DIVISION MATERIALS AND PRODUCTION CHEMICAL PROBLEM SOLVING

Calculation of uncertainty intervals for skewed distributions - Application in chemical analysis with large uncertainties

Eskil Sahlin, RISE Research Institutes of Sweden (Borås, Sweden)

Bertil Magnusson, Trollboken (Göteborg, Sweden)

Thomas Svensson, Ingenjörsstatistik (Borås, Sweden)

RISE Report 2021:07

Calculation of uncertainty intervals for skewed distributions - Application in chemical analysis with large uncertainties

Eskil Sahlin, RISE Research Institutes of Sweden (Borås, Sweden)

Bertil Magnusson, Trollboken (Göteborg, Sweden)

Thomas Svensson, Ingenjörsstatistik (Borås, Sweden)

Abstract

Calculation of uncertainty intervals for skewed distributions - Application in chemical analysis with large uncertainties

A measurement result x is normally reported, with an expanded uncertainy U at a stated condidifence level, as $x \pm U$. When the distribution of results are skewed the result x will be reported with a skewed interval as $x - U_l$ to $x + U_u$ where U_l and U_u are the lower and upper limits of uncertainty. It is concluded that skewness needs to be taken into account in order to report more correct uncertainty interval for results at relative standard deviations exceeding approximately 15 to 20 %. A power transformation, x^B, that will transfer (many) measurement results that have a skewed distribution to an approximate normal distribution is suggested in order to report more correct uncertainty intervals. The parameter B needs to be optimized and the optimized value depends on the distribution of the measurement results. The transformation is characterized and studied using Monte Carlo simulations. Optimization of B can be performed based on modelling of results, on judgement based on experience or on experimental results. Optimization based on experimental results is difficult since a very large data set is needed to get a reliable value of B. Two important B values are B equal to 1 that corresponds to an approximate normal distribution of the original measurement results, and B approaching 0 that corresponds to an approximate log-normal distribution of the original measurement results. An expression for calculation of uncertainty intervals when using transformation based on x^B with an optimized B is given, and compared with other types of uncertainty intervals where it is assumed that measurement results have a normal or a log-normal distribution. It is also suggested how to combine uncertainty contributions with different skewness. Implementation of the working approach is demonstrated with three examples from chemical analysis.

Key words:
normal distribution
log-normal distribution
probability distribution
skewness
asymmetry
power transformation
measurement uncertainty
uncertainty interval
chemical analysis

RISE Research Institutes of Sweden AB

RISE Report 2021:07
ISBN: 978-91-89167-89-6
Borås 2021

Content

1 Introduction

The uncertainty of results from many types of measurements is most often expected and assumed to be described by a normal distribution i.e. a symmetric distribution. Measurement uncertainties are typically given as $\pm U$ where U is the expanded uncertainty (often at 95 % confidence level) calculated assuming a normal distribution of the result. The assumption of a normal distribution is based on the central limit theorem [1, 2] that states that the probability distribution for A, when calculated according to a measurement model:

$$A = A_1 + A_2 + ...+A_n \quad (1)$$

where A_1, A_2, ... A_n are independent random variables, will approach a normal distribution when n increases. Hence, it is assumed that the result of the measurement is calculated mainly by addition and subtraction of a number of variables. However, in chemical analysis, as well as in many other types of measurements, multiplication and division are important mathematical steps. A logarithmic transformation will convert multiplications in the model equation to additions (divisions to subtractions) and the central limit theorem will apply for the logarithmic data. The probability distribution for A, when calculated according to a measurement model with only multiplications

$$A = A_1 \times A_2 \times .....\times A_n \quad (2)$$

will approach a log-normal distribution when n increases. This distribution is asymmetric as shown to the left in Fig. 1. When taking the logarithm of a log normal distribution the transformed data will be a normal, symmetric distribution as shown to the right in Fig. 1 with the $\log_{10}$ scale.

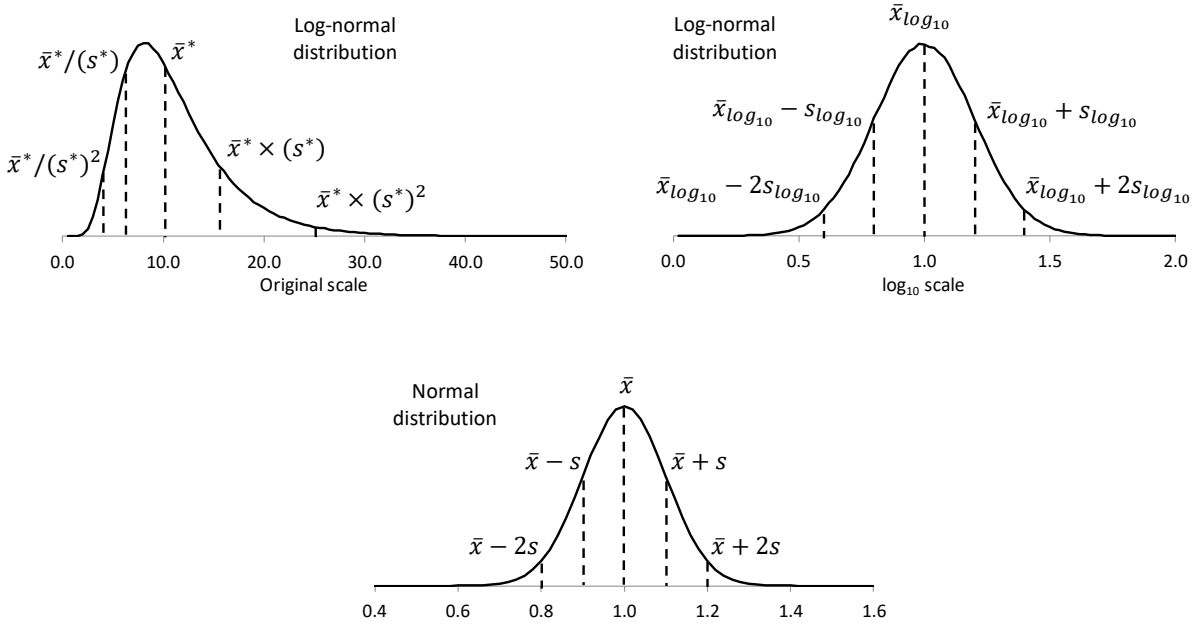

Figure 1. Fundamental characteristics of the normal and log-normal probability distributions. The log-normal distribution is skewed (left) but becomes symmetric (right) after taking the logarithm of the values. For comparison a normal distribution, that is symmetric, is included in the figure.

The log-normal distribution is skewed but becomes symmetric after taking the logarithm of the values. For comparison a normal distribution, that is symmetric, is included in the figure. For a normal distribution, the coverage probability of different intervals are [3, 4]:

$\bar{x} - s$ to $\bar{x} + s$	68.3 %
$\bar{x} - 2s$ to $\bar{x} + 2s$	95.5 %
$\bar{x} - 3s$ to $\bar{x} + 3s$	97.3 %

where $\bar{x}$ is the mean and s is the standard deviation. For a log-normal distribution the corresponding intervals in the original space are [3, 4]:

$\bar{x}^*/s^*$ to $\bar{x}^* \times s^*$	68.3 %
$\bar{x}^*/(s^*)^2$ to $\bar{x}^* \times (s^*)^2$	95.5 %
$\bar{x}^*/(s^*)^3$ to $\bar{x}^* \times (s^*)^3$	97.3 %

where $\bar{x}^* = 10^{\tilde{x}_{log_{10}}}$ where $\tilde{x}_{log_{10}}$ is the median of $\log_{10} x$ and $s^* = 10^{s_{log_{10}}}$ where $s_{log_{10}}$ is the standard deviation of $\log_{10} x$. A more detailed description of the characteristics of the two distributions and their importance is available in the literature [3, 4].

In the following discussion the standard deviation, s, is assumed to be equal to the standard uncertainty, u [1, 2]. Often a log-normal distribution can be approximated with a normal distribution. The coefficient of variation (CV) is often used to decide if this approximation is valid. For CV below 20 % the difference in shape and skewness between the two distributions is small [3], and a normal distribution approximation can be used. At larger CV, or relative standard uncertainties, the skewness of the results needs to be taken into account, and this is typically performed by assuming a log-normal distribution [5-9]. However, sometimes the $\log_{10} x$ transformation is not sufficient to obtain symmetry and we here propose a more general transformation. Though relative standard uncertainties are often smaller than 20 %, it is possible to encounter larger standard uncertainties. For instance, sampling in chemical analysis can contribute substantially to the overall uncertainty and to skewness in the probability distribution of the result [5, 6, 8, 9].

Distributions of these large uncertainties are rarely addressed in the literature [7]. In "*Evaluation of measurement data - Guide to the expression of uncertainty in measurement*" (well-known as "GUM") [1, 2] only additive measurement errors are considered, and it is assumed that the probability distribution can be approximated with a normal distribution (or a t-distribution). It is also stated that the GUM uncertainty framework might not be satisfactory when the probability distribution for the output quantity is either asymmetric, or not a Gaussian or a t-distribution[1]. In a Eurachem guide [10] and a GUM supplement [11], however, Monte Carlo methods that can be used to study the distribution of the output quantity

[1] In section G5.2 of GUM asymmetry is discussed: *The alternative is to give an interval that is symmetric in probability (and thus asymmetric in U): the probability that Y lies below the lower limit $y - U_-$ is equal to the probability that Y lies above the upper limit $y + U_+$. But in order to quote such limits, more information than simply the estimates y and $u_c(y)$ [and hence more information than simply the estimates x_i and $u(x_i)$ of each input quantity X_i] is needed.*

when all input quantities can be well described are presented. Example is given of an output asymmetric probability distribution and reporting of an expanded asymmetric interval. Hence, it is somewhat doubtful to consider large asymmetric uncertainties to be covered by the scope of GUM.

In this work is studied how skewness in measurement results can be handled with focus on chemical analysis with large uncertainties. A transformation often used to stabilize variance [12] is suggested in the following section that 1) will transform skewed distributions to a symmetric distribution that can be assumed to be normal, and 2) using an expression for back-transformation an asymmetric uncertainty interval can be calculated with a correct confidence interval, e.g. 95 %. For comparison, results obtained when using no transformation and transformation using $\log_{10} x$ are also included.

2 List of symbols and abbreviations

A	Variable with a normal probability distribution
B	Parameter in transformation (x^B)
B_{opt}	Optimized B
C	Concentration
CV	Coefficient of variation (in %)
CV_{trans}	Coefficient of variation in transformed space using x^B transformation
$F_{2.5\,\%}$	Fraction of data points below $\bar{x} - 1.96s$ (%)
$F_{97.5\,\%}$	Fraction of data points above $\bar{x} + 1.96s$ (%)
k	Coverage factor for a given probability
n	Number of data
s_{rel}	Relative standard deviation
$s_{rel,trans}$	Relative standard deviation in transformed space using x^B transformation
s	Sample standard deviation
$s_{log_{10}}$	Sample standard deviation after transformation using $\log_{10} x$
s_{log_e}	Sample standard deviation after transformation using $\log_e x$
x	Data in the original space
x_{trans}	Transformed data
$\bar{x}$	Average (sample mean)
$\tilde{x}$	Median
U_l	Lower limit of an uncertainty interval for a measurement result x
U_u	Upper limit of an uncertainty interval for a measurement result x
${}^{F}U$	Uncertainty factor
γ	Skewness
μ_A	Mean of random variable A
σ_A	Standard deviation of variable A

3 Model

Transformation of original data is performed according to

$$x_{trans} = x^B \tag{3}$$

where x_{trans} and x are the transformed and original data, respectively, and B is a parameter that is optimized with the goal that the transformed data should have a symmetric distribution, i.e. having a skewness that is close to 0. Skewness, γ, is here calculated as

$$\gamma = \frac{n}{(n-1)\times(n-2)} \sum_{i=1}^{n} \left(\frac{x_{trans,i} - \bar{x}_{trans}}{s_{trans}} \right)^3 \tag{4}$$

where n is the number of data, $x_{trans,i}$ is the transformed individual data, $\bar{x}_{trans}$ is the mean of transformed data, and s_{trans} is the standard deviation of the transformed data. The optimized B will be denoted B_{opt}.

Back-transformation of transformed data is obtained by

$$x = {x_{trans}}^{1/B_{opt}} \tag{5}$$

Note that for $B_{opt} < 0$, the order of data in the transformed space will be opposite to the order of data in the original space. Hence, when calculating a confidence interval the lower limit of the interval in the transformed space corresponds to the upper limit in the original space.

Power transformation according Eq. 3 and similar equations (e.g. equation used in Box-Cox transformation) are well-known to stabilize variances and transfer data to more normal distributed data. For instance, transformation according to Eq. 3 has been used in variance stabilizing transformation where B is adjusted to give a minimal dependence of variance on x [12]. The Box-Cox transformation is often written as

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & if\ \lambda \neq 0 \\ \ln y & if\ \lambda = 0 \end{cases} \tag{6}$$

and is used to transform skewed data in many applications prior to use of statistical analysis tools where normal distributed data is needed [13]. This transformation is constructed to obtain the limit $\ln y$ when λ approaches zero. However, for our purposes the simple power transformation is regarded as sufficient.

In order to illustrate and characterize the transformation procedure, probability distributions of transformed data with different values of B for an original normal distribution and an original log-normal distribution are shown in Fig. 2(a) and 2(b), respectively.

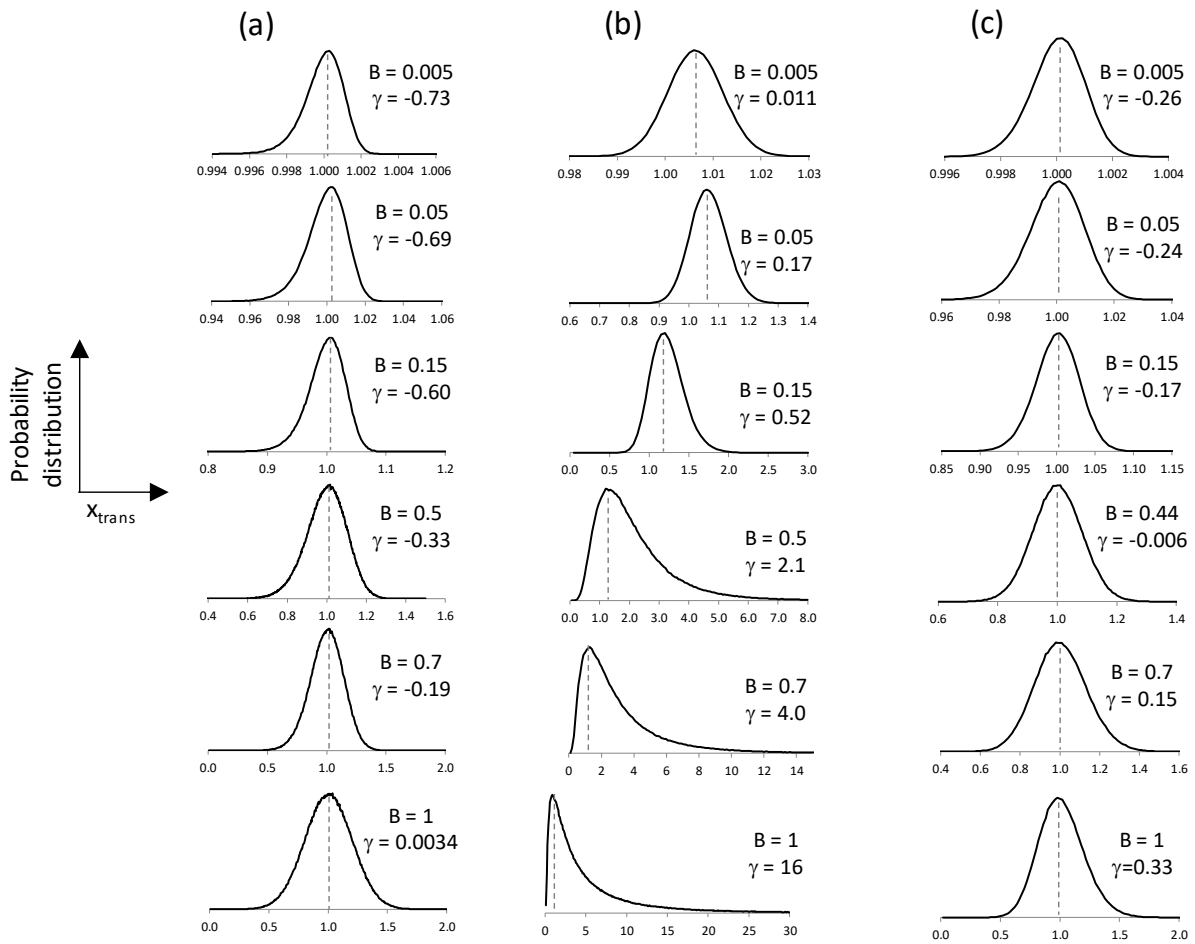

Figure 2. Probability distributions of transformed data using different values of B for data that have an original (a) normal distribution, (b) log-normal distribution, and (c) distribution somewhere "between" a normal and a log-normal distribution. In order to emphasize the symmetry and asymmetry in the diagrams a dashed line has been added to mark the probability distribution maximum. In addition, values for skewness (γ) are also given for each distribution. B = 1 shows the original distributions.

Note that for a normal distribution the skewness (γ) will approach 0 when B approaches 1, and x_{trans} will approach x. For a log-normal distribution γ will approach 0 when B approaches 0. However, for B=0, the standard deviation will be infinitely small since x_{trans} will be equal to 1 for all data. Hence, for data that originally has a distribution that is somewhere "between" a normal and a log-normal distribution it can be expected that $0 < B_{opt} \leq 1$ when B has been optimized. A probability distribution of transformed data for an original distribution that is somewhere "between" a normal and a log-normal distribution with different values of B is shown in Fig. 2(c). The data in Fig. 2(c) were generated using an arbitrary chosen equation

$$A_1 \times \frac{A_2+A_3}{A_4+A_5} \times A_6 + A_7 - A_8 \qquad (7)$$

where the variables A_1, A_2, ... and A_8 have a normal probability distribution with mean values equal to 1, and standard deviations σ_{A_1}, σ_{A_2}, ..., σ_{A_6} = 0.1 and σ_{A_7}, σ_{A_8} = 0.05. These values were arbitrary chosen to generate data that have a distribution "between" a normal and a log-

normal distribution suitable for illustrating how the transformation works. For the transformed data, γ will be close to 0 (corresponding to a symmetric distribution) when $B = 0.44$. For B values < 0.44, γ will be negative and γ will become more negative when B approaches 0. For B values > 0.44, γ will be positive and γ will increase when B approaches 1. Note that in Fig. 2(a) to (c), distributions where $B = 1$ will be equal to the original distributions.

Clearly, by optimizing the value B for Eq. 3, data can be transformed to data that have a symmetric distribution.

4 Calculations

All calculations are performed using Excel software (Office 365, Microsoft). Random data with a normal probability distribution were generated using Excel function NORM.INV(RAND();mean;standard deviation). Random data with a log-normal probability distribution were generated as $10^{\text{NORM.INV(RAND();mean;standard deviation)}}$. Random data with a rectangular distribution were generated using RAND(). Random data with a probability distribution somewhere "between" a normal and log-normal probability distribution were generated by multiplying, dividing, adding or subtracting random data with normal probability distributions generated as described above. All simulations were based on 10^6 data if not otherwise mentioned.

Optimization of B was obtained by utilizing Solver (Excel add-in program) with the constraint $B \geq 0.0001$ to prevent that B will reach 0 in the optimization. Start value for B is not critical, and here 0.5 was used as a start value. If optimization resulted in $B = 0.0001$, a second optimization step was performed with the constraint $B \leq -0.0001$ and a start value of -0.5. Settings used in the Solver optimization are given in Table 1.

Table 1. Settings used in Solver.

Set objective	Cell containing value for absolute skewness
To	Min
By changing variable cells	Cell containing value for parameter $B^{(1)}$
Subject to the constraints	Cell containing value for parameter $B^{(1)}$: >= 0.0001 or Cell containing value for parameter $B^{(1)}$: <= -0.0001 (Second setting is used if optimization using the first setting results in $B^{(1)}$ = 0.0001).
Make unconstrained variables non-negative	Not active
Select a solving method	GRG (Generalized reduced gradient) Non-linear
Precision (All methods)	0.000001
Convergens (GRG Non-linear)	0.0001
Derivatives (GRG Non-linear)	Forward
(1) B is a parameter in an equation used to transform data.	

Some of the calculations below, including random data generation, were also performed using the software R (ver. 4.0.0) [14]. Identical results were obtained showing that the Excel calculations have adequate accuracy.

Analysis of variance (ANOVA) was performed using RANOVA2 (a stand-alone program running in Microsoft Excel) available from Royal Society of Chemistry (RSC) website [15].

5 Results

5.1 Characterization of the transformation procedure

In order to demonstrate the applicability of the transformation procedure, several data sets were processed, and the results are given in Table 2. Parameters describing the distributions for the original data i.e. without transformation are given first, namely, CV, γ, fraction of data points below $\bar{x} - 1.96s$ (denoted as $F_{2.5\,\%}$), and fraction of data points above $\bar{x} + 1.96s$ (denoted as $F_{97.5\,\%}$). Then, parameters describing the distribution after transformation using $x^{B_{opt}}$

including B_{opt}, skewness (γ), $F_{2.5\,\%}$, and $F_{97.5\,\%}$ are given. For comparison, parameters describing the distribution after transformation using $\log_{10} x$ are also given. The different data sets included 10^6 data points and were generated as described in Table 2 where the variables A_1, A_2, ... where randomly generated data that had a normal probability distribution with means μ_{A_1}, μ_{A_2}, ... and standard deviations σ_{A_1}, σ_{A_2}, ... The parameters for simulations are chosen to only generate positive values.

Table 2. Parameters describing the probability distribution 1) original data i.e. without transformation, 2) after transformation using $x^{B_{opt}}$, and 3) after transformation using $\log_{10} x$ for different data sets.

Data set	**Way to generate data set**	**Original data i.e. without transformation**				**After transformation using $x^{B_{opt}}$**			**After transformation using $\log_{10} x$**	
		Distribution of data set	**CV of data set (%)**	**γ**	**$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)**	**B_{opt}**	**γ**	**$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)**	**γ**	**$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)**
1	A_1 $\mu_{A_1} = 1, \sigma_{A_1} = 0.1$	Normal distribution	10	6×10^{-3}	2.49 2.51	0.98	-5×10^{-8}	2.51 2.49	-0.30	3.27 1.65
2	10^{A_1} $\mu_{A_1} = 1, \sigma_{A_1} = 0.1$	Log-normal distribution	23	0.72	0.56 4.01	-0.004	5×10^{-9}	2.50 2.49	3×10^{-3}	2.49 2.50
3	$A_1 \times A_2 \times A_3 \times \ldots\ldots \times A_{20}$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_{20}} = 1$ $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_{20}} = 0.02$	"Between" normal and log-normal distribution	9.0	0.25	1.78 3.13	0.065	-1×10^{-9}	2.52 2.49	-2×10^{-2}	2.57 2.45
4	$A_1 \times A_2 \times A_3 \times \ldots\ldots \times A_{20}$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_{20}} = 1$ $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_{20}} = 0.05$	"Between" normal and log-normal distribution	23	0.65	0.69 3.90	0.055	-1×10^{-7}	2.51 2.49	-4×10^{-2}	2.61 2.39
5	$A_1 \times A_2 \times A_3 \times \ldots\ldots \times A_{20}$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_{20}} = 1$ $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_{20}} = 0.1$	"Between" normal and log-normal distribution	47	1.4	0 4.61	0.053	-3×10^{-8}	2.51 2.51	-7×10^{-2}	2.71 2.30
6	$A_1 \times A_2$ $\mu_{A_1}, \mu_{A_2} = 1$ $\sigma_{A_1}, \sigma_{A_2} = 0.1$	"Between" normal and log-normal distribution	14	0.21	1.87 3.04	0.51	-2×10^{-7}	2.48 2.50	-0.22	3.06 1.87

Data set	Way to generate data set	Original data i.e. without transformation				After transformation using $x^{B_{opt}}$			After transformation using $\log_{10} x$	
		Distribution of data set	CV of data set (%)	γ	$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)	B_{opt}	γ	$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)	γ	$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)
7	$A_1 \times A_2 \times A_3 \times \ldots\ldots \times A_5$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_5} = 1$ $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_5} = 0.1$	"Between" normal and log-normal distribution	23	0.54	0.92 3.74	0.20	-3×10^{-8}	2.51 2.50	-0.14	2.88 2.11
8	$A_1 \times A_2 \times A_3 \times \ldots\ldots \times A_{10}$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_{10}} = 1$ $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_{10}} = 0.1$	"Between" normal and log-normal distribution	32	0.89	0.21 4.25	0.10	-7×10^{-9}	2.50 2.51	-0.1	2.76 2.22
9	$\frac{A_1 \times A_2 \times \ldots\ldots \times A_{12}}{A_{13} \times A_{14} \times \ldots\ldots \times A_{20}}$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_{20}} = 1$ $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_{20}} = 0.1$	"Between" normal and log-normal distribution	48	1.5	0 4.59	0.0098	-1×10^{-8}	2.49 2.50	-0.013	2.52 2.46
10	$\frac{A_1 \times A_2}{A_3 \times A_4 \times \ldots\ldots \times A_{20}}$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_{20}} = 1$ $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_{20}} = 0.1$	"Between" normal and log-normal distribution	48	1.7	0 4.62	-0.040	3×10^{-8}	2.51 2.50	0.055	2.35 2.66
11	$A_1 \times \frac{A_2+A_3}{A_4+A_5} \times A_6 + A_7 - A_8$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_8} = 1$ $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_6} = 0.1$ and $\sigma_{A_7}, \sigma_{A_8} = 0.05$	"Between" normal and log-normal distribution	19	0.33	1.58 3.30	0.44	-2×10^{-8}	2.52 2.52	-0.27	3.15 1.81

Data set	Way to generate data set	Original data i.e. without transformation				After transformation using $x^{B_{opt}}$			After transformation using $\log_{10} x$	
		Distribution of data set	CV of data set (%)	γ	$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)	B_{opt}	γ	$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)	γ	$F_{2.5\,\%}$ $F_{97.5\,\%}$ (%)
12	$A_1 - (A_2 \times A_3 \times \ldots\ldots \times A_{11})$ $\mu_{A_1} = 3$ $\sigma_{A_1} = 0.1$ $\mu_{A_2}, \mu_{A_3}, \ldots\ldots, \mu_{A_{11}} = 1$ $\sigma_{A_2}, \sigma_{A_3}, \ldots., \sigma_{A_{11}} = 0.05$	"Between" normal and log-normal distribution and negative skewness	9.4	-0.26	3.16 1.78	1.89	-6×10^{-8}	2.50 2.48	-0.59	3.74 0.98
13	$100 + 10^{A_1}$ $\mu_{A_1} = 1, \sigma_{A_1} = 0.1$	log-normal distribution added to a constant	Not relevant	0.71	0.55 3.99	-9.8	2×10^{-8}	2.41 2.43	0.64	0.69 3.90
14	$100 - 10^{A_1}$ $\mu_{A_1} = 1, \sigma_{A_1} = 0.1$	log-normal distribution subtracted from a constant	Not relevant	-0.72	4.01 0.55	9.7	-3×10^{-9}	2.42 2.40	-0.82	4.11 0.39

Clearly, for all the data, transformed data subsequent to optimization according to equation 3 will have a skewness close to 0. The fraction of data below $\bar{x} - 1.96s$ and above $\bar{x} + 1.96s$ will both be 2.5 % making it possible to give a correct confidence interval. From this it can be concluded that the transformed data of all the datasets, subsequent to optimization of B, can be approximated with a normal distribution. Here follows a discussion of the different datasets.

Dataset 1 and 2 - For an original normal distribution (data set 1), transformation using $x^{B_{opt}}$ will have no effect on the data and B_{opt} will be close to 1. For an original log-normal distribution (data set 2), transformation will have the same effect as using $\log_{10}$ and optimized B will be close to 0.

Dataset 3 to 5 - Data with distributions described as "between" a normal and a log-normal distribution obtained by multiplying 20 identical variables. Note that for this construction, the skewness will increase with increasing standard deviation of the variables, although the number of multiplication steps is the same. Hence, the difference between the distribution of the data and a normal distribution will increase with increasing standard deviation of the variables. For equations containing multiplication and division steps, the distribution of the result can thus be approximated with a normal distribution when standard deviations of the variables are sufficiently small. Note that the value of B_{opt} will be similar although the skewness of the data increases.

Dataset 5 to 8 – Data with increasing number of multiplication steps from 1 to 19. The value of B_{opt} will decrease from 0.51 to 0.053 when increasing the number of multiplication steps. Hence, the value of B_{opt} will reflect how the original data is related to a normal distribution and a log-normal distribution, and not the skewness of the data. For instance, for original data with a small standard deviation, the skewness will be small, but the value of B_{opt} can still be close to 0 indicating that the data has a distribution similar to a log-normal distribution.

Dataset 9 to 10 - When including division steps transformed data can also be approximated with a normal distribution.

For data that has been obtained by using a somewhat more complex equation consisting of multiplication, division, addition and subtraction (data set 11), transformation optimization will again result in data that can be approximated with a normal distribution. Without transformation, or if using $\log_{10} x$ as transformation, the data will be skewed. The equation and the standard deviations of the different variables in the equation used to generate the data were chosen in order to generate data that had a distribution "between" a normal and a log-normal distribution. The same data was used in Fig. 2(c) above.

Typically, probability distributions for the results in chemical analysis will have a negligible or positive skewness. A negative skewness will seldom be encountered. If a negative skewness exist this will result in B_{opt} values > 1. An example is given above (data set 12). Optimization will result in a B_{opt} value of 1.89, and transformed data can again be approximated with a normal distribution.

Data set 13 and 14 exemplify results obtained by adding a number with a relatively small or zero uncertainty to data that are negatively or positively skewed. In chemical analysis this can occur for instance when a measurand is calculated as the residual (for instance the copper content in weight- % in brass can be calculated as 100 %-sum of determined contents of other elements). Optimized B values will be far outside the interval 0 to 1 (in this case -9.8 and 9.7, respectively)

but transformed data will be highly symmetric. Fractions of data below $\bar{x} - 1.96s$ and above $\bar{x} + 1.96s$ will be around 2.4 %, i.e. somewhat lower than 2.5 %, suggesting that transformed data are not fully normal distributed but close to normal distributed. Transformation using $\log_{10} x$ will not handle asymmetry well in either of the two data sets.

The Excel add in program Solver has here been used for optimization of B with the goal that the skewness of the transformed data should be 0. The optimization process is illustrated in Fig. 3(a) to (c) showing the relationship between B and absolute skewness for an original normal distribution (data set 1), an original log-normal distribution (data set 2), and an original distribution that is "between" a normal and a log-normal distribution (data set 6).

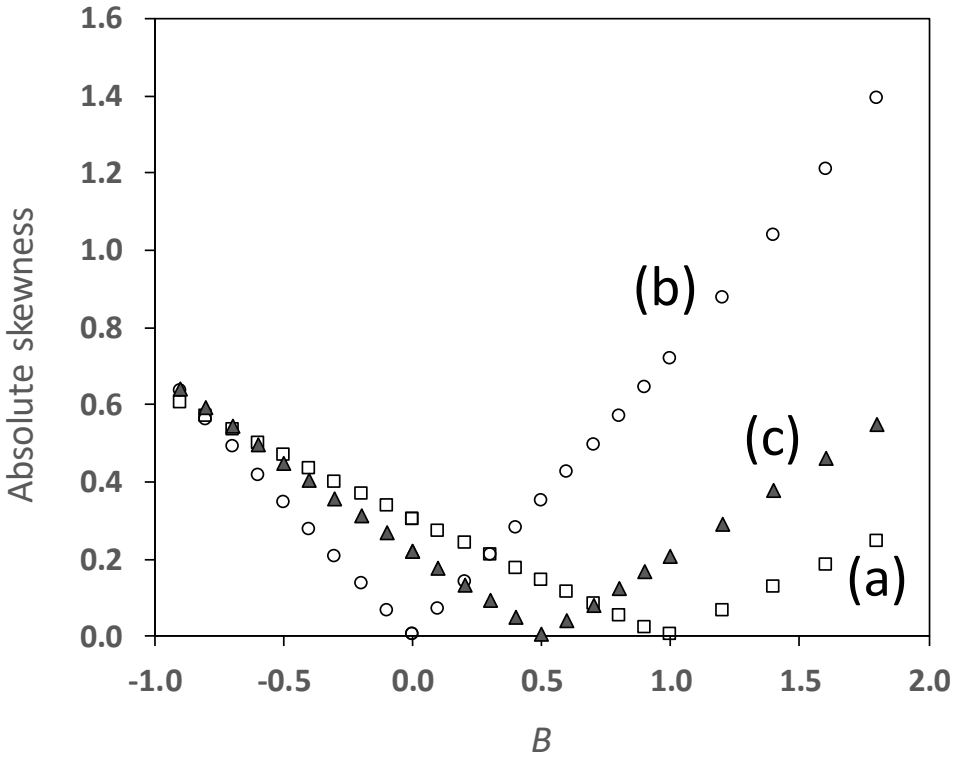

Figure 3. Optimization of B to obtain skewness close to 0. Relationship between B and absolute skewness for (a) an original normal distribution (data set 1 in Table 3), (b) an original log-normal distribution (data set 2 in Table 3), and (c) an original distribution that is "between" a normal and a log-normal distribution (data set 6 in Table 3).

As can been seen, there is a clear minimum in absolute skewness at B equal to 1 for an original normal distribution (Fig. 3(a)), at B close to zero for an original log-normal distribution (Fig. 3(b)), and at a B value between 0 and 1 for an original distribution that is "between" a normal and log-normal distribution (Fig. 3(c)).

It should be pointed out that some of the data in Table 2 have an asymmetry that is negligible in reality when evaluating measurement uncertainties. Intervals for original data comprising 95 % of the data in a data set, with 2.5 % of the data below and above the interval, can be obtained by calculating $\bar{x} \pm 1.96 \times s$ for transformed and optimized data, followed by back-transformation of the interval to the original space using Eq. 5. Such intervals are given in

Table 3 for the data sets in Table 2, together with the value of $\bar{x}$ back transformed to the original space. Also included are the corresponding intervals and values of $\bar{x}$ obtained without transformation and with transformation using $\log_{10} x$. In addition, CV for the original data set has been included.

Table 3. Average values ($\bar{x}$) and intervals $\bar{x} \pm 1.96 \times s$ calculated for transformed data followed by back-transformation to the original space, for the data in Table 2 with transformation using $x^{B_{opt}}$ of B, without transformation, and with transformation using $\log_{10} x$. Also included is the coefficient or variation (CV) for the original data.

			Back transformed value of $\bar{x}$ of transformed data **Back transformed intervals $\bar{x} \pm 1.96 \times s$ of transformed data**		
Data set	**CV for original data (%)**	**γ for original data**	**Without transformation**	**After transformation using $x^{B_{opt}}$.**	**After transformation using $\log_{10} x$**
1	10	6×10^{-3}	1.000 0.804 – 1.196	1.000 0.804 – 1.196	0.995 0.816 – 1.213
2	23	0.72	10.27 5.57 – 14.97	9.998 6.37 – 15.71	9.999 6.37 – 15.70
3	9.0	0.25	1.000 0.824 – 1.176	0.996 0.835 – 1.186	0.996 0.836 – 1.187
4	23	0.65	1.001 0.557 – 1.444	0.977 0.626 - 1.509	0.976 0.629 – 1.514
5	47	1.4	1.000 0.081 – 1.919	0.909 0.366 – 2.161	0.904 0.372 – 2.195
6	14	0.21	1.000 0.722 – 1.278	0.995 0.737 – 1.292	0.990 0.748 – 1.311
7	23	0.54	1.000 0.557 – 1.443	0.980 0.616 – 1.498	0.975 0.625 – 1.520
8	32	0.89	1.000 0.366 – 1.634	0.956 0.500 – 1.754	0.951 0.508 – 1.781
9	48	1.5	1.086 0.073 – 2.100	0.981 0.402 – 2.375	0.980 0.403 – 2.382

			Back transformed value of $\bar{x}$ of transformed data Back transformed intervals $\bar{x} \pm 1.96 \times s$ of transformed data		
Data set	***CV* for original data (%)**	**γ for original data**	**Without transformation**	**After transformation using $x^{B_{opt}}$.**	**After transformation using $\log_{10} x$**
10	48	1.7	1.202 0.062 – 2.342	1.080 0.452 – 2.664	1.084 0.447 – 2.633
11	19	0.33	1.005 0.634 – 1.376	0.995 0.663 – 1.403	0.987 0.678 – 1.436
12	9.4	-0.26	2.000 1.631 – 2.369	2.008 1.608 – 2.347	1.991 1.648 – 2.406
13	Not relevant	0.71	110.3 105.6 – 115.0	110.0 106.3 – 115.8	110.2 105.7 – 115.0
14	Not relevant	-0.72	89.7 85.0 – 94.4	90.0 84.2 – 93.7	89.7 85.1 – 94.6

Note that intervals obtained by transformation using $x^{B_{opt}}$ will be the true intervals for the simulated data sets. From this it can be seen that skewness become practically important at CV > approximately 15 to 20 %. Furthermore, transformation using $\log_{10} x$ will in many cases handle skewness sufficiently well. However, the described mathematical process will be more general compared to assuming either a normal or a log-normal distribution. In addition, it can also handle negative skewness, as well as skewness after addition of a number with no or small uncertainty to the results. Measurements giving rise to CV > 20 % within laboratories can sometimes be found in some chemical analyses. However, CV for reproducibility for some methods can be in the range 15-30 %. CV for proficiency testing schemes can also be > 20 %.

Furthermore, when considering sampling of heterogeneous samples such as some types of wastes and contaminated soil, CV well above 20 % are common. The cause of these large CV are typically not well-understood and difficult to model. Furthermore, in microbiology it is common that CV are > 20 % for food analysis.

5.2 Calculation of uncertainty intervals at different measurand levels

In order to calculate uncertainty intervals originating from precision for samples, using a known B_{opt}, it is necessary to determine the measurand level dependence. It is not straightforward how to generate data that will reflect the measurand level dependence. However, it is suggested that a relevant approach will be to use different values for the mean value of A_1, i.e. μ_{A_1}, keeping CV for A_1 constant for data generated using expressions described as

$$A_1 \times f(A_2, \ldots, A_n) \quad (8)$$

For such data, both B_{opt} and CV for data after transformation will be independent of the value μ_{A_1}. This is exemplified in Table 4 and 5 showing obtained B_{opt}, standard deviation (s) and CV after transformation and optimization of B, for data generated as

$A_1 \times A_2 \times A_3 \times \ldots\ldots \times A_{20}$

(i.e. only using multiplication) in Table 4 and

$\frac{A_1 \times A_2}{A_3 \times A_4 \times \ldots\ldots \times A_{20}}$

(i.e. also including division) in Table 5, where μ_{A_2}, ..., $\mu_{A_{20}}$ = 1 and σ_{A_2}, ..., $\sigma_{A_{20}}$ = 0.1, and μ_{A_1} is varied between 0.001 and 1000 while keeping CV for A_1 constant at 10 %. Also included are s and CV for the data without transformation and after transformation using $\log_{10} x$.

Table 4. Obtained B_{opt}, standard deviation (s) and CV after transformation using $x^{B_{opt}}$. s and CV obtained without transformation. s and CV obtained after transformation using $\log_{10} x$. Data generated at different levels (using multiplication) as described in the text.

	Without transformation		**Transformed data using $x^{B_{opt}}$**			**Transformed data using $\log_{10} x$**	
μ_{A_1}	s	CV (%)	B_{opt}	s	CV (%)	s	CV (%)
0.001	0.0005	46.9	0.052	0.016	2.36	0.197	-6
0.01	0.005	46.9	0.051	0.018	2.31	0.197	-10
0.1	0.05	46.9	0.054	0.021	2.44	0.197	-19
1	0.5	46.9	0.053	0.024	2.39	0.197	-446
10	5	46.9	0.053	0.027	2.39	0.197	21
100	47	46.9	0.052	0.030	2.35	0.197	10
1000	469	46.9	0.051	0.033	2.30	0.197	7

Table 5. Obtained B_{opt}, standard deviation (s) and CV after transformation using $x^{B_{opt}}$. s and CV obtained without transformation. s and CV obtained after transformation using $\log_{10} x$. Data generated at different levels (using multiplication and division) as described in the text.

	Without transformation		**Transformed data using $x^{B_{opt}}$**			**Transformed data using $\log_{10} x$**	
μ_{A_1}	s	CV (%)	B_{opt}	s	CV (%)	s	CV (%)
0.001	0.00058	48.5	-0.042	0.025	1.89	0.197	-7
0.01	0.0058	48.4	-0.042	0.023	1.90	0.197	-10
0.1	0.058	48.4	-0.042	0.021	1.88	0.197	-20
1	0.58	48.5	-0.042	0.019	1.90	0.197	561
10	5.8	48.5	-0.041	0.017	1.87	0.197	19
100	58	48.4	-0.041	0.015	1.84	0.197	10
1000	584	48.5	-0.041	0.014	1.88	0.197	6.5

This points out that after transformation using $x^{B_{opt}}$, B_{opt} and CV for transformed data will be independent of the value of μ_{A_1}. This contrasts to using $\log_{10} x$ transformation where the standard deviation (s) of transformed data is independent of μ_{A_1}.

If measurement results can be approximated with a normal distribution and standard deviation, s, is independent of the measurand, a confidence interval for the uncertainty of a measurement result, x, can be calculated as

$$x - k \times s \quad \text{to} \quad x + k \times s \tag{9}$$

where k is the coverage factor.

If measurement results can be approximated with a normal distribution and the relative standard deviation, s_{rel}, is independent of the measurand, a confidence interval for the uncertainty of a measurement result, x, will be asymmetric since s will be different at the lower and the upper limit. For small s_{rel} the asymmetry is often neglected and the interval is calculated as

$$x - k \times s_{rel} \times x \quad \text{to} \quad x + k \times s_{rel} \times x \tag{10}$$

However, the lower limit, U_l, can be calculated as $U_l = x - k \times s_{rel} \times U_l$ which can be rearranged to $U_l = x/(1 + k \times s_{rel})$. Likewise, the upper limit, U_u, can be calculated as $U_u = x + k \times s_{rel} \times U_u$ which can be rearranged to $U_u = x/(1 - k \times s_{rel})$. Hence, the interval will be calculated as [7]

$$\frac{x}{1+k\times s_{rel}} \quad \text{to} \quad \frac{x}{1-k\times s_{rel}} \tag{11}$$

This expression is valid for $k \times s_{rel} < 1$.

Applying Eq. 9 on data transformed using $\log_{10} x$ will give a confidence interval of a measurement result in the transformed space. After back-transformation to the original space, a confidence interval around a measurement x can be calculated as

$$\frac{x}{10^{k \times s_{log_{10}}}} \quad \text{to} \quad x \times 10^{k \times s_{log_{10}}} \tag{12}$$

where $s_{log_{10}}$ is the standard deviation of transformed data. As an alternative, Eq. 12 can be expressed using the uncertainty factor [5, 6] [2].

Applying Eq. 11 on data transformed using $x^{B_{opt}}$ will give a confidence interval, after back-transformation to the original space, that is

$$\frac{x}{\left(1+k \times s_{rel,trans}\right)^{1/B_{opt}}} \quad \text{to} \quad \frac{x}{\left(1-k \times s_{rel,trans}\right)^{1/B_{opt}}} \tag{13}$$

where $s_{rel,trans}$ is the relative standard deviation of transformed data.

A confidence interval is constructed (Eq. 12 and 13) in the transformed space, symmetric around the mean value, which here, because of symmetry, is the same as the median. After back-transformation the quantiles represented by median and confidence limits are transformed to the same quantiles in the original space. Therefore, the back-transformed confidence interval covers the median with the intended probability and not the mean.

The skewness originated in possible measurement errors also introduces a bias, i.e the mean value obtained due to possible error-influenced-measurements may differ from the true value. The magnitude and sign of the bias can not be properly estimated without knowledge of the skewed distribution.

5.3 Implementation on small data sets representing experimental observations

In reality the number of data available for estimating the probability distribution experimentally is typically very limited compared to what ideally is needed. In chemical analysis it is very common to use control samples and control charts [16]. These charts are an important tool in the internal quality work. Furthermore, data for control samples will provide an estimate of the within-laboratory reproducibility that can be used to estimate measurement uncertainty in so called top-down approaches [10, 17, 18]. The number of data in typical control charts are after a few years in the order of 10^2 or more. Data from control charts can therefore be of value when estimating and handling skewness in probability distributions of the results. In order to

[2] It has also been suggested that Eq. 12 can be written as

$$\frac{x}{{}^{F}U} \quad \text{to} \quad x \times {}^{F}U$$

where ${}^{F}U$ is called the expanded uncertainty factor calculated as ${}^{F}U = 10^{k \times s_{log_{10}}}$ or ${}^{F}U = e^{k \times s_{log_e}}$ where s_{log_e} is the standard deviation of transformed data using the natural logarithm ($\log_e$).

investigate how well the described mathematical approach is applicable to small data sets, data sets with different distributions containing 10^2 data points were generated and transformed followed by optimization of B. Assuming t-distribution with n=10^2, intervals comprising 95 % of the data were then calculated as $\bar{x} \pm 1.98 \times s$ and back transformed to the original space, resulting in lower and upper limits of the intervals in the original space (denoted as $\bar{x} - 1.98s$ and $\bar{x} + 1.98s$, respectively). This was repeated 100 times for each distribution and the average and standard deviation of the quantiles $\bar{x} - 1.98s$ and $\bar{x} + 1.98s$ were calculated (denoted as $\bar{x}_{\bar{x}-1.98s}$, $\bar{x}_{\bar{x}+1.98s}$, $s_{\bar{x}-1.98s}$, and $s_{\bar{x}+1.98s}$, respectively). The results are summarized in Table 6 below giving the range of B_{opt} values obtained, and $\bar{x}_{\bar{x}-1.98s}$, $\bar{x}_{\bar{x}+1.98s}$, $s_{\bar{x}-1.98s}$, and$s_{\bar{x}+1.98s}$ for six different distributions.

Table 6. Performance of the transformation and optimization procedure when applied to small data (10^2 data points).

Data set	Way to generate data	Distribution of data	After transformation using $x^{B_{opt}}$ and back-transformation				
			Small data set (10^2 data points) (repeated 100 times)			Large data set (10^6 data points)	
			Range of B_{opt} values obtained	$\bar{x}_{\bar{x}-1.98s}$(1) $s_{\bar{x}-1.98s}$(3) (n=100)	$\bar{x}_{\bar{x}+1.98s}$(2) $s_{\bar{x}+1.98s}$(4) (n=100)	$\bar{x}-1.96s$ $F_{2.5\,\%}$ (%)	$\bar{x}+1.96s$ $F_{97.5\,\%}$ (%)
15	A_1 μ_{A_1} = 1, σ_{A_1} = 0.01	Normal distribution	-18 to 19	0.980 0.0021	1.020 0.0021	0.980 2.50	1.020 2.49
16	A_1 μ_{A_1} = 1, σ_{A_1} = 0.2	Normal distribution	-0.2 to 2.6	0.608 0.045	1.400 0.044	0.608 2.50	1.391 2.50
17	10^{A_1} μ_{A_1} = 1, σ_{A_1} = 0.01	Log-normal distribution	-8.0 to 8.5	9.549 0.049	10.468 0.053	9.559 2.51	10.462 2.51
18	10^{A_1} μ_{A_1} = 1, σ_{A_1} = 0.15	Log-normal distribution	-0.8 to 0.8	5.07 0.37	19.73 1.52	5.09 2.49	19.69 2.50
19	$A_1 \times A_2 \times A_3 \times \ldots\ldots \times A_5$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_5}$ = 1 $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_5}$ = 0.01	"Between" normal and log-normal distribution	-8.3 to 8.7	0.956 0.0039	1.045 0.005	0.957 2.49	1.044 2.48
20	$A_1 \times A_2 \times A_3 \times \ldots\ldots \times A_5$ $\mu_{A_1}, \mu_{A_2}, \ldots\ldots, \mu_{A_5}$ = 1 $\sigma_{A_1}, \sigma_{A_2}, \ldots., \sigma_{A_5}$ = 0.1	"Between" normal and log-normal distribution	-0.6 to 1.4	0.617 0.035	1.497 0.070	0.616 2.50	1.497 2.49

(1) Average of $\bar{x} - 1.98s$ for 100 data sets each containing 10^2 data points
(2) Average of $\bar{x} + 1.98s$ for 100 data sets each containing 10^2 data points
(3) Standard deviation of $\bar{x} - 1.98s$ for 100 data sets each containing 10^2 data points
(4) Standard deviation of $\bar{x} + 1.98s$ for 100 data sets each containing 10^2 data points

The six different distributions comprise two normal distributions with low and high standard deviations, two log-normal distributions with low and high standard deviations, and two "between" normal and log-normal distributions with low and high standard deviations. Table 6 also contains lower and upper limits of corresponding intervals (back transformed to the original space) for data containing 10^6 data points (denoted as $\bar{x} - 1.96s$ and $\bar{x} + 1.96s$, respectively). In addition, fraction of data points below $\bar{x} - 1.96s$ (denoted as $F_{2.5\,\%}$), and fraction of data points above $\bar{x} + 1.96s$ (denoted as $F_{97.5\,\%}$) are also given.

Several conclusions can be made from the results. In all cases (data set 15-20), $\bar{x}_{\bar{x}-1.984s}$ and $\bar{x}_{\bar{x}+1.984s}$ for 100 repeated data sets, each with 10^2 data points, will be equal to $\bar{x} - 1.96s$ and $\bar{x} + 1.96s$ calculated for a large data set with 10^6 data points as can be expected. Since the fraction of data below $\bar{x} - 1.96s$ and above $\bar{x} + 1.96s$ will both be 2.5 % after transformation for all six cases, the fraction of data below $\bar{x}_{\bar{x}-1.98s}$ and above $\bar{x}_{\bar{x}+1.98s}$ will both also be 2.5 % after transformation for all six cases.

With these small data sets (containing 10^2 data points), B_{opt} values will vary a lot from data set to data set, and B_{opt} values < 0 and > 1 can be obtained. Especially for data with small standard deviations (data set 15, 17 and 19) i.e. when normal and log-normal distributions are very similar, large variations in B_{opt} values will occur. Hence, obtained B_{opt} from small data will not reflect the real B_{opt} value although the estimation will improve with increasing standard deviation. Indeed, it has been pointed out in the literature that departure from normality have to be quite large in order to demonstrate non-normality [19].

Today this issue is typically handled by making a judgement if data can be considered to have a normal distribution or a log-normal distribution [5]. Looking at a histogram of the data can help making such judgement.

Skewness of within-laboratory reproducibility of real data is illustrated in Example *6.1 Study of the distribution of results from determination of sulfur in gas samples using gas chromatography and chemiluminescence*. Here data sets containing around 700 data points are used given a somewhat better estimate of the real B_{opt} of the method. CV is 15 % which is on the border when skewness is becoming important to consider.

5.4 Comparison of transformations using x^B with B approaching 0 and $\log_{10} x$

From the results and the discussion above it appears as transformation using x^B with B approaching 0 and transformation using $\log_{10} x$ will be analogous transformations. This is demonstrated in two examples below.

In the first example confidence intervals were calculated according to

$$\frac{x}{(1+k\times s_{rel,trans})^{1/B}} \quad \text{to} \quad \frac{x}{(1-k\times s_{rel,trans})^{1/B}} \tag{14}$$

with B approaching 0 and compared with confidence intervals calculated according to Eq. 12:

$$\frac{x}{10^{k\times s_{log10}}} \quad \text{to} \quad x \times 10^{k\times s_{log10}} \tag{12}$$

using k equal to 1.96 for data in data set 1 to 14 in Table 2. For all data sets identical intervals were obtained when B approaches 0. This is further illustrated in Fig. 4 showing the ratio of the lower limits of the two confidence intervals and the ratio of the upper limits of the confidence intervals when B approaches 0 using data in data set 7.

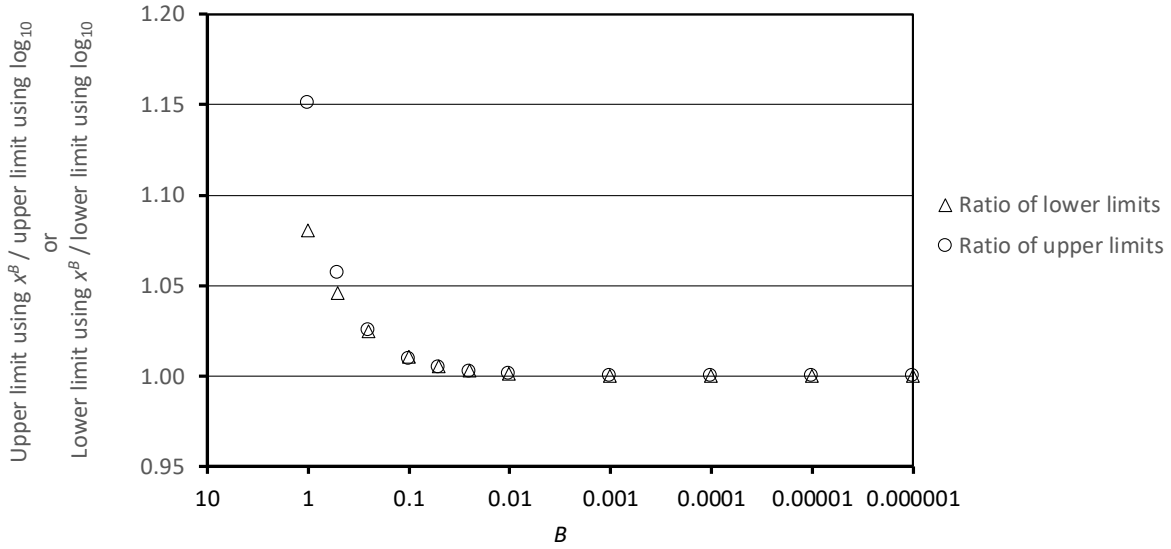

Figure 4. Ratio of the lower limits of the two confidence intervals and the ratio of the upper limits of the confidence intervals when B approaches 0 using data in data set 7.

Hence, identical confidence intervals will be obtained with Eq. 14 with B approaching 0 and Eq. 12, i.e. the two different transformations will be analogous with B approaching 0.

Another example is given below in Example *6.2 Calculation of sampling uncertainty using the "duplicate" method and ANOVA*.

From these examples it is apparent that log-normal distributed data can be processed using transformation by $\log_e$ (or $\log_{10}$) as well as by x^B with B approaching 0 (for instance using B equal to 0.0001).

5.5 Possibilities to obtain B_{opt} from modelling

As discussed above, it is typically not feasible to obtain values for B_{opt} from experimental data since these data sets typically contain too few data. As an alternative, it is here suggested to obtain B_{opt} from modelling of the uncertainty where large data sets can be simulated. This is demonstrated below in Example *6.3 Calculation of measurement uncertainty for determination of organophosporus pesticides in bread*. Hence, it is needed to have a model equation with all all input quantities well described.

5.6 Adding additional uncertainty contributions

It can be of interest to be able to combine two different uncertainty components with different distributions. As an example, an uncertainty contribution handling bias or sampling can be added to an uncertainty describing precision. Sometimes bias is included in the uncertainty instead of correcting the result [20-22]. The handling of bias is still under discussion and different opinions exist [20, 21], but this issue will not be discussed further here.

Asymmetric distributions cannot be combined as naturally as standard uncertainties in the first-order Taylor series approximation (GUM 5.1.2) [1]. Often asymmetry can be cured by log transformation followed by standard treatment and back transformation and in cases when this is unsufficient the B-transformation presented here can solve more general cases. These transformation methodologies are not treated in GUM, the only solution for combining assymmetric distributions that is presented is Monte Carlo simulation techniques (GUM supplement JCGM 101 [11]).

An example of a procedure for adding an uncertainty that can be assumed to have a normal distribution to an uncertainty with a skewed distribution is illustrated schematically in Fig. 5.

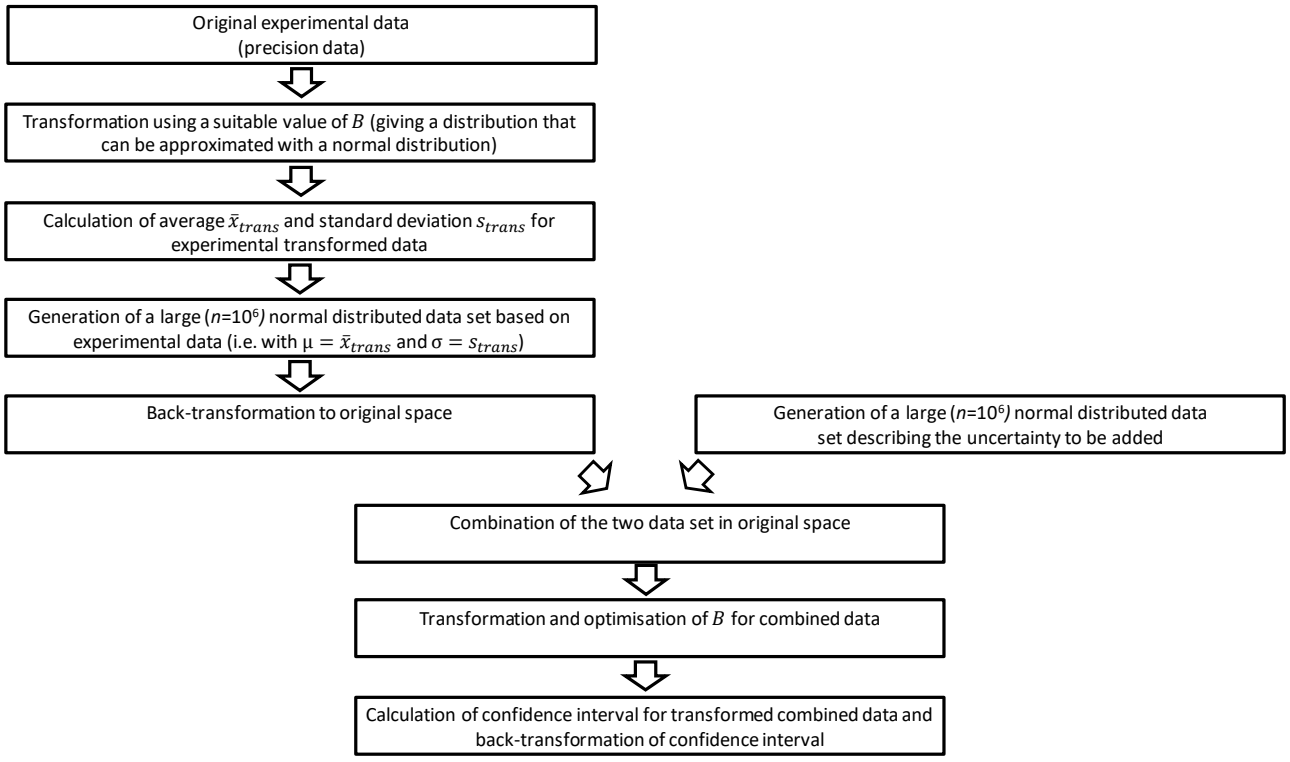

Figure 5. Procedure for adding an additional uncertainty component to precision uncertainty.

Based on experimental data, a large normal distributed data set (n=10^6) is generated in the transformed space. This data set is then back-transformed to the original space. For the uncertainty to be added, a large normal distributed data set (n=10^6) with μ=1 and σ equal to the standard uncertainty is generated in the original space. Data from the two data sets are then multiplied or added based on judgements how they influence the measurand (in the original space) giving a new

data set for the combined uncertainty. Finally, transformation and optimization of B is performed for the combined data set.

Uncertainty components with skewed distributions can also be added. In this case, a B_{opt}-value should be obtained that transforms the uncertainty distribution to be added to a symmetric distribution. A large normal distributed data set is then generated in this transformed space and transformed back to the original space, where it is combined with a large data set representing the original experimental data.

6 Examples

Different applications of transformation using x^B when evaluating measurement uncertainties in chemical analysis are given in three examples below. An overview of the examples is given in Table 7.

Table 7. Overview of three different examples demonstrating application of transformation using x^B.

Example	Title	Issues that are illustrated
6.1	Study of the distribution of results from determination of sulfur in gas samples using gas chromatography and chemiluminescence	Skewness of within-laboratory reproducibility data. Comparison of transformations using x^B with B approaching 0 and $\log_{10} x$.
6.2	Calculation of sampling uncertainty using the "duplicate" method and ANOVA	Transformation prior to ANOVA calculations. Comparison of transformations using x^B with B approaching 0 and $\log_{10} x$. Comparison of uncertainty intervals calculated using different transformations.
6.3	Calculation of measurement uncertainty for determination of organophosporus pesticides in bread	Possibility to obtain B from modelling. Comparison of confidence intervals calculated using different transformations.

6.1 Study of the distribution of results from determination of sulfur in gas samples using gas chromatography and chemiluminescence

Application (type of data): Data from control samples reflecting within-laboratory reproducibility.

Introduction: In order to illustrate skewness of data from real measurements, data for two control samples used when determining sulfur (S) in gas samples using gas chromatography and chemiluminescence were utilized. Hence, these data reflect within-laboratory reproducibility.

Calculations: The two data sets were transformed using x^B using an optimized B. Confidence intervals (95 %) were then calculated in the transformed space and back-transferred to the original space. The results are shown in Table 8. For comparison, results when transforming using $\log_{10} x$ and without transformation are given.

Table 8. Study of skewness of within-laboratory reproducibility data [1)] for determination of sulfur (S) in gas samples using gas chromatography and chemiluminescence.

Concentration level (mg/kg)	**Number of data**	**Empirical 2.5 and 97.5 percentile in original space (mg/kg)**	**Skewness in the original space**	**CV (%) in the original space**	**Optimized B**	**Confidence interval (95 %) calculated in original space (mg/kg)**	**Confidence interval (95 %) calculated in $x^{B_{opt}}$-transformed space and back-transformed to original space (mg/kg)**	**Confidence interval (95 %) calculated in $\log_{10}$-transformed space and back-transformed to original space (mg/kg)**
9	744	6.6 11.9	0.42	15.0	0.41	6.3 11.6	6.5 11.8	6.6 11.9
19	685	14.2 25.3	0.18	14.9	0.69	13.6 24.9	13.8 25.1	14.1 25.6

1) New control samples were prepared when the previous control sample was finished. The measured concentrations have been corrected to account for the difference in nominal concentrations between the control samples.

Discussion: CV in the original space is 15 % for both concentration levels i.e. on the border when skewness is becoming important to consider. The two optimized B values are 0.41 and 0.69. Considering the difficulty in estimating B_{opt} from small data sets (here around 700 data are used), this indicates that a real B_{opt} is around 0.5, i.e. the distribution is somewhere "between" a normal and a log-normal distribution. Confidence intervals (95 %) obtained without transformation, and with transformation using $x^{B_{opt}}$ and $\log_{10} x$ are fairly similar, and compares well with the empirical 2.5 and 97.5 percentiles in the original space. This confirms the rule of thumb that for CV up to 15-20 % the asymmetry is not critical.

6.2 Calculation of sampling uncertainty using the "duplicate" method and ANOVA

Application (type of data): Replicate data from sampling of heterogeneous sampling targets.

Introduction: In order to evaluate measurement uncertainty for the sampling step, it is often possible to assume that the sampling uncertainty is dominated by the repeatability of the sampling step [5]. The sampling repeatability and the analysis repeatability can be obtained from measurements of duplicate samples using ANOVA. This is sometimes called the "duplicate method" and is described in the Eurachem/CITAC Guide Measurement uncertainty arising from sampling - A guide to methods and approaches [5]. When sampling solid material, a log-normal distribution is sometimes encountered or assumed, and the results are therefore transformed using $\log_e$ (or $\log_{10}$) subsequent to evaluation using ANOVA. This has been exemplified using results from measurements of the lead (Pb) content in contaminated top soil (see Example A2 in the Eurachem/CITAC Guide). An elaborated description and discussion of the experiments, and calculations and results when using transformation based on $\log_e$, are available in the Eurachem/CITAC Guide.

Calculations: In the original example, the between-target variability was found to have a positive skewness similar to a log-normal distribution. It was argued that the sampling variability and the between-target variability were controlled mainly by heterogeneity of the analyte. Hence, sampling variability could also be assumed to have a log-normal distribution, and this motivated the use of log-transformation above. A histogram illustrating the between-target variability that has a CV of 138 % is shown in Fig. 6.

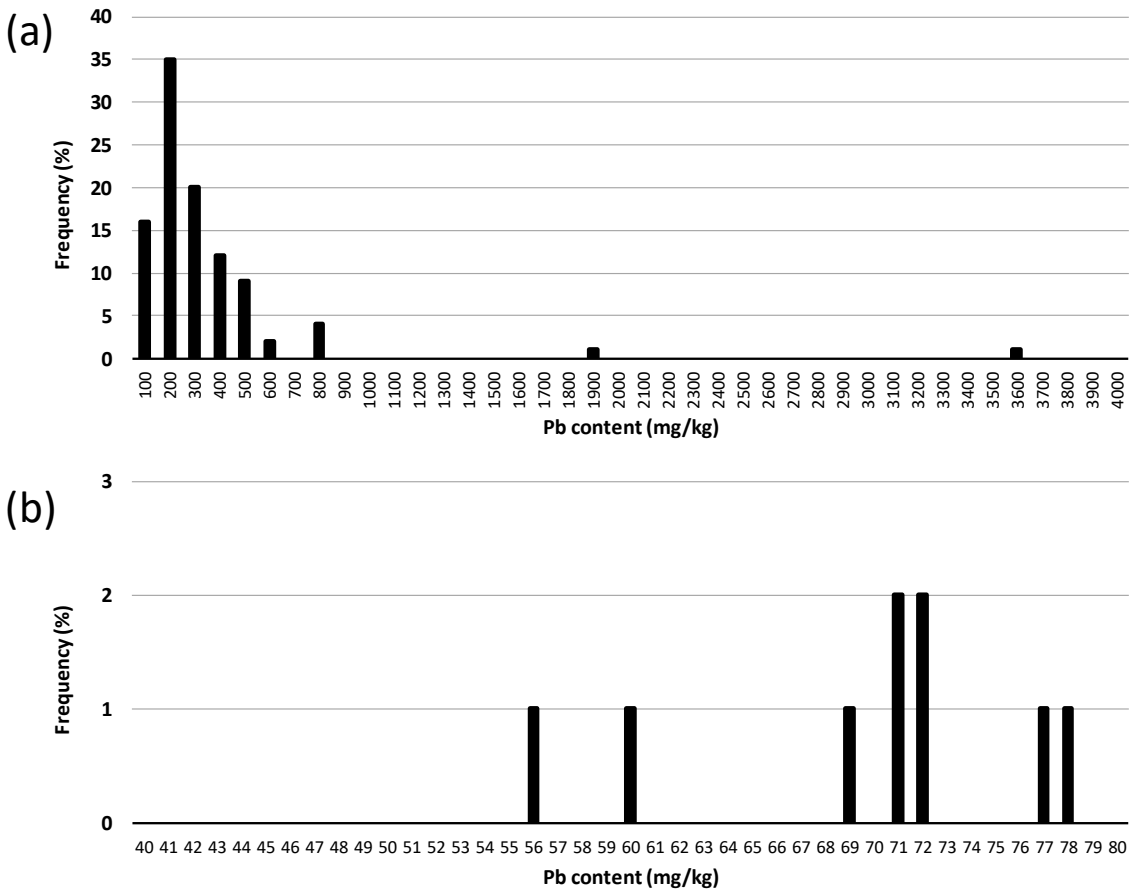

Figure 6. (a) Histogram showing the between-target variability of Pb content in contaminated top soil (100 samples). (b) Enlarge lower part of the histogram.

Here an optimized B-value of -0.306 was obtained for the data set. Confidence intervals covering 95 % of the data were then calculated in the 1) original space, 2) in the $x^{B_{opt}}$-transformed space and back-transformed to the original space, and 3) in the $\log_{10}$-transformed space and back-transformed to the original space. The results are given in Table 9. Also included are the empirical 2.5 and 97.5 percentile in the original space.

Table 9. Empirical 2.5 and 97.5 percentile in the original space, and confidence intervals calculated in original space, and in $x^{B_{opt}}$-transformed space and back-transformed to original space, and in $\log_{10}$-transformed space and back-transformed to original space.

Empirical 2.5 and 97.5 percentile in the original space (mg/kg)		**Confidence interval (95 %) calculated in original space (mg/kg)**	**Confidence interval (95 %) calculated in $x^{B_{opt}}$-transformed space and back-transformed to original space (mg/kg)**	**Confidence interval (95 %) calculated in $\log_{10}$-transformed space and back-transformed to original space [1] (mg/kg)**
2.5 percentile	**97.5 percentile**			
60-69	800-1900	-508 to 1091	61 to 1133	49 to 893

1) Confidence interval will be identical if calculated in $\log_e$-transformed space and back-transformed to original space.

Repeatability data (available in the Eurachem/CITAC guide) were then transformed using $x^{B_{opt}}$ and ANOVA was used to obtain standard deviations for the sampling step, analysis step and complete measurement in the transformed space. Finally, uncertainty intervals in the original space for the sampling step, analysis step and complete measurement around a nominal value of 300 mg/kg were calculated according to Eq. 14:

$$\frac{x}{\left(1+k\times s_{rel,trans}\right)^{1/B}} \quad \text{to} \quad \frac{x}{\left(1-k\times s_{Srel,trans}\right)^{1/B}} \qquad (14)$$

with B equal to -0.306 i.e. B_{opt}. The results are given in Table 10.

Table 10. Comparison of uncertainty intervals (mg/kg) using a coverage factor of 2 around a nominal measured value of 300 mg/kg for sampling step, analysis step and complete measurement evaluated using transformation based on $x^{B_{opt}}$ with B_{opt} equal to -0.306, $x^{0.0001}$, and $\log_e$. It is here assumed that repeatability is the dominating contribution to uncertainty both regarding sampling and analysis.

	Transformation based on		
	$x^{B_{opt}}$ with B_{opt} equal to -0.306	**x^B with B equal to 0.0001**	**$\log_e x$**
Sampling step	88 to 730	115 to 781	115 to 781
Analysis step	263 to 341	268 to 336	268 to 336
Complete measurement	87 to 734	114 to 786	114 to 786

For comparison, uncertainty intervals are also given in Table 10 when using transformations based on x^B with B equal to 0.0001 and $\log_e$ (as in the Eurachem/CITAC Guide). In the latter case uncertainty intervals are calculated according to

$$\frac{x}{10^{k \times s_{log_e}}} \quad \text{to} \quad x \times 10^{k \times s_{log_e}} \tag{15}$$

Discussion: In the Eurachem/CITAC guide, it was argued that the sampling variability and the between-target variability were controlled mainly by heterogeneity of the analyte. Hence, sampling variability could also be assumed to have the same distribution, as the between target variability. For between-target data it was found that transformation using $x^{B_{opt}}$ with B_{opt} equal to -0.306 will result in a confidence interval that best corresponds to the empirical 2.5 and 97.5 percentile.

The uncertainty intervals calculated using transformation based on $x^{-0.306}$ are somewhat different compared to using $\log_e$. Note also that transformation using $\log_e$ and $x^{0.0001}$ results in the same uncertainty intervals confirming that the two different transformations will be analogous with B approaching 0.

6.3 Calculation of measurement uncertainty for determination of organophosporus pesticides in bread

Application (type of data): Large measurement uncertainty evaluated based on a measurement model

Introduction: Determination of pesticides in many sample types is known to have large measurement uncertainties. Calculation of measurement uncertainty for determination of organophosporus pesticides in bread based on a modelling approach is described in the Eurachem/CITAC Guide CG4 Quantifying uncertainty in analytical measurement (Example A4) [10] giving a relative expanded uncertainty of 68 %.

Calculations: The modelling equation for the concentration of pesticide, C, is given by

$$C = \frac{I_p \times C_{ref} \times V_{dil}}{I_{ref} \times m \times R} \times F_{hom} \times F_I \tag{16}$$

and the input quantities are defined in Table 11.

Table 11. List of input quantities, their distribution, value, and standard deviation or halfwidth.

Input quantity	Definition	Distribution	Mean	Standard deviation (s) or halfwidth (hw)
I_p	Intensity of sample peak height	Normal	1	0.005 (s)
I_{ref}	Intensity of reference standard peak height	Normal	1	0.005 (s)
C_{ref}	Concentration of reference standard	Rectangular	1	0.1 (hw)
m	Mass of sub-sample	Normal	1	0.0005 (s)
V_{dil}	Final volume of extract	Normal	1	0.0025 (s)
R	Recovery	Rectangular	1	0.25 (hw)
F_{hom}	Correction factor for sample heterogeneity	Normal	1	0.2 (s)
F_I	Correction factor for within-laboratory reproducibility (intermediate precision)	Normal	1	0.2 (s)

In order to calculate B_{opt} for this uncertainty model, a data set with 10^6 data was generated. For simplicity, the means of all input quantities have been set to 1. Standard deviations (for normal distributions) or halfwidths (for rectangular distributions) have been set to reasonable values aiming to give a relative expanded uncertainty close to the evaluated uncertainty in the original example. (In the original example data were not given for all the individual input quantities.) This resulted in B_{opt} equal to 0.32. Uncertainty interval around a nominal value of 1 was then calculated using Eq. 13 and this is given in Table 12 below.

Table 12. Comparison of uncertainty intervals (mg/kg) using a coverage factor of 1.96 around a nominal measured value of 1.

	Interval calculated as $\pm k \times u_{c,rel} \times x$	**Without transformation**	**Transformation based on**	
			$x^{B_{opt}}$ with B_{opt} equal to 0.32	**$\log_{10} x$**
Uncertainty interval	0.37 to 1.63	0.61 to 2.83	0.55 to 2.08	0.51 to 1.94

For comparison, uncertainty interval based on transformation using $\log_{10} x$ calculated using Eq. 12, and without transformation using Eq. 11 are also included. Also given is uncertainty interval calculated as $\pm k \times u_{c,rel}$ where $u_{c,rel}$ is the combined relative standard uncertainty of the input quantities.

Discussion: When having a model equation including knowledge about the distributions of the input quantities, a reliable B_{opt} value can be calculated using simulations. This B_{opt} can then be used for new data and is not dependent on empirical data set size. A B_{opt} value of 0.32 suggests that the distribution of results is somewhere "between" a normal and a log-normal distribution. There is a clear difference in the obtained uncertainty intervals when a transformation procedure is used and when not used. There is also a small but apparent difference in the obtained uncertainty intervals when transformations based on $\log_{10} x$ and $x^{B_{opt}}$ are used.

7 Summary and conclusions

The GUM document *Evaluation of measurement data – Guide to the expression of uncertainty in measurement* [1, 2], is the fundamental reference document when evaluating measurement uncertainty. Confidence of uncertainty intervals is here based on the Central Limit Theorem. This is based on a linear combination of random variables resulting in a normal symmetric distribution of results. The expanded uncertainty, U is then given as a symmetric interval based on a normal distribution around the measured value, x[3]. However, this theorem is not applicable to multiplicative combination of random variables that is common in many measurements.

In many measurements multiplicative combinations are present resulting in asymmetric distributions of the results. At relative standard uncertainties less than 15 %[4], these distributions can

[3] In section G5.2 of GUM asymmetry is discussed: *The alternative is to give an interval that is symmetric in probability (and thus asymmetric in U): the probability that Y lies below the lower limit $y - U_-$ is equal to the probability that Y lies above the upper limit $y + U_+$. But in order to quote such limits, more information than simply the estimates y and $u_c(y)$ [and hence more information than simply the estimates x_i and $u(x_i)$ of each input quantity X_i] is needed.*

[4] The value is approximate 15 – 20 %.

typically be approximated with a normal distribution to find proper confidence intervals. At larger uncertainties, the skewness needs to be considered in order to evaluate the uncertainty interval of a measurement result. Handling of such large asymmetric uncertainties are outside the scope of GUM, and it is not straightforward how to handle them. It is common to transform data by using $\log_e$ or $\log_{10}$. Transformation by taking the square root is sometimes used for instance in microbiology [23]. Using a normal distribution approximation, an expanded uncertainty can then be estimated in the transformed space, and an uncertainty interval for a measurement result can be calculated around the transformed value. This interval is then transformed back to the original space giving an asymmetric uncertainty interval.

In this project a broader range of transformations is investigated by using the transformation

$$x_{trans} = x^B \tag{3}$$

where x_{trans} and x are the transformed and original data, respectively, and B is a parameter that is chosen for each specific distribution. Many skewed distributions can be transformed to a more symmetric distribution by selecting a value of B that gives minimum skewness. Skewness indicates the direction and relative magnitude of a distribution's deviation from a symmetric distribution, i.e. is a measure of asymmetry. The transformed symmetric data is then assumed to be normal and can be processed with standard statistical procedures, and symmetric confidence intervals expressing measurement uncertainty of a result can be calculated. The interval is then transformed back to the original space giving an asymmetric uncertainty interval. Using different values of B different distributions will be transformed to symmetric distributions as illustrated in Fig. 7.

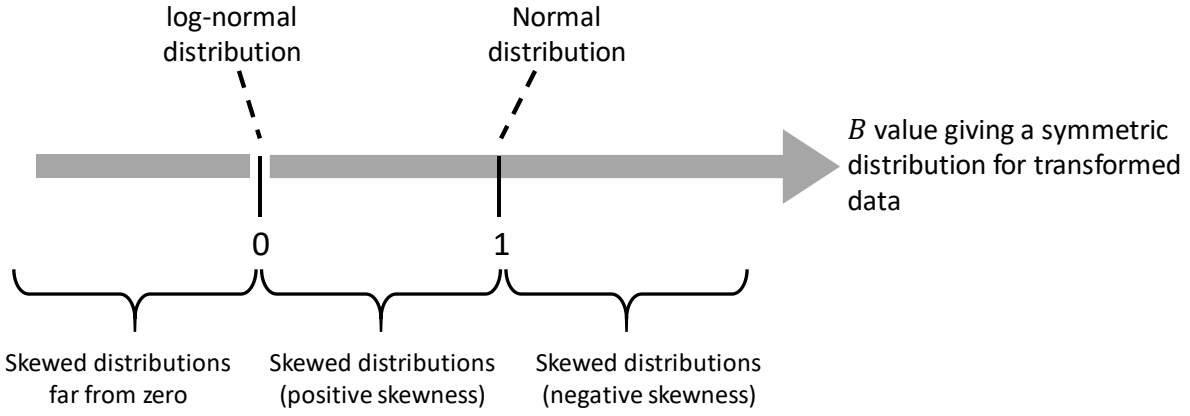

Figure 7. Illustration of B values used to transform different distributions to a symmetric distribution using the transformation x^B

Two important values of B that can serve as reference points on a "B scale" are 0 and 1. With B equal to 1 no transformation will occur. Using B close to zero will be analogous to taking the logarithm of the values. With B values somewhere between 0 and 1, skewness of other types of distributions can be transformed to a normal distribution approximation. Note that using B equal to 0.5 will be equivalent to taking the square root that is sometimes used in microbiology [23]. Skewed distributions far from zero is best transformed with B less than zero.

To summarize the simulation results:

- with $B > 1$ distributions with a negative skewness will be transformed to symmetric distributions
- with $0 < B < 1$ distributions with a positive skewness will be transformed to symmetric distributions
- with $B < 0$ positively skewed distributions far from zero will be transformed to symmetric distributions

For existing data, estimation of B that will result in transformation to a symmetric distribution can be performed by searching for a B value that will minimize skewness of transformed data. The estimation of an optimized B value will improve with an increased number of data, and an increased relative standard deviation of the data. However, for small data series (for instance with n in the order of 10^2), it is not possible to obtain a relevant estimation of an optimized B. Hence, in reality it is difficult to estimate an optimized B from experimental data and use it for new data. However, the obtained optimized B can still be used to process the existing experimental data.

For measurements where the model equation can well describe the output results Monte Carlo simulation can generate big data sets. From these data sets reliable B values can be estimated.

When a relevant measurement model cannot be obtained for asymmetric distributions it seems reasonable to use judgement and assume $B = 0.5$ that is equal to a square root transformation or assume B close to 0 (for instance 0.0001) that is analogous to a logarithmic transformation. Alternatively, a general agreed value of B could be used for a specific measurement procedure.

Without any other information of a proper value for B, it seems as it is most reasonable to assume B equal to 1 when CV is less than 15 % (i.e. no transformation of the data is performed). For these relatively low CV the skewness is not critical for the measurement uncertainty intervals.

Uncertainty intervals for a result x (in the original space) when transformation is performed using $x^{B_{opt}}$ can be calculated as:

$$\frac{x}{\left(1+k\times s_{rel,trans}\right)^{1/B_{opt}}} \quad \text{to} \quad \frac{x}{\left(1-k\times s_{rel,trans}\right)^{1/B_{opt}}} \tag{13}$$

where $s_{rel,trans}$ is the relative standard deviation of transformed data, k is the coverage factor taken from the t-distribution giving the required confidence level, and B_{opt} is a B that will result in a symmetric distribution of the transformed data. It is here assumed that the relative standard deviation in the original space, s_{rel}, is independent of the measurand level which is often the case for many measurements at levels well above the limit of quantification.

This approach can be compared to other ways of calculating uncertainty intervals when s_{rel} is independent of the measurand level and a summary is given in Table 13.

Table 13. Different ways of calculating uncertainty intervals for a measurement result, x, when the relative standard deviation in the original space, s_{rel}, is independent of the measurand level.

Using no transformation i.e. neglecting skewness in measurement results		**Using transformation to handle skewness in measurement results**	
Neglecting that *s* will be different at lower and upper limit	**Taking into account that *s* will be different at lower and upper limit (giving an assymetric interval)**	**Transformation using $x^{B_{opt}}$**	**Transformation using $\log_{10} x$**
$x - k \times s_{rel} \times x$ to $x + k \times s_{rel} \times x$	$\frac{x}{1+k \times s_{rel}}$ to $\frac{x}{1-k \times s_{rel}}$ (valid for $k \times s_{rel} < 1$)	$\frac{x}{(1+k \times s_{rel,trans})^{1/B_{opt}}}$ to $\frac{x}{(1-k \times s_{rel,trans})^{1/B_{opt}}}$	$\frac{x}{10^{k \times s_{log_{10}}}}$ to $x \times 10^{k \times s_{log_{10}}}$

s = standard deviation in the original space

s_{rel} = relative standard deviation in the original space

$s_{rel,trans}$ = relative standard deviation in the transformed space after transformation using $x^{B_{opt}}$

$s_{log_{10}}$ = standard deviation in the transformed space after transformation using $\log_{10} x$

k = coverage factor

It is also possible to combine uncertainty contributions with different asymmetry. This is performed by 1) finding an optimized *B* for each uncertainty contribution, 2) generation of a large normal distributed data set in the transformed space for each uncertainty contribution, 3) back-transformation of the data sets to the original space, and 4) combination of the data sets in the original space. Finally, transformation and optimization of *B* for the final data can be calculated.

8 Acknowledgements

Vinnova Sweden's Innovation Agency is greatly acknowledged for financial support.

Borealis AB, Stenungsund (Sweden), is acknowledged for kindly providing large data sets from control sample measurements.

9 References

[1] JCGM 100:2008, Evaluation of measurement data - Guide to the expression of uncertainty in measurement Joint Committee for Guides in Metrology 2008
[2] JCGM 104:2009, Evaluation of measurement data - An introduction to the “Guide to the expression of uncertainty in measurement” and related documents, Joint Committee for Guides in Metrology 2009.
[3] E. Limpert, W.A. Stahel, M. Abbt, Log-normal Distributions across the Sciences: Keys and Clues, BioScience, 51 (2001) 341-352.
[4] E. Limpert, W.A. Stahel, Problems with Using the Normal Distribution – and Ways to Improve Quality and Efficiency of Data Analysis, PLOS ONE, 6 (2011) e21403.
[5] M.H. Ramsey, Why do we need the uncertainty factor?, Analytical Methods, 11 (2019) 2105-2107.
[6] M.H. Ramsey, S.L.R. Ellison, Combined uncertainty factor for sampling and analysis, Accreditation and Quality Assurance, 22 (2017) 187-189.
[7] A. Williams, Calculation of the expanded uncertainty for large uncertainties using the lognormal distribution, Accreditation and Quality Assurance, 25 (2020) 335-338.
[8] Eurachem / CITAC Guide Measurement uncertainty arising from sampling - A guide to methods and approaches, M.H. Ramsey, S.L.R. Ellison, P. Rostron (Eds.), 2019 (available at www.eurachem.org).
[9] B. Magnusson, M. Krysell, E. Sahlin, T. Näykki, Uncertainty from sampling - A Nordtest handbook for sampling planners on sampling quality assurance and uncertainty estimation (NT TR 604) 2020 (available at www.nordtest.info/wp/).
[10] Eurachem / CITAC CG4 Guide Quantifying Uncertainty in Analytical Measurement, S.L.R. Ellison, A. Williams (Eds.), 2012 (available at www.eurachem.org).
[11] JCGM 101:2008, Evaluation of measurement data - Supplement 1 to the "Guide to the expression of uncertainty in measurement" - Propagation of distributions using a Monte Carlo method Joint Committee for Guides in Metrology 2008
[12] G.E.P. Box, J.S. Hunter, W.G. Hunter, Statistics for Experimenters: Design, Innovation, and Discovery, Wiley 2005.
[13] Engineering statistics handbook NIST/SEMATECH https://www.itl.nist.gov/div898/handbook/, 2012.
[14] R.D.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2011.
[15] RANOVA2 (A stand-alone program running in Microsoft Excel), Royal Society of Chemistry.
[16] B. Magnusson, H. Hovind, M. Krysell, U. Lund, I. Mäkinen, Internal Quality Controll - Handbook for Chemical Laboratories (NT TR 569), 2018 (available at www.nordtest.info/wp/).
[17] B. Magnusson, T. Näykki, H. Hovind, M. Krysell, E. Sahlin, Calculation of measurement uncertainty in environmental laboratories (NT TR 537), 2017 (available at www.nordtest.info/wp/).
[18] Eurolab Technical Report No. 1/2007 Measurement uncertainty revisited: Alternative approaches to uncertainty evaluation, 2007 (available at www.eurolab.org/pubs-techreports).
[19] J.N. Miller, Are my data normal?, Analytical Methods, 9 (2017) 5847-5850.
[20] B. Magnusson, S.L.R. Ellison, Treatment of uncorrected measurement bias in uncertainty estimation for chemical measurements, Analytical and Bioanalytical Chemistry, 390 (2008) 201-213.
[21] R. Frenkel, I. Farrance, T. Badrick, Bias in analytical chemistry: A review of selected procedures for incorporating uncorrected bias into the expanded uncertainty of analytical measurements and a graphical method for evaluating the concordance of reference and test procedures, Clinica Chimica Acta, 495 (2019) 129-138.

[22] JCGM 200:2012, International vocabulary of metrology – Basic and general concepts and associated terms (VIM), Joint Committee for Guides in Metrology 2008
[23] T. Slapokas, Proficency Testing - Drinking Water Microbiology, Swedish Food Agency Uppsala, Sweden, 2020 (available at www.livsmedelsverket.se/).

Through our international collaboration programmes with academia, industry, and the public sector, we ensure the competitiveness of the Swedish business community on an international level and contribute to a sustainable society. Our 2,800 employees support and promote all manner of innovative processes, and our roughly 100 testbeds and demonstration facilities are instrumental in developing the future-proofing of products, technologies, and services. RISE Research Institutes of Sweden is fully owned by the Swedish state.

I internationell samverkan med akademi, näringsliv och offentlig sektor bidrar vi till ett konkurrenskraftigt näringsliv och ett hållbart samhälle. RISE 2 800 medarbetare driver och stöder alla typer av innovationsprocesser. Vi erbjuder ett 100-tal test- och demonstrationsmiljöer för framtidssäkra produkter, tekniker och tjänster. RISE Research Institutes of Sweden ägs av svenska staten.

RISE Research Institutes of Sweden AB
Box 857, 501 15 BORÅS, SWEDEN
Telephone: +46 10-516 50 00
E-mail: info@ri.se, Internet: www.ri.se

Chemical Problem Solving
RISE Report 2021:07
ISBN: 978-91-89167-89-6