Contents lists available at ScienceDirect

Journal of Structural Geology

journal homepage: http://ees.elsevier.com

Which average, how many grains, and how to estimate robust confidence intervals in unimodal grain size populations

Marco A. Lopez-Sanchez

Géosciences Montpellier, Université de Montpellier & CNRS, Place E. Bataillon, Montpellier, 34095, cedex 5, France

ARTICLE INFO

Keywords
Grain size
Averages
Lognormal distribution
Confidence intervals

ABSTRACT

This paper provides guidelines for determining average grain sizes with an optimal level of statistical adequacy and determine the best methods for estimating confidence intervals in grain size studies. I used Monte Carlo simulations and relative efficiency curves to test which central tendency measure (average) and which method of confidence interval estimation perform best depending on the lognormal shape and sample size. Results from synthetic populations were then compared and validated using actual grain size populations and random sampling methods. Overall, the geometric mean and the confidence intervals for the geometric mean provide the best balance between efficiency and robustness. The median is the best alternative when data contamination is an issue. Confidence intervals for the arithmetic mean in actual grain size populations are problematic regardless of the method used, especially for small sample sizes ($n < 100$). Lastly, we warn against the use of the area-weighted mean or the root mean square as both averages perform poorly.

1. Introduction

Grain size unimodal populations usually result from a single process and can be fully described using a measure of central tendency (i.e. an average) and a measure of dispersion (the standard deviation or others depending on the average used and/or the model assumed). Although both are important parameters for microstructural characterization (e.g. Ter Heege et al., 2004; Herwegh et al., 2014; Soleymani et al. 2020), we commonly use average grain size or apparent grain size values in constitutive equations (e.g. paleopiezometry and flow laws) for simplicity. In this situation, the following questions arise: which average to use? How to estimate the minimum error associated with the average estimate? How many grains are needed to achieve a required level of statistical adequacy? or what is the required level of statistical adequacy? The answer to these questions lies in two central statistical concepts, the margin of error and the degree of certainty. Both ultimately allow a confidence interval to be established. For example, a typical target might be to estimate the minimum sample size needed so that the sample mean has a margin of error of ±5% with 95% certainty.

In geosciences and more specifically in the structural geology community there are no standardized guidelines for determining average grain sizes. Previous studies have covered specific topics such as (1) the use of the median over the arithmetic mean and the estimation of error margins in lognormal populations (Ranalli, 1984), (2) the use of area- or volume-weighted means (Berger et al., 2011), or (3) how different sources of bias affect different common averages (Lopez-Sanchez and Llana-Fúnez, 2015). Only the first paper dealt with the question of how to estimate error margins. Unfortunately, Ranalli (1984) refers to the median and the geometric mean as the same central tendency estimator leading to confusion. Indeed, he advocates the use of the median over the arithmetic mean, but provides formulas to estimate the error margin for the geometric mean. Both measures of central tendency have the same value in a perfect lognormal distribution, hence the confusion, but this never occurs in real grain size populations (see section 2), which can result in incorrect confidence intervals. Besides, the formulas provided for estimating the error margins in Ranalli (1984) did not take into account the degree of certainty.

None of the papers above address the minimum sample size needed. Some rules of thumb based on experience are provided in metallurgical literature, such as a minimum of 200 grain sections (e.g. Humphreys, 2001). The American Society for Testing and Materials (ASTM) advises measuring a minimum of 500 grain sections in fully recrystallized polycrystalline materials in the standard E2627 (ASTM International, 2013a). Similar empirical-based values have been provided in geosciences literature (e.g. Higgins, 2006). In a recent discussion in the Journal of Structural Geology, Stipp (2018) pointed out the limitations of minimum sample sizes based on experience and proposed using the minimum sample of 433 grains provided in Lopez-Sanchez and Llana-Fúnez (2016) as an alternative. However, the minimum sample size provided in Lopez-Sanchez and Llana-Fúnez (2016), based on random sampling methods, strictly applies to the sample from

E-mail address: marco-antonio.lopez-sanchez@umontpellier.fr (M.A. Lopez-Sanchez)

https://doi.org/10.1016/j.jsg.2020.104042
Received 19 December 2019; Received in revised form 5 March 2020; Accepted 11 March 2020
Available online xxx
0191-8141/© 2020.

which this minimum value was estimated and specifically for the arithmetic mean. This value is not valid for populations with different distribution shapes or average even when the error margin and the confidence interval considered remains the same.

The key to the problem portrayed above is that there is no such thing like a universal minimum sample size. Error margins depend on features such as the shape and the spread of the population, the average considered, or the certainty. The same applies for the minimum sample size needed as this is the opposite problem of estimating a confidence interval and it depends on the same factors. Empirical-based minimum sample sizes are useful as a preliminary reference value but will fall large or short in most cases, sometimes causing severe failure. A superior approach is provided in the ASTM norm E112 (Standard Test Methods for Determining Average Grain Size, ASTM International, 2013b), where the focus is on estimating a sample mean with an error margin below a certain limit (10%) at a fixed certainty (95%) to be considered acceptable. This approach specifies nothing about the minimum sample size needed, so if the error margin obtained is above the acceptable limit the sample size must be increased. Although this approach is more robust than considering any attempt of a universal minimum sample size, the method depends on estimating reliable confidence intervals. For this, the E112 norm uses the standard error margin formula based on the entral limit theorem (e.g. Davis, 2002, see also section 2.3). This approach, however, assumes that the underlying distribution is normally distributed and this elicits the question as to whether this method produces reliable error margins in lognormal or asymmetric unimodal populations, typically of grain size populations.

This paper aims to examine all the questions posed above and provide guidelines for determining average grain sizes in unimodal lognormal-like populations based on robust statistical principles. The study focuses on lognormal distributions because recrystallization processes usually result in lognormal grain size distributions (see Lopez-Sanchez and Llana-Fúnez, 2016 and references there in). With robust statistics, we mean that (1) small deviations from the model assumptions should slightly affect the performance of the average or the confidence interval and (2) larger deviations from the model should not cause total failure. Although the paper is focused on grain size measures performed on thin sections, i.e. apparent grain size distributions, the same statistical principles apply for three-dimensional measures.

2. Scope, methods, and datasets

The results of this study are based entirely on Monte Carlo simulations. The procedure consists of generating a large number of grain size populations, either synthetic or from actual samples via random sampling (bootstrapping) methods, and assesses how the sample size and the shape of the distribution affect the different averages or confidence interval methods.

2.1. Lognormal distributions and datasets

Lognormal distributions have been described in detail elsewhere (Limpert et al., 2001). Briefly, a lognormal distribution has three features: (1) a random variable x is lognormally distributed if $y = log(x)$ is normally distributed, (2) only positive values are possible, and (3) distribution skews to the *right* on a linear scale. The probability density function of a lognormal distribution can be fully described using two parameters, either the mean (μ_y) and the standard deviation (σ_y) of the log-transformed values or, alternatively, in linear scale with the geometric mean (μ_g) and the multiplicative or geometric standard deviation (σ_g), also known as the scale and shape parameters respectively (Fig. 1) (e.g. Limpert et al., 2001; Lopez-Sanchez and Llana-Fúnez, 2016).

The study of the performance of different averages and confidence interval methods via Monte Carlo simulations requires limiting the number of potential lognormal populations. For this, we use as reference the multiplicative standard deviation as this is a measure of the shape of the lognormal population independent of their scale (Fig. 1b). In nature, multiplicative standard deviation (MSD) values mostly fall within the range of 1.4–3.0 (Limpert et al., 2001). Establishing a reliable range of MSD values in grain size studies is yet to be done, although values within the ~1.3 and ~2.7 range seems to be common (e.g. Lopez-Sanchez and Llana-Fúnez, 2016). Lognormal synthetic populations were generated using the Numpy v.1.15 random Python library (Van Der Walt et al., 2011; see codes in Supplementary material) with a fixed geometric mean $(\mu_g = 3)$ and variable shapes (σ_g) and sample sizes. Specifically, we use a range of MSDs from 1.2 to 2.5 and sample sizes from 3 up to 10000.

2.1.1. Actual grain size populations

To augment the validity of the results using synthetic populations, we used three datasets that come from actual samples via random sampling methods (bootstrapping). The datasets considered have different features (acquisition methods, distribution shapes and scales, source of bias) and stand out for their extremely large sample sizes (Table 1, Fig. 2). Hence, they are considered representative samples suitable for resampling methods.

Dataset MAL05 comes from natural deformed granite (Lopez-Sanchez and Llana-Fúnez, 2015, 2016, 2018). The quartz-rich domains (35% in volume) appear almost fully recrystallized due to dynamic recrystallization possibly modified by late static recrystalliza-

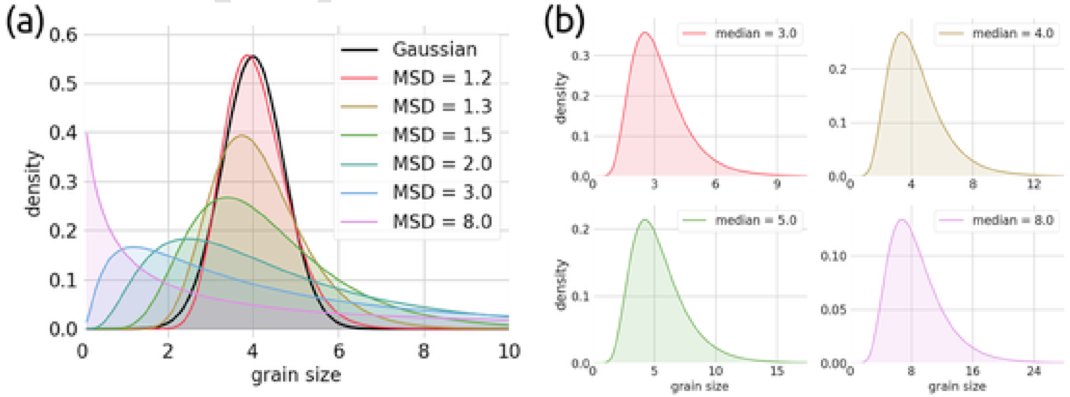

Fig. 1. Probability density functions (PDF) of different lognormal distributions. (a) PDFs with different multiplicative or geometric SD (MSD) and the same geometric mean and median (4.0). The MSD parameter controls the shape and thus the asymmetry of the lognormal distribution. Note that lognormal populations with MSD values around 1.2 and below are difficult to distinguish from normal distributions. (b) Pdfs with different geometric mean and median values (4.0) but the same MSD (1.5). A change in the geometric mean, also known as the scale of the distribution, affects the scaling in horizontal -the grain size range- and vertical directions, but the shape of the PDF remains the same.

Table 1
Pilot sample statistical features.

sample	size (n)	arith. mean	geo. mean	median	RMS	weighted	mode (KDE)	MSD
MAL05	12298	35.79	31.45	32.82	39.99	53.26	27.15	1.69
W1051	9051	4.69	4.38	4.20	5.11	7.69	3.38	1.42
W1051 [a]	9044	4.67	4.38	4.20	5.01	6.35	3.35	1.41
UO2	4264	9.39	8.29	8.86	10.41	13.48	8.21	1.68

[a] Sample W1051 with outliers removed.

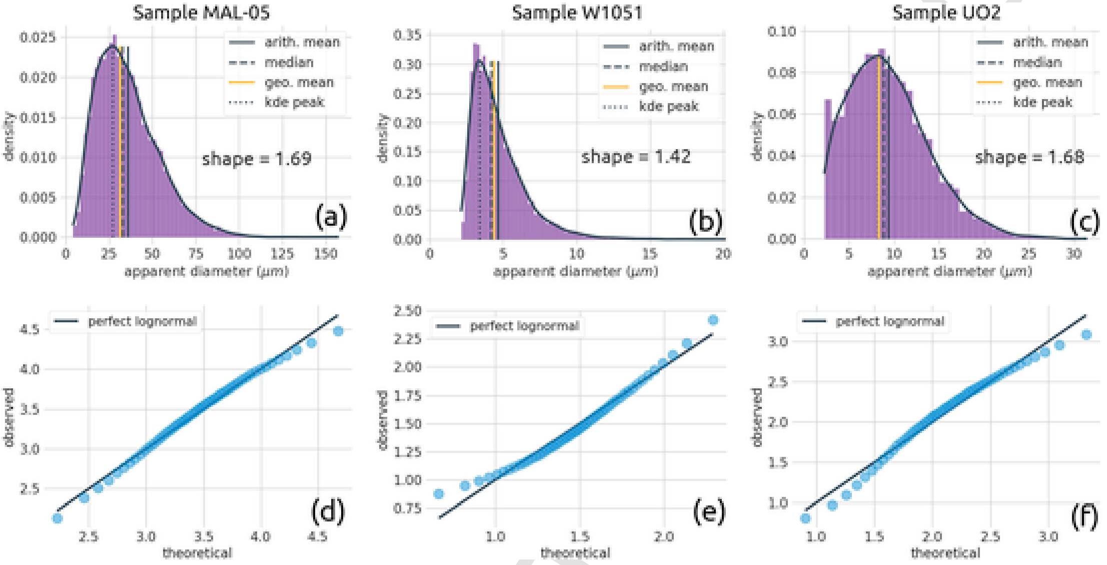

Fig. 2. Grain size population features of the pilot samples used for bootstrapping. (a, b, c) Apparent grain size populations of samples MAL-05, W1051, and UO2, respectively. The position of main averages and the MSD (shape) is indicated. (d, e, f) Quantile-quantile plots comparing the actual distributions against a theoretical lognormal distribution.

tion. The dataset consists of four different grain maps located in the recrystallized quartz areas. Grain maps were made via manual grain segmentation from light optical images. The pixel sizes of the optical images were within the range 0.4–0.5 μm, 1/80 times the mean grain size. For further details, see Lopez-Sanchez and Llana-Fúnez (2016).

Dataset W1051 comes from an experimentally deformed quartzite (Cross et al., 2017). Grain map was made from electron backscatter diffraction (EBSD) data with a size of ∼2000 × 560 μm (step size of 0.95 μm). Grains were segmented using a Voronoi decomposition algorithm with the threshold misorientation set to 10°. The apparent grain size distribution is bimodal due to the mixture of original non-crystallized and recrystallized grains. Recrystallized grains were separated on the basis of a grain orientation spread (GOS) threshold, for details in the protocol see Cross et al. (2017). Interestingly, the grain segmentation procedure as put in Cross et al. (2017) introduces a few non-recrystallized grains in the final dataset with much larger sizes than those produced during recrystallization (see Supplementary material). To check the effect of these grains on the estimation of averages and confidence intervals, we generated another dataset with these grains removed based on a section area threshold. Another notable feature is that the distribution of apparent grain sizes decays somewhat abruptly on the left side (Fig. 2b), locally departing from a lognormal distribution (Fig. 2e). This is probably an artefact due to the presence of grains below the resolution (i.e., step size) of the EBSD map, and the imposed removal of grains containing fewer than five pixels.

Dataset UO2 comes from an annealed aggregate of uranium dioxide with a foam-like microstructure (Depriester and Kubler, 2019). Grain map was made from electron backscatter diffraction (EBSD) data with a size of 1200 × 320 μm (step size of 2 μm). Grains were segmented using a Voronoi decomposition algorithm with a threshold misorientation of 5°. The shape of the distribution is similar to that of the dataset MAL05 but the scale, the mean grain size (4.38 μm in UO2, versus 35.79 μm in MAL05), and the acquisition and grain segmentation methods differ markedly. The distribution of apparent grain sizes ends abruptly towards the left side, showing an abnormal local density peak (Fig. 2c). The former is probably an artefact due to the relative similarity between the EBSD step size (2 μm) and mean grain size (9.39 μm) and the latter might due to not removing grain sections below a minimum value. Overall, this produces a local deviation from a theoretical lognormal population in both extremes (Fig. 2f).

Subsampling was performed using the Numpy v.1.15 random Python library (Van Der Walt et al., 2011). More specifically using the random choice generator with the replace option enabled, i.e. bootstrapping (see codes in Supplementary material). A potential limitation of this approach is that for very small sample sizes the method might generate subsamples that depart from real ones as the distribution of grain sizes in recrystallized rocks seems to occur with some degree of clustering (Kidder et al., 2015).

2.2. Central tendency measures (averages) considered

The arithmetic mean and the median are the most widely used averages in geosciences and thus main targets. Two other means are of common use in the structural geology, the area-weighted mean (e.g. Berger et al., 2011) and the root mean square (RMS) (e.g. Stipp and Tullis, 2003; Cross et al., 2017). The geometric mean, although not as used, is a natural choice for lognormal distributions and was also considered. Indeed, although it is defined as the nth root of the product of *n* numbers:

$$\mu_g = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n} \tag{1}$$

It is usually estimated by computing the arithmetic mean of the log-transformed values and taking the antilogarithm and, in theory, it represents a better measure of central tendency in lognormal distributions than the arithmetic mean (i.e. in perfect lognormal distributions the geometric mean lies between the higher and lower half of a data sample). For the sake of simplicity, we discard other means (e.g. the harmonic mean) or estimators that require complex operations such as the Method of Moments, Serfling or Finney estimators (Ginos, 2009).

The mode or frequency peak is also of common use in structural geology. However, the way it is calculated varies across studies. It can be estimated using distribution-free methods such as the histogram or the kernel density estimator (KDE). Lopez-Sanchez and Llana-Fúnez (2015) proved that KDE-based modes are superior to those based on the histogram and thus only the KDE-based method is considered here. Alternatively, there are parametric approaches. For example, if we assume that the population follows a lognormal distribution, the mode can be estimated using the mean and the standard deviation of the log-transformed data as follows:

$$mode = exp\left(\mu_y - \sigma_y^2\right) \tag{2}$$

Based on (2), the error margins of this mode will depend on the error margins of μ_y and σ_y and hence it would be better to directly consider the back-transformed value of μ_y, which is the geometric mean. Another drawback is that the mode estimated in such way inherits all the problems related to the estimation of the means and standard deviations such as the low resistance to outliers (e.g. Lopez-Sanchez and Llana-Fúnez, 2015). Lastly, parametric-based approaches depend on the underlying model and this could result in severe bias when the actual grain size distribution departs from the model. Thus, we discarded this estimator.

Lastly, we avoid any stereological correction. First, stereological methods are built on ill-posed geometric assumptions (e.g. grains are non-touching spheres) that do not hold in polycrystalline rocks. This means that one obtains, at best, a fair approximation of the actual 3D distribution. The use of universal correction factors between apparent and actual grain size averages (e.g. Fullman, 1953) lacks a mathematical basis for random polydisperse grain size distributions and are discouraged (Heilbronner and Barret, 2014). Saltykov-type methods lack a formulation for estimating errors during the unfolding procedure, and the precision of any average based on the unfolded 3D distribution will be much poorer and prone to suffer from severe bias than those directly obtained from the apparent grain size distributions.

2.3. Margin of error and confidence intervals

The methods available for estimating error margins and confidence intervals depend on the distribution features and the average considered. Setting the confidence level is the responsibility of the researcher. The most common value is 95%, meaning that the chance of obtaining a result outside the expected margin of error is one in 20. This value, however, is neither better nor worse than any other common values such as 90 or 99%, and ultimately this will depend on the degree of certainty required by the study. It should be noted however that small increments in the degree of certainty above 95% can result in large variations in the length of the confidence interval or the minimum sample size needed (e.g. Lopez-Sanchez and Llana-Fúnez, 2015). Accordingly, we consider here a certainty value of 95%.

The usual approach to estimate a confidence interval based on the central limit theorem uses the standard error formula (e.g. Davis, 2002) that when rearranged becomes:

$$\mu \pm z \frac{\sigma}{\sqrt{n}} \tag{3}$$

where μ is the arithmetic mean, σ is the standard deviation of the sample, n is the sample size, and z is the so-called *z-score*, which is a scalar value that multiplies the standard deviation to meet the desired certainty. Note that error margins depend on the square root of the sample size. Table 2 presents commonly used *z-scores* for confidence intervals at 90, 95, and 99%. Unfortunately, for small sample sizes, neither the standard deviation nor the error margin can be reliably estimated. In this situation, the use of the *t-score* is common practice since it provides safer error intervals taking into account the degrees of freedom (Table 2):

$$\mu \pm t \frac{\sigma}{\sqrt{n}} \tag{4}$$

Equation (4) is the recommended procedure to estimate an error margin when calculating a mean grain size by the American Society for Testing and Materials in the norm E112 (ASTM International, 2013b). As previously mentioned, this approach has the limitation of assuming that the grain size populations follow a normal distribution, which is rarely the case. The use of (4) might lead to estimates of confidence intervals with wrong coverages or underestimating the minimum sample size in lognormal-like distributions and needs to be tested.

2.3.1. Margin of error estimation in lognormal distributions

Since by definition the log-transformed values of a lognormal population follow a normal distribution, the same approach can be applied upon the log-transformed data and then estimate the back-transformed values. Accordingly, the formulae for estimating the error margins in lognormal distributions become based on (4):

$$\mu_g = exp\left(\mu_y\right) \tag{5}$$

$$UpperCI = exp\left(\mu_y + t\frac{\sigma_y}{\sqrt{n}}\right) \tag{6}$$

$$LowerCI = exp\left(\mu_y - t\frac{\sigma_y}{\sqrt{n}}\right) \tag{7}$$

Several things need to be noted. First, these confidence intervals are valid for the geometric $\left(\mu_g\right)$, not the arithmetic mean, which otherwise is (Aitchison and Brown, 1957):

$$\mu = exp\left(\mu_y + 0.5\sigma_y^2\right) \tag{8}$$

Table 2
Typical z- and t-scores for confidence intervals of 90, 95, and 99%.

confidence interval	z-score [a]	t-score (3) [b]	t-score (10)	t-score (30)	t-score (100)
90%	1.645	2.353	1.812	1.697	1.66
95%	1.96	3.182	2.228	2.042	1.965
99%	2.576	5.841	3.169	2.75	2.586

[a] Two and three are also common z-score values in literature, usually referred to as 2-sigma and 3-sigma errors. These values correspond with confidence intervals of 95.5 and 99.7%, respectively.

[b] Values in parenthesis are degrees of freedom or DOF (n - 1). The higher the DOF, the similar the z- and t-score values.

Second, the error margins are asymmetric around the geometric mean and therefore the formulae split up into upper and lower confidence intervals. This approach is similar to that proposed in Ranalli (1984) but using a factor (the *t-score*) to set the certainty of the error margin. Third, these equations can also be used to estimate the minimum sample size required, provided that the standard deviation of the log-transformed population is known (e.g. Hale, 1972; Hewett, 1995). Estimating a confidence interval for the arithmetic mean in a lognormal population is however not so straightforward (see a summary of methods and references in Appendix A) but important to consider as it is still the most widely used average.

Lastly, distribution-free estimators such as the median or the mode do not allow to estimate standard errors and confidence intervals based on entrenched statistical proof. Instead, there are rules of thumb to estimate approximate confidence intervals for the median (see Appendix A) or random sampling methods such as bootstrapping.

2.4. Monte Carlo simulations and relative efficiency curves

Monte Carlo simulations will serve for three different purposes:

1. Generate tables with reference values for error margins and minimum sample size depending on distribution features.
2. Estimate the efficiency and robustness of the different central tendency estimators via relative efficiency curves.
3. Compare the different methods to estimate error margins in simulated and real grain size populations using bootstrapping.

Monte Carlo simulations applied to synthetic samples uses 10000 trials for each defined distribution, with the exception of the KDE-based mode which use 5000 trials due to its larger computational and time cost. Samples sizes increase from 1 to 100 in unit increments, from 100 to 2500 in increments of three, and from 2500 to 10000 squarely. The Monte Carlo simulations using bootstrap generate 10000 random subsamples for each sample size considered (2000 in the case of the KDE-based mode). Sample sizes increase from 3 to 100 in unit increments and from 100 to 2002 in increments of three. Accordingly, the minimum sample sizes reported in the tables are exact for $n \leq 100$ and have a precision of ±1 for the $100 < n \leq 2500$ interval.

The performance of the different averages is presented using absolute and relative efficiency curves. This is, plotting sample size vs spread of the averages and estimating their relative efficiency (RE) with respect to the arithmetic mean as follows:

$$RE(\theta,\ \phi) = \frac{SD(\ \phi)}{SD\ (\theta)} \tag{9}$$

Where θ refers to the arithmetic mean, ϕ to any other central tendency estimator and SD to the standard deviation of the estimator. RE values less than 1 indicate that a given central tendency estimator is performing better than the arithmetic mean, and *vice versa*.

2.4.1. Confidence interval methods

Based on previous literature, we test the reliability and performance of six different methods for estimating confidence intervals. Specifically, for the arithmetic mean we consider the one described in the ASTM standard E112 (ASTM International, 2013b), the modified Cox (Armstrong, 1992; Zou et al., 2009), and the generalized confidence interval (GCI) (Krishnamoorthy and Mathew, 2003) methods. For the geometric mean, we consider the method based on the central limit theorem (CLT method), equations (4)–(6), and a Bayesian approach (Oliphant, 2006). For the median, we test the method proposed in Hollander and Wolfe (1999). For evaluating their adequacy, we use three criteria over the Monte Carlo simulation output:

- The coverage, which estimates how many times the known average value is within the confidence interval. The confidence interval is set to a certainty of 95% for all methods. Accordingly, using 50000 Monte Carlo trials in the simulation means that suitable coverages must lie within the 94.8 and 95.2% interval (e.g. Zou et al., 2009).
- The average length of the confidence interval.
- The balance between the lower and upper tail errors, i.e. how many times the confidence interval lies below or above the actual average.

Among the three, coverage is the most critical parameter and thus the first-order criterion. Coverages below 94.8% will be considered as inadequate, while those above 95.2% suitable but an indication of the method being too conservative. Provided that the coverage obtained is safe, small average length intervals and a good balance between lower and upper tail errors will be considered better qualities.

3. Results

3.1. Average performance depending on distribution asymmetry and sample size

Figs. 3 and 4 shows that the performance (i.e., relative efficiency) of the different average estimators depends largely on the shape of the distribution. Overall, the coefficient of variation and the differences between the performance of the different averages increase with distribution asymmetry for all estimators. The exception to this trend is the KDE-based mode, which starts to decrease their coefficient of variation somewhere in the range of MSD values between 1.8 and 2.0.

3.1.1. Performance of the different means

For large sample sizes and taking the arithmetic mean as a reference, the geometric mean performs the better, being the difference in performance insignificant for MSD values less than 1.2 (i.e. almost a normal distribution) and improving monotonically with increasing asymmetry (Fig. 3c). The RMS mean performs fairly similar to the arithmetic mean for low MSD values, but its performance decreases dramatically as the asymmetry increases. For example, for an MSD value of two, its performance compared to the arithmetic mean is almost two times worse (Fig. 3c).

For small samples sizes ($n < 100$), the trend for the means remain the same (Fig. 4). For MSD values above 1.6, the RMS mean yields the worst efficiency of all those averages. Indeed, in the range between 1.8 and 2.5, its performance is ~1.5–~2.5 times worse than the arithmetic mean. Overall, the geometric mean performs best, although for MSD values less than 1.4 its performance with respect to the arithmetic mean is fairly similar. In highly asymmetric distributions (≥2.5), the performance of the geometric mean relative to the arithmetic mean is almost twice as good (efficiency ~0.5).

3.1.2. Median and mode (frequency peak) performance

Similar to means, the performance of the median decrease with asymmetry. However, it increases at a slow pace relative to the arithmetic mean, resulting in better performance for MSD values above ~1.7 whatever the sample size (Figs. 3 and 4). The median also outperforms the RMS mean at MSD values above ~1.4. In contrast, it performs always worse than the geometric mean.

The behaviour of the KDE-based mode is complex and overall it performs poorly compared to means and the median (Figs. 3 and 4). As pointed out elsewhere (Lopez-Sanchez and Llana-Fúnez, 2015), the only advantage of KDE-based mode is its robustness against the resolution limit effect (either physical or imposed due to a protocol) as it is the only central tendency measure not affected by this problem.

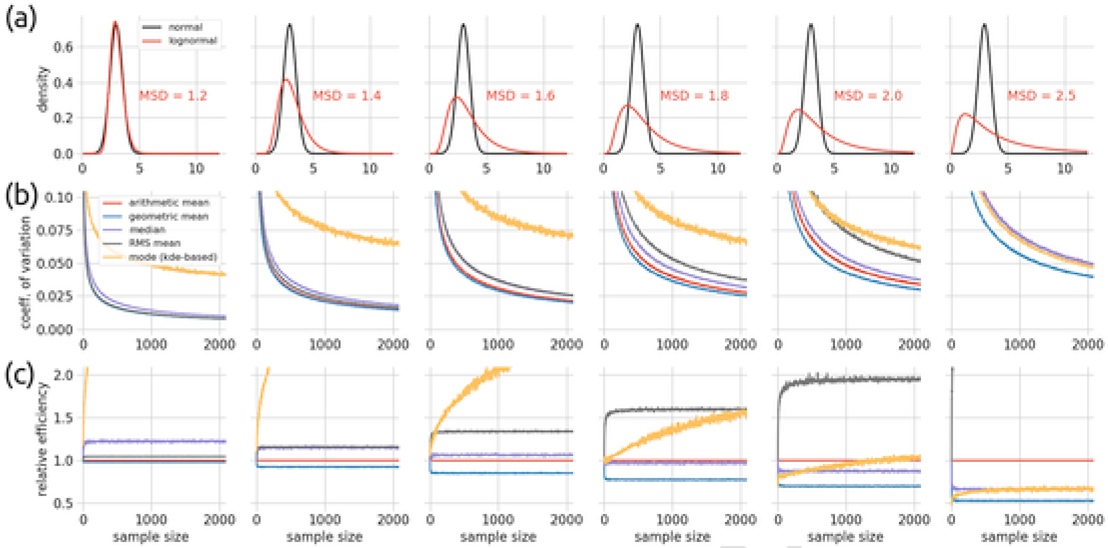

Fig. 3. Performance of the different average estimators relating to the shape of the lognormal distribution (MSD value) and the sample size (up to n = 2000). Top row, lognormal shapes considered. Middle row, variation in the estimation of the different averages with the sample size using the coefficient of variation. Bottom row, the efficiency of the different averages relative to the arithmetic mean. Values below one indicate better efficiency than the arithmetic mean and vice-versa. See full sample size ranges (up to 10000) and between 200 and 500 in the supplementary material.

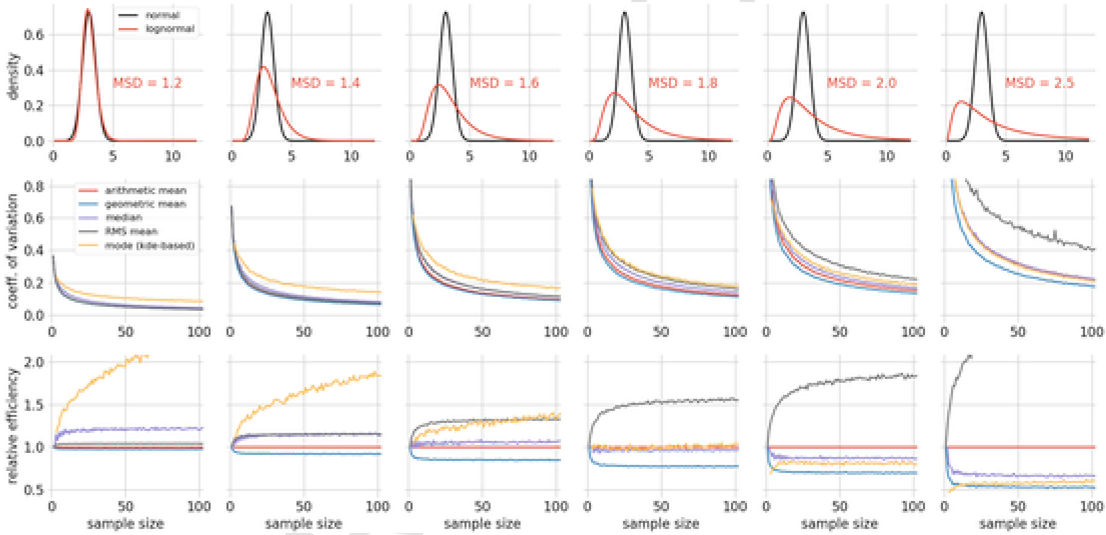

Fig. 4. Same as Fig. 3 but for the sample size interval 2 to 100.

3.1.3. How many grains are needed for an error of 5% (at 95% certainty)

Fig. 5 and Table 3 show that except for the mode, the minimum sample size needed increases monotonically with the asymmetry of the distribution. Regarding means and medians, the increase is more notable for the arithmetic mean and particularly for the RMS which grow exponentially. The mode behaves in a complex way, requiring very large sample sizes for distributions with low to moderate asymmetries and close to other averages for large asymmetries. If 10% error is used, the KDE-based mode behaves more in line with the other estimators. This suggests that a confidence interval of 5% (c.i. 95%) for the KDE-based mode is too strict or close to its maximum asymptotic performance for some cases.

3.2. Average performances on real (pilot) datasets

The behaviour of the different averages in samples MAL05 and UO2 follow similar trends to those observed for the synthetic sam-

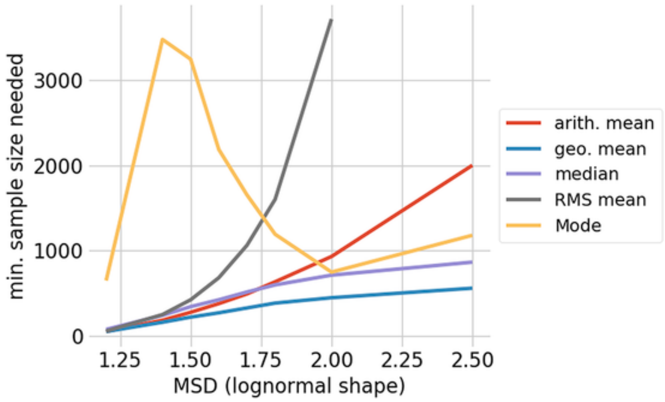

Fig. 5. The minimum sample size needed upon lognormal shape normalized to an error margin of 5% of the arithmetic mean at 95% certainty. See values in Supplementary material.

Table 3
The minimum sample size needed based on lognormal shape for a normalized 5 and 10% error margins respect to the arithmetic mean.

estimator	1.2	1.4	1.5	1.6	1.7	1.8	2	2.5
arith. mean	51	184	277	379	493	635	932	2003
geo. mean	51	175	253	337	427	524	737	1277
median	80	268	400	536	689	836	1145	2003
RMS mean	54	223	358	536	794	1139	2180	> 10000
KDE mode	815	6561	8836	9409	8836	6889	4225	5184
arith. mean [a]	14	46	69	92	127	160	238	509
geo. mean [a]	13	44	64	86	109	136	184	322
median [a]	20	68	98	136	175	205	289	503
RMS mean [a]	14	54	88	139	199	277	545	2309
KDE mode [a]	60	337	469	515	524	499	503	911

[a] Coefficient of variation ≤10%.

ples (Table 4, Fig. 6). The geometric mean performs best although the efficiency relative to the arithmetic mean is lower than expected. For a lognormal MSD value of ~1.7, common for both samples (Table 4), the theoretical difference in efficiency should fall between 0.77 and 0.85 (Figs. 3 and 4), but it varies between 0.98 and 0.93. These are relative efficiency values typical of lognormal populations with 1.2–1.4 MSD values. The median performs slightly worse than the arithmetic mean on both cases, ~1.19–1.26 times less efficiently, again showing relative efficiencies typical of populations with 1.2–1.4 MSD values. The area-weighted mean performs the worst of all, being of about two times less efficient for n > 30. The relative efficiency of the KDE-based mode decreases monotonically with sample size without stabilizing, being between 3 and 4 times less efficient for sample sizes around 500.

The behaviour of the different averages in sample W1051 bring in distinctive trends compared to both synthetic and the other pilot samples (Fig. 6). The geometric mean performs the best with a notable relative efficiency better than expected (0.76) for a population with an MSD value of ~1.4 (cf. Figs. 3, 4 and 6). It is also noteworthy that the median outperforms the arithmetic mean (0.92) needing a small sample size to reach the same performance (Table 4). This trend in the median disappears when cleaning the extra-large grains produced during grain segmentation, yielding in this case almost the same performance (Fig. 6, Table 4). Yet, even in this case, the relative performance of the median remains better than expected compared to the performance difference obtained in synthetic populations with similar MSD values. The performance of the area-weighted mean is dramatically worse in this case, up to more than 35 times worse respect to the arithmetic mean. For small sample sizes (< 100) the arithmetic mean, the median and the KDE-based mode perform approximately the same (Fig. 6).

Overall, there are some differences in the behaviour of the different averages between synthetic lognormal and natural samples. The main difference is the introduction of biases and the deviation from the lognormal distribution in the natural datasets. This explains the differences between the expected and the actual relative efficiencies of the different averages despite the similar lognormal shape. In any case, the geometric mean still provides the best balance between performance and reliability. Lastly, the difference in results between regular and the fully-cleaned W1051 datasets indicate that the inclusion of a few extra-large grains severely affects the performance of the area-weighted mean, makes the median to perform better than the arithmetic mean, and increases the relative performance of the geometric mean compare to the arithmetic mean.

Table 4
Minimum sample size needed normalized to the same absolute error margin (5% respect to the arithmetic mean at 95% certainty).

ref.	MSD	abs. Err.	arith. mean	geo. mean	median	weighted	mode (kde)
MAL05	1.69	0.91	382	331	601	1885	> 2000
UO2	1.68	0.24	355	325	496	1237	> 2000
W1051	1.42	0.12	286	169	253	> 2000	445
W1051 [a]	1.41	0.12	241	166	259	> 2000	328

[a] Cleaned population.

3.3. Comparison of methods for estimating confidence intervals

When comparing the different error estimation methods on synthetic lognormal populations the following features emerge (Table 5). For the arithmetic mean, the modified Cox and the GCI method systematically provide adequate coverages within the sample size interval considered (30–1000). Confidence interval lengths are rather similar but the modified Cox sometimes provide small interval lengths in small sample sizes (< 100) while the CGI method provides more balanced tail errors (*cf.* Zou et al., 2009). Regarding the method recommended for the ASTM, for larger samples sizes (~1000), the coverage is adequate irrespective of the lognormal shape within the MSD interval 1.2–2.0, but the tail errors are severely unbalanced to the right compared to the Cox and GCI methods. In populations close to normal ones (e.g. MSD ≤ 1.2), it provides acceptable coverages for the sample sizes above 100 but fails at small samples sizes. This trend worsens with increased asymmetry. For example, for lognormal shapes equal or above 1.7 samples with sizes equal or below 500 yield unacceptable coverages and, therefore, it is not suitable for estimating confidence intervals.

For the geometric mean, the method CLT and the Bayesian-based methods provide similar coverages, confidence intervals lengths, and tail error balances. Coverages are all adequate within the sample size and lognormal shapes intervals considered. The most significant feature is that when populations depart from normal-like ones (MSD > 1.2), the confidence interval lengths obtained for the geometric mean are smaller than those obtained for the arithmetic mean and the median and thus they perform better.

The rule of thumb method proposed for the median provide in all cases coverages higher than expected (≫ 95.2%), indicating that it is too conservative. Indeed, the confidence interval lengths are larger than those obtained for the means and the tail errors are also severely unbalance to the left. Although overly conservative, it is remarkable that the confidence intervals remain robust at any condition tested even for small sample sizes (~30).

3.3.1. Confidence intervals in actual datasets

Confidence intervals for the geometric mean provide acceptable coverages within the sample size range (30–1000) with two excep-

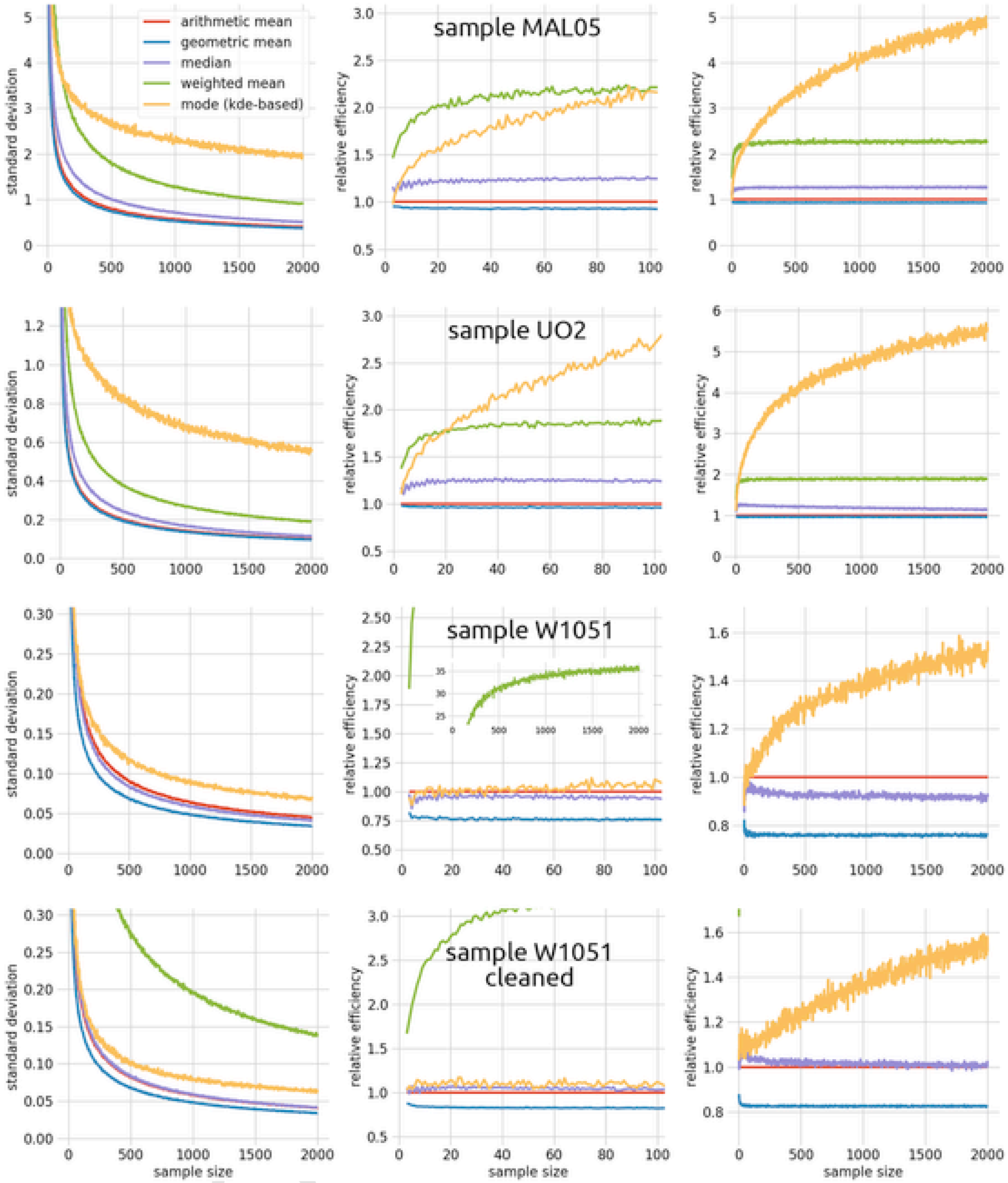

Fig. 6. Performance of the different average estimators with sample size (up to n = 2000) in samples MAL05, UO2, and W1051. Left column, the variation of standard deviation with sample size within the interval 0–2000. Standard deviations are Bessel corrected. Middle column, the efficiency of the different averages relative to the arithmetic mean for the sample size interval 0–100. Values below one indicate better efficiency than the arithmetic mean and vice-versa. The inset in sample W1051 shows the path of the area-weighted mean. Right column, the same for the sample size interval 0–2000.

tions, in samples UO2 and W1051 (cleaned) the coverage was slightly less than expected for $n = 30$ (Table 6). The interval lengths are smaller and better balanced than those estimated for the arithmetic mean and the median, therefore, outperforming them. Another significant note is that the Bayesian-based method worked slightly better than the CLT method for small samples sizes ($n \leq 100$) in sample W1051 and, unlike this, it provides a suitable coverage in sample W1051 (cleaned) for $n = 30$ (Table 6).

The confidence intervals for the arithmetic mean behave very differently from that observed in synthetic samples and vary across samples. The method recommended in the ASTM standard E112 provides adequate coverages for large samples sizes (>250), excepting for sample W1051 (uncleaned). This indicates that the introduction of outliers on the right side of the population (i.e. grains much larger than average) or small sample sizes (≪ 250) can lead to unacceptable confidence intervals when using this method. The GCI and modified Cox methods behave unexpectedly. They provide adequate coverages in MAL05 and UO2 samples within the sample size interval considered (30–500) but fail with larger sample sizes. Additionally, they provide higher than expected coverages and severely unbalanced tail errors. On the other hand, both methods provide inadequate coverages in sample W1051 irrespective of whether it is the regular or the fully cleaned version without outliers. Overall, this indicates that the modified Cox and CGI methods fail when the grain size distribution deviates from the lognormal distribution.

Table 5
Comparison of methods for estimating confidence intervals in synthetic lognormal populations.

shape/sample size	ASTM arith. mean				Modified Cox				GCI method				CLT geo. mean				Bayesian geo. mean				median rule of thumb			
	Cov [a] (%)	length	AL(%) [b]	AR(%) [b]	Cov.	length	AL	AR	Cov.	length	AL	AR	Cov.	length	AL	AR	Cov.	length	AL	AR	Cov.	length	AL	AR
shape = 1.2																								
30	**94.2**	0.41	2.03	3.76	94.8	0.42	2.22	2.99	94.8	0.42	2.68	2.52	94.8	0.40	2.57	2.64	94.8	0.41	2.55	2.62	97.7	0.64	2.11	0.23
100	94.8	0.22	2.23	2.93	95.0	0.22	2.39	2.58	95.0	0.22	2.69	2.36	95.0	0.22	2.60	2.45	95.0	0.22	2.59	2.44	96.2	0.29	2.85	0.95
250	94.9	0.14	2.32	2.78	95.0	0.14	2.44	2.59	94.9	0.14	2.64	2.44	95.0	0.14	2.58	2.39	95.0	0.14	2.58	2.39	96.0	0.18	2.62	1.35
500	95.0	0.10	2.29	2.73	95.0	0.10	2.39	2.61	95.0	0.10	2.50	2.54	94.9	0.10	2.50	2.56	94.9	0.10	2.50	2.56	95.6	0.12	2.65	1.76
1000	94.8	0.07	2.57	2.61	94.8	0.07	2.65	2.52	94.8	0.07	2.73	2.43	94.8	0.07	2.74	2.47	94.8	0.07	2.74	2.47	95.3	0.09	2.69	2.04
shape = 1.5																								
30	**93.6**	0.99	1.32	5.12	94.9	1.02	1.79	3.33	94.9	1.05	2.82	2.30	95.0	0.90	2.51	2.48	95.0	0.91	2.49	2.46	97.6	1.46	2.17	0.24
100	**94.4**	0.54	1.65	3.93	94.8	0.55	2.10	3.08	94.9	0.55	2.68	2.46	94.9	0.48	2.46	2.61	94.9	0.48	2.46	2.61	96.1	0.66	2.86	1.02
250	94.9	0.34	1.89	3.24	95.0	0.34	2.28	2.73	95.0	0.34	2.59	2.39	94.9	0.30	2.59	2.51	94.9	0.30	2.59	2.51	96.2	0.41	2.44	1.36
500	94.8	0.24	2.07	3.11	94.9	0.24	2.34	2.73	94.9	0.24	2.57	2.54	94.9	0.21	2.56	2.58	94.9	0.21	2.56	2.58	95.5	0.28	2.79	1.76
1000	95.0	0.17	2.18	2.87	95.0	0.17	2.37	2.64	95.0	0.17	2.56	2.45	94.9	0.15	2.63	2.44	94.9	0.15	2.63	2.44	95.4	0.19	2.65	1.92
shape = 1.7																								
30	**93.0**	1.40	0.97	6.06	94.8	1.48	1.58	3.60	94.9	1.55	2.78	2.27	94.9	1.19	2.54	2.54	95.0	1.19	2.52	2.53	97.6	1.96	2.18	0.25
100	**94.1**	0.77	1.44	4.45	94.9	0.78	2.03	3.12	94.9	0.79	2.75	2.36	94.8	0.63	2.70	2.50	94.8	0.63	2.70	2.50	96.0	0.86	2.96	1.04
250	**94.7**	0.49	1.74	3.52	95.1	0.49	2.17	2.76	95.1	0.49	2.64	2.31	95.0	0.40	2.56	2.40	95.0	0.40	2.56	2.40	96.3	0.53	2.41	1.33
500	**94.7**	0.34	1.87	3.48	94.9	0.34	2.24	2.84	94.9	0.34	2.57	2.53	95.0	0.28	2.50	2.52	95.0	0.28	2.50	2.52	95.4	0.36	2.80	1.81
1000	94.9	0.24	2.03	3.09	95.0	0.24	2.29	2.73	95.0	0.24	2.53	2.51	95.0	0.20	2.53	2.44	95.0	0.20	2.53	2.44	95.4	0.25	2.61	2.03
shape = 2.0																								
30	**91.7**	2.09	0.65	7.64	94.8	2.25	1.31	3.85	94.9	2.44	2.80	2.25	95.0	1.57	2.51	2.48	95.0	1.57	2.49	2.47	97.5	2.65	2.21	0.28
100	**93.7**	1.16	1.08	5.20	94.9	1.18	1.82	3.28	94.9	1.20	2.73	2.35	94.9	0.83	2.57	2.54	94.9	0.83	2.57	2.54	96.1	1.14	2.88	1.04
250	**94.3**	0.74	1.42	4.24	94.8	0.74	2.07	3.13	94.8	0.74	2.67	2.50	94.8	0.52	2.70	2.52	94.8	0.52	2.70	2.52	96.1	0.70	2.53	1.32
500	**94.5**	0.52	1.67	3.84	94.9	0.52	2.21	2.92	94.8	0.52	2.66	2.55	94.9	0.37	2.51	2.63	94.9	0.37	2.51	2.63	95.3	0.47	2.77	1.91
1000	95.0	0.37	1.75	3.24	95.2	0.37	2.17	2.60	95.2	0.37	2.44	2.36	95.0	0.26	2.47	2.53	95.0	0.26	2.47	2.53	95.3	0.33	2.70	2.02

[a] Cov - coverage. In bold, unsuitable coverages (<94.8).

[b] AL - the average lies below (left) de confidence interval, AR - the average lies above (right) the confidence interval.

Table 6
Comparison of methods for estimating confidence intervals in actual grain size populations.

shape/sample size	ASTM arith. mean				Modified Cox				GCI method				CLT geo. mean				Bayesian geo. mean				median rule of thumb			
	Cov [a](%)	length	AL(%) [b]	AR(%) [b]	Cov.	length	AL	AR	Cov.	length	AL	AR	Cov.	length	AL	AR	Cov.	length	AL	AR	Cov.	length	AL	AR
MAL-05																								
30	**94.0**	12.92	1.67	4.34	96.6	15.21	1.69	1.69	96.1	15.95	2.90	0.97	94.8	12.35	3.18	2.01	94.8	12.36	3.16	2.00	97.6	20.90	2.07	0.28
100	94.8	7.02	1.81	3.42	96.8	8.07	2.06	1.15	96.4	8.17	2.84	0.81	95.0	6.57	2.75	2.23	95.0	6.57	2.75	2.23	96.2	9.46	2.78	1.00
250	94.9	4.43	2.06	3.03	96.4	5.06	2.85	0.76	96.0	5.09	3.45	0.58	95.0	4.13	2.70	2.29	95.0	4.13	2.69	2.29	96.2	5.96	2.42	1.38
500	95.0	3.13	2.16	2.84	95.7	3.57	3.77	0.54	95.2	3.58	4.33	0.48	95.2	2.91	2.62	2.22	95.2	2.91	2.62	2.22	95.5	4.07	2.66	1.83
1000	95.0	2.21	2.29	2.71	**94.1**	2.52	5.63	0.30	**93.6**	2.52	6.17	0.28	95.1	2.06	2.58	2.35	95.1	2.06	2.58	2.35	95.3	2.84	2.75	1.92
UO2																								
30	**94.2**	3.25	1.87	3.95	97.0	3.93	1.79	1.24	96.4	4.12	2.91	0.70	**94.7**	3.21	3.43	1.85	**94.7**	3.22	3.42	1.85	97.6	5.27	2.14	0.29
100	94.8	1.77	1.91	3.25	97.1	2.09	2.04	0.85	96.6	2.11	2.84	0.56	95.0	1.71	2.90	2.06	95.0	1.71	2.90	2.06	96.2	2.43	2.75	1.05
250	94.9	1.11	2.03	3.05	96.5	1.31	2.92	0.58	96.0	1.32	3.56	0.47	94.9	1.07	2.68	2.39	94.9	1.07	2.68	2.39	96.2	1.49	2.46	1.37
500	95.0	0.79	2.14	2.81	95.6	0.92	4.10	0.32	95.0	0.93	4.70	0.26	95.1	0.76	2.61	2.26	95.1	0.76	2.61	2.26	95.5	1.00	2.68	1.81
1000	94.8	0.56	2.45	2.77	**92.9**	0.65	6.92	0.16	**92.3**	0.65	7.52	0.14	94.9	0.54	2.75	2.35	94.9	0.54	2.75	2.35	95.3	0.69	2.79	1.87
W1051																								
30	**92.7**	1.37	0.84	6.46	**91.9**	1.26	1.73	6.36	**92.1**	1.29	2.74	5.13	94.8	1.14	1.74	3.49	94.8	1.15	1.74	3.47	97.6	1.92	2.13	0.25
100	**94.0**	0.76	1.05	4.96	**91.4**	0.67	2.08	6.48	**91.7**	0.67	2.58	5.70	94.9	0.61	1.98	3.09	94.9	0.61	1.98	3.09	96.2	0.84	2.75	1.08
250	**94.2**	0.49	1.28	4.55	**90.6**	0.42	1.96	7.43	**90.9**	0.42	2.28	6.86	94.8	0.38	2.22	3.03	94.8	0.38	2.22	3.03	96.2	0.51	2.49	1.34
500	94.8	0.35	1.34	3.87	**90.1**	0.30	1.59	8.30	**90.4**	0.30	1.75	7.82	95.2	0.27	2.17	2.65	95.2	0.27	2.17	2.65	95.7	0.34	2.69	1.62
1000	**94.7**	0.25	1.58	3.73	**88.4**	0.21	1.26	10.36	**88.7**	0.21	1.33	9.96	95.1	0.19	2.27	2.62	95.1	0.19	2.27	2.62	95.5	0.24	2.57	1.90
W1051 clean																								
30	**93.0**	1.31	0.96	6.00	**91.9**	1.21	1.68	6.41	**92.4**	1.24	2.52	5.10	**94.6**	1.11	1.90	3.55	94.9	1.13	1.72	3.35	97.7	1.91	2.06	0.22
100	**94.4**	0.72	1.40	4.20	**92.2**	0.65	2.06	5.73	**92.4**	0.66	2.55	5.01	94.8	0.60	2.13	3.09	94.9	0.60	2.09	3.03	96.1	0.84	2.85	1.02
250	94.8	0.45	1.64	3.54	**92.0**	0.41	1.93	6.11	**92.1**	0.41	2.20	5.67	95.0	0.38	2.09	2.93	95.0	0.38	2.06	2.91	96.2	0.51	2.46	1.33
500	94.9	0.32	1.91	3.23	**91.6**	0.29	1.79	6.64	**91.8**	0.29	1.97	6.19	95.0	0.27	2.26	2.78	95.0	0.27	2.24	2.77	95.6	0.34	2.64	1.71
1000	94.9	0.23	2.06	3.08	**90.7**	0.21	1.47	7.82	**91.0**	0.21	1.57	7.45	95.1	0.19	2.31	2.59	95.1	0.19	2.30	2.58	95.3	0.24	2.72	2.01

[a] Cov - coverage. In bold, unsuitable coverages (<94.8).

[b] AL - the average lies below (left) de confidence interval, AR - the average lies above (right) the confidence interval.

Regarding the median, the rule of thumb method provides adequate coverage for the entire sample size range considered (Table 6). As in the case of synthetic populations, the method yields higher than expected coverages and longer interval lengths compare to means, but it is the only method that provided reliable confidence intervals for the full range of tested cases.

4. Concluding remarks

The performance of the different averages and confidence interval methods and the minimum sample size needed for a given performance depends on the quality of the data (i.e. the sources of bias and their magnitude), the sample size, and the distribution shape. In general, the efficiency and robustness of the different averages and confidence interval methods trade-off against each other, ultimately preventing a universal answer as to which average or confidence interval method is better. Despite this, some guidelines emerged from the study. Concluding remarks are separated into three subsections. To promote standardized procedures in grain size studies, all these methods and guidelines have been incorporated within the GrainSizeTools script v.3 (Lopez-Sanchez, 2018), a free, open-source, and cross-platform script written in Python.

4.1. Which average to use (and which not)

- Use the geometric mean. It performs better than other central tendency measures in lognormal-like populations regardless of asymmetry and sample size. A note of caution using this average is that the data cannot contain values equal to zero, which should never occur in grain size populations.
- The median is also a wise choice and more backwards-compatible average than the geometric mean due to their common use in the past. The median is more robust to outliers than means and, according to Shih and Binkowitz (1987), it can perform better than the geometric mean if data contamination (outliers or observations from mixed distributions) is above 10%. This is, therefore, the preferred option when data contamination could be an issue.
- The arithmetic mean is the most backwards-compatible average due to their common use in the past. Theoretically, it outperforms the median in low to moderately skewed distributions (MSD < 1.7) as long the presence of extreme values (outliers) remains small. In contrast, the estimation of error margins remains problematic in real samples (see section 4.2 below), and the geometric mean or the median is preferred.
- The behaviour of the KDE-based mode is complex and, overall, it performs worse than other central tendency measures. This average is, however, the only one that remains robust in some specific situations that are likely to occur in grain size studies such as notable different resolution limits and size cut-offs. The estimation of reliable confidence interval requires bootstrapping methods and hence large sample sizes.
- Avoid the use of root mean squared (RMS) or the area-weighted mean, both perform poorly.
- Average grain size will be measured under diverse conditions. Accordingly, experimental calibrations using grain size averages (e.g. piezometers) should provide the four averages considered above for future-proof comparisons, or alternative, provide the entire dataset of grain sizes in the supplementary data.

4.2. Reliable confidence intervals

- The confidence intervals for the geometric mean outperforms the others. They systematically provide acceptable coverages with smaller confidence interval lengths. The Bayesian approach performs slightly better than the CLT method for small datasets (<100) in a few cases and is therefore preferred when n < 100.
- The confidence intervals for the median is overly conservative providing larger confidence interval lengths compare to those estimated for the means. Even so, the method remains very robust even when the population severely departs from lognormal shapes.
- The estimation of confidence intervals for the arithmetic mean in lognormal-like populations is problematic. As expected, the GCI and the modified Cox methods outperform the one recommended by the ASTM in synthetic lognormal populations. However, deviations from a logarithmic shape may cause these methods to fail severely. Yet, the method recommended by the ASTM provides unreliable coverages for small samples sizes which in turn depends on the asymmetry of the distribution and the outliers contained in the dataset. As a result, the ASTM recommended method should never be used in data sets with less than 100 grain sections and with caution for larger sample sizes. The ASTM method can be preferred over the modified Cox or GCI when the distribution deviates from lognormal. No clear recommendation arises for the arithmetic mean.

4.3. Minimum number of (section) grains needed

The difficulty of estimating reliable confidence intervals in real datasets makes it difficult to estimate reliable minimum sizes in grain size studies, especially if the target is the arithmetic mean. Equations (6) and (7) allows estimating reliable minimum sample sizes for the geometric mean by solving for *n*. The same principle can be applied to any confidence interval estimation method. This approach, however, requires knowing the standard deviation of the population in advance and thus a pilot sample. In any event, it seems unlikely to find a situation in which all the samples to be studied will have the same lognormal standard deviation to apply this approach. Under this situation, we consider a better strategy to use minimum sample size reference values provided in tables once the range of lognormal shape is known approximately and report the confidence interval. If a specific margin of error is required, it would be necessary to proceed iteratively by increasing the sample size until the error desired is reached.

Declaration of competing interest

No conflict of interest exists.

Acknowledgements

This work was supported by a postdoctoral fellowship co-funded by the European Union and the Government of the Principality of Asturias through the "Clarín-Cofund" program [grant number ACA17-32] within the Marie Skłodowska-Curie COFUND (FP7) Actions. I thank S. Kidder and A.J. Cross for their insightful comments and grammatical corrections, and the editorial guidance of C.W. Passchier.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsg.2020.104042.

Appendix A. Formulae to estimate confidence intervals in lognormal distributions for the arithmetic mean and the median

A.1. Arithmetic mean

Many methods have been proposed over time for estimating confidence intervals for the arithmetic mean in lognormal distributions (e.g. Land, 1971, 1972; Parkin et al., 1990; Armstrong, 1992; Krishnamoorthy and Mathew, 2003; Endo, 2009; Zou et al., 2009

). Here, we only detail the first method that provided exact error margins and the two methods tested in this study. The reason for limiting the number of methods is due to some of them have the sole objective of simplifying the calculations or reducing the calculation time without providing quantitative improvements in the estimation of confidence intervals. The criterion for choosing which methods to test was based on finding a balance between their simplicity (i.e. easy to implement), its efficiency (i.e. fast in modern processors), and its widespread use.

A.1.1. Land's method The first method that provides exact solutions is that of Land (1971). It uses the following equations:

$$LowerCI = exp\left(\mu_y + 0.5\sigma_y^2 + C_L\frac{\sigma_y}{\sqrt{(n-1)}}\right) \tag{A1}$$

$$UpperCI = exp\left(\mu_y + 0.5\sigma_y^2 + C_U\frac{\sigma_y}{\sqrt{(n-1)}}\right) \tag{A2}$$

where C_L and C_U are factors provided in a series of tables (Land, 1975) calculated from a function that depends on the sample size (n), the standard deviation of the log-transformed values (σ_y), and the relative error (α). When exact values are not in the tables it requires interpolation. Although the method provides exact confidence intervals for lognormal populations, the use of tables makes the process tedious and/or computationally intensive. Due to this, several authors proposed simpler and faster alternatives (e.g. Parkin et al., 1990; Armstrong, 1992; Krishnamoorthy and Mathew, 2003; Endo, 2009; Zou et al., 2009).

A.1.2. Modified Cox method Originally appeared in Land (1971) and later modified by Armstrong (1992) introducing the use of the *t-score*.

$$LowerCI = exp\left(\mu_y + 0.5\sigma_y^2 - t\left(\frac{\sigma_y}{\sqrt{n}}\right)\sqrt{1+\frac{\sigma_y^2 n}{2(n+1)}}\right) \tag{A3}$$

$$UpperCI = exp\left(\mu_y + 0.5\sigma_y^2 + t\left(\frac{\sigma_y}{\sqrt{n}}\right)\sqrt{1+\frac{\sigma_y^2 n}{2(n+1)}}\right) \tag{A4}$$

This is a widely used method as an alternative to the Land's and it is fairly proved that the original Cox method provides satisfactory confidence interval estimates as long as the sample size is large (n > 60) (Parkin et al., 1990). The modified Cox version (Armstrong, 1992) plus the use of the Bessel corrected standard deviation, as implemented in this paper, provide satisfactory results even for smaller samples.

A.1.3. Generalized Confidence Intervals (GCI) method This is a Monte Carlo-type algorithm proposed by Krishnamoorthy and Mathew (2003) that uses the following procedure:

1. For a given dataset $x_1, \ldots, x_n$ estimate the log-transformed population $y_i = ln(x_i)$
2. Compute the arithmetic mean and the variance of the log-transformed population

$$\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{A5}$$

$$\sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2 \tag{A6}$$

4. Generate a large number of random values of the non-central chi-square distribution with $n-1$ degrees of freedom $U^2 \sim \chi^2_{n-1}$ (10000 random values were generated).
5. Generate a large number of random values from a normal distribution with mean equal zero and a standard deviation of one $Z \sim N(0,1)$ (10000 random values were generated).
6. Compute the statistic T using m Monte Carlo iterations by random sampling the generated values of Z and U^2 and estimate:

$$T = \mu_y - \frac{Z}{U/\sqrt{n-1}}\frac{\sigma_y}{\sqrt{n}} + 0.5\frac{\sigma_y^2}{U^2/(n-1)} \tag{A7}$$

7. Sort the T values obtained.
8. Estimate the corresponding percentile values according to the significance level required.

It is proved that this method provides satisfactory confidence intervals for n > 15 and that it performs better than the modified Cox within the range of MSD values 1.1 to 7.4 and sample sizes within the range 10–50 (Zou et al., 2009).

A.2. Median

This is a rule-of-thumb method proposed in Hollander and Wolfe (1999) using the following procedure:

1. Sort the population values
2. Estimate the position of the confidence intervals using

$$upper = 1 + \frac{n}{2} + z\frac{\sqrt{n}}{2} \tag{A8}$$

$$lower = \frac{n}{2} - z\frac{\sqrt{n}}{2} \tag{A9}$$

where z corresponds with the *z-score* value desired (e.g. 1.96 for a certainty of 95%) and *n* the sample size.

3. Extract the values that correspond to that positions. When the position obtained is not an integer, the procedure implemented use the floor and ceiling value instead of the rounded one.

Uncited references

Pérez and Lefante, 1997.

References

Aitchison, J, Brown, J A C, 1957. The Log-Normal Distribution. Cambridge University Press, Cambridge (UK).

Armstrong, B G, 1992. Confidence intervals for arithmetic means of lognormally distributed exposures. Am. Ind. Hyg. Assoc. J. 53, 481–485. doi:10.1080/15298669291360003.

ASTM International, 2013. Standard Practice for Determining Average Grain Size Using Electron Backscatter Diffraction (EBSD) in Fully Recrystallized Polycrystalline Materials. ASTM E2627 – 13. doi:10.1520/E2627-13.

ASTM International, 2013. Standard Test Methods for Determining Average Grain Size. ASTM E112-13. doi:10.1520/E0112-13.

Berger, A, Herwegh, M, Schwarz, J O, Putlitz, B, 2011. Quantitative analysis of crystal/grain sizes and their distributions in 2D and 3D. J. Struct. Geol. 33, 1751–1763. doi:10.1016/j.jsg.2011.07.002.

Cross, A J, Prior, D J, Stipp, M, Kidder, S, 2017. The recrystallized grain size piezometer for quartz: an EBSD-based calibration. Geophys. Res. Lett. 44, 6667–6674. doi:10.1002/2017GL073836.

Davis, J C, 2002. Statistics and Data Analysis in Geology. J. Wiley.

Depriester, D, Kubler, R, 2019. Resolution of the Wicksell's equation by minimum distance estimation. Image Anal. Stereol. 213–226. doi:10.5566/ias.2133.

Endo, Y, 2009. Estimate of confidence intervals for geometric mean diameter and geometric standard deviation of lognormal size distribution. Powder Technol. 193, 154–161. doi:10.1016/j.powtec.2008.12.019.

Fullman, R L, 1953. Measurement of particle size in opaque bodies. Trans. Metall. Soc. AIME 197, 447–452.

Ginos, B, 2009. Parameter Estimation for the Lognormal Distribution Theses and Dissertations.

Hale, W E, 1972. Sample size determination for the log-normal distribution. Atmos. Environ. 6, 419–422. doi:10.1016/0004-6981(72)90138-2.

Heilbronner, R, Barret, S, 2014. Image Analysis in Earth Sciences. Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-10343-8.

Herwegh, M, Poulet, T, Karrech, A, Regenauer-Lieb, K, 2014. From transient to steady state deformation and grain size: a thermodynamic approach using elasto-visco-plastic numerical modeling. J. Geophys. Res.: Solid Earth 119 (2), 900–918. doi:10.1002/2013JB010701.

Hewett, P, 1995. Sample size formulae for estimating the true arithmetic or geometric mean of lognormal exposure distributions. Am. Ind. Hyg. Assoc. J. 56, 219–225. doi:10.1080/15428119591017042.

Higgins, M D, 2006. Quantitative Textural Measurements in Igneous and Metamorphic Petrology. Cambridge University Press.

Hollander, M, Wolfe, D A, 1999. Nonparametric Statistical Methods, Solutions Manual. second ed. Wiley Series in Probability and Statistics.

Humphreys, F J, 2001. Grain and subgrain characterisation by electron backscatter diffraction. J. Mater. Sci.. doi:10.1023/A:1017973432592.

Kidder, S, Hirth, G, Avouac, J-P, Behr, W M, 2015. The influence of stress history on the grain size and microstructure of experimentally deformed quartzite. J. Struct. Geol. 83, 194–206. doi:10.1016/j.jsg.2015.12.004.

Krishnamoorthy, K, Mathew, T, 2003. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. J. Stat. Plann. Inference 115, 103–121. doi:10.1016/S0378-3758(02)00153-2.

Land, C E, 1971. Confidence intervals for linear functions of the normal mean and variance. Ann. Math. Stat. 42, 1187–1205. doi:10.1214/aoms/1177693235.

Land, C E, 1972. An evaluation of approximate confidence interval estimation methods for lognormal means. Technometrics 14, 145. doi:10.2307/1266926.

Land, C E, 1975. Tables of confidence limits for linear functions of the normal mean and variance. In: Statistics, I. of M (Ed.), Selected Tables in Mathematical Statistics, 3. American Mathematical Society, Providence, RI, pp. 385–419.

Limpert, E, Stahel, W A, Abbt, M, 2001. Log-normal distributions across the sciences: keys and clues. Bioscience 51, 341. doi:10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2.

Lopez-Sanchez, M A, 2018. GrainSizeTools: a Python script for grain size analysis and paleopiezometry based on grain size. J. Open Source Softw. 3, 863. doi:10.21105/joss.00863.

Lopez-Sanchez, M A, Llana-Fúnez, S, 2015. An evaluation of different measures of dynamically recrystallized grain size for paleopiezometry or paleowattometry studies. Solid Earth 6, 475–495. doi:10.5194/se-6-475-2015.

Lopez-Sanchez, M A, Llana-Fúnez, S, 2016. An extension of the Saltykov method to quantify 3D grain size distributions in mylonites. J. Struct. Geol. 93, 149–161. doi:10.1016/j.jsg.2016.10.008.

Lopez-Sanchez, M A, Llana-Fúnez, S, 2018. A cavitation-seal mechanism for ultramylonite formation in quartzofeldspathic rocks within the semi-brittle field (Vivero fault, NW Spain). Tectonophysics 745, 132–153. doi:10.1016/j.tecto.2018.07.026.

Oliphant, T, 2006. In: A Bayesian Perspective on Estimating Mean, Variance, and Standard-Deviation from Data, 278. Faculty Publications.

Parkin, T B, Chester, S T, Robinson, J A, 1990. Calculating confidence intervals for the mean of a lognormally distributed variable. Soil Sci. Soc. Am. J. 54, 321–326. doi:10.2136/sssaj1990.03615995005400020004x.

Pérez, A, Lefante, J J, 1997. On sample size estimation of the arithmetic mean of a lognormal distribution with and without type I. Rev. Colomb. Estadística 18, 2389–8976.

Ranalli, G, 1984. Grain size distribution and flow stress in tectonites. J. Struct. Geol. 6, 443–447. doi:10.1016/0191-8141(84)90046-4.

Shih, W J, Binkowitz, B, 1987. Median versus geometric mean for lognormal samples. J. Stat. Comput. Simulat. 28, 81–83. doi:10.1080/00949658708811013.

Soleymani, H., Kidder, S. B., Hirth, G., Garapic, G., (in press). The effect of cooling during deformation on recrystallized grain-size piezometry, Geology.https://doi.org/10.1130/G46972.1.

Stipp, M, 2018. Comment on “Quartz grainsize evolution during dynamic recrystallization across a natural shear zone boundary” by Haoran Xia and John P. Platt. J. Struct. Geol.. doi:10.1016/j.jsg.2018.09.009.

Stipp, M, Tullis, J, 2003. The recrystallized grain size piezometer for quartz. Geophys. Res. Lett. 30, 1–5. doi:10.1029/2003GL018444.

Ter Heege, J H, De Bresser, J H P, Spiers, C J, 2004. Composite flow laws for crystalline materials with log-normally distributed grain size: theory and application to olivine. J. Struct. Geol. 26 (9), 1693–1705. doi:10.1016/j.jsg.2004.01.008.

Van Der Walt, S, Colbert, S C, Varoquaux, G, 2011. The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. 13, 22–30. doi:10.1109/MCSE.2011.37.

Zou, G Y, Huo, C Y, Taleban, J, 2009. Simple confidence intervals for lognormal means and their differences with environmental applications. Environmetrics 20, 172–180. doi:10.1002/env.919.