Confidence Interval for Common Mean in Interlaboratory Comparisons with Systematic Laboratory Biases[1]

Viktor Witkovský[a], Gejza Wimmer[a,b,c,d]

[a]Institute of Measurement Science, Slovak Academy of Sciences
Dúbravská cesta 9, 841 04 Bratislava, Slovak Republic
E-mail: witkovsky@savba.sk

[b]Faculty of Natural Sciences, Matej Bel University
Tajovského 40, 974 01 Banská Bystrica, Slovak Republic
[c]Mathematical Institute, Slovak Academy of Sciences
Štefánikova 49, 814 73 Bratislava, Slovak Republic
[d]Faculty of Science, Masaryk University
Janáčkovo nám. 2a, 662 95 Brno, Czech Republic
E-mail: wimmer@mat.savba.sk

Abstract. *We consider the problem of evaluation of the measurement results from the interlaboratory comparisons in metrology. It is assumed that the laboratories have either normally, uniformly, or triangularly distributed systematic errors (biases). We propose an approximate interval estimator for the common mean, i.e. the true value of the measurand. The empirical coverage probabilities of the suggested interval estimator were estimated and compared by large Monte Carlo simulations for different experimental designs. The suggested approach is based on a metrological methodology and is fully consistent with the Supplement 1 to the Guide to the Expression of Uncertainty in Measurement - Propagation of Distributions Using a Monte Carlo Method, [3].*

Keywords: Common mean, interlaboratory comparisons, key comparisons; key comparison reference value KCRV, expanded uncertainty, confidence interval

1. Introduction

In metrology, key comparisons are specific interlaboratory comparisons carried out by the national metrology institutes (NMIs). The purpose of such measurement comparisons between NMIs is to check whether measurements performed in the participating countries are consistent, taking into account the uncertainties assigned to the measurements. In simple key comparisons the participating laboratories measure repeatedly and independently the same physical quantity (measurand) of stable value during the comparison. The uncertainty of the measurement process is influenced by the measurement errors of the participating laboratories and by the systematic laboratory biases, for more details see [1, 4]. One part of the problem, studied by the key comparisons, is determination of the unknown quantity of the measurand. From statistical point of view the problem is known as the common mean problem. It has been studied for a long time, see e.g. [9, 6, 5], however, the metrological specifications, such as the heteroscedasticity

[1] The work was supported by the Slovak Research and Development Agency, grant RPEU-0008-06, by the Slovak Scientific Grant Agency VEGA, grant 1/3016/06, and by the Ministry of Education, Youth and Sports of Czech Republic, LC06024.

k	Laboratory	Country	$\bar{y}_i$	n_i	s_i	$\sigma_{(B),i}$
1	PTB	Germany	0.12662	9	0.0000429	0.0000617
2	BNM-CESTA	France	0.12690	5	0.0005477	0.0003164
3	CSIRO-NML	Australia	0.12670	5	0.0000837	0.0001864
4	CMI	Czech Republic	0.12670	5	0.0002321	0.0003260
5	CSIR-NML	South Africa	0.12710	5	0.0000837	0.0003795
6	CENAM	Mexico	0.12657	5	0.0000826	0.0003142
7	NRC	Canada	0.12650	5	0.0002688	0.0002650
8	KRISS	Korea	0.12659	6	0.0000361	0.0002274
9	NMIJ	Japan	0.12655	4	0.0000818	0.0003137
10	VNIIM	Russia	0.12694	5	0.0001140	0.0002746
11	NIST	United States	0.12640	5	0.0002000	0.0001954
12	Nmi-VSL	The Netherlands	0.12662	5	0.0001171	0.0001560

Table 1: Sample means, number of replications, and corresponding Type A and Type B uncertainties of charge sensitivity measurements of the back-to-back accelerometer for 500 Hz. The systematic errors of the laboratories are assumed to be independent and uniformly distributed with mean values $\beta_i = 0$, for all $i = 1, \ldots, k$, and known standard deviations $\sigma_{(B),i}$.

of measurements, systematic laboratory biases, and the Type A and Type B evaluation of uncertainties in measurement, see [1], leads to new challenges also from statistical perspective, see e.g. [7]. The problem of determination of the appropriate confidence interval for the common mean was not fully resolved until now, especially in situation where the involved laboratories are subject to the systematic errors (biases).

In this paper we study the behavior of an approximate confidence interval for the common mean suggested in [16] and further studied in [17, 18], and consistent with the metrological approach [3, 8]. In metrology, it is referred to as the coverage interval, see [3], and/or key comparison reference value (KCRV) and its expanded uncertainty. The goal of this paper is to study, by the Monte Carlo simulations, the frequentist properties (i.e. the coverage probability) of the suggested interval estimator for the true value of the measurand, say μ.

Let $k \geq 2$ be the number of laboratories, the participants of the key comparisons. We will assume that each laboratory measures the same quantity μ (the true value of the measurand) repeatidly and independently n_i times, $n_i \geq 2$, $i = 1, \ldots, k$. We will consider the following model (structural equations) for the measurement process:

$$Y_{ij} = \mu + B_i + \varepsilon_{ij}, \tag{1}$$

where Y_{ij} denotes the jth measurement in the ith laboratory, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$; $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{(A),i})$ represent the mutually independent normally distributed measurement errors, $\sigma_{(A),i}$ being the unknown standard deviations to be estimated by the standard statistical (sample) methods, i.e. by the Type A evaluation of uncertainties, as defined in [1]. The random variables B_i represent the (unobservable) laboratory systematic effects. The particular statistical approach that is appropriate for the estimation of μ depends on what assumptions are made about the laboratory biases B_i. Here, we will assume that B_i are random variables distributed independently, according to the specified distributions

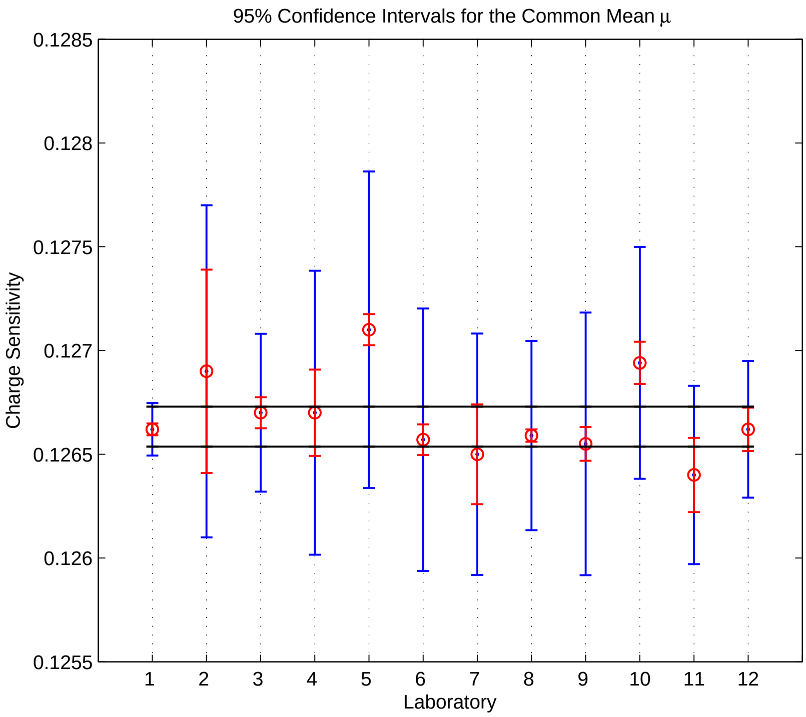

Figure 1: Charge sensitivity measurements, $\bar{y}_i \pm 2 \times (s_i^2/n_i + \sigma^2_{(B),i})^{1/2}$ plotted together with the 95% confidence interval (solid horizontal lines) for μ calculated by the suggested method and given by the equation (8). Numerically, the confidence interval is given by 0.1266327 ± 0.9628e-004.

(e.g. normal, uniform or triangular) with mean values β_i and standard deviations $\sigma_{(B),i}$, to be determined by the non-statistical methods postulated based on scientific judgment, i.e. by the Type B evaluation of uncertainties, see [1]. From this point of view, the parameters β_i and $\sigma_{(B),i}$, $i = 1, \ldots, k$, are considered here to be known (in practical situations they are evaluated by qualified expert's judgment). This case is equivalent to the model described in the International Organization for Standardization⁄s Guide to the Expression of Uncertainty in Measurement [1], see also [3]. The distributions of B_i are usually referred to as Type B distributions, for more details see the Model 3 defined in [7].

The measurement outcome of the key comparisons is given by the laboratory sample means and sample variances, $\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ and $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$, and by full description of the Type B distributions. Note that $\bar{Y}_i$ and S_i^2, for $i = 1, \ldots, k$, are mutually independent variables.

2. Examples

For illustration purposes, we present two examples with output from interlaboratory comparisons. In Table 1 we present the data taken from the *Final report on key comparison CCAUV.V-K1*, see [11]. The key interlaboratory comparisons were taken by 12 National Metrolgy Institutes (NMIs) in the area of vibration (quantity of acceleration) on the measurements of the charge sensitivity of the accelerometer standards (back-to-back accelerometer) at different frequencies and acceleration amplitudes. The resulted 95% confidence intervals are presented in Figure 1.

The second example illustrates the application of the suggested method for data from comparison measurements of time, see [2]. A practical scale of time for world-wide use has two essential elements: a realization of the unit of time and a continuous temporal reference. The reference used is International

k	Standard Clock	Laboratory	$\bar{d}_i$	n_i	s_i	$\sigma_{(B),i}$
1	PTB-CS1	Braunschweig	-15.0	35	5.0	8.0
2	PTB-CS2	Braunschweig	0.2	35	3.0	12.0
3	SYRTE-FOM	Paris	2.1	30	0.1	1.0
4	SYRTE-FOM	Paris	3.8	10	0.2	1.3
5	SYRTE-FOM	Paris	1.0	15	0.2	1.1
6	SYRTE-JPO	Paris	3.8	35	0.8	6.3
7	PTB-CSF1	Braunschweig	1.2	15	1.0	1.0
8	NIST-F1	Boulder	4.1	15	0.3	0.7
9	NPL-CsF1	Teddington	8.2	30	0.7	1.9
10	NICT-CsF1	Tokyo	4.5	15	1.0	1.9

Table 2: Sample means, number of replications, and corresponding uncertainties of the deviations of the International Atomic Time (TAI) frequency with that of the given individual Primary Frequency Standards (Standard Clocks). The systematic errors of the laboratories are assumed to be independent and uniformly distributed with mean values $\beta_i = 0$, for all $i = 1, \ldots, k$, and known standard deviations $\sigma_{(B),i}$.

Atomic Time (TAI), a time scale calculated at the BIPM (Bureau International des Poids et Mesures) using data from some two hundred atomic clocks in over fifty national laboratories. TAI is a realization of coordinate time. The Table 2 gives the mean deviations $\bar{d}_i$, $i = 1, \ldots, k$, obtained on the given period of estimation by comparison of the TAI frequency with that of the given individual Primary Frequency Standards (Standard Clocks). In the table s_i is the uncertainty originating in the instability of the standard and $\sigma_{(B),i}$ is the uncertainty from systematic effects. All values are expressed in 10^{-15} of second.

Based on the available measurements over the considered period (September 26 - October 31, 2007), and taking into account their individual uncertainties, the BIPM estimate of the mean deviation d and its computed expanded standard uncertainty u was stated as $d = 3.1 \times 10^{-15}$ and $u = 0.5 \times 10^{-15}$. On the other hand, the 95% confidence interval for the mean deviation d, calculated by the suggested method (8), and further assuming uniform distribution of the systematic effects, is equal to $\langle 2.1642 \times 10^{-15},\ 3.7483 \times 10^{-15} \rangle$, with the estimated mean deviation $d = 2.9563 \times 10^{-15}$ and with standard uncertainty $u = 0.4078 \times 10^{-15}$. Notice, that in order to be able to compare numerical results with other candidate methods, here we present the resulted values with more significant digits than necessary.

3. Key comparison reference value and its expanded uncertainty

Recently, several papers have been published that suggested methods for deriving KCRV (the estimator of μ) and its expanded uncertainty, under the assumption that the measurement process is influenced by the systematic laboratory effects, fully characterized by their Type B distributions, see e.g. [8, 7, 12] among others. However, only few of them have analyzed the frequentist statistical properties of the suggested interval estimators for μ. Such example is e.g. [7], where the suggested generalized confidence interval estimator was conservative, but in no instance in the considered simulated cases did the empirical coverage probability creep below the nominal rate of 95%. Here we consider an approximate confidence interval for μ, originally suggested in [16], based on the conditional statistical inference consistent with the metrological approach suggested in ISO GUM [1, 3]. It assumes that the measurement process follows the model (1). It combines the posterior information about the true value of the measurand μ, given the observed data from each of the laboratories. In the following paragraph we briefly describe the suggested approach.

Let $\mu_i = \mu + b_i$ denotes the value of the measurand drifted by the systematic laboratory effect (b_i represents the realization of the random variable B_i which is, however, unobservable). If we know the true value of the ith laboratory mean μ_i, then our knowledge about the true value of the measurand μ is given by the probability distribution of the random variable $\tilde{\mu}_{(i)} = \mu_i - B_i$. The value of the parameter μ_i is unknown and could be estimated by the ith laboratory sample mean $\bar{Y}_i$ together with its sample standard deviation $\sqrt{S_i^2/n_i}$. Under the model assumptions (1) the random variable $T_i = (\bar{Y}_i - \mu_i)/\sqrt{S_i^2/n_i}$ has the Student's t distribution with $n_i - 1$ degrees of freedom. Given the observed values $\bar{y}_i$ of $\bar{Y}_i$, and s_i^2 of S_i^2, our knowledge about the true value of the parameter μ_i is given by the distribution of the random variable

$$\tilde{\mu}_i = \bar{y}_i - \sqrt{\frac{s_i^2}{n_i}}T_i, \tag{2}$$

where $\bar{y}_i$ and s_i^2 are considered to be given constants and $T_i \sim t_{n_i-1}$ is a random variable with the Student's t distribution with $n_i - 1$ degrees of freedom. From that, we can express our knowledge about the true value of the measurand μ (based on the information from the ith laboratory) by the distribution of the random variable

$$\tilde{\tilde{\mu}}_{(i)} = \tilde{\mu}_i - B_i = \bar{y}_i - \sqrt{\frac{s_i^2}{n_i}}T_i - B_i. \tag{3}$$

By combining the random variables we can finally express our knowledge about the true value of the measurand μ (based on the information from the k laboratories) by the probability distribution of the random variable

$$\tilde{\tilde{\mu}} = \sum_{i=1}^{k} w_i \tilde{\tilde{\mu}}_{(i)} = \sum_{i=1}^{k} w_i \bar{y}_i - \sum_{i=1}^{k} w_i \sqrt{\frac{s_i^2}{n_i}}T_i - \sum_{i=1}^{k} w_i B_i, \tag{4}$$

where w_i, $\sum_{i=1}^k w_i = 1$, are properly chosen weights, in [16] suggested as

$$w_i = \left(1\Big/\left(\sqrt{\frac{s_i^2}{n_i}}\sqrt{\frac{s_p^2}{n_i}}\frac{n_i-1}{n_i-3} + \sigma^2_{(B),i}\right)\right) \Big/ \left(\sum_{l=1}^{k} 1\Big/\left(\sqrt{\frac{s_l^2}{n_l}}\sqrt{\frac{s_p^2}{n_l}}\frac{n_l-1}{n_l-3} + \sigma^2_{(B),l}\right)\right), \tag{5}$$

where s_p^2 is the pooled variance estimate, $s_p^2 = \sum_{i=1}^k (n_i - 1)s_i^2/(\sum_{i=1}^k n_i - k)$.

Here is the reasoning for selection of the weights given by (5): The goal was to select such weights that in a specific situation with $\sigma^2_{(B),i} = 0$, $i = 1, \ldots, k$, the proposed interval estimator would be the exact $(1 - \alpha) \times 100\%$ confidence interval for μ. In particular, let us consider the model without systematic errors, namely $\bar{Y}_i = \mu + \bar{\varepsilon}_i$, $i = 1, \ldots, k$, $\bar{\varepsilon}_i \sim N(0, \sigma^2_{(A),i}/n_i)$. Note that under given assumption $T_i = (\bar{Y}_i - \mu)/\sqrt{(S_i^2/n_i)} \sim t_{n_i-1}$. Let $W = \sum_{i=1}^k u_i T_i$ where u_i are non-stochastic constants. Based on the above assumptions Fairweather in [5] derived the exact confidence interval for μ of the form

$$\frac{\sum_{i=1}^k \sqrt{n_i/S_i^2}\, u_i \bar{Y}_i}{\sum_{i=1}^k \sqrt{n_i/S_i^2}\, u_i} \pm \frac{q_{1-\alpha/2}}{\sum_{i=1}^k \sqrt{n_i/S_i^2}\, u_i}. \tag{6}$$

where the quantile $q_{1-\alpha/2}$ is implicitly defined by the equation $\Pr\left(|\sum_{i=1}^k u_i T_i| < q_{1-\alpha/2}\right) = 1 - \alpha$. In this set-up, let $w_i = \sqrt{n_i/S_i^2}\, u_i/(\sum_l^k \sqrt{n_l/S_l^2}\, u_l)$ denote the weights of the weighted mean of sample

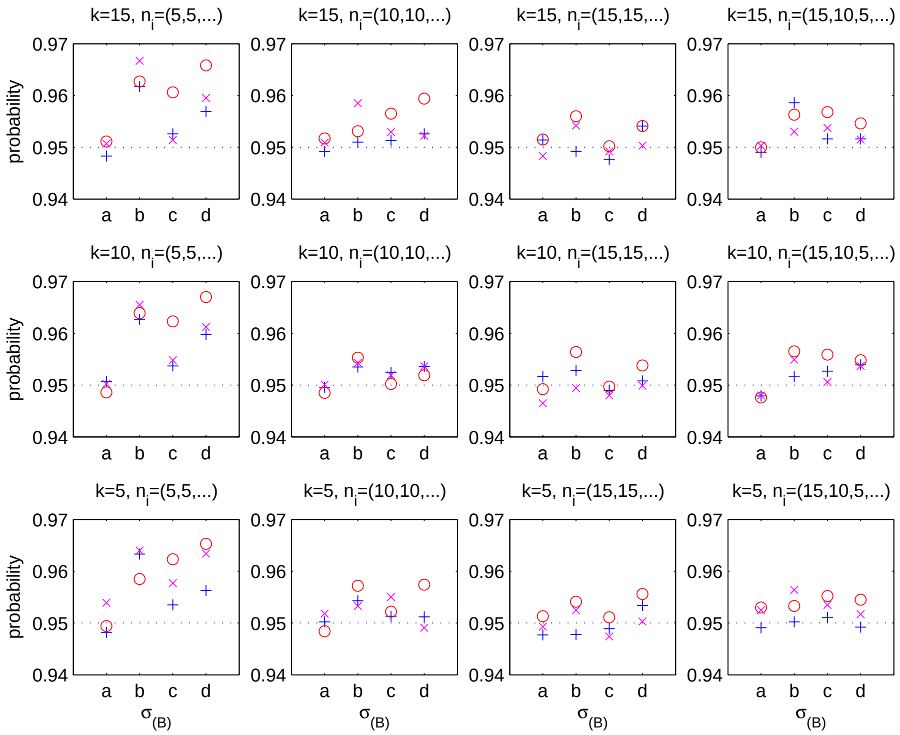

Figure 2: The empirical coverage probabilities of the interval estimator (8) for μ. Here, the systematic errors are assumed to be independent with normal distributions with $B_i \sim \mathcal{N}(0, \sigma^2_{(B),i})$.

means. The confidence interval (6) is exact for any non-stochastic coefficients u_i. It was shown in [15] that if we choose the natuaral weights (i.e. inversly proportional to the sample variances), i.e. $w_i \overset{prop.}{\sim} n_i/S_i^2$, the interval (6) is no more the exact confidence interval on the given significance level α. If $\sigma^2_{(B),i} = 0$, the weights given by (5) are of the form $w_i = \sqrt{n_i/S_i^2}\, u_i/(\sum_l^k \sqrt{n_l/S_l^2}\, u_l)$, where u_i are the non-stochastic coefficients, and so has the property that the interval (6) is an exact $(1-\alpha)\times 100\%$ confidence interval.

So, given the weights (5), the observed value of the key comparison reference value (KCRV) is given as the expected value of the distribution of the random variable $\tilde{\tilde{\mu}}$ and its associated standard uncertainty is given as the standard deviation of $\tilde{\tilde{\mu}}$, see [3], i.e.

$$\mu_{KCRV} = \sum_{i=1}^{k} w_i(\bar{y}_i - \beta_i), \qquad u_{KCRV} = \sqrt{\sum_{i=1}^{k} w_i^2 \left(\frac{s_i^2}{n_i} \frac{n_i - 1}{n_i - 3} + \sigma^2_{(B),i} \right)}. \tag{7}$$

This uncertainty is well defined if $n_i > 3$ for all $i = 1, \ldots, k$. The expanded uncertainty of KCRV is defined as the half-length of the $(1-\alpha)\times 100\%$ coverage interval for μ, see [3], i.e.

$$\left\langle \mu_{KCRV} + q_{\alpha/2},\ \mu_{KCRV} + q_{1-\alpha/2} \right\rangle, \tag{8}$$

where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are the quantiles of the distribution of the random variable $\tilde{\tilde{\mu}} - \mu_{KCRV}$. The quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ could be evaluated approximately by Monte Carlo simulations, or exactly by the algorithm `tdist`, briefly explained in the following section. For more details see also [14].

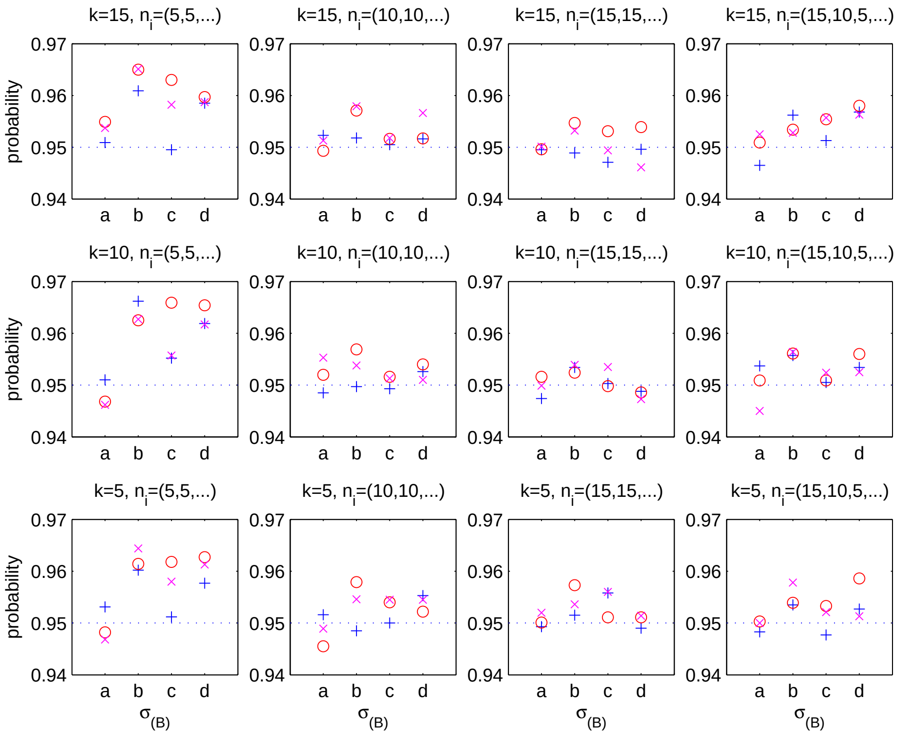

Figure 3: The empirical coverage probabilities of the 95% confidence interval (8) for μ. Here, the systematic errors are assumed to be independent with uniform distributions with $B_i \sim \mathcal{U}(-\delta_i, \delta_i)$, $\delta_i = \sqrt{3}\sigma_{(B),i}$.

4. Algorithm `tdist` for computing the exact distribution of a linear combination of independent random variables

The algorithm `tdist` numerically evaluates the exact distribution of a linear combination of independent random variables with the standard normal, uniform, triangular, and Student's t distributions. The first version of the algorithm (calculates the distribution of a linear combination of independent Student's t random variables) was implemented in MATLAB and R (the environment for statistical computing), and is publicly available at `http://www.mathworks.com/matlabcentral/fileexchange/`, file object Id 4199 (Matlab version), and at `http://cran.r-project.org/src/contrib/Descriptions/tdist.html` (R version).

Here we briefly describe the priciple and the method used in the algorithm `tdist`, for more details see [14]. First, consider a random variable $T = \sum_{i=1}^{k} \lambda_i T_i$, i.e. a linear combination of independent Student's t random variables with ν_i, $i = 1, \ldots, k$, degrees of freedom. Let $\phi_{T_i}(t)$ denote the characteristic function of T_i. The characteristic function of T is

$$\phi_T(t) = \phi_{T_1}(\lambda_1 t) \cdots \phi_{T_k}(\lambda_k t). \tag{9}$$

where

$$\phi_{T_i}(\lambda_i t) = \frac{1}{2^{\frac{\nu_i}{2}-1}\Gamma(\frac{\nu_i}{2})} \left(\nu_i^{\frac{1}{2}} |\lambda_i t|\right)^{\frac{\nu_i}{2}} K_{\frac{\nu_i}{2}} \left\{\nu_i^{\frac{1}{2}} |\lambda_i t|\right\}, \tag{10}$$

where ϕ_{T_i} denotes the characteristic function of the random variable T_i with Student's t distribution with ν_i degrees of freedom, and $K_\alpha\{z\}$ denotes the modified Bessel function of the second kind. For detailed derivation see [13]. Note that the characteristic function of the Student's t random variable is a real

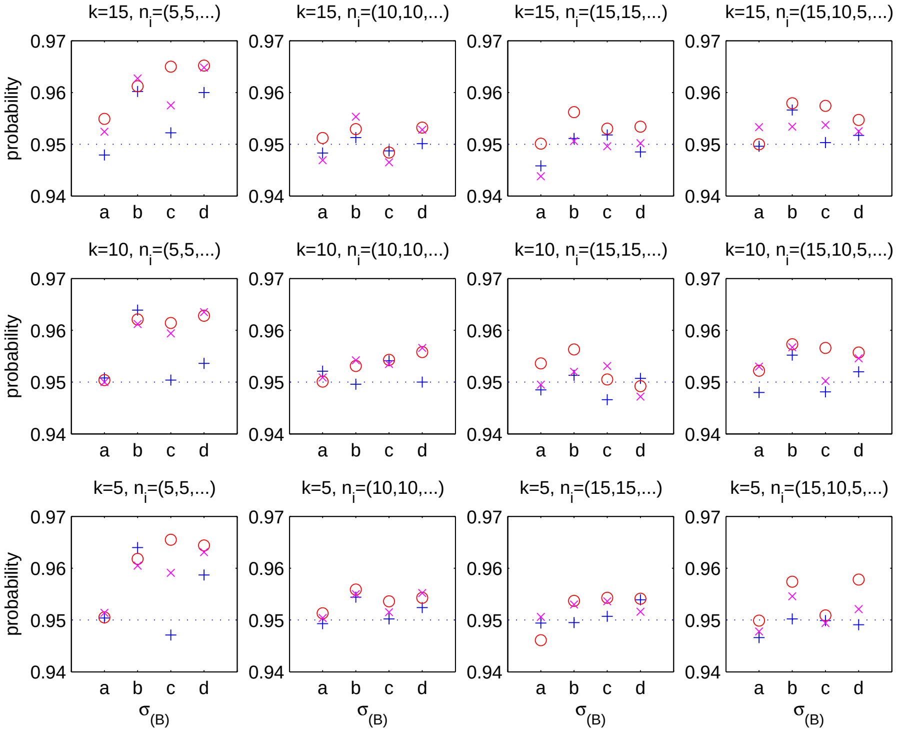

Figure 4: The empirical coverage probabilities of the 95% confidence interval (8) for μ. Here, the systematic errors are assumed to be independent with triangular distributions $B_i \sim \Delta(-\delta_i, \delta_i)$, $\delta_i = \sqrt{6}\sigma_{(B),i}$.

function. Then, the distribution function $F_T(x) = \Pr\{T \leq x\}$ is given by

$$F_T(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \Im\left(\frac{e^{-itx}\phi_T(t)}{t}\right) dt = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{\sin(tx)\phi_T(t)}{t}\, dt, \tag{11}$$

and the probability density function $f_T(x)$ of T is given by

$$f_T(x) = \frac{1}{\pi}\int_0^\infty \Re\left(e^{-itx}\phi_T(t)\right) dt = \frac{1}{\pi}\int_0^\infty \cos(tx)\phi_T(t)\, dt. \tag{12}$$

The algorithm `tdist` evaluates the integrals in (11) and (12) by multiple p-points Gaussian quadrature over the real interval $t \in (0, 10\pi)$ which involves base points b_{ij} and the weight factors w_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, m$.

$$F_T(x) \approx \frac{1}{2} + \frac{1}{\pi}\sum_{j=1}^{m}\sum_{i=1}^{p}\frac{\sin(b_{ij}x)}{b_{ij}}W_{ij}, \tag{13}$$

$$f_T(x) \approx \frac{1}{\pi}\sum_{j=1}^{m}\sum_{i=1}^{p}\cos(b_{ij}x)W_{ij}, \tag{14}$$

with $W_{ij} = w_{ij}\phi_T(b_{ij})$. Notice, that for evaluation of $F_T(x)$ and $f_T(x)$ in many different points the algorithm requires only one evaluation of the weights W_{ij}, which directly depend on the characteristic function $\phi_T(\cdot)$ and does not depent on x.

The algorithm could be easily generalized to other symmetric distributions with known characteristic functions. Here we consider independent random variables $Z \sim \mathcal{N}(0,1)$ (standard normal distribution), $U \sim \mathcal{U}(-1,1)$ (uniform distribution over $(-1,1)$), and $\Delta \sim \mathcal{T}(-1,1)$ (triangular distribution over $(-1,1)$) with

$$\phi_Z(\lambda t) = \exp\left\{-\frac{(\lambda t)^2}{2}\right\}, \tag{15}$$

$$\phi_U(\lambda t) = \frac{\sin(\lambda t)}{\lambda t}, \tag{16}$$

$$\phi_\Delta(\lambda t) = \frac{(2 - 2\cos(\lambda t))}{(\lambda t)^2}, \tag{17}$$

where λ stands for arbitrary multiplication coefficient.

5. Simulation study on empirical coverage probability of the approximate confidence interval for μ

In order to check the frequentist statistical properties of the interval estimator (8) we have performed a large simulation study. We have considered three types of the distribution for the systematic laboratory effects B_i. The normal (gaussian) distribution $B_i \sim \mathcal{N}(0, \sigma^2_{(B),i})$, the uniform (rectangular) distribution $B_i \sim \mathcal{U}(-\delta_i, \delta_i)$, $\delta_i = \sqrt{3}\sigma_{(B),i}$, and the triangular distribution $B_i \sim \Delta(-\delta_i, \delta_i)$, $\delta_i = \sqrt{6}\sigma_{(B),i}$, $i = 1, \ldots, k$. Without loss of generality we have assumed that $\mu = 0$, and that the distributions of the laboratory effects are centered at zero, i.e. $\beta_i = 0$. We have considered the following specific designs of the interlaboratory comparison experiment: $k \in \{5, 10, 15\}$, $n_i = 5$, $n_i = 10$, $n_i = 15$, and $n_i \in \{15, 10, 5, 15, 10, 5, 15, 10, 5, 15, 10, 5, 15, 10, 5\}$, $i = 1, \ldots, k$. Figure 2 presents the empirical coverage probabilities of the interval estimator (8) for the true value of the measurand μ, based on 10000 Monte Carlo simulations from the model (1) with $B_i \sim \mathcal{N}(0, \sigma^2_{(B),i})$, $i = 1, \ldots, k$, for each specific design with the nominal significance level $\alpha = 0.05$. By a we denote the designs with $\sigma_{(B),i} = 0$, by b we denote the designs with $\sigma_{(B),i} = 1$, by c we denote the designs with $\sigma_{(B),i} = 5$, and by d we denote the designs with $\sigma_{(B),i} \in \{1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5\}$, $i = 1, \ldots, k$. The coverage probabilities for designs with $\sigma_{(A),i} = 1$, $i = 1, \ldots, k$, are plotted by the symbol $+$, for designs with $\sigma_{(A),i} = 5$, $i = 1, \ldots, k$, are plotted by the symbol $\circ$, and for designs with $\sigma_{(A),i} \in \{1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5\}$, $i = 1, \ldots, k$, are plotted by the symbol $\times$. Similarly, Figure 3 presents the empirical coverage probabilities with $B_i \sim \mathcal{U}(-\delta_i, \delta_i)$, $\delta_i = \sqrt{3}\sigma_{(B),i}$, $i = 1, \ldots, k$, and Figure 4 presents the empirical coverage probabilities with $B_i \sim \Delta(-\delta_i, \delta_i)$, $\delta_i = \sqrt{6}\sigma_{(B),i}$, $i = 1, \ldots, k$, respectively.

6. Conclusions

The general conclusion of the present simulation study is the observation that the interval estimator (8) shows good statistical properties with the empirical coverage probabilities close to the nominal 95% level for all three considered types of distributions of the laboratory systematic effects. For designs with smaller number of replication ($n_i = 5$, $i = 1, \ldots, k$) the confidence interval (8) is slightly conservative with the values of the emprirical coverage probabilities growing up to 96.5%.

References

[1] BIPM, IEC, IFCC, ISO, IUPAC, OIML (1995). *Guide to the expression of uncertainty in measurement*. 2nd ed., Sèvres: Bureau International des Poids et Mesures.

[2] BIPM (2007). CIRCULAR T 238. NOVEMBER 14, 2007. Bureau International des Poids et Mesures, ISSN 1143-1393.

[3] Joint Committee for Guides in Metrology (2006). *Evaluation of Measurement Data -, Supplement 1 to the Guide to the Expression of Uncertainty in Measurement -, Propagation of Distributions Using a Monte Carlo Method*. Final draft, September 2006.

[4] CIPM (1999). *Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Institutes.* Sèvres: Bureau International des Poids et Mesures.

[5] W.R. Fairweather (1972). A method of obtaining an exact confidence interval for the common mean of several normal populations. *Applied Statistics*, 21, 229–233.

[6] F.A. Graybill and R.B. Deal (1959). Combining unbiased estimators. *Biometrics*, 15, 543–550.

[7] H.K. Iyer, C.M.J. Wang, and T. Mathew (2004). Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, 99, 1060–1071.

[8] R.N. Kacker, R.U. Datla and A.C. Parr (2003). Statistical interpretation of key comparison reference value and degrees of equivalence. *Journal of Research of the National Institute of Standards and Technology*, 108, No. 6, 439–446.

[9] P. Meier (1953). Variance of a weighted mean. *Biometrics*, 9, 59–73.

[10] B. Toman (2006). Linear statistical models in the presence of systematic effects requiring a Type B evaluation of uncertainty. *Metrologia*, 43, 27,-33.

[11] H.J. von Martens,C. Elster, A. Link, a. Täubner, and W. Wabinski (2003). Final report on key comparison CCAUV.V-K1. *Metrologia*, 40, (2003), Tech. Suppl. 09001.

[12] C.M. Wang and H.K. Iyer (2006). A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties. *Measurement*, 39, 856-,863.

[13] V. Witkovský (2001). On the exact computation of the density and of the quantiles of linear combinations of t and F random variables. *Journal of Statistical Planning and Inference*, 94 (1), 2001, 1-13.

[14] V. Witkovský (2004). Matlab algorithm TDIST: The distribution of a linear combination of Student's t random variables. *In Jaromir Antoch (Ed.): COMPSTAT. Proceedings in Computational Statistics. 16th Symposium Held in Prague, Czech Republic, 2004*, Physica-Verlag, 1995–2002.

[15] V. Witkovský (2005). Comparison of some exact and approximate interval estimators for common mean. In: *Measurement Science Review*, 5 (1), 2005, 19-22.

[16] V. Witkovský and G. Wimmer (2006). Exact and approximate confidence intervals for the comparison reference value. In: *PROBASTAT 2006 The Fifth International Conference on Probability and Mathematical Statistics, Abstracts*. Smolenice Castle, Slovak Republic, June 5-9, 2006, p.49.

[17] V. Witkovský and G. Wimmer (2007). Method for evaluation of the key comparison reference value and its expanded uncertainty based on metrological approach. In: *I. Frollo, J. Maňka, and V. Juráš (Eds.): MEASUREMENT 2007, Proceedings of the 6th International Conference on Measurement Smolenice, Slovakia*, May 20-24 2007, 26-29.

[18] V. Witkovský and G. Wimmer (2007). Key comparison reference value and its expanded uncertainty under normally, uniformly and triangularly distributed laboratory biases. In: *Proceedings ISI 2007 - Lisboa. Bulletin of the International Statistical Institute, 56th Session. Lisboa*, 22-29 August 2007, 2007, CD-ROM.