On evaluating the efficiency of the delta-lognormal mean estimator and predictor

Aubry Philippe

MethodsX 9 (2022) 101830

Contents lists available at ScienceDirect

MethodsX

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

Method Article

On evaluating the efficiency of the delta-lognormal mean estimator and predictor

Philippe Aubry

OFB - Office français de la biodiversité - Direction surveillance, évaluation, données - Unité données et appui méthodologique, Saint Benoist, BP 20, Le Perray-en-Yvelines F-78612, France

A B S T R A C T

A variable taking positive values from a lognormal distribution and null values with a given probability is distributed according to the so-called delta-lognormal distribution. Two situations arise depending on whether the data are regarded as a random sample from an infinite population (superpopulation) or from a finite population, itself considered as a random sample from a superpopulation. In the case of an infinite population, estimating the mean can be accomplished using a uniformly minimum-variance unbiased estimator (UMVUE). Likewise, the prediction of the mean in the case of a finite population may be based on the UMVUE. In both cases, one expects a gain in precision when taking into account the shape of the distribution by relying on the UMVUE rather than on the sample mean, which is a nonparametric estimator (or predictor).

1. For the infinite population case, the relative efficiency results presented in this article are more complete and more accurate than those published so far.
2. The article fills a gap regarding the question of relative efficiency in the case of a finite population.
3. Calculations were performed using the exact expression for the variance of the UMVUE of the mean, expressed in terms of the confluent hypergeometric limit function.

A R T I C L E I N F O

Method name: UMVU-based estimation or prediction of the mean for delta-lognormal data
Keywords: Delta-lognormal distribution, Superpopulation, Point estimation, Finite population, Mean model, Empirical predictor, Prediction error variance, Relative efficiency, Uniformly minimum-variance unbiased estimator (UMVUE), Confluent hypergeometric limit function
Article history: Received 23 June 2022; Accepted 17 August 2022; Available online 23 August 2022

E-mail address: philippe.aubry@ofb.gouv.fr

https://doi.org/10.1016/j.mex.2022.101830

Specifications table

Subject area:	Probability distributions
More specific subject area:	Relative efficiency of statistical estimators or predictors
Method name:	UMVU-based estimation or prediction of the mean for delta-lognormal data
Name and reference of original method:	• J. Aitchison, J. Brown, The lognormal distribution, Cambridge University Press, Cambridge, UK, 1957. • K. Shimizu, Point estimation, in: E. L. Crow, K. Shimizu (Eds.), Lognormal distributions: theory and applications, Marcel Dekker, New York, USA, 1988, pp. 27-86. • S. Smith, Evaluating the efficiency of the Δ-distribution mean estimator, Biometrics 44 (1988) 485-493.
Resource availability:	*not applicable*

Method details

Let $\mathcal{U}$ be a finite population of sampling units, unambiguously identifiable by integer labels $i = 1, 2, \ldots, N$. Let y be a variable of interest measured or observed on the sampling units, and the total $t_{\mathcal{U}} = N\bar{y}_{\mathcal{U}}$, with $\bar{y}_{\mathcal{U}}$ the mean of y defined over $\mathcal{U}$. A sample $s \subseteq \mathcal{U}$ of size n is drawn from $\mathcal{U}$ by an ignorable selection mechanism. Classical (frequentist) statistics assume that the y_i ($i \in \mathcal{U}$) are random variables of joint distribution ξ. Said another way, $\mathcal{U}$ is itself a random sample drawn from an infinite set of populations sharing the same general statistical properties (i.e., a superpopulation), described by stochastic model ξ. In this article, we refer to the situation where, from the sample s at hand, the purpose is either to estimate the expectation of y in the model ξ, or to predict the finite population mean $\bar{y}_{\mathcal{U}}$ (or equivalently, the total $t_{\mathcal{U}}$).

We consider here a variable of interest y taking nonnegative values ($y \geq 0$). The sample s can be partitioned into $s = s_0 \cup s_1$, $s_0 \cap s_1 = \emptyset$, with s_0 of size n_0 having zero values ($y = 0$) and s_1 of size n_1 having positive values ($y > 0$). If $n_1 = 1$, we note the unique positive value y_{s_1}. A characteristic of such data is that they may exhibit a high proportion of zero values. To take this into account in a sufficiently flexible manner, one approach is to use a two-component mixture model. There are two possibilities: (i) increasing the probability of zero values from a distribution defined for $y \geq 0$ (*zero-inflated distributions*); (ii) introducing a dichotomy between $y = 0$ and $y > 0$ in a mixture model with two separately estimable parts (*hurdle-at-zero, conditional, two-part, and delta distributions* designate the same thing). If the second possibility is adopted, then a suitable model for nonnegative values is written as:

$$G(y; p_0, \boldsymbol{\theta}) = \begin{cases} p_0 & y = 0 \\ p_0 + (1 - p_0)F(y; \boldsymbol{\theta}) & y > 0 \end{cases} \tag{1}$$

where $F(y; \boldsymbol{\theta})$ is a cumulative distribution with parameters $\boldsymbol{\theta}$, corresponding to a positive distribution, either discrete or continuous; here we consider the lognormal distribution.

Lognormal distribution

Let z be a random variable distributed according to the standard normal distribution (i.e., $z \sim \mathrm{Norm}(0, 1)$) of probability density:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} \qquad z \in \mathbb{R} \tag{2}$$

Then, $y = \exp(\mu + \sigma z)$ follows a lognormal distribution that is completely specified by μ and σ^2. The probability density function of the lognormal distribution is written as [1, p. 8, Eq. (2.5)]:

$$f(y; \mu, \sigma^2) = \frac{1}{y\sigma}\phi((\ln y - \mu)/\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln y - \mu}{\sigma}\right)^2\right\} \qquad y \in \mathbb{R}_+^* \tag{3}$$

Delta-lognormal distribution

The lognormal distribution is no longer appropriate when zero values must be accounted for. This leads to using the delta-lognormal distribution [2], often also called the Δ-*distribution* [1] and occasionally the *Bernoulli-lognormal two-part model* (e.g., [3, p. 703]).

The delta-lognormal distribution results from a mixture of a Dirac mass at 0 with probability p_0 and a lognormal distribution with probability $(1 - p_0)$, that is:

$$g(y; p_0, \mu, \sigma^2) = p_0\, \delta(y) + (1 - p_0)\, f(y; \mu, \sigma^2) \tag{4}$$

where $\delta(y)$ is a Dirac distribution that concentrates a unit mass at 0.

The first three cumulants (i.e., expectation, variance, and third central moment) of the distribution (4) are written as [1, p. 95, Eq. (9.43)–(9.45)]:

$$\kappa_1 = (1 - p_0)\, \alpha \tag{5}$$

$$\kappa_2 = (1 - p_0)\, \alpha^2 \{\beta - (1 - p_0)\} \tag{6}$$

$$\kappa_3 = (1 - p_0)\, \alpha^3 \left\{\beta^3 - 3(1 - p_0)\, \beta + 2(1 - p_0)^2\right\} \tag{7}$$

with:

$$\alpha = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \tag{8}$$

$$\beta = \exp\left(\sigma^2\right) \tag{9}$$

Even if it is not necessarily the best possible definition, in this article, skewness is classically defined as:

$$\gamma_1 = \mathrm{E}\left\{\left(\frac{y - \kappa_1}{\kappa_2^{1/2}}\right)^3\right\} = \frac{\kappa_3}{\kappa_2^{3/2}} \tag{10}$$

The skewness γ_1 of the delta-lognormal distribution increases dramatically as σ^2 increases. For small values of σ^2, as p_0 increases, γ_1 first decreases and then increases. For $\sigma^2 > 1$ approximately, as p_0 increases, γ_1 only increases (Fig. 1).

Relative efficiency assessement

In the context of a finite population, one can indifferently consider the mean $\bar{y}_{\mathcal{U}}$ or the total $t_{\mathcal{U}}$ as the quantity of interest. For the finite population that has actually been sampled, these quantities have fixed values that one would be able to know exactly if $s = \mathcal{U}$ (ignoring possible measure or observation errors). Under a superpopulation model ξ, these statistics are random variables whose values one wants to predict. In an infinite population (superpopulation), one is interested in estimating the expectation $\mathrm{E}(y) = \kappa_1$.

For delta-lognormal data, depending on the level of skweness of the distribution, the question arises of the gain in precision that can be achieved by relying on the *uniformly minimum-variance unbiased estimator* (UMVUE) compared to using the sample mean $\bar{y}_s$, either for estimating κ_1 or for predicting $\bar{y}_{\mathcal{U}}$ (or equivalently, $t_{\mathcal{U}}$). In other words, one may compare the situation where the shape of the distribution is known, to the situation where it is unknown (or known but not taken into account), first in the case of an infinite population (estimation context), then in that of a finite population (prediction context).

In the estimation context, Aitchison and Brown [1, p. 98, Fig. 9.1] provided relative efficiency results only for $p_0 = 0.5$ and for the degenerate case of the lognormal distribution ($p_0 = 0$), using a variance approximation (see the validation section). By doing so, the sample size n is disregarded in

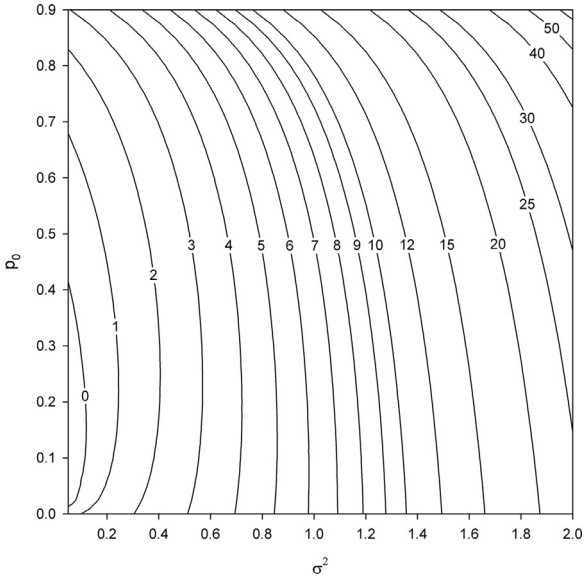

Fig. 1. Skewness γ_1 for the delta-lognormal distribution as a function of p_0 and σ^2. For maximum bounds of the parameter space considered on the figure ($p_0 = 0.9, \sigma^2 = 2$), the skewness is $\gamma_1 \simeq 64.5$.

the relative efficiency assessment. Shimizu [2] did not document the relative efficiency in the case of the delta-lognormal distribution. Smith [4] considered the relative efficiency for $p_0 = 0.1$ and $p = 0.5$, for very small sample sizes, using exact or approximate variance. To our knowledge, the relative efficiency assessment in the prediction context has not been documented yet.

In this technical article, after providing a compendium of fundamental formulas for UMVU estimation in the case of the delta-lognormal distribution, we document the relative efficiency more thoroughly than in the past by considering both the estimation and prediction contexts, taking into account the sample size (and the finite population size in the prediction context), and varying the probabilitiy of getting a zero value up to $p_0 = 0.9$. In all cases we use the exact expression of the variance of the estimator (or predictor).

Estimation context

Unknown shape of the distribution

Let s be a sample of size n drawn by random sampling from an infinite population of unknown shape. The unbiased estimator of κ_1 is the sample mean $\bar{y}_s$ and its sampling variance is written as:

$$\mathrm{V}(\bar{y}_s) = \frac{\kappa_2}{n} \tag{11}$$

where κ_2 is estimated without bias by:

$$S_s^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)^2 \tag{12}$$

The sampling variance is then estimated without bias by:

$$\widehat{\mathrm{V}}(\bar{y}_s) = \frac{S_s^2}{n} \tag{13}$$

Known shape of the distribution

When y is distributed according to a delta-lognormal distribution, κ_1 can be estimated by the UMVUE [1, p. 97, Eq. (9.54)] (typo corrected); [4]:

$$\hat{\kappa}_1 = \begin{cases} (1-\hat{p}_0)\exp(\overline{y}_{\ln})\, g_{n_1-1}\left(\frac{1}{2}s^2_{\ln}\right) & n_1 > 1 \\ \frac{y_{s_1}}{n} & n_1 = 1 \\ 0 & n_1 = 0 \end{cases} \tag{14}$$

with:

$$\overline{y}_{\ln} = \frac{1}{n_1}\sum_{i\in s_1} \ln y_i \tag{15}$$

$$s^2_{\ln} = \frac{1}{n_1 - 1}\sum_{i\in s_1} (\ln y_i - \overline{y}_{\ln})^2 \tag{16}$$

and $g_m(t)$ an infinite series introduced by Finney [5, Eq. (10)], which can be written as [6, Eq. (1.2)], [7, Eq. (3)]:

$$g_m(t) = \sum_{j=0}^{\infty} \frac{1}{j!}\frac{m+2j}{m}\left(\frac{m}{m+1}t\right)^j \prod_{i=1}^{j}\frac{m}{m+2i} \tag{17}$$

The exact variance of $\hat{\kappa}_1$ (14) was provided by Smith [4, Eq. (6)]. The function $g_m(t)$ belongs to the class of generalized hypergeometric functions and can be written as a particular instance of the confluent hypergeometric limit function (here denoted as ${}_0F_1$) as [6, Eq. (2.1)]:

$$g_m(t) = {}_0F_1\left(\frac{m}{2}; \frac{m^2 t}{2(m+1)}\right) \tag{18}$$

with [8, Eq. (4)], [9, p. 333, Eq. (16.3.1)]:

$${}_0F_1(a; z) = \sum_{j=0}^{\infty} \frac{z^j}{j!\,(a)_j} \tag{19}$$

where $(a)_j$ is the notation used in special function theory for the rising factorial:

$$(a)_j = \begin{cases} 1 & j = 0 \\ \prod_{i=0}^{j-1}(a+i) & j \geq 1 \end{cases} \tag{20}$$

Note that the numerical evaluation of special functions ${}_0F_1(a; z)$ and $g_m(t)$ is addressed later in the article.

Using the confluent hypergeometric limit function, according to Shimizu [2, p. 50], estimators $\hat{\kappa}_1$ and $\hat{\kappa}_2$ can be written, respectively, as:

$$\hat{\kappa}_1 = \begin{cases} (1-\hat{p}_0)\exp(\overline{y}_{\ln})\, {}_0F_1\left(\frac{n_1-1}{2}; \frac{(n_1-1)^2}{4n_1}s^2_{\ln}\right) & n_1 > 1 \\ \frac{y_{s_1}}{n} & n_1 = 1 \\ 0 & n_1 = 0 \end{cases} \tag{21}$$

and

$$\hat{\kappa}_2 = \begin{cases} (1-\hat{p}_0)\exp(2\overline{y}_{\ln})\Big[{}_0F_1\Big(\frac{n_1-1}{2};\frac{(n_1-1)^2}{n_1}s^2_{\ln}\Big) & \\ \qquad -\Big(\frac{n_1-1}{n-1}\Big){}_0F_1\Big(\frac{n_1-1}{2};\frac{(n_1-1)(n_1-2)}{2n_1}s^2_{\ln}\Big)\Big] & n_1 > 1 \\ \frac{y^2_{s_1}}{n} & n_1 = 1 \\ 0 & n_1 = 0 \end{cases} \tag{22}$$

The exact variance of $\hat{\kappa}_1$ is [10, Remark 3.1], [2, pp. 50-51]:

$$\mathrm{V}(\hat{\kappa}_1) = \alpha^2\left[\frac{1}{n^2}\sum_{j=2}^{n}\binom{n}{j}(1-p_0)^j\, p_0^{n-j}\, j^2 \exp\left(\frac{\sigma^2}{j}\right) \times {}_0F_1\left(\frac{j-1}{2};\frac{(j-1)^2}{4j^2}\sigma^4\right) - (1-p_0)^2\right] \tag{23}$$

Equivalent approximations of $\mathrm{V}(\hat{\kappa}_1)$ were given by Shimizu [2, p. 51] and Aitchison and Brown [1, p. 99, Eq. (9.58)] (see the validation section); in this article we use the exact expression (23), which translates into Algorithm 1, using Algorithm 2 [11].

Algorithm 1 Exact variance $\mathrm{V}(\hat{\kappa}_1)$ (23) with n the sample size, p, m, v the values for parameters p_0, μ, σ^2, respectively.

```
function ExactVariance(n, p, m, v)
    q ← 1 − p
    x ← q
    y ← p^(n−1)
    S ← 0
    for j = 2 to n do
        x ← x × q
        y ← y/p
        z ← ((j − 1)/(2 × j) × v)^2
        S ← S + Binomial(n, j) × x × y × j^2 × exp(v/j) × Hypergeometric0F1((j − 1)/2, z)
    end for j
    r ← exp(2 × m + v) × (S/n^2 − q^2)
    [ return r ]
end function
```

Pennington [12] has provided the UMVUE for the variance $\mathrm{V}(\hat{\kappa}_1)$ expressed with the function $g_m(t)$ [12, Eq. (4)]; for the sake of homogeneity, we present it here expressed using the confluent hypergeometric limit function:

$$\widehat{\mathrm{V}}(\hat{\kappa}_1) = \begin{cases} (1-\hat{p}_0)\exp(2\overline{y}_{\ln})\Big[(1-\hat{p}_0)\,{}_0F_1^2\Big(\dfrac{n_1-1}{2};\dfrac{(n_1-1)^2}{4n_1}s^2_{\ln}\Big) & \\ \qquad -\Big(\frac{n_1-1}{n-1}\Big){}_0F_1\Big(\frac{n_1-1}{2};\frac{(n_1-1)(n_1-2)}{2n_1}s^2_{\ln}\Big)\Big] & n_1 > 1 \\ \Big(\dfrac{y_{s_1}}{n}\Big)^2 & n_1 = 1 \\ 0 & n_1 = 0 \end{cases} \tag{24}$$

Relative efficiency

We compare the precision of the estimator for κ_1 according to whether the distribution shape is unknown or known. It is expected that taking into account the knowledge of the distribution shape will lead to a gain in precision. The relative efficiency is defined in the same way as that used by

Algorithm 2 Iterative computation of the binomial coefficient.

```
function Binomial(n, k)
    p ← n − k
    if k < p then
        p ← k
        k ← n − p
    end if
    if p < 0 then                                   ▷ error in the arguments
        [ return 0 ]
    end if
    if p = 0 then                                   ▷ trivial case
        [ return 1 ]
    end if
    r ← k + 1
    for i = 2 to p do
        r ← r × (k + i))/i
    end for i
    [ return r ]
end function
```

Aitchison and Brown [1, p. 99, Eq. (9.62)]:

$$\text{eff}_1 = \frac{\text{V}(\hat{\kappa}_1)}{\text{V}(\bar{y}_s)} \tag{25}$$

that is, taking into account the shape of the distribution leads to a gain in precision, which is higher when eff_1 is low ($\text{V}(\hat{\kappa}_1) < \text{V}(\bar{y}_s)$). Hence, we quantify the gain in precision by expressing it as a function of p_0, σ^2 (or, equivalently, σ) and n. The parameter μ vanishes through the elimination of the α^2 term that appears in the numerator and denominator of the relative efficiency.

We vary $p_0 = 0(0.025)0.9$, $\sigma = 0.05(0.05)2$ and $n = 50, 100, 500, 1\,000$. The results show that the gain in precision increases substantially (eff_1 decreases) as σ increases and increases more weakly — for a fixed (p_0, σ) point — as n increases (Fig. 2). For a fixed σ value, the gain decreases (eff_1 increases) as p_0 increases, and this is all the more pronounced when n is small (see Fig. 2a). This phenomenon diminishes as n increases (see Fig. 2b–d). Finally, we note a small difference between $n = 500$ (Fig. 2c) and $n = 1\,000$ (Fig. 2d), which suggests a convergence of eff_1 toward its asymptotic limit.

To illustrate the speed of convergence of eff_1 toward its asymptotic limit, we set $\sigma^2 = 2$, and we repeat the calculations for $p_0 = 0.0(0.1)0.90$ and $50 \leq n \leq 1\,000$. The asymptotic limit and the speed of convergence of eff_1 depend on p_0; the lower p_0 is, the higher the efficiency gain and the speed of convergence (Fig. 3).

Prediction context

By denoting $r = \mathcal{U} - s$, the total defined on the population can be written as:

$$t_{\mathcal{U}} = \sum_{i \in \mathcal{U}} y_i = \underbrace{\sum_{i \in s} y_i}_{t_s} + \underbrace{\sum_{i \in r} y_i}_{t_r} \tag{26}$$

that is the sum of the totals defined over the sample (t_s) and the remaining part in the population (t_r). A predictor of $t_{\mathcal{U}}$ can be written as:

$$\tilde{t}_{\mathcal{U}} = \sum_{i \in s} y_i + \underbrace{\sum_{i \in r} \tilde{y}_i}_{\tilde{t}_r} \tag{27}$$

Consider a simple mean model where the random variables y_i $(i = 1, \ldots, N)$ have the same expectations and variances and are not correlated, that is:

$$\mathrm{E}(y_i) = \kappa_1 \tag{28a}$$

$$\mathrm{V}(y_i) = \kappa_2 \tag{28b}$$

$$\mathrm{Cov}(y_i, y_j) = 0 \qquad (i \neq j) \tag{28c}$$

The empirical predictor is written as:

$$\tilde{t}_{\mathcal{U}} = \sum_{i \in s} y_i + \underbrace{(N-n)\widehat{\mathrm{E}}(y_i)}_{\tilde{t}_r} \tag{29}$$

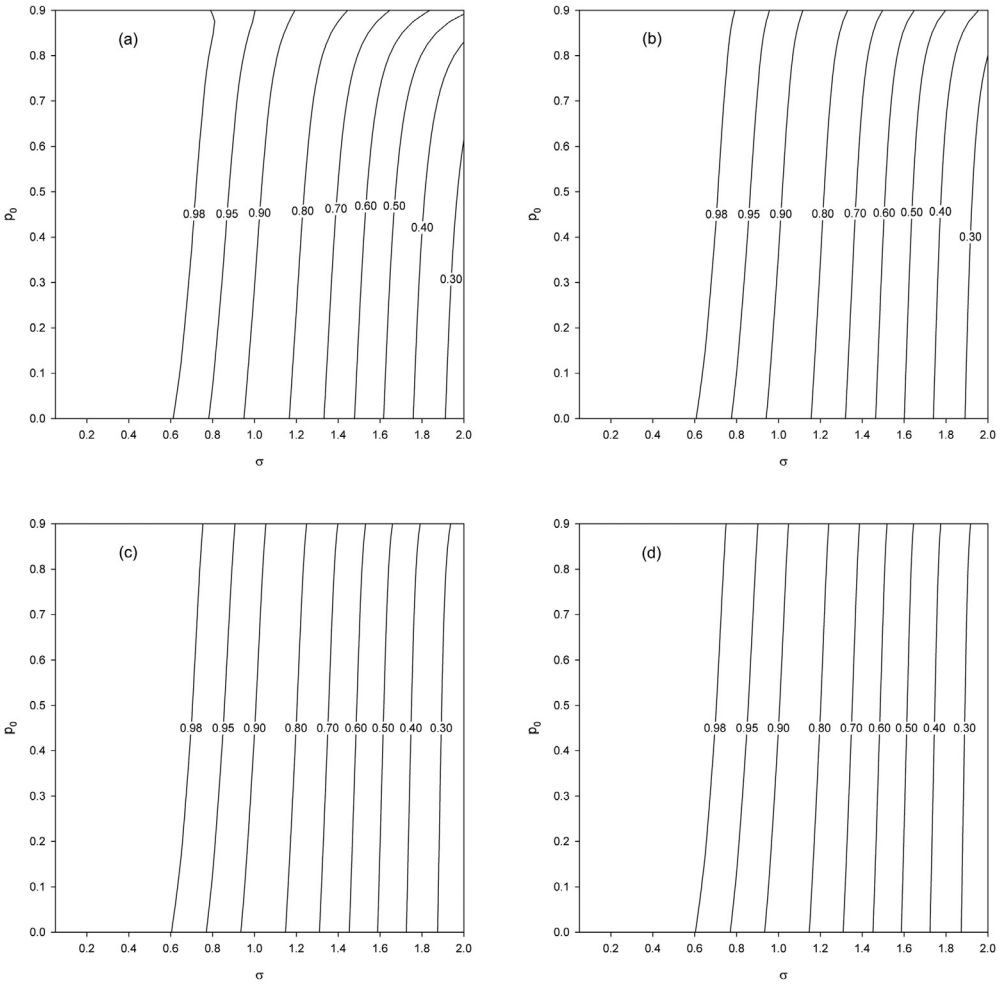

Fig. 2. Relative efficiency eff_1 as a function of p_0 and σ for different sample sizes (n). (a) $n = 50$. (b) $n = 100$. (c) $n = 500$. (d) $n = 1\,000$.

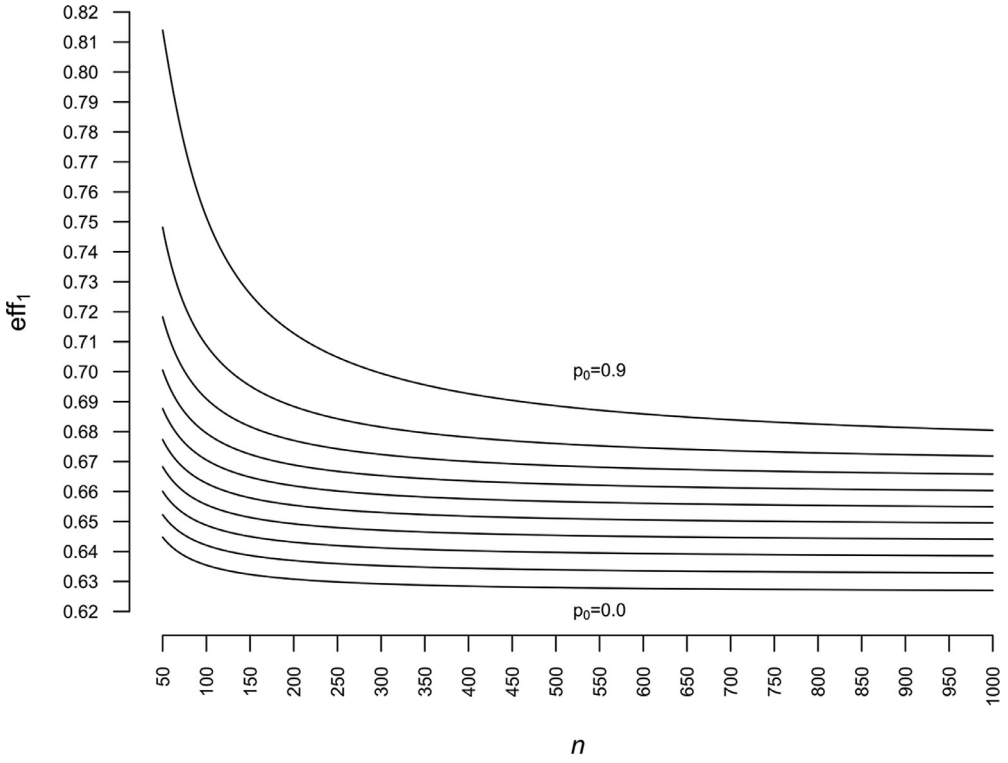

Fig. 3. Relative efficiency eff_1 as a function of sample size (n) for $\sigma^2 = 2$ and $p_0 = 0.0(0.1)0.9$.

Predictor (29) is model-unbiased; that is, if the model is correct, then $\text{E}(\tilde{t}_{\mathcal{U}} - t_{\mathcal{U}}) = 0$. The variance of the prediction error is obtained as:

$$\text{V}(\tilde{t}_{\mathcal{U}} - t_{\mathcal{U}}) = \text{V}(\tilde{t}_r - t_r) = \text{V}(\tilde{t}_r) + \text{V}(t_r) - 2\underbrace{\text{Cov}(\tilde{t}_r, t_r)}_{0} \tag{30}$$

The ξ-covariance between $\tilde{t}_r$ and t_r is zero since: (i) $\tilde{t}_r$ is a function of the set of values in s ($\{y_i, i \in s\}$), not of the set of values in r ($\{y_i, i \in r\}$) which we do not know; and (ii) the two sets of values $\{y_i, i \in s\}$ and $\{y_i, i \in r\}$ are uncorrelated under the model (see (28c)).

Distribution of unknown shape

Predictor (29) can be written as:

$$\tilde{t}_{\mathcal{U}}^{\text{exp}} = \sum_{i \in s} y_i + (N-n)\bar{y}_s = n\bar{y}_s + (N-n)\bar{y}_s = N\bar{y}_s \tag{31}$$

which we can designate as an *expansion predictor* [13]. From relation (30), the prediction error variance of $(\tilde{t}_{\mathcal{U}}^{\text{exp}} - t_{\mathcal{U}})$ is obtained as:

$$\text{V}(\tilde{t}_{\mathcal{U}}^{\text{exp}} - t_{\mathcal{U}}) = (N-n)^2[\text{V}(\bar{y}_s) + \text{V}(\bar{y}_r)] \tag{32}$$

$$= (N-n)^2\left(\frac{1}{n} + \frac{1}{N-n}\right)\kappa_2 \tag{33}$$

$$= N^2\left(1 - \frac{n}{N}\right)\frac{\kappa_2}{n} \tag{34}$$

The predictor of the mean is $\bar{y}_s$, and its prediction error variance is $\mathrm{V}(\bar{y}_s - \bar{y}_\mathcal{U}) = (1 - n/N)\kappa_2/n$. It follows the limit:

$$\lim_{N\to\infty} \mathrm{V}(\bar{y}_s - \bar{y}_\mathcal{U}) = \mathrm{V}(\bar{y}_s) \tag{35}$$

For prediction error variance (34), an unbiased estimator is obtained by substituting estimator $\hat{\kappa}_2$ (22) for parameter κ_2.

Distribution of known shape

Predictor (29) can be written as:

$$\tilde{t}_\mathcal{U}^{\mathrm{mvu}} = \sum_{i\in s} y_i + (N-n)\hat{\kappa}_1 \tag{36}$$

From relation (30), the prediction error variance $(\tilde{t}_\mathcal{U}^{\mathrm{mvu}} - t_\mathcal{U})$ is obtained as:

$$\mathrm{V}(\tilde{t}_\mathcal{U}^{\mathrm{mvu}} - t_\mathcal{U}) = (N-n)^2[\mathrm{V}(\hat{\kappa}_1) + \mathrm{V}(\bar{y}_r)] \tag{37}$$

$$= (N-n)[(N-n)\,\mathrm{V}(\hat{\kappa}_1) + \kappa_2] \tag{38}$$

where the variance $\mathrm{V}(\hat{\kappa}_1)$ is given by expression (23). The mean predictor is $\tilde{\bar{y}}_\mathcal{U} = N^{-1}\tilde{t}_\mathcal{U}^{\mathrm{mvu}}$ (e.g., [14, Eq. (3)]), and its prediction error variance is $\mathrm{V}(\tilde{\bar{y}}_\mathcal{U} - \bar{y}_\mathcal{U}) = N^{-2}\,\mathrm{V}(\tilde{t}_\mathcal{U}^{\mathrm{mvu}} - t_\mathcal{U})$. It follows the limits:

$$\lim_{N\to\infty} \tilde{\bar{y}}_\mathcal{U} = \lim_{N\to\infty} \left[\frac{1}{N}\sum_{i\in s} y_i + \frac{(N-n)}{N}\hat{\kappa}_1\right] = \hat{\kappa}_1 \tag{39}$$

$$\lim_{N\to\infty} \mathrm{V}(\tilde{\bar{y}}_\mathcal{U} - \bar{y}_\mathcal{U}) = \lim_{N\to\infty} \left[\frac{(N-n)^2}{N^2}\,\mathrm{V}(\hat{\kappa}_1) + \frac{(N-n)}{N^2}\kappa_2\right] = \mathrm{V}(\hat{\kappa}_1) \tag{40}$$

For prediction error variance (38), an unbiased estimator is obtained by substituting estimators $\widehat{\mathrm{V}}\left(\hat{\kappa}_1\right)$ (24) and $\hat{\kappa}_2$ (22) for $\mathrm{V}\left(\hat{\kappa}_1\right)$ and κ_2, respectively.

Relative efficiency

We compare the precision of the $t_\mathcal{U}$ prediction depending on whether the shape of the distribution is unknown or known. In addition to p_0, σ^2 and n, we must also vary the finite population size N. The relative efficiency is defined as:

$$\mathrm{eff}_2 = \frac{\mathrm{V}(\tilde{t}_\mathcal{U}^{\mathrm{mvu}} - t_\mathcal{U})}{\mathrm{V}(\tilde{t}_\mathcal{U}^{\mathrm{exp}} - t_\mathcal{U})} = \frac{\mathrm{V}(\tilde{\bar{y}}_\mathcal{U} - \bar{y}_\mathcal{U})}{\mathrm{V}(\bar{y}_s - \bar{y}_\mathcal{U})} \tag{41}$$

From (35) and (40), we obtain:

$$\lim_{N\to\infty} \mathrm{eff}_2 = \mathrm{eff}_1 \tag{42}$$

First, we examine the effect of the population size N for a fixed sample size n, which is equivalent to examining the effect of the sampling fraction $f = n/N$. We set $n = 50$ and vary the sampling fraction as $f = 0.4, 0.2, 0.1, 0.05$ ($N = 125, 250, 500, 1\,000$). As before, we vary $p_0 = 0(0.025)0.9$ and $\sigma = 0.05(0.05)2$. The obtained results (Fig. 4) show a smaller gain in precision than in the case of an infinite population (Fig. 2a). This finite population effect is less pronounced as f tends toward 0 ($N \to \infty$) since in that scenario the situation tends toward the asymptotic result of the case considered here, which corresponds to Fig. 2a.

To illustrate the speed of convergence of eff_2 toward its asymptotic limit eff_1, we set $\sigma^2 = 2$, and we repeat the calculations for $p_0 = 0.0(0.1)0.90$, $n = 50$ and $500 \le N \le 10\,000$. As in the infinite population case, the asymptotic limit and the speed of convergence of eff_2 depend on p_0. The lower p_0 is, the higher the efficiency gain (eff_2 decreases). In contrast, the speed of convergence becomes less important when p_0 decreases (Fig. 5). In practice, we can consider that we are almost at convergence

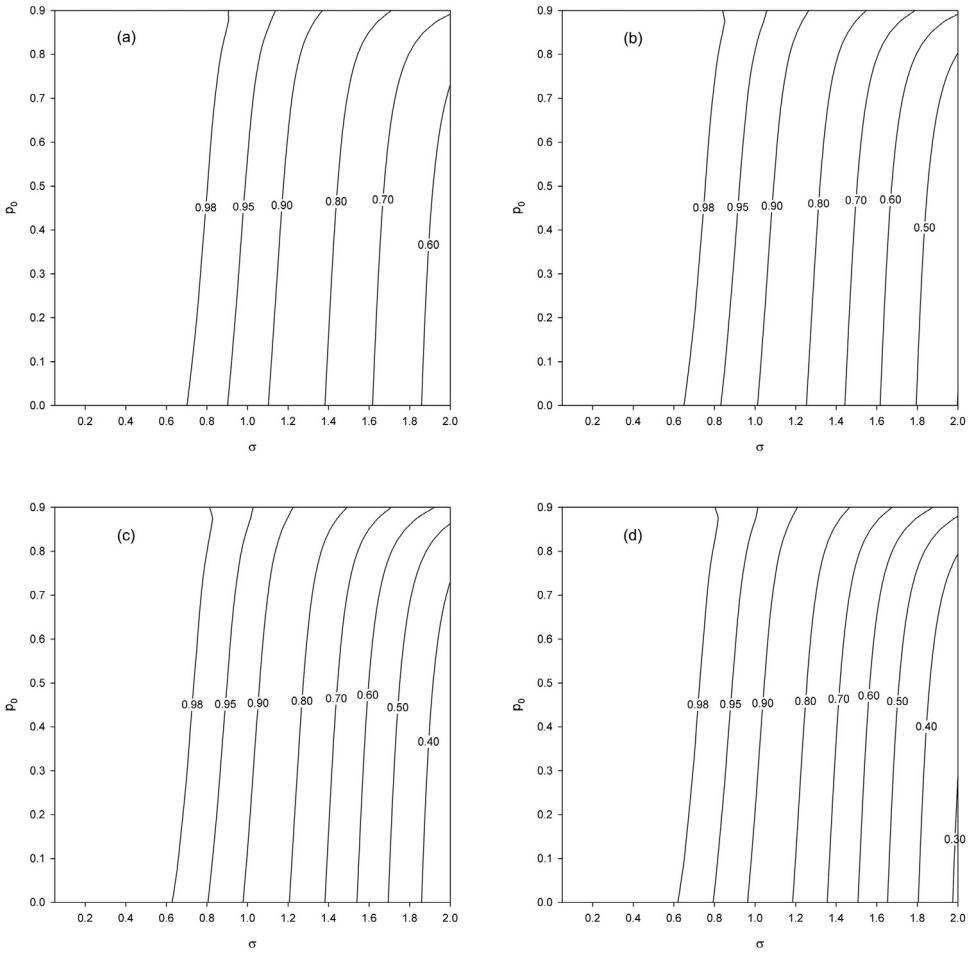

Fig. 4. Relative efficiency eff_2 as a function of p_0 and σ for a fixed sample size ($n = 50$) and different finite population sizes (N) and hence different sampling fractions (f). (a) $N = 125$, $f = 0.4$. (b) $N = 250$, $f = 0.2$. (c) $N = 500$, $f = 0.1$. (d) $N = 1\,000$, $f = 0.05$.

as soon as $N = 5\,000$ (or $N = 10\,000$ if one is more conservative). Note that limit values correspond to the values for $n = 50$ in Fig. 3 and are represented by dotted half lines in Fig. 5.

Finally, we examine the gain in precision when n and N increase jointly, keeping the sampling fraction constant. For $f = 0.1$, we increase the population size as $N = 500, 1\,000, 5\,000, 10\,000$ ($n = 50, 100, 500, 1\,000$). The obtained results (Fig. 6) show a gain in precision that increases (eff_2 decreases) as N and n jointly increase. Note that Fig. 6a is the same as Fig. 4c ($N = 500$, $n = 50$). There is only a small difference between the case $N = 5\,000, n = 500$ (Fig. 6c) and the case $N = 10\,000, n = 1\,000$ (Fig. 6d), which suggests the convergence of eff_2 toward its asymptotic limit (in the sense that n and N jointly increase and for $f = 0.1$).

Numerical evaluation of the generalized hypergeometric functions

In this article, we need to numerically evaluate ${}_0F_1(a; z)$ ($a \in \mathbb{R}_+^*$ and $z \in \mathbb{R}_+^*$) — or equivalently $g_m(t)$ — to calculate $\text{V}(\hat{\kappa}_1)$ (23), which is directly involved in the relative efficiency eff_1 (25) and through formula (38) in the relative efficiency eff_2 (41).

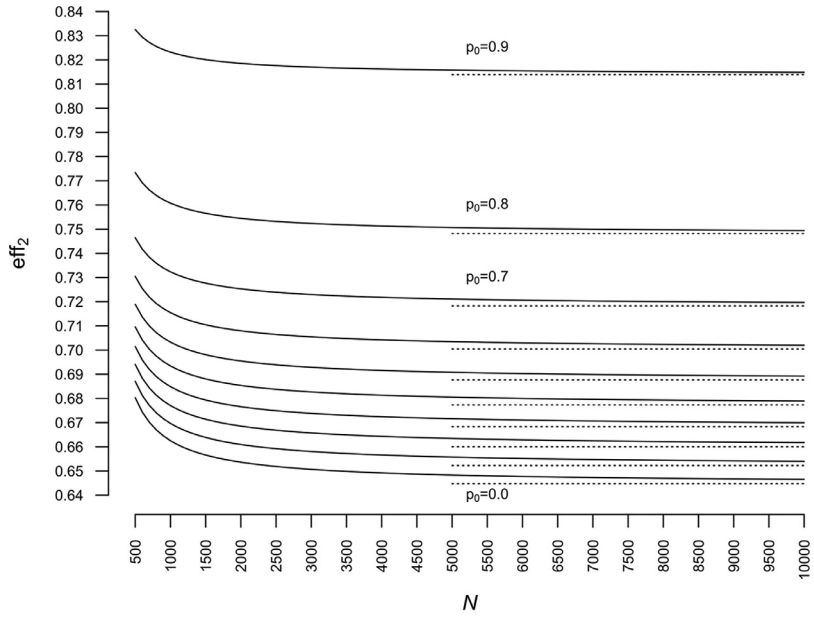

Fig. 5. Relative efficiency eff_2 as a function of population size (N) for a fixed sample size ($n = 50$), for $\sigma^2 = 2$ and $p_0 = 0.0(0.1)0.9$. Details in the text.

Evaluating ${}_0F_1(a; z)$ by recurrence

The confluent hypergeometric limit function (19) can be written as:

$$ {}_0F_1(a; z) = 1 + \frac{z}{a} + \sum_{j=2}^{\infty} \frac{1}{\prod_{i=0}^{j-1}(a+i)} \times \frac{z^j}{j!} \tag{43} $$

which gives the recurrence relation:

$$ \begin{aligned} A_1 &= \frac{z}{a} & S_1 &= 1 + A_1 \\ A_j &= A_{j-1} \times \frac{1}{a+j-1} \times \frac{z}{j} & S_j &= S_{j-1} + A_j \qquad \text{for } j = 2, 3, \ldots, k \end{aligned} \tag{44} $$

and translates into Algorithm 3.

The series is infinite, but in practice, it is only required to compute the sum until it reaches a level of convergence considered sufficient; for example, with the convergence criterion $|S_j - S_{j-1}| < \epsilon$, we used $\epsilon = 10^{-10}$. Another way to proceed is to determine the number of terms necessary to reach the precision allowed by the computer at hand (see [15, pp. 88-89]). Computing successive terms of the series by means of the recurrence relation is a computational method often used by default (e.g., [16, p. 99]). For other methods, the interested reader is referred to [17,18].

Evaluating $g_m(t)$ by recurrence

One can proceed in the same way as previously for the function $g_m(t)$ (17), which can be written as (see, for example, [4, Eq. (1)], with $m = n_1 - 1$):

$$ g_m(t) = 1 + \frac{m}{m+1} \times t + \sum_{j=2}^{\infty} \frac{m^{2j-1}}{(m+1)^j \prod_{i=2}^{j}(m + 2(i-1))} \times \frac{t^j}{j!} \tag{45} $$

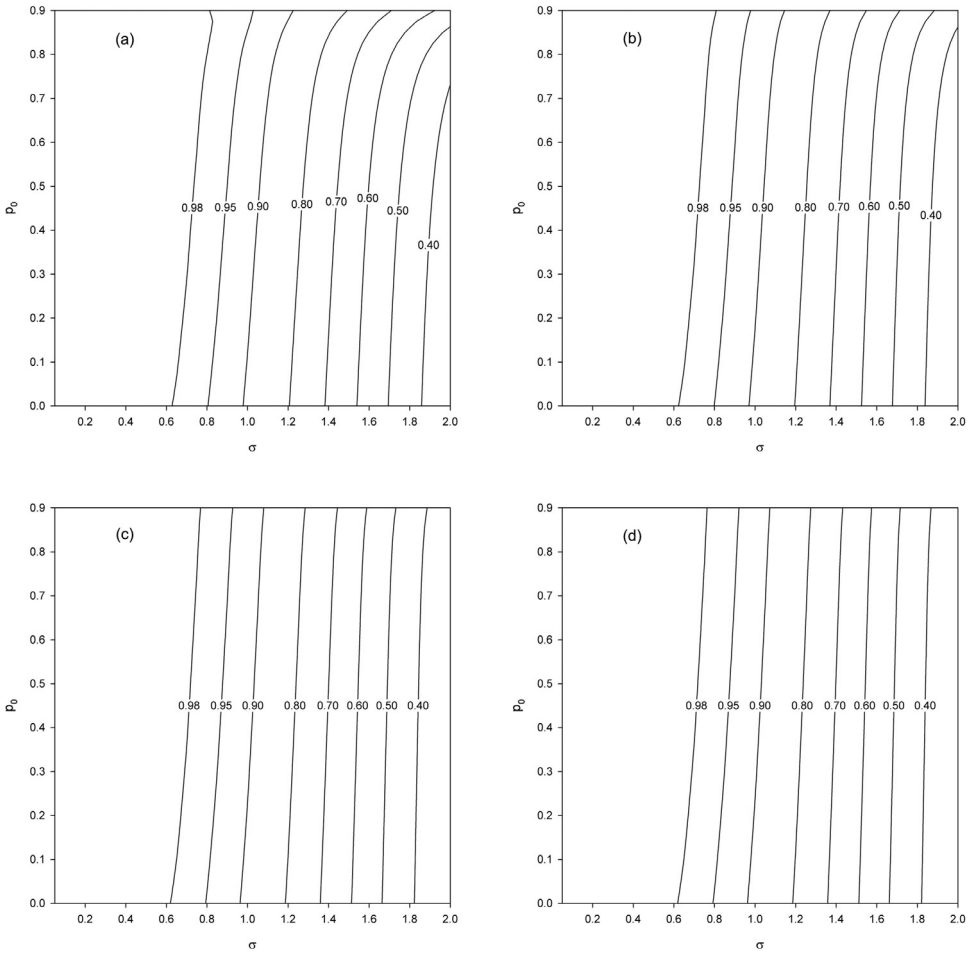

Fig. 6. Relative efficiency eff_2 as a function of p_0 and σ for a fixed sampling fraction $f = 0.1$ and increasing finite population (N) and sample (n) sizes. (a) $N = 500$ ($n = 50$). (b) $N = 1\,000$ ($n = 100$). (c) $N = 5\,000$ ($n = 500$). (d) $N = 10\,000$ ($n = 1\,000$).

Algorithm 3 Confluent hypergeometric limit function ${}_0F_1$.

function Hypergeometric0F1(a, z, ϵ)
  $A \leftarrow z/a$
  $R \leftarrow 1 + A$
  $j = 2$
  loop
    $A \leftarrow A \times 1/(a + j - 1) \times z/j$
    $S \leftarrow R + A$
    if $|S - R| < \epsilon$ **then**
      [return S]
    end if
    $j \leftarrow j + 1$
    $R \leftarrow S$
  end loop
end function

which gives the recurrence relation:

$$\begin{aligned} A_1 &= \frac{m}{m+1} \times t & S_1 &= 1 + A_1 \\ A_j &= A_{j-1} \times \frac{m^2}{(m+1)(m+2(j-1))} \times \frac{t}{j} & S_j &= S_{j-1} + A_j \qquad \text{for } j = 2, 3, \ldots, k \end{aligned} \tag{46}$$

and translates into Algorithm 4.

Algorithm 4 Finney's generalized hypergeometric function $g_m(t)$

```
function Finney(m, t, ϵ)
    A ← m/(m + 1) × t
    B ← A × m
    R ← 1 + A
    j = 2
    loop
        A ← A × B/(m + 2 × (j − 1))/j
        S ← R + A
        if |S − R| < ϵ then
            [ return S ]
        end if
        j ← j + 1
        R ← S
    end loop
end function
```

Validation

In this section: (i) we check the calculation accuracy of the confluent hypergeometric limit function ${}_0F_1(a; z)$ required for computing $\mathrm{V}(\hat{\kappa}_1)$ using Algorithm 1; (ii) we assess the correctness of Algorithm 1 by comparing its results against those of Monte Carlo simulations; (iii) we justify the use of the exact expression for the variance $\mathrm{V}(\hat{\kappa}_1)$ (23) for accurately documenting the relative efficiency when considering $\hat{\kappa}_1$ in the estimation or prediction contexts.

Precision of ${}_0F_1(a; z)$ evaluated by recurrence

We verify that the calculations performed in this article — using double-precision floating-point arithmetic — by implementing the computation of ${}_0F_1(a; z)$ through Algorithm 3 with $\epsilon = 10^{-10}$ have sufficient numerical accuracy. To do this, we first examine the ranges of variation of parameter a and argument z and then compare the results of our procedure with reference values.

The examination of $\mathrm{V}(\hat{\kappa}_1)$ (23) shows that the minimal value that a can take in this expression is $a = 0.5$ and that for the maximal value $n = 1000$ used in this article, we have $a \simeq 500$. Besides, we may consider $0 < z \leq 1\,000$ for covering the situations addressed in [19] and similar future examples.

For $a = 0.5$, it is possible to use a particular representation of ${}_0F_1(a; z)$ [20, p. 594, Table 7.13.1 (row 5, column 2)]:

$${}_0F_1(0.5; z) = \cosh(2\sqrt{z}) \tag{47}$$

with $\cosh = (\exp(z) + \exp(-z))/2$. For $z = 0.01(0.01)1$, we obtain an absolute error less than 10^{-12} (at most $k = 8$ iterations are required). For $z = 1(1)1000$, we obtain a relative error less than 10^{-12}. To gain an idea of the order of magnitude, ${}_0F_1(0.5; 1000) \simeq 1.466103955241 \times 10^{27}$ (at most $k = 69$ iterations are required).

For $a > 0.5$, we use as reference values those returned by the function `Hypergeometric0F1[a,z]` of Mathematica®. For $a = 5(5)500$ and $z = 0.01(0.01)1$, we obtain an absolute error less than 10^{-12} (at most $k = 7$ iterations are required). For $a = 5(5)500$ and $z = 10(10)1000$, we obtain a relative error less than 10^{-11} (at most $k = 67$ iterations are required).

Table 1
Results of the Monte Carlo simulations to assess the correctness of Algorithm 1. Details in the text.

p_0	n	K	κ_1	$E_{MC}(\hat{\kappa}_1)$	$V(\hat{\kappa}_1)$	$V_{MC}(\hat{\kappa}_1)$
0.1	50	10^8	29.8039	29.8039	268.1132	268.2039
0.1	100	10^8	29.8039	29.8027	126.5944	126.5823
0.1	500	10^7	29.8039	29.8032	24.1667	24.1690
0.1	1000	10^7	29.8039	29.8028	12.0127	12.0134
0.5	50	10^8	16.5577	16.5584	167.2448	167.2557
0.5	100	10^8	16.5577	16.5568	75.8297	75.8418
0.5	500	10^7	16.5577	16.5574	13.9909	13.9922
0.5	1000	10^7	16.5577	16.5565	6.9244	6.9199
0.9	50	10^8	3.3115	3.3113	61.8555	61.3674
0.9	100	10^8	3.3115	3.3115	22.3205	22.3108
0.9	500	10^7	3.3115	3.3108	3.1210	3.1220
0.9	1000	10^7	3.3115	3.3112	1.4862	1.4867

Monte Carlo simulations

We simulate K random samples of size n from a delta-lognormal distribution of parameters p_0, $\mu = 1.5$ and $\sigma = 2$, and we compute the UMVUE (21) for each sample. This results in K values $\hat{\kappa}_1$ from which we calculate Monte Carlo approximations of expectation (denoted as $E_{MC}(\hat{\kappa}_1)$) and variance (denoted as $V_{MC}(\hat{\kappa}_1)$). Then, we can compare these Monte Carlo approximations to their theoretical values, that is, κ_1 — since $\hat{\kappa}_1$ is unbiased — and $V(\hat{\kappa}_1)$, respectively. We vary $n = 50, 100, 500, 1000$ and $p_0 = 0.1, 0.5, 0.9$. For accurate Monte Carlo approximations, we set $K = 10^8$ for $n = 50, 100$ and $K = 10^7$ for $n = 500, 1000$. Table 1 shows a very good agreement between the Monte Carlo approximations and the calculated theoretical values. We conclude that, for $50 \leq n \leq 1000$ and $p_0 \leq 0.9$, one can trust Algorithm 1.

Variance approximation of $V(\hat{\kappa}_1)$

An approximation of $V(\hat{\kappa}_1)$ in $O(n^{-2})$ is [2, p. 51]:

$$V_{app}(\hat{\kappa}_1) = \frac{1}{n}(1-p_0)\left[p_0 + \frac{1}{2}\sigma^2(\sigma^2+2)\right]\exp(2\mu+\sigma^2) \tag{48}$$

Approximation (48) is algebraically equivalent to the approximation given by Aitchison and Brown [1, p. 99, Eq. (9.58)]. Aitchison & Brown indicated that this approximation is given for large n and p_0 substantially less than 1, without further details. The quality of approximation (48) depends on the sample size n and parameters p_0 and σ^2. To specify the domains of use of the approximation, we compute the ratio $V_{app}(\hat{\kappa}_1)/V(\hat{\kappa}_1)$ for $p = 0(0.025)0.9$, $\sigma = 0.05(0.05)2.0$ and $n = 50, 100, 500, 1000$ (Fig. 7). The approximation improves when σ and p_0 are small, and for the same fixed (p_0, σ) point, the accuracy increases as the sample size n increases. Thus, accurate calculations for great p_0 (i.e., here up to $p_0 = 0.9$) and σ^2 values requires using formula (23) (translated into Algorithm 1) instead of approximation (48), unless n is sufficiently large.

Additional information

Pennington [12] introduced the use of the delta-lognormal distribution in marine biology to estimate mean abundance more efficiently than can be achieve with the sample mean in the case of highly skewed distributions. The article by Pennington [12] is widely cited in the literature (603 citations according to Google Scholar, at the time of writing this article). Regarding the robustness of this approach, the reader is referred to Myers and Pepin [21,22], Syrjala [23] and Christman [24]. In the context of the mean (or total) prediction, see the recent contribution [19].

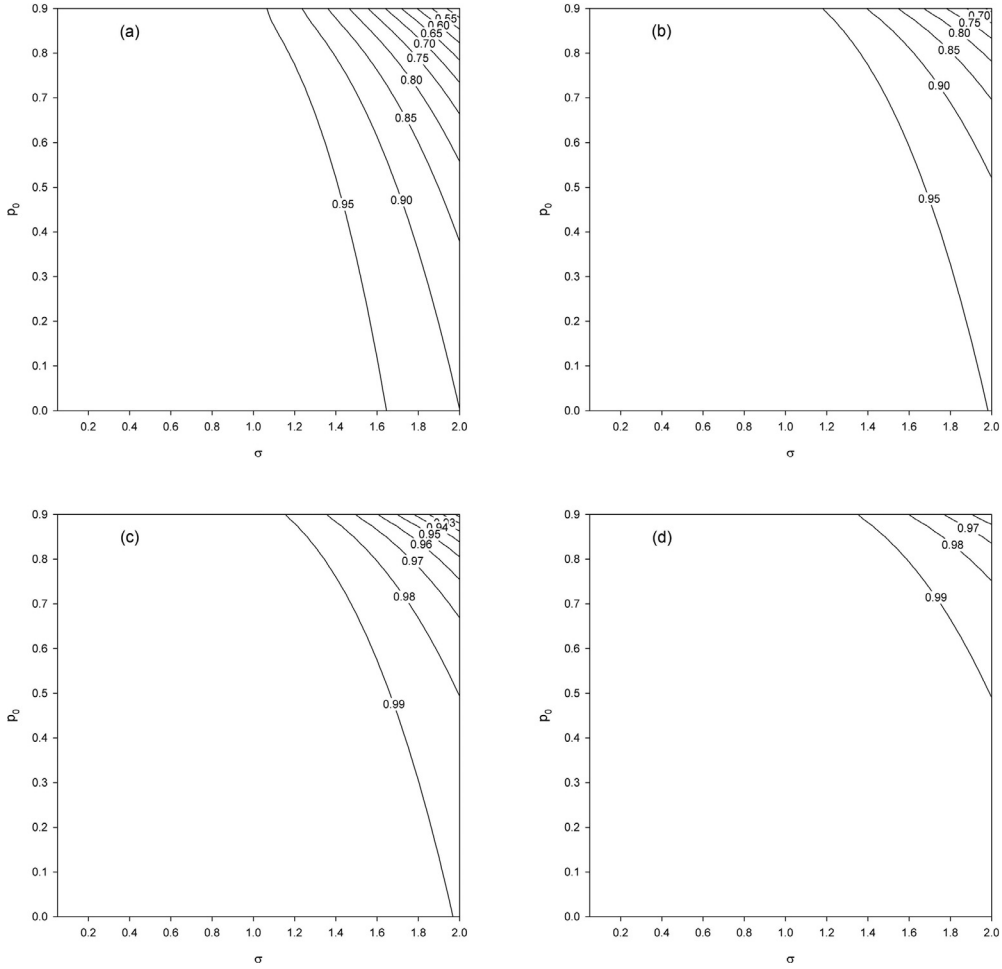

Fig. 7. Ratio $V_{app}\left(\hat{\kappa}_1\right)/V(\hat{\kappa}_1)$ as a function of p_0 and σ for different sample sizes (n). (a) $n = 50$. (b) $n = 100$. (c) $n = 500$. (d) $n = 1\,000$.

Declaration of Competing Interest

The author declare that he has no known competing for financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I thank American Journal Expert (AJE) and my colleague Matthieu Guillemain for checking the English of this article.

References

[1] J. Aitchison, J. Brown, The Lognormal Distribution (Reprinted from 1957 Edition), Cambridge University Press, Cambridge, UK, 1969.
[2] K. Shimizu, Point estimation, in: E.L. Crow, K. Shimizu (Eds.), Lognormal distributions: theory and applications, Marcel Dekker, New York, USA, 1988, pp. 27–86.

[3] B. Neelon, A. O'Malley, Two-part models for zero-modified count and semicontinuous data, in: A. Levy, S. Goring, C. Gatsonis, B. Sobolev, E. van Ginneken, R. Busse (Eds.), Health services evaluation, Springer, New York, USA, 2019, pp. 695–716.
[4] S. Smith, Evaluating the efficiency of the Δ-distribution mean estimator, Biometrics 44 (1988) 485–493.
[5] D. Finney, On the distribution of a variate whose logarithm is normally distributed, Supp. J. Roy. Stat. Soc. 7 (1941) 155–161.
[6] D. Ebbeler, A note on estimation in log-normal linear models, J. Stat. Comput. Simul. 2 (1973) 225–231.
[7] X.-H. Zhou, S. Gao, Estimation of the log-normal mean, Stat. Med. 17 (1998) 2251–2264.
[8] G. Taraldsen, A precise estimator for the log-normal mean, Stat. Methodol. 2 (2005) 111–120.
[9] A. Cuyt, V. Petersen, B. Verdonk, H. Waadeland, W. Jones, Handbook of Continued Fractions for Specials Functions, Springer, Dordrecht, The Netherlands, 2008.
[10] K. Shimizu, K. Iwase, Uniformly minimum variance unbiased estimation in lognormal and related distributions, Commun. Stat. - Theory Methods 10 (1981) 1127–1147.
[11] M. Wolfson, H. Wright, Algorithm 160. Combinatorial of M things taken N at a time, Commun. ACM 6 (1963) 161.
[12] M. Pennington, Efficient estimators of abundance, for fish and plankton surveys, Biometrics 39 (1983) 281–286.
[13] H. Bolfarine, S. Zacks, Prediction Theory for Finite Populations, Springer, New York, USA, 1992.
[14] S. Smith, Use of statistical models for the estimation of abundance from groundfish trawl survey data, Can. J. Fish. Aquat. Sci. 47 (1990) 894–903.
[15] R. Forrey, Computing the hypergeometric function, J. Comput. Phys. 137 (1997) 79–100.
[16] J. Pearson, Computation of hypergeometric functions, Master's thesis, University of Oxford, UK, 2009.
[17] F. Johansson, Computing hypergeometric functions rigorously, ACM Trans. Math. Softw. 45 (2019) 1–26.
[18] G. Navas-Palencia, High-precision computation of uniform asymptotic expansions for special functions, Polytechnic University of Catalonia, Barcelona, Spain, 2019 Ph.D. thesis.
[19] P. Aubry, C. Francesiaz, On comparing design-based estimation versus model-based prediction to assess the abundance of biological populations, Ecol. Indic. (2022). (submitted)
[20] A. Prudnikov, Y. Brychkov, O. Marichev, Integrals and Series. Volume 3. More Special Functions, Gordon and Breach Science Publishers, New York, USA, 1990.
[21] R. Myers, P. Pepin, The robustness of lognormal-based estimators of abundance, Biometrics 46 (1990) 1185–1192.
[22] R. Myers, P. Pepin, Rejoinder to the letter to the editors from M. Pennington, "On testing the robustness of lognormal-based estimators", Biometrics 47 (1991) 1623–1624.
[23] S. Syrjala, Critique on the use of the delta distribution for the analysis of trawl survey data, ICES J. Mar. Sci. 57 (2000) 831–842.
[24] M. Christman, Review of Estimation Methods for Parameters of the Delta-Lognormal Distribution, Technical Report, MCC Statistical Consulting LLC, Gainesville, Gainesville, FL, USA, 2019.