irrelevant and uninformative.

6. Conclusions

In sampling a large group of non-experts (a "crowd") for the solution to a quantitative problem, there is no guarantee (e.g. by some principle of probability or statistics) that the answer provided by the crowd will be correct or accurate. What usable information the crowd may provide is encoded in the distribution of responses, which the analyst can observe empirically (e.g. as a histogram) or try to deduce theoretically (as in Part 1) by modeling the reasoning process of an informed and incentivized crowd. The distribution function provides the means for obtaining the mean, median, mode, variance, and higher-order moments of the hypothetical population of which the sampled crowd is an approximate representation. Without knowledge of the distribution, statistical measures of uncertainty cannot be interpreted probabilistically.

The antecedent paper [1], showed that crowdsourced solutions to problems involving products (or sums of products) of random variables—as in the case of image analysis and counting of physical objects—led to a log-normal distribution. The log-normal $\Lambda\left(m, s^2\right)$ is a two-parameter distribution with location parameter m and scale parameter s. The present paper has shown that maximum likelihood and Bayesian estimation methods applied to the log-normal distribution yield the same expression for m, but different expressions for s that become identical in the limit of an infinitely large sample. For most practical purposes, the asymptotic limit is attained in sample sizes of a few hundred to a thousand and possibly even as low as on the order of tens.

In applications where the analyst receives only the mean response of the crowd and a measure of its uncertainty, the principle of maximum entropy shows that the most probable distribution compatible with this information is either a Gaussian (for outcomes that span the real axis) or a truncated Gaussian (for non-negative outcomes). It is possible, however, that the equations for the parameters of the maximum entropy distribution lead to no solution given the prior information. In such a case, as illustrated by the coin-estimation experiment, the sample mean of the crowd, irrespective of its value, is not a reliable statistic, since, without an underlying single-run distribution, no confidence limits can be assigned to the uncertainty of the sample mean.

The foregoing problem is in all cases avoidable if the analyst utilizes the complete set of responses from the crowd to obtain the sample distribution, either empirically or by appropriate modeling.

Conflicts of Interest

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

References

[1] 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

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

[3] Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) Introduction to the Theory of Statistics. 3rd Edition, McGraw-Hill, New York, 271-273, 339-343.

[4] Jaynes, E.T. (2003) Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, 314-317.

[5] Carlin, B.P. and Louis, T.A. (2000) Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall/CRC, London, 4-10. https://doi.org/10.1201/9781420057669

[6] Kempthorne, O. and Folks, L. (1971) Probability, Statistics, and Data Analysis. Iowa State University Press, Ames, 345-346.

[7] Feller, W. (1957) An Introduction to Probability Theory and Its Applications. Wiley, New York, Vol. 1, 2nd Edition, 113-114; (1966) Vol. 2, 55-56.

[8] Silverman, M.P. (2014) A Certain Uncertainty: Nature's Random Ways. Cambridge University Press, Cambridge, 54-66, 74-83. https://doi.org/10.1017/CBO9781139507370

[9] Hald, A. (1952) Statistical Theory with Engineering Applications. Wiley, New York, 72-77.

[10] Martin, B.R. (1971) Statistics for Physicists. Academic Press, New York, 72-84.

[11] Jeffreys, H. (1946) An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society A*, **186**, 453-461. https://doi.org/10.1098/rspa.1946.0056

[12] Parzen, E. (1960) Modern Probability Theory and Its Applications. Wiley, New York, 180-181.

[13] Kendall, M.G. and Stuart, A. (1963) The Advanced Theory of Statistics Vol. 1 Distribution Theory. Hafner, New York, 193-195.

[14] Dawid, A.P., Stone, M. and Zidek, J.V. (1973) Marginalization Paradoxes in Bayesian and Structural Inference. *Journal of the Royal Statistical Society B*, **35**, 189-233. https://doi.org/10.1111/j.2517-6161.1973.tb00952.x

[15] Berger, J.O., Bernardo, J.M. and Sun, D. (2009) The Formal Definition of Reference Priors. *The Annals of Statistics*, **37**, 905-938. https://doi.org/10.1214/07-AOS587

[16] Jaynes, E.T. (1980) Marginalization and Prior Probabilities. In: Rosenkrantz, R.D., Ed., *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, Kluwer Academic Publishers, Dordrecht, 337-375. https://doi.org/10.1007/978-94-009-6581-2_12

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

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

[19] Wikipedia (2019) Principle of Maximum Entropy. https://en.wikipedia.org/wiki/Principle_of_maximum_entropy

[20] Wu, N.J. (1997) The Maximum Entropy Method. Springer, New York. https://doi.org/10.1007/978-3-642-60629-8

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

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

[23] Jaynes, E.T. (1978) Where Do We Stand on Maximum Entropy? In: Rosenkrantz,

R.D., Ed., *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, Kluwer Academic Publishers, Dordrecht, 210-314. https://doi.org/10.1007/978-94-009-6581-2_10

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

[25] Callan, H.B. (1985) Thermostatics and an Introduction to Thermostatistics. Wiley, New York, 379-385.

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

[27] Davidson, N. (1962) Statistical Mechanics. McGraw-Hill, New York, 86-90.

[28] Wilson, A.H. (1966) Thermodynamics and Statistical Mechanics. Cambridge University Press, London, 99-101.

[29] Arfken, G.B. and Weber, H.J. (2005) Mathematical Methods for Physicists. 6th Edition, Elsevier Academic Press, New York, 136-137.

[30] Grandy, W.T. (2012) Entropy and the Time Evolution of Macroscopic Systems. Oxford University Press, Oxford, 10-11.

[31] Reif, F. (1965) Fundamentals of Statistical and Thermal Physics. McGraw-Hill, New York, 122-124.

[32] Wikipedia (2019) Boltzmann's Entropy Formula. https://en.wikipedia.org/wiki/Boltzmann%27s_entropy_formula

Appendix 1.

Maximum Likelihood Solution to the Maximum Entropy Distribution of Coin Estimates

A general consequence of probability theory cited in Section 5.2 is that maximizing the entropy subject to constraints on the first and second moments is equivalent to maximizing the likelihood function over the manifold of sampling distributions selected by maximum entropy. The significance of this is that one can use the sample mean and sample mean square to obtain the first and second moments as prior information with which to derive the maximum entropy distribution. This equivalence is demonstrated below for the coin-estimation experiment, which is an archetype for problems whereby the outcomes are non-negative numbers.

The likelihood function for the set $\{z_k\}$, $k=1,\cdots,n$, of estimates of the number of coins is given by

$$L=\prod_{k=1}^{n}p\left(z_k\,|a,b\right)=\frac{2^{n/2}\exp\left(-\sum_{k=1}^{n}\left(z_k-a\right)^2\Big/2b^2\right)}{\pi^{n/2}b^n\left(1-\operatorname{erf}\left(-a/\sqrt{2}\,b\right)\right)^n}, \tag{94}$$

where the form of the maximum-entropy PDF derived on the basis of prior information (60)-(62) is given by Equation (76). The log-likelihood function is then

$$\mathscr{L}\left(\{z_k\}|a,b\right)=-n\ln\left(b\right)-n\ln\left(1-\operatorname{erf}\left(-a/\sqrt{2}\,b\right)\right)-\sum_{k=1}^{n}\left(z_k-a\right)^2\Big/2b^2\,, \tag{95}$$

where only terms involving parameters a and b were included.

The ML equations for the parameters are

$$\frac{\partial\mathscr{L}}{\partial a}=0\;\Rightarrow\;a=\overline{Z}-bq\left(a,b\right), \tag{96}$$

$$\frac{\partial\mathscr{L}}{\partial b}=0\;\Rightarrow\;b^2=\sum_{k=1}^{n}\left(z_k-a\right)^2+abq\left(a,b\right), \tag{97}$$

where

$$q\left(a,b\right)=\frac{\sqrt{2/\pi}\,\mathrm{e}^{-a^2/2b^2}}{1+\operatorname{erf}\left(a/\sqrt{2}\,b\right)} \tag{98}$$

was defined previously in Equation (79), and

$$\overline{Z}=n^{-1}\sum_{k=1}^{n}\left(z_k\right) \tag{99}$$

is the sample mean.

Comparison of Equations (96) and (77) shows that the two equations are equivalent if the expectation value $\langle Z\rangle$ is estimated by the sample mean (99). Furthermore, replacement of a in Equation (97) by the right hand side of Equation (96) and comparison with Equation (83) leads to the equivalence of the expectation value $\langle Z^2\rangle$ and the sample mean-square

$$\overline{Z^2} = n^{-1}\sum_{k=1}^{n}\left(z_k^2\right). \tag{100}$$

Thus, the ML and PME equations lead to the same distribution parameters when the first and second moments in the maximum entropy equations are estimated by the sample moments obtained by the method of maximum likelihood.

Appendix 2.

Gaussian Approximation to a Log-Normal Distribution

The PDF of a general log-normal as defined in Equation (5) is repeated below for convenience

$$p\left(z|m,s\right) = \frac{1}{\sqrt{2\pi}sz}\exp\left(-\left(\ln\left(z\right)-m\right)^2/2s^2\right). \tag{101}$$

Transformation of the location parameter m

$$\mu_0 = \mathrm{e}^m, \tag{102}$$

and change of variable

$$z = \mu_0 + \varepsilon, \tag{103}$$

where $\varepsilon \ll 1$, lead to the form

$$p\left(z|m,s\right) = \frac{1}{\sqrt{2\pi}s\left(\mu_0+\varepsilon\right)}\exp\left(-\left(\ln\left(1+\frac{\varepsilon}{\mu_0}\right)\right)^2\Big/2s^2\right). \tag{104}$$

Neglect of ε in the denominator for $\mu_0 > 1$ and expansion of the exponential in Equation (104) to first power in ε leads to the Gaussian function

$$\begin{aligned} p\left(z|m,s\right) &\sim \frac{1}{\sqrt{2\pi}\mu_0 s}\exp\left(-\varepsilon^2/2\left(\mu_0 s\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi}\mathrm{e}^m s}\exp\left(-\left(z-\mathrm{e}^m\right)^2/2\left(\mathrm{e}^m s\right)^2\right), \end{aligned} \tag{105}$$

with mean e^m and standard deviation $\mathrm{e}^m s$. Note that e^m is the median of the log-normal distribution.