Accreditation and Quality Assurance (2020) 25:107–119
https://doi.org/10.1007/s00769-019-01417-4

GENERAL PAPER

Confidence intervals for robust estimates of measurement uncertainty

Peter D. Rostron[1] · **Tom Fearn**[2] · **Michael H. Ramsey**[1]

Received: 5 April 2019 / Accepted: 2 December 2019 / Published online: 4 February 2020

Abstract
Uncertainties arising at different stages of a measurement process can be estimated using analysis of variance (ANOVA) on duplicated measurements. In some cases, it is also desirable to calculate confidence intervals for these uncertainties. This can be achieved using probability models that assume the measurement data are normally distributed. However, it is often the case in practice that a set of otherwise normally distributed measurement values is contaminated by a small number of outlying values, which may have a disproportionate effect on the variances calculated using the 'classical' form of ANOVA. In this case, robust ANOVA methods are able to provide variance estimates that are much closer to the parameters of the underlying normal distributions. A method using bootstrapping to calculate confidence intervals from robust estimates of variances is proposed and evaluated and is shown to work well when the number of outlying values is small. The method has been implemented in a visual basic program.

Keywords Measurement uncertainty · Bootstrap · Confidence interval · Confidence limit · Robust ANOVA · Duplicate method

Introduction

The importance of the estimation of measurement uncertainty, including the uncertainty from sampling, by either modelling or empirical methods, is now well established [1]. One empirical method, recommended because of its cost-effectiveness and simple application, is the *duplicate method*. In this method, the processes of sampling and analysis are duplicated following pre-determined protocols applied to a number of different sampling targets. This allows the variances at each of these two stages to be separated and estimated by analysis of variance (ANOVA).

In some cases, it is useful for the analyst or researcher to quantify the reliability of variance estimates, expressing this reliability as confidence intervals for the variance estimates from the ANOVA. For example, the researcher might estimate uncertainties of measurements on the same targets using more than one analytical method. Confidence limits (i.e. the extremes of the confidence interval) on these uncertainties would indicate whether the uncertainty estimates themselves were significantly different between different analytical methods. A further potential application is the comparison of analyte heterogeneity in materials, which can also be estimated using the duplicate method [2].

The uncertainties and CI's can be estimated using a probability model, if the nature of the distribution of measurements can be assumed. An example of this approach is provided by Lyn et al. [3]. This study demonstrated two important characteristics of uncertainties estimated by the duplicate method:

1. As would be expected, the width of the confidence intervals around the uncertainty estimates decreases as the number of sampling targets (n) increases.
2. It is suggested that for many applications the decrease in widths of the confidence intervals obtained when $n > 8$ may not justify the costs of obtaining the additional duplicated measurements.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00769-019-01417-4) contains supplementary material, which is available to authorized users.

✉ Peter D. Rostron
pr52@outlook.com

Michael H. Ramsey
m.h.ramsey@sussex.ac.uk

[1] School of Life Sciences, University of Sussex, Falmer, Brighton BN1 9QG, UK

[2] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

In this case, the validity of estimates of the uncertainties and their CI's from the classical ANOVA method depends on the assumption that the data are normally distributed. However, many data sets obtained by experiment contain a small proportion of outlying values that may have a disproportionately large effect on these estimates. One solution to this problem is the use of robust statistical methods, where estimates are adjusted to accommodate up to 10 % of outlying values [4]. Robust ANOVA has been widely used for the estimation of measurement uncertainty, especially that arising from the primary sampling process [5, 6].

Previously, there has not been a method available to calculate the confidence interval on a value of uncertainty estimated using robust ANOVA. The overall purpose of this paper is to propose such a method, using a bootstrapping approach, applied with a new computer program CI-RANOVA, and to evaluate its performance for both normal and contaminated data. One practical problem with doing this arises from the fact that CI-RANOVA is, for ease of use by the analytical community, implemented in Excel. This is fine for routine use but too slow to permit the extensive simulations needed to accurately estimate coverage probabilities for the confidence intervals. To enable this to be done, the method was also implemented in Matlab. Having two independent, in the sense that they were coded on different platforms by different programmers, implementations also enabled a validation of the CI-RANOVA coding, and to this end simulations were run using both versions.

The objectives of this work are as follows:

1. Validate the confidence limits produced by a bootstrapping method against those calculated by analytical (mathematical) formulas from the results of a classical ANOVA, using multiple simulations of normally distributed data.
2. Validate the implementation of the bootstrapping method in CI-RANOVA, by comparing the confidence limits produced by CI-RANOVA with those produced by the Matlab implementation.
3. Validate the bootstrapping method applied by CI-RANOVA when up to 10 % outlying values are included.

The analytical method for estimating CI's is shown in the next section, with examples of its application to normally distributed data and further demonstrating the breakdown of the method when outliers are present. The bootstrapping method for estimating CI's is described in the "Methods" section. The "Results" section is in three parts: (1) validation of the bootstrapping method with normally distributed data (uncontaminated by outliers); (2) validation of the Excel implementation (CI-RANOVA); (3) further validation of the bootstrapping method with data that include outlying values.

The duplicate method and the analysis of the resulting data by classical ANOVA

The simplest form of the full three-tiered balanced experimental design is illustrated in Fig. 1. A number of sampling targets (n) are chosen at random, where ideally $n \geq 8$ [3], from a wide selection of such targets. Two samples are acquired from each of these targets by independent duplication of the sampling protocol. These two samples are then treated individually and are both subjected to the same preparation procedures. Two test portions are then drawn from each of these test samples and analysed individually. This method enables variance estimates to be extracted by ANOVA for each of the following levels:

1. Between-target variance;
2. Between-sample variance;
3. Between-analysis variance (an estimate of analytical repeatability).

The method is described in more detail in the Eurachem guide [1].

If the variability at each of these three levels can be assumed to be normally distributed, then a confidence interval at the bottom (analysis) level can be derived from a Chi-squared distribution, and Williams [7] provides a method of calculating approximate confidence limits at the top 2 (target and sample) levels. Details can be found in Graybill [8]. The equations below give the confidence intervals for a nested experimental design of size $I \times J \times K$ where I is the number of targets, assumed to be drawn from a normal distribution with mean μ and variance σ^2_{Target}, J is the number of samples per target, with sampling variance σ^2_{Sample}, and K is the number of analyses per sample, with analytical variance $\sigma^2_{\text{Analysis}}$. MS_{Sample} is the mean square of the middle (sampling) level

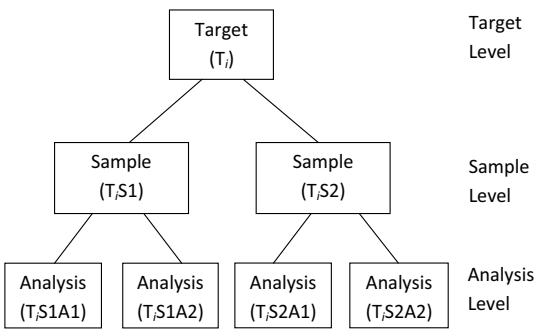

Fig. 1 The simplest form of the fully balanced experimental design for the evaluation of measurement uncertainty with n targets T_i where $1 <= i <= n$. The fully balanced design requires a minimum of two samples per target and two analyses per sample (referred to as an $n \times 2 \times 2$ experimental design)

from the ANOVA. A circumflex denotes an estimate, e.g. $\hat{\sigma}^2_{\text{Target}}$ is an estimate of the target variance.

$F_{p,\,\nu 1,\,\nu 2}$ is the inverse cumulative distribution function (cdf) of the F probability distribution with degrees of freedom $\nu 1,\nu 2$ for a probability p, and $\chi^2_{p,\nu}$ is the inverse cdf of the Chi-squared distribution with degrees of freedom ν for probability p.

Target level

$$\left(\frac{I-1}{\chi^2_{\alpha/2,I-1}}\right)\left[\hat{\sigma}^2_{\text{Target}}+\frac{MS_{\text{Sample}}}{\text{JK}}\left(1-F_{\alpha/2,\ I-1,\ I(J-1)}\right)\right]\le\sigma^2_{\text{Target}}\le\left(\frac{I-1}{\chi^2_{1-\alpha/2,I-1}}\right)\left[\hat{\sigma}^2_{\text{Target}}+\frac{MS_{\text{Sample}}}{\text{JK}}\left(1-F_{1-\alpha/2,\ I-1,\ I(J-1)}\right)\right] \quad (1)$$

Sample level

$$\left(\frac{I(J-1)}{\chi^2_{\alpha/2,I(J-1)}}\right)\left[\hat{\sigma}^2_{\text{Sample}}+\frac{\hat{\sigma}^2_{\text{Analysis}}}{K}\left(1-F_{\alpha/2,\ I(J-1),\ IJ(K-1)}\right)\right]\le\sigma^2_{\text{Sample}}\le\left(\frac{I(J-1)}{\chi^2_{1-\alpha/2,I(J-1)}}\right)\left[\hat{\sigma}^2_{\text{Sample}}+\frac{\hat{\sigma}^2_{\text{Analysis}}}{K}\left(1-F_{1-\alpha/2,\ I(J-1),\ IJ(K-1)}\right)\right] \quad (2)$$

Analysis level

$$\hat{\sigma}^2_{\text{Analysis}}\left(\frac{IJ(K-1)}{\chi^2_{\frac{\alpha}{2},IJ(K-1)}}\right)\le\sigma^2_{\text{Analysis}}\le\hat{\sigma}^2_{\text{Analysis}}\left(\frac{IJ(K-1)}{\chi^2_{1-\frac{\alpha}{2},IJ(K-1)}}\right) \quad (3)$$

These equations give confidence intervals for variances. To obtain intervals for the standard deviations, as reported below, one simply takes the square root of each of the limits of the intervals for the variances. At the target and sample levels, the ANOVA estimates of variance are obtained by subtraction, and it is possible for the lower limits or even the variance estimates themselves to be negative. The standard practice of replacing negative estimates or confidence limits by zero, given that the true value of the variance cannot possibly be in the negative part of the interval, is followed here.

A numerical example showing the effect of an outlier

The data in Table 1 were simulated using normal distributions with input parameters $\mu = 26.3$, $\sigma_{\text{Target}} = 8.9$, $\sigma_{\text{Sample}} = 3.0$, $\sigma_{\text{Analysis}} = 1.2$.

The classical ANOVA has produced estimates close to the true values of the standard deviations, and the confidence intervals include these true values, as they should do in 95 % of examples.

Table 2 shows the same simulated data but with the inclusion of 1 large outlying value in the final position (Target number 10, S2A2). It can be seen that this single analytical outlier has had a large effect on both the variances calculated by classical ANOVA and the associated confidence limits. These are no longer representative of the bulk of the data and differ significantly from the input parameters used in the original simulation.

An alternative is to use Robust ANOVA. In brief, the robust mean μ_r is initially estimated as the classical mean, and the robust standard deviation σ_r as the median of the absolute differences between duplicated measurements. Any values that are found to exceed $\mu_r + c\ \sigma_r$ are replaced with $\mu_r + c\ \sigma_r$, and any values that fall below $\mu_r - c\ \sigma_r$ are replaced with $\mu_r - c\ \sigma_r$, where c is a factor between 1 and 2 (typically set to 1.5). The robust statistics μ_r and σ_r are then recalculated, and the process is repeated multiple times, until μ_r converges to an acceptable level of accuracy [4]. Robust ANOVA on the data in Table 2 using the program RANOVA2 [9] yields the following results: $\hat{\sigma}_{\text{Target}} = 9.8$, $\hat{\sigma}_{\text{Sampling}} = 4.3$, $\hat{\sigma}_{\text{Analysis}} = 1.1$. These are much closer to both the input parameters and also to the standard deviations calculated by classical ANOVA

Table 1 Estimated standard deviations from classical ANOVA, with 95 % confidence intervals from Eqs. (1) to (3), on data *without* outliers

Target	Simulated data, normally distributed				Estimated standard deviations ($\hat{\sigma}$) from classical ANOVA, confidence intervals from Eqs. (1) to (3)			
	S1A1	S1A2	S2A1	S2A2	Level	TRUE σ	Estimated $\hat{\sigma}$	Confidence Interval
1	31.2	29.8	29.8	28.2	Target	8.9	8.7	5.1, 16.5
2	26.7	26.2	29.1	27.6	Sample	3.0	3.9	2.6, 6.8
3	12.5	13.5	13.2	10.3	Analysis	1.2	1.0	0.7, 1.4
4	22.2	22.6	35.9	34.8				
5	28.4	28.8	28.2	30.7				
6	37.2	34.4	41.5	42.2				
7	32.7	31.7	24.3	24.8				
8	16.9	15.7	20.6	20.6				
9	16.9	17.9	19.3	18.2				
10	43.7	44.6	40.4	39.9				

Table 2 Standard deviation estimates and confidence intervals calculated from the data in Table 1 with the *inclusion of a single analytical outlier* in Target number 10

Target	Data including 1 analytical outlier				Estimated standard deviations ($\hat{\sigma}$) from classical ANOVA, confidence intervals from Eqs. (1) to (3)			
	S1A1	S1A2	S2A1	S2A2	Level	TRUE σ	Estimated $\hat{\sigma}$	Confidence Interval
1	31.2	29.8	29.8	28.2	Target	8.9	20.2	0.0, 57.4
2	26.7	26.2	29.1	27.6	Sampling	3.0	0.0	0.0, 59.2
3	12.5	13.5	13.2	10.3	Analysis	1.2	56.7	43.4, 81.9
4	22.2	22.6	35.9	34.8				
5	28.4	28.8	28.2	30.7				
6	37.2	34.4	41.5	42.2				
7	32.7	31.7	24.3	24.8				
8	16.9	15.7	20.6	20.6				
9	16.9	17.9	19.3	18.2				
10	43.7	44.6	40.4	**399.0**				

on the original normal data (Table 1). They are also more representative of the main body of data in Table 2 than those calculated using classical methods. The high value outlier in this example is the original value (39.9) multiplied by 10 and demonstrates the ability of robust ANOVA to cope with extreme outliers that might be encountered from, for example, a transcription error. Ideally, values resulting from this type of mistake would not be included in uncertainty estimation [10], but practically they are not always identified and corrected or removed.

An extreme outlying value such as shown in Table 2 needs serious consideration. Its presence and the method used to treat it should be reported. The presence of such an outlying value may affect the inferences drawn from the experiment. Less extreme outlying values can occur for a number of reasons, e.g. deficiencies in the experimental method, operator error, or a simple typing error. They might also be due to genuine variations in the property being measured and should not be ignored. One potential use of robust statistics in this application is to draw attention to the presence of outliers in experimental data. For example, the difference between the robust and classical standard deviations derived from the data in Table 2 is a clear indication that these data are not normally distributed, and one or more outlying values may be present.

Robust ANOVA provides reasonable estimates of the underlying variances, i.e. the variances of the "good" data, in data that include a small proportion of outlying values. It can be easily applied using the program RANOVA2 [9]. However, the calculation of confidence intervals in the context of a robust ANOVA is far from straightforward. It cannot be achieved via formulas based on assumed probability distributions, because it is precisely when these assumptions break down that we need to use the robust approach. An alternative is to use bootstrapping methods. In this computer-based approach, a large number B of independent bootstrap samples are generated. A bootstrap sample is a data set, of the same size and structure as the observed one, generated by random sampling with replacement from the observed data set. The statistic of interest (e.g. a variance) is calculated for each bootstrap sample. Confidence intervals can then be derived from the empirical distribution of these results [11].

Method: Estimating confidence limits on variances from robust ANOVA

The existing program RANOVA2 [9] was modified to provide confidence intervals on variances calculated by both classical and robust ANOVA. It is intended that the new program CI-RANOVA will also be made available on the AMC website.

The data simulation, robust ANOVA and CI estimation were reproduced in Matlab software supplied by MathWorks. This is much more efficient than Excel when processing multiple data arrays, allowing a greater number of simulations to be created and analysed within a practical time frame, and in particular allowing the accurate estimation of coverage probabilities, requiring 50 000 simulations. Versions of the two programs were produced independently by different researchers based in different institutions.

Both programs calculate confidence limits on variances produced by classical ANOVA using a mathematical method based on Eqs. (1) to (3). Confidence limits on the variances produced by robust ANOVA are estimated using a bootstrapping method. During development, a problem was encountered if a large number of bootstrap samples are generated on data containing outlying values. The bootstrapping method generates samples by selecting means and differences (at every level) at random with replacement. Some of these bootstrap samples therefore contain fewer outliers than the original data set, while some contain more. When a large

number (e.g. $B = 2000$) of bootstrap samples are generated, some of these samples are likely to contain many repetitions of very large outlying differences. The robust routines used in RANOVA2 are only intended to accommodate up to 10 % of outlying values. It was found in practice that a small proportion of the bootstrap samples includes too many large outlying differences for the robust ANOVA to cope with. For this reason, a winsorization process (Fig. 2) was incorporated into both programs. This is applied to the input data matrices after the initial robust analysis but prior to generating the bootstrap samples.

The winsorization process brings in large outliers in a very similar way to how they are dealt with in the robust analysis, but using wider limits. The idea, due to [12], is that this will change the results of the robust analysis very little, because the outliers are still outside the limits used in that analysis, but it will limit the damage that they can cause when they occur in large numbers and in particular will avoid the breakdown of the robust analysis that can occur in this situation. The limit used here, as suggested by Singh [12], is 1.5 times the limit used in the robust analysis, which itself is 1.5 [4], hence the 1.5^2 in the algorithm in Fig. 2.

An example of the winsorization process is shown in Table 3. In this case, the bootstrapping method randomly produced three top-level outlying values on rows 1, 2 and 4 (Table 3a), even though the input data (from Table 2) included only 1 outlying value. This results in a spuriously high robust estimate of the target standard deviation, due to breakdown of the robust algorithm. Applying the winsorization method prior to robust ANOVA reduces the magnitude of the outlying values sufficiently that they can be

1. Create tables of target means, and sample and analytical differences. There are n items in each of the rows below, where n is the number of targets:
    1. Target Level Mean (μ_{Target}) [(TnS1A1+TnS1A2+TnS2A1+TnS2A2)/4]
    2. Sample Level Difference (Δ_{Sample}) [(TnS1A1+TnS1A2)/2 – (TnS2A1+TnS2A2)/2]
    3. Analytical Level Difference1 ($\Delta_{Analysis1}$) [TnS1A1 – TnS1A2]
    4. Analytical Level Difference2 ($\Delta_{Analysis2}$) [TnS2A1 – TnS2A2]

2. Set truncation limits (TL), see text for an explanation of the 1.5^2 :
    1. $TL_{Target} = 1.5^2 * \sqrt{(1\text{-}1/n)}$
    2. $TL_{Sample} = 1.5^2 * \sqrt{(1\text{-}1/2)}$
    3. $TL_{Analysis} = 1.5^2 * \sqrt{(1\text{-}1/2)}$

3. Convert robust variances to scale factors (SF):
    1. $SF_{Target} = \sqrt{(\sigma^2_{Target} + \sigma^2_{Sample}/2 + \sigma^2_{Analysis}/4)}$
    2. $SF_{Sample} = \sqrt{(\sigma^2_{Sample} + \sigma^2_{Analysis}/2)}$
    3. $SF_{Analysis} = \sqrt{(\sigma^2_{Analysis})}$

4. For $i = 1$ to n, create working tables (W) of target level (using robust mean μ_{Robust}) and sample and analytical differences, with truncated outliers :
    1. $W_{Target}(i) = [\mu_{Robust} + Max(-TL_{Target} * SF_{Target}, Min(TL_{Target} * SF_{Target}, \mu_{Target}(i) - \mu_{Robust})]$
    2. $W_{Sample}(i) = [Max(-2 * TL_{Sample} * SF_{Sample}, Min(2 * TL_{Sample} * SF_{Sample}, \Delta_{Sample}(i))]$
    3. $W_{Analysis1}(i) = [Max(-2 * TL_{Analysis} * SF_{Analysis}, MIN(2 * TL_{Analysis} * SF_{Analysis}, \Delta_{Analysis1}(i))]$
    4. $W_{Analysis2}(i) = [Max(-2 * TL_{Analysis} * SF_{Analysis}, MIN(2 * TL_{Analysis} * SF_{Analysis}, \Delta_{Analysis2}(i))]$

5. For $i = 1$ to n, construct winsorised data matrix of $n*2*2$ (1st, 2nd, 3rd, 4th) analytical measurements:
    1. Table entry (1^{st}) = $(W_{Target}(i) - W_{Sample}(i) / 2) – (W_{Analysis1}(i) / 2)$
    2. Table entry (2^{nd}) = $(W_{Target}(i) - W_{Sample}(i) / 2) + (W_{Analysis1}(i) / 2)$
    3. Table entry (3^{rd}) = $(W_{Target}(i) + W_{Sample}(i) / 2) – (W_{Analysis2}(i) / 2)$
    4. Table entry (4^{th}) = $(W_{Target}(i) + W_{Sample}(i) / 2) + (W_{Analysis2}(i) / 2)$

Fig. 2 Winsorization process used to limit the effect of multiple occurrences of outlying observations prior to bootstrapping, for an $n*2*2$ experimental design (See Fig. 1)

Table 3 Example of the winsorization method of accommodating outlying values on an individual bootstrap sample created from the data in Table 2

(a) Data and robust ANOVA for individual bootstrap sample *without* prior winsorization

S1A1	S1A2	S2A1	S2A2	Level	TRUE SD	Robust SD
130.0	**129.5**	**134.8**	**133.3**	Target	8.9	56.2
132.2	**130.8**	**− 47.0**	**311.6**	Sample	3.0	6.1
36.2	37.2	42.5	39.6	Analysis	1.2	1.1
125.4	**125.5**	**138.1**	**138.6**			
25.0	23.7	29.2	31.7			
32.5	33.2	26.6	23.8			
19.2	19.2	17.5	17.9			
32.2	32.6	44.8	45.8			
32.0	30.4	24.1	23.1			
32.7	31.7	25.3	23.8			

(b) Data and robust ANOVA for individual bootstrap sample *with* prior winsorization. Adjusted values are shown in bold type

S1A1	S1A2	S2A1	S2A2	Level	TRUE SD	Robust SD
47.2	**46.8**	**52.1**	**50.6**	Target	8.9	11.7
49.5	**48.0**	**47.8**	**51.3**	Sample	3.0	6.1
36.2	37.2	42.5	39.6	Analysis	1.2	1.1
42.7	**42.7**	**55.4**	**55.9**			
25.0	23.7	29.2	31.7			
32.5	33.2	26.6	23.8			
19.2	19.2	17.5	17.9			
32.2	32.6	44.8	45.8			
32.0	30.4	24.1	23.1			
32.7	31.7	25.3	23.8			

Bold values in (a) indicate outlying values produced by bootstrapping
Bold values in (b) indicate adjusted values

accommodated by the robust algorithm (Table 3b). Note that all other data values and standard deviations remain unchanged.

Following winsorization, a large number B of bootstrap samples ($B = 2000$) are generated. The method of producing the bootstrap samples for an $n * 2 * 2$ experimental design (Fig. 1) is detailed in Fig. 3.

Variances are calculated for each of the B bootstrap samples using robust ANOVA. The variance at the target (top) level needs to be multiplied by $n/(n-1)$ to compensate for a bias caused by sampling with replacement from the n targets [11]. The bootstrap samples at the lower levels were based on the differences, so no such correction is needed for the other two variances. The variances for each level are sorted into numerical order. The 2.5 and 97.5 percentiles of the sorted variances could then be considered as bootstrap estimates of the 95 % confidence limits. However, this simple approach is known not to work particularly well for skewed distributions, especially when the data being bootstrapped are small, and Efron and Tibshirani [11] recommend what they call the Bca method in this case. This is described in Fig. 4, where the constant a has been chosen to optimise the procedure for the Chi-squared distributions that one would expect to see with normally distributed data.

Validation of the bootstrapping method was performed by simulating 50 000 normally distributed data sets based on five input parameters: (a) the number of targets n; (b) the mean target value μ_{Target}; (c) between-target standard deviation σ_{Target}; (d) between-sample standard deviation σ_{Sample}; (e) between-analyses standard deviation σ_{Analysis}. In the case of contaminated data, outliers were introduced at one of the three levels in each simulation. These were created by randomly selecting a target or targets from the normally distributed data sets and adding a constant, in this case 500, to either all four measurements for that target, or both measurements for one of its samples, or one single analytical measurement (similar to the method described by Ramsey et al. [5]). Tests were also performed to investigate the effects of different severities of the outlying value by varying the magnitudes of the added constant between 0 and 1000. An example data set ($n = 20$) with all three types of contamination is shown in Table 4.

Results and discussion

The underlying calculations, including the averaging of the results of different simulations, were performed using variances, because variances from classical ANOVA are unbiased estimators of the population variances, whereas their square roots are not unbiased estimators of the corresponding standard deviations. However, the tables in this section present the results as standard deviations because these have the same units as the original data and are more readily interpretable.

Objective 1: Validation of the bootstrap confidence intervals with normally distributed data

Validation of the bootstrapping method of estimating confidence limits at three different levels was performed by comparison with those produced by the mathematical method. Two values of n were used ($n = 10$ and $n = 100$). For each n, variances and confidence intervals were estimating using the Matlab program, with B set to 2000. This allowed 50 000 simulated data matrices to be analysed in a practical time frame to give coverage probabilities accurate to approximately 0.1 %. Coverage percentages were estimated by counting the number of times the CI contained the true

1. The bootstrapping routine initially creates 3 working tables using the winsorised data. There are *n* items in rows 1 and 2 below and 2*n* in row 3, where *n* is the number of targets:
    1. Target Level Mean [(TnS1A1+TnS1A2+TnS2A1+TnS2A2)/4]
    2. Level Difference [(TnS1A1+TnS1A2)/2 – (TnS2A1+TnS2A2)/2]
    3. Analytical Level Differences (x2) [TnS1A1 – TnS1A2] and [TnS2A1 – TnS2A2]

A large number (e.g. 2000) of bootstrapped measurement-datasets are then generated:

2. Generate *n* x 2 table of samples. For each paired (1st and 2nd) entry into this table:
    1. Select a Target Level Mean at random with replacement
    2. Select a Sample Level Difference at random with replacement
    3. Table entry: (1st) = Target Level Mean + Sample Level Difference/2, (2nd) = Target Level Mean – Sample Level Difference/2.

3. Generate *n* x 4 table of analytical measurements. For each sample pair in the table generated in step 2:
    1. Select an Analytical Level Difference at random with replacement.
    2. Table entry (1st) = Sample Level Pair (1st) + Analytical Level Difference/2
    3. Table entry (2nd) = Sample Level Pair (1st) - Analytical Level Difference/2
    4. Select a new Analytical Level Difference at random with replacement
    5. Table entry (3rd) = Sample Level Pair (2nd) + Analytical Level Difference/2
    6. Table entry (4th) = Sample Level Pair (2nd) - Analytical Level Difference/2

Fig. 3 The procedure used to generate bootstrap samples for an $n*2*2$ experimental design (See Fig. 1)

The Bca (Bias-Corrected and Accelerated) intervals adjust the percentiles used to construct confidence limits to reflect the shape of the distribution of the variances over the bootstrap samples. Two numbers z_0 (Bias-Correction) and a (Acceleration) are calculated as follows:

$$z_0 = \Phi^{-1}\left(\frac{\#\ \mathrm{v_i} < \mathrm{v}}{\mathrm{B}}\right)$$

The numerator of this equation is the number of variances from the bootstrap samples that are less than their mean value, B is the total number of bootstraps, and Φ^{-1} is the inverse normal cdf.

$$a = \frac{1}{3}\sqrt{\frac{2}{\upsilon}}$$

Where υ is the degrees of freedom in the line of the ANOVA table corresponding to the level of interest. The corrected percentiles are then calculated as follows:

$$p_1 = \Phi\left(z_0 + \frac{z_0 - z}{1 - a\,(z_0 - z)}\right)$$

$$p_2 = \Phi\left(z_0 + \frac{z_0 + z}{1 - a\,(z_0 + z)}\right)$$

Where z is the z-score for the coverage probability, e.g. for a coverage probability of 95 %, $z = 1.96$, and Φ is the standard normal cdf.

Fig. 4 Bca (Bias-corrected and accelerated) correction for 95 % confidence limits

Table 4 Example illustrating the three types of contamination used in the simulations. Only one type was used in any one simulation

Target	S1A1	S1A2	S2A1	S2A2
1	4.0	−6.7	−29.7	−32.8
2	**580.5**	**618.1**	**562.6**	**598.4**
3	19.2	−18.7	−45.4	−20.6
4	−44.9	−55.8	−18.1	1.6
5	104.3	118.3	−16.9	40.3
6	19.0	47.9	**600.5**	**548.9**
7	43.0	43.4	102.8	59.0
8	149.1	133.4	180.7	173.8
9	−34.0	11.3	−5.5	49.9
10	143.5	167.3	102.9	**632**

Outlying values are shown in bold

Table 5 Input parameters for simulations

Mean target value (μ_{Target})	75.8
Between-target standard deviation (σ_{Target})	73.2
Between-sample standard deviation (σ_{Sample})	27
Between-analyses standard deviation ($\sigma_{Analysis}$)	20.4

value of the input parameter. Input parameters were the same for all simulations, based on real values from experimental data (Table 5). Results of these trials are shown in Table 6 ($n=10$) and Table 7 ($n=100$).

For both $n=10$ and $n=100$, the means and component standard deviations at the three different levels calculated by classical ANOVA are good approximations of the input parameters. For $n=10$, the component standard deviations calculated by robust ANOVA are slightly higher than the true values, especially at the two upper levels (Table 6). This is a consequence of the approximations involved in the bias correction of the robust estimates in the case of the simultaneous estimation of both mean and variance.

The coverage probabilities calculated by the mathematical method in Tables 6 and 7 are close to the expected 95 %, although in most cases they appear to be slightly conservative: the intervals are a little wider than is required for true 95 % confidence. With $n=10$, the intervals calculated using the bootstrapping method are too narrow resulting in < 95 % coverage (Table 6). This is consistent with previous findings where bootstrapping methods have been applied to estimate confidence intervals on variances (See p. 181–183 in [11]). For $n=100$, these coverage percentages are very close to the expected 95 % (Table 7), indicating that a higher value of n enables better estimates of the confidence intervals on the component standard deviations.

Objective 2: Validation of the implementation using CI-RANOVA with a smaller number of simulations

The same trials were performed with CI-RANOVA using a smaller number (1000) of simulated data matrices (Tables 8, 9). It is not possible to calculate precise coverage percentages with this number of simulations; however, comparison of the CI's with those in Tables 6 and 7 shows there is good agreement between the two programs.

Table 6 Comparison between average standard deviations and confidence limits calculated by classical and robust ANOVA on 50000 simulated data matrices with no outlying values ($n=10$), also showing coverage probability (cov, %)

(μ_{Target}=75.8)		Classical ANOVA ($\hat{\mu}_{Target}=75.8$)			Robust ANOVA ($\hat{\mu}_{Target}=75.8$)		
	σ	$\hat{\sigma}$	Math. CI	cov, %	$\hat{\sigma}$	Bootstrap CI	cov, %
Target	73.2	73.1	(43.7, 137.7)	95.7	74.8	(37.2, 122.6)	92.4
Sample	27.0	27.0	(13.3, 52.0)	96.2	28.0	(16.2, 48.8)	89.0
Analysis	20.4	20.4	(15.6, 29.5)	95.0	20.7	(15.5, 30.5)	92.0

Table 7 Comparison between average standard deviations and confidence limits calculated by classical and robust ANOVA on 50 000 simulated data matrices with no outlying values ($n=100$), also showing coverage probability (cov, %)

(μ_{Target}=75.8)		Classical ANOVA ($\hat{\mu}_{Target}=75.8$)			Robust ANOVA ($\hat{\mu}_{Target}=75.8$)		
	σ	$\hat{\sigma}$	Math. CI	cov, %	$\hat{\sigma}$	Bootstrap CI	cov, %
Target	73.2	73.2	(62.9, 86.2)	95.9	73.3	(61.8, 88.7)	95.4
Sample	27.0	27.0	(22.3, 32.6)	95.9	27.1	(22.1, 33.6)	94.5
Analysis	20.4	20.4	(18.6, 22.6)	95.1	20.4	(18.3, 23.1)	94.8

Table 8 The results shown in Table 6, $n=10$ reproduced using CI-RANOVA with a smaller number (1000) of simulated data matrices

(μ_{Target}=75.8)		Classical ANOVA ($\hat{\mu}_{Target}$=74.5)		Robust ANOVA ($\hat{\mu}_{Target}$=74.4)	
	σ	$\hat{\sigma}$	Mathematical CI	$\hat{\sigma}$	Bootstrap CI
Target	73.2	73.0	(43.9, 137.5)	74.7	(37.2, 122.4)
Sample	27.0	26.4	(12.6, 51.0)	27.4	(15.8, 47.7)
Analysis	20.4	20.4	(15.6, 29.5)	20.6	(15.4, 30.6)

Table 9 The results shown in Table 7. $N=100$ reproduced using CI-RANOVA with a smaller number (1000) of simulated data matrices

(μ_{Target}=75.8)		Classical ANOVA ($\hat{\mu}_{Target}$=75.7)		Robust ANOVA ($\hat{\mu}_{Target}$=75.7)	
	σ	$\hat{\sigma}$	Mathematical CI	$\hat{\sigma}$	Bootstrap CI
Target	73.2	73.4	(63.1, 86.5)	73.4	(62.0, 89.0)
Sample	27.0	27.0	(22.3, 32.6)	27.1	(22.0, 33.6)
Analysis	20.4	20.4	(18.6, 22.7)	20.5	(18.4, 23.1)

Objective 3: Validation of the robust confidence intervals on data including outliers

Further simulations were created with outlying values added to the data. Outlying values were generated using the three scenarios shown in Table 4. Each outlying value was applied by randomly selecting a target (without replacement), and adding 500 to the original simulated value(s) to create one or more extreme outlier(s). Each independently simulated data matrix was then analysed using CI-RANOVA.

Summaries of the average variances (expressed as standard deviations) and their associated confidence limits for 1000 simulations are shown in Table 10 for $n=100$. The simulation input parameters ('true' values) have been included to simplify comparisons with the average values. This table also shows coverage percentages calculated from 50 000 simulations using the Matlab program. This program and CI-RANOVA implement the same method and give consistent results, so the coverage probabilities can be taken to apply to CI-RANOVA as well.

Note that in the simulations with contaminated data, the coverage probabilities are still calculated as the percentage of intervals that include the parameters used to generate the data before contamination.

Robust confidence interval estimates with different types of outlier

Comparisons of the classical and robust standard deviations for the three different levels (Table 10) confirm that the robust method produces values that are much closer estimates of the input parameter 'true' values than the classical method, especially for the level at which outliers have been added. This is consistent with earlier work [4, 5, 13]. With a small number of outlying values (2) the robust estimates show a relatively small positive bias (<4 %) compared to the true values. These biases are most evident at the levels where outliers have been added, and also at any levels above. For example, adding two sample outliers result in a positive bias at the sample and target levels. The same pattern is seen for all of the outlier scenarios.

With two outliers, the coverage percentages calculated using the Matlab program range between 92 and 95 %, indicating that the bootstrapping method is producing accurate confidence intervals when there are a small number of outlying values. The coverage percentages are slightly below the nominal 95 %. This is not because the intervals are too narrow, but because the robust estimates have a slight upwards bias compared to the true values for the uncontaminated data.

Increasing the number of outlying values to 4 increases the biases in robust standard deviation estimates of the affected levels, by up to 9 % for 4 outliers at the sample level. Coverage percentages are consequently lower (minimum 86 %). The unaffected levels still show coverage percentages that are very close to the nominal 95 %. This latter observation remains true when the number of outlying values is increased to 10, although in this scenario the biases in the robust estimates increase to 23 to 24 % at the sample level, and the coverage percentages at the affected levels are reduced in some cases to <50 %.

Overall, these results suggest that the bootstrapping method produces reasonable estimates of the confidence intervals of the robust standard deviations when there are a small number of outlying values (i.e. <5 %). When a larger number of outliers are present (e.g. at 10 % of targets), the coverage calculated from the true value is seriously reduced below the nominal 95 %.

Lower values of *n* (*n* <100)

The $n=100$ scenario discussed above suggests that the bootstrapping method produces good estimates of confidence intervals for lower numbers of outliers. This value of n would be fairly unusual in practice, due to the financial cost of obtaining that quantity of duplicated measurements. It is therefore appropriate to repeat the experiments for lower n values. Tables 11 and 12 show the averaged standard deviations, CIs and coverage percentages for $n=20$ and $n=10$, respectively.

The widths of the robust confidence intervals decrease as n increases (Fig. 5). Again the robust estimates of standard deviations are closer to the true values than are the classical

Table 10 Average classical and robust standard deviations and robust confidence interval estimates with three different outlier scenarios

(μ_{Target} =75.8)	$\hat{\mu}_{Target}$ classical	$\hat{\mu}_{Target}$ robust	Level	σ True	$\hat{\sigma}$ Classical	$\hat{\sigma}$ Robust	Robust CI	Cov, %[a]
(a) 2 Outliers Target Level	85.7	78.3	Target	73.2	101.7	75.8	(63.7, 92.7)	92.9
			Sample	27.0	27.1	27.3	(22.3, 33.9)	94.6
			Analysis	20.4	20.4	20.4	(18.3, 23.0)	94.7
(a) 2 Outliers Sample Level	80.7	78.5	Target	73.2	73.1	75.7	(63.7, 92.5)	93.6
			Sample	27.0	56.8	28.1	(22.9, 35.1)	91.8
			Analysis	20.4	20.4	20.5	(18.4, 23.1)	94.7
(a) 2 Outliers Analysis Level	78.2	77.8	Target	73.2	73.1	74.5	(62.6, 90.4)	94.7
			Sample	27.0	27.0	28.0	(22.8, 35.1)	92.2
			Analysis	20.4	40.7	20.8	(18.6, 23.5)	93.4
(b) 4 Outliers Target Level	95.5	81.2	Target	73.2	122.8	78.9	(66.3, 97.9)	86.4
			Sample	27.0	27.0	27.2	(22.2, 33.7)	94.5
			Analysis	20.4	20.4	20.4	(18.3, 23.0)	94.8
(b) 4 Outliers Sample Level	86.2	81.8	Target	73.2	73.0	78.6	(65.8, 97.3)	88.6
			Sample	27.0	75.6	29.4	(24.1, 37.2)	85.7
			Analysis	20.4	20.4	20.5	(18.4, 23.1)	94.8
(b) 4 Outliers Analysis Level	80.8	80.0	Target	73.2	73.1	75.9	(63.6, 92.3)	92.1
			Sample	27.0	27.1	29.2	(23.7, 37.0)	87.4
			Analysis	20.4	54.0	21.1	(18.9, 23.9)	90.2
(c) 10 Outliers Target Level	125.6	92.3	Target	73.2	167.3	90.7	(75.7, 125.4)	39.3
			Sample	27.0	27.0	27.2	(22.1, 33.7)	94.6
			Analysis	20.4	20.4	20.5	(18.4, 23.2)	94.9
(c) 10 Outliers Sample Level	100.5	91.5	Target	73.2	68.6	87.9	(72.5, 116.0)	53.5
			Sample	27.0	115.2	33.6	(27.4, 44.7)	46.6
			Analysis	20.4	20.3	20.4	(18.3, 23.0)	94.7
(c) 10 Outliers Analysis Level	87.7	86.2	Target	73.2	72.1	79.6	(66.4, 97.4)	84.7
			Sample	27.0	27.3	33.1	(26.8, 44.4)	55.3
			Analysis	20.4	81.6	22.1	(19.7, 25.4)	71.2

Outliers have been applied to (a) 2 targets, (b) 4 targets, (c) 10 targets. Calculated with $n = 100$, using Excel CI-RANOVA with 1000 simulations. Raw values are provided in the electronic supplement

[a]Coverage% is based on the number of times the true value of the input parameter occurs within the robust CI in 50 000 simulations created by the Matlab program

estimates for the levels at which outliers have been included. For $n = 20$ with 1 outlying value (Table 11), the biases in the robust standard deviations range between 9 and 14 % above the true values for the levels with outlying values and any levels above. Coverage percentages for the true values are a minimum of 87 % for the sample level with 1 sample outlier. Generally, the sample level appears to show the highest bias, and also the lowest percentage coverage, in all scenarios where this level has been affected by outliers either in the level itself, or in the analysis level below it.

One consequence of narrowing CI's with increasing n is that the lower confidence limits for high n become close to the true values. This is illustrated in Fig. 5, where it can be seen that in one case, with outliers added to 10 targets, the true value μ_{Target} lies just outside the average CI for $n = 100$.

In the cases where outlying values have been added to 10 % of the targets (the 2-outlier scenarios in Table 11, and all single outlier scenarios in Table 12), there is up to ~ 25 % bias at the affected levels, as previously seen in Table 9 where 10 outliers were added to 100 targets. This has caused a loss of coverage. This loss arises because of the bias in the robust estimate, which results in the true value lying below the lower confidence limit for a larger number of simulations.

The coverage percentage drops to 75 % at the sample level for $n = 10$ (Table 12), which may be thought to imply that the bootstrapping method of estimating confidence intervals is unreliable with this value of n. However, these experiments suggest that the problem lies with the bootstrap estimates themselves, which show some upwards bias when compared with the parameters of the underlying uncontaminated distribution. The bootstrapping is able to accurately represent the random variability in the robust estimates but cannot correct for this bias, and hence the coverage falls below the nominal

Table 11 Average results of standard deviation and confidence limits (as Table 9) calculated with $n = 20$, with outliers applied to (a) 1 target and (b) 2 targets. Raw values are provided in the electronic supplement

($\mu_{Target} = 75.8$)	$\hat{\mu}_{Target}$ Classical	$\hat{\mu}_{Target}$ Robust	Level	σ True	$\hat{\sigma}$ Classical	$\hat{\sigma}$ Robust	Robust CI	Cov, %[a]
(a) 1 Outlier Target Level	101.7	83.9	Target	73.2	132.6	80.8	(55.2, 130.7)	91.3
			Sample	27.0	26.8	27.3	(17.7, 42.9)	92.6
			Analysis	20.4	20.4	20.5	(16.5, 27.3)	93.8
(a) 1 Outlier Sample Level	88.1	82.5	Target	73.2	73.5	80.1	(53.4, 127.9)	93.0
			Sample	27.0	83.7	30.9	(20.8, 52.3)	87.0
			Analysis	20.4	20.5	20.7	(16.6, 27.4)	93.9
(a) 1 Outlier Analysis Level	81.8	80.7	Target	73.2	72.5	76.7	(50.4, 116.3)	94.4
			Sample	27.0	28.1	30.4	(20.2, 51.9)	88.0
			Analysis	20.4	59.6	21.4	(17.2, 29.1)	91.2
(b) 2 Outliers Target Level	125.0	91.2	Target	73.2	170.9	92.1	(63.8, 157.5)	78.1
			Sample	27.0	26.8	27.4	(17.9, 43.1)	92.6
			Analysis	20.4	20.4	20.4	(16.4, 27.3)	93.8
(b) 2 Outliers Sample Level	100.2	90.5	Target	73.2	71.5	88.5	(58.4, 145.8)	85.8
			Sample	27.0	115.1	33.9	(23.5, 61.7)	73.5
			Analysis	20.4	20.3	20.5	(16.4, 27.3)	94.0
(b) 2 Outliers Analysis Level	87.8	86.2	Target	73.2	72.5	80.5	(52.1, 123.0)	93.4
			Sample	27.0	28.2	33.1	(22.2, 61.0)	76.8
			Analysis	20.4	81.7	22.3	(17.8, 31.4)	85.9

[a]Coverage% is based on the number of times the true value of the input parameter occurs within the robust CI in 50 000 simulations created by the Matlab program

Table 12 Average results of standard deviation and confidence limits (as Table 9) calculated with $n = 10$, with outliers applied to 1 target. Raw values are provided in the electronic supplement

($\mu_{Target} = 75.8$)	$\hat{\mu}_{Target}$ classical	$\hat{\mu}_{Target}$ Robust	Level	σ True	$\hat{\sigma}$ Classical	$\hat{\sigma}$ Robust	Robust CI	Cov, %[a]
1 Outlier Target Level	126.1	91.4	Target	73.2	174.5	91.6	(52.9, 178.0)	85.7
			Sample	27.0	27.5	28.4	(16.6, 49.8)	89.5
			Analysis	20.4	20.3	20.7	(15.5, 30.3)	91.9
1 Outlier Sample Level	102.3	92.1	Target	73.2	74.2	88.6	(47.1, 163.1)	91.3
			Sample	27.0	115.1	33.9	(21.9, 72.2)	75.3
			Analysis	20.4	20.4	20.7	(15.4, 30.5)	92.0
1 Outlier Analysis Level	90.1	88.4	Target	73.2	73.4	82.0	(39.3, 137.8)	94.5
			Sample	27.0	30.2	33.9	(21.6, 72.6)	77.8
			Analysis	20.4	81.8	22.5	(16.9, 35.8)	86.4

[a]Coverage% is based on the number of times the true value of the input parameter occurs within the robust CI in 50 000 simulations created by the Matlab program

value. Thus, when this robust method is used on data that includes outlying values, caution is needed in the interpretation of both the bootstrap point estimates and the bootstrap CI's as they relate to the parameters of any supposed underlying ('uncontaminated') distribution. This is particularly the case for smaller values of n (i.e. $n < 100$) or for large n values with more than a small percentage (e.g. > 2 %) of outlying values.

The simulations used input parameters that were derived from a single set of real experimental values (Table 4). Experiments using substantially different input parameters support these findings.

The effects of changing the severity of the outlying value by adding different amounts (the perturbation) to the original simulated values are shown in Figs. 6 and 7. Figure 6 shows that the robust estimate of standard deviation is affected by the outlier, but the effect is bounded as the perturbation increases, whereas the classical estimate increases in an unbounded way and is severely affected once the size of the perturbation increases beyond 200, or approximately 3 × the

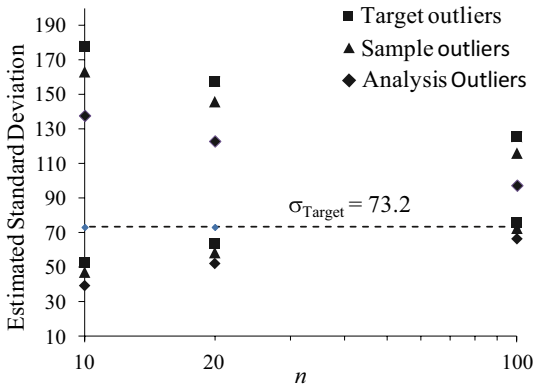

Fig. 5 Decreasing widths of the confidence intervals as n increases, for robust calculation of target level standard deviation where outliers have been applied to 10 % of the targets

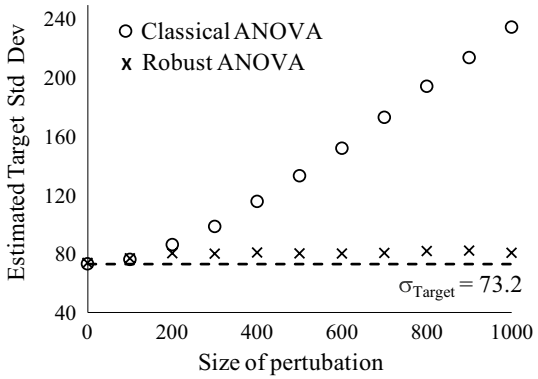

Fig. 6 The effect on estimated classical and robust standard deviations when the size of the perturbation to a target outlier is increased from 0 to 1000 for $n = 20$

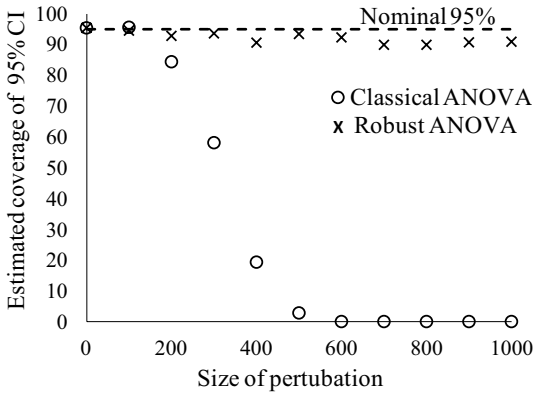

Fig. 7 Coverage percentages of the 95 % CI's when the size of the perturbation to a target outlier is increased from 0 to 1000 for $n = 20$

standard deviation of the target. Figure 7 shows the effect on coverage of the 95 % CI's. Coverage percentages were estimated by counting the number of times the estimated CI contained the true value of the standard deviation of the target. Coverage of the estimated robust 95 % CI falls slightly below 95 % due to the outlier, but stays fairly constant as the perturbation increases. The classical interval is severely affected once the size of the perturbation increases beyond 3 × the standard deviation of the target.

Conclusion

Confidence limits on uncertainties estimated using the duplicate method are potentially useful to the researcher. A mathematical method can be used for normally distributed data. When outlying values are present, the robust ANOVA method consistently gives closer estimates of the parameters of the underlying distribution than classical ANOVA. A bootstrapping method has been devised and incorporated into a new computer program (CI-RANOVA) that allows the confidence limits of robust ANOVA to be estimated in the presence of outliers, for a 3-tier nested experimental design. When data are normally distributed within levels, the CI's produced by the bootstrapping method from robust ANOVA compare well with an established mathematical method for a high n value ($n = 100$), with coverage percentages close to the nominal 95 %, ranging between 94.5 % and 95.4 %. When a low value of n is used ($n = 10$), the coverage percentages are lower at 89.0 % to 92.4 %. This is consistent with previously reported limitations of using the bootstrapping method to estimate variances (See p. 181–183 in [11]).

When data with outlying values are analysed using robust ANOVA, the widths of the confidence intervals decrease with increasing n, as would be expected. If outlying values are deliberately added to a known (underlying) distribution, they result in biases in the robust estimates of variances against the known variances of the underlying distribution. This occurs both at the level where the outliers are present, and also at any levels above this. The biases tend to increase as the proportion of outlying values increases; however, the robust estimates are still much closer estimates of the parameters of the underlying distribution than those obtained from classical ANOVA. The simulated scenarios presented here suggest that the bootstrapping method as described produces estimates of confidence limits that are accurate representations of the random variability in the robust variance estimates. Caution should be used when interpreting these intervals with respect to the parameters of a known or theoretical underlying distribution. This is particularly the case with a low n value and/or proportions of outlying values that exceed 2 %.

References

1. Ramsey MH, Ellison SLR (eds) (2007) Measurement uncertainty arising from sampling: a guide to methods and approaches. Eurachem, EUROLAB, CITAC, Nordtest and the RSC Analytical Methods Committee. https://www.eurachem.org/images/stories/Guides/pdf/UfS_2007.pdf. Accessed 1 Mar 2019
2. Rostron P, Ramsey MH (2017) Quantifying heterogeneity of small test portion masses of geological reference materials by PXRF: implications for uncertainty of reference values. Geostand Geoanalytical Res 41(3):459–473
3. Lyn JA, Ramsey MH, Coad DS, Damant AP, Wood R, Boon KA (2007) The duplicate method of uncertainty estimation: are eight targets enough? Analyst 132:1147–1152
4. AMC (1989) Robust statistics—how not to reject outliers. Part 1, basic concepts. Analyst 114:1693–1697
5. Ramsey MH, Thompson M, Hale M (1992) Objective evaluation of precision requirements for geochemical analysis using robust analysis of variance. J Geochem Explor 44:23–36
6. Rostron P, Ramsey MH (2012) Cost effective, robust estimation of measurement uncertainty from sampling using unbalanced ANOVA. Accred Qual Assur 17:7–14
7. Williams JS (1962) A confidence interval for variance components. Biometrika 49:278–281
8. Graybill FA (1976) Theory and application of the linear model. Duxbury Press, Boston
9. AMC (2014) RANOVA2 computer program. http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/RANOVA2.asp. Accessed 1 Mar 2019
10. JCGM 100 (2008) Evaluation of measurement data—guide to the expression of uncertainty in measurement (GUM 2008). Joint Committee for Guides in Metrology (Sèvres Cedex). http://www.bipm.org/en/publications/guides/gum.html. Accessed 1 Mar 2019
11. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall Inc, Routledge
12. Singh K (1998) Breakdown theory for bootstrap quantiles. Ann Stat 26(5):1719–1732
13. AMC (1989) Robust statistics—how not to reject outliers. Part 2, inter-laboratory trials. Analyst 114:1699–1702