 DOI: 10.1051/metrology/201712003

Analysis of interlaboratory comparison when the measurements are not normally distributed

Alexandre Allard[1,*], *Soraya* Amarouche[1*]

[1]Laboratoire National de métrologie et d'Essais, 1 rue Gaston Boissier 75724 Paris Cedex 15, France

Abstract. Testing laboratories are more and more concerned with the characterization of their measurement processes. In particular, the standard ISO 17025[2] requires the accredited laboratories to participate to interlaboratory comparisons to evaluate their proficiency to realize the measurement. Different statistical methods are available to exploit the results of this type of comparisons. In our study, first we have evaluated the reference value and the proficiency standard deviation with ISO 5725-2 standard [1] and in second part the calculation of statistical indicator Z-score with ISO13528 [5] standard to assess of proficiency of laboratory with the first estimated parameters. However, these statistical methods rely on the assumption that the measurement results are normally distributed. Based on measurements expressed in dBµV/m, which is a log transformation of an electric field level expressed in µV/m, this paper aims at the comparison between the statistical analysis of data expressed in the two different units and relates these results to the assumed statistical assumptions.

1. Introduction

1.1. The objectives of an interlaboratory comparison

The interlaboratory comparisons are defined as the organization, the execution and the exploitation of measurements, testing or calibrations on similar items (samples, standards, reference solutions) by at least two different laboratories in predetermined conditions. The implementation of an interlaboratory comparison has different objectives (cf.Fig 1):

-Evaluation of the performance of the laboratories. The objective consists to estimate and to demonstrate the proficiency of laboratories to realize the measurement. Each participant implemented his routinely measurement method.

-Estimation of accuracy (trueness and precision) of measurement method. The objective consists to evaluate the performance of the measurement method through repeatability and reproducibility standard deviation. Each participant implemented the same measurement method.

* Corresponding authors: alexandre.allard@lne.fr; soraya.amarouche@lne.fr

Article available at http://cfmetrologie.edpsciences.org or https://doi.org/10.1051/metrology/201712003

-Attribution of a consensual value to a characteristic of a material. The objective consists to assign a reference value to a material. The participating laboratories must be specialized in the determination of the concerned characteristic.

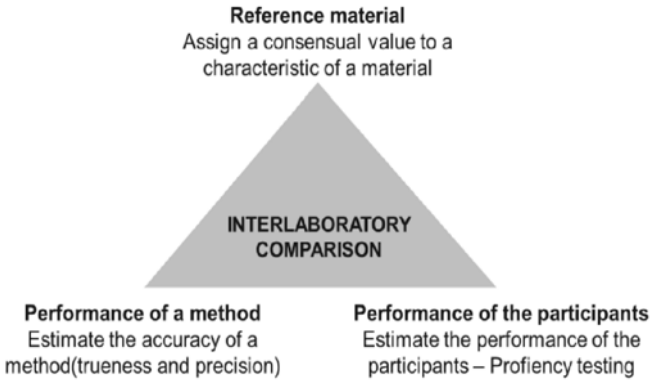

Fig.1. Objectives of interlaboratory comparisons

In this study, several objectives are realized: The performance of the method, the performance of the participants and the measurement uncertainty evaluation. This is possible with the implementation of the same testing method by all participants described in the next section.
First, with the results of participants, we have applied the method describe in ISO 5725 part2 standard [1] to evaluate the overall mean, repeatability standard deviation and reproducibility standard deviation. Second, using the statistical parameters above, the performance of each participant is evaluated with a Zscore, a statistical indicator from ISO 17043[4]. At the end, in accordance with ISO 21748[6] standard, the evaluation of uncertainty of measurement are calculated with reproducibility standard deviation.

However, all these statistical methods are conditioned to some assumptions which are often made without being checked, in particular the Gaussian behaviour of the observations.

1.2. Testing method of all participants

In order to ensure quality control, the participants in the Eurolab France (a professional association of laboratory) dedicated to Electromagnetic Compatibility (EMC) regularly organise interlaboratory comparison scheme. For each scheme, a protocol is defined for a specific measurand. In this paper, we consider the measurement of the electric field emitted by an electronic device according to the standard EN 55016-2-3:2010[9]. To this extent, the device (a comb generator coupled with an omnidirectional antenna) shown in Fig.2 was circulated between the participants who performed the required measurement within their own facilities.

Fig.2. Illustration of the Device under testing

The measurement is performed in an anechoic room to avoid electromagnetic perturbations of the surroundings. The device is positioned at a distance d = 3 m from the reference point of the antenna, which is at a height denoted as h (see Fig.3).

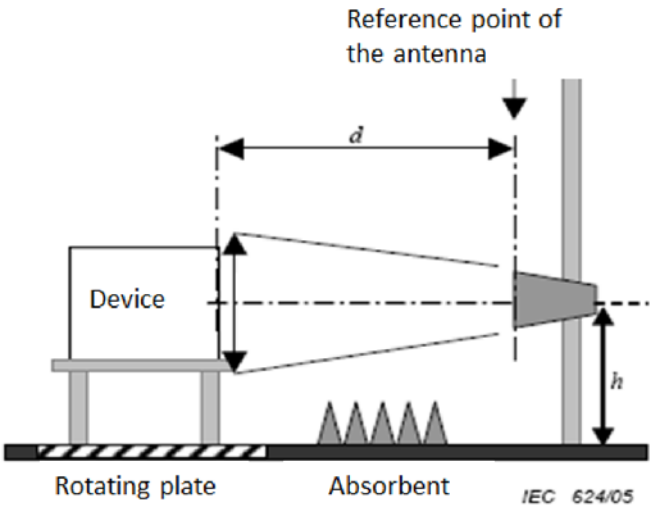

Fig. 3. Representation of the measurement setup

The participants are asked to perform the measurement both in vertical and horizontal polarisations. For each polarisation, a set of 9 frequencies is chosen and the maximum value for the electric field is reported for each of the 9 frequencies.

The measurement results are expressed in dBµV/m, which is a nonlinear transformation of the corresponding SI unit for an electric field: µV/m.
Let y be the measurement expressed in dBµV/m and x the one expressed in µV/m:

$$\begin{cases} y = 20\,log(x) \\ x = 10^{\frac{y}{20}} \end{cases}$$

In order to evaluate the repeatability and reproducibility of the measurement method, 4 independent and repeated measurements are performed for each polarization and each frequency.

As a consequence, the statistical analysis is performed for each polarization and each frequency: 18 levels are available for the comparison. In this paper, we will focus on the results for some of these 18 levels to illustrate our purpose.

2. Implementation of the interlaboratory comparison

This interlaboratory comparison follows a process defines in a plan of campaign that described the collect of the participants' results, the statistical methods of exploitation and the parameters to be estimated. This process is realized on the mesurande in its two expressions: dBµV/m and µV/m.

The final objective aims at raising awareness of the underlying assumptions when using a statistical method and describes the implications of their inappropriate use. Finally, we discuss the choice of the suitable unit of the measurand to apply the statistical analysis for our example.

2.1. Results of interlaboratory comparison organized by Eurolab

The results of the interlaboratory comparison were expressed in dBµV/m, which is not a SI unit, but a convenient working unit in the field of EMC. However, a request was made to perform the analysis in µV/m. In this section, we present the results when considering the data in both units and we will conclude in the discussion regarding the best choice for the purpose of the statistical analysis.

2.1.1.Data

Each of the 22 participating laboratories performed a set of 4 repeated measurements for 9 frequencies of the measurement domain and in 2 polarizations. For clarity, we only present in this paper the results obtained in the horizontal polarization, for the frequency 2.25 GHz (cf. Table 1).

 DOI: 10.1051/metrology/201712003

Table 1. : Results in horizontal polarization, at 2.25 GHz, in dBµV/m and in µV/m.

	dBµV/m				**µV/m**			
Lab 1	64.6	64.4	64.2	64.8	1704.1	1665.3	1621.8	1729.8
Lab 2	56.5	57.5	57.9	56.4	671.4	746.4	780.7	657.7
Lab 3	52.6	52.9	48.2	44.0	426.6	441.6	257.0	158.5
Lab 4	60.7	65.5	65.6	64.3	1083.9	1883.6	1905.5	1640.6
Lab 5	54.8	55.3	58.5	59.5	551.4	582.1	841.4	940.8
Lab 6	57.7	53.4	52.8	55.2	766.1	468.4	434.2	573.3
Lab 7	61.8	62.1	60.9	60.6	1230.3	1273.5	1109.2	1071.5
Lab 8	51.0	41.1	42.2	38.3	354.8	113.5	128.8	82.2
Lab 9	55.4	54.5	55.2	54.4	588.8	530.9	575.4	524.8
Lab 10	49.8	49.7	49.9	50.0	309.0	305.5	312.6	316.2
Lab 11	61.7	63.7	63.4	61.3	1216.2	1531.1	1479.1	1161.4
Lab 12	63.0	62.6	63.3	63.2	1404.6	1349.7	1455.6	1439.5
Lab 13	56.1	54.4	60.5	55.1	638.3	524.8	1059.3	568.9
Lab 14	57.0	57.9	57.3	57.9	707.9	785.2	732.8	785.2
Lab 15	50.0	54.7	56.7	54.0	317.0	543.9	680.8	501.8
Lab 16	58.4	58.4	57.8	58.1	831.8	831.8	776.2	803.5
Lab 17	46.7	45.9	46.7	48.2	216.3	196.6	215.8	256.4
Lab 18	68.2	67.3	68.1	68.9	2570.4	2317.4	2541.0	2786.1
Lab 19	52.1	52.5	53.7	52.4	402.7	421.7	484.2	416.9
Lab 20	59.5	60.3	59.2	57.9	944.1	1035.1	912.0	785.2
Lab 21	69.2	68.1	67.8	69.0	2884.0	2541.0	2454.7	2818.4
Lab 22	51.9	49.1	41.9	47.3	393.6	285.1	124.5	231.7

First, it can be observed that the data can be considered as normally distributed when expressed in dBµV/m, but not when expressed in µV/m, as pointed out by the two statistical tests for normality in Table 2 : Lilliefors and Anderson-Darling [7,8].

Table 2. Result of the normality tests.

Normality Tests	***Lilliefors***	***Anderson-Darling***
p-value in dBµV/m	0.50	0.70
p-value in µV/m	**< 0.001**	**< 0.0005**

A p-value lowers than 0.05 indicates a significant deviation of the sample from the Gaussian assumption.

2.2. Evaluation of the performance of a measurement method (ISO 5725-2)

2.2.1.Statistical procedures

In order to evaluate the performance of a measurement method, guidance is provided in the standard ISO 5725-2. Before exploiting the results of the participants, it is necessary to make sure that the results arise from the same process of measurement by applying test of homogeneity. This homogeneity tests are performed to detect potential outliers among the results.

First, the Cochran's test must conclude to homogeneity of the variances of the participants. If not, this means that one of the variances associated with a laboratory is considered as significantly different from the others. In this case, the repeated measurements of the laboratory are investigated: if they are consistent, then the laboratory is removed from the statistical analysis (because of a too high variance) whereas if an outlier is found, only this outlier value is discarded and the other results are kept in the analysis with updated mean value and standard deviation for the laboratory. The same procedure is then applied until all the variances are homogeneous.

In a second step, a Grubbs' test is performed to ensure that all the mean values from the different laboratories are consistent. If not, outlier laboratories are also discarded from the statistical analysis. The goal of this procedure is to avoid the impact of outliers on the estimation of the overall mean value and the repeatability and reproducibility standard deviations.

2.2.2.Cochran's homogeneity test for the variances

Let p denote the number of participants, the principle of Cochran's test is to test the assumption $H_0 : \sigma_1^2 = \cdots = \sigma_i^2 = \cdots = \sigma_p^2$ against the assumption $H_1 : \max(\sigma_i^2) > \sigma_j^2, j \neq i$.
To this extent, the following Cochran's statistic is obtained thanks to the results of the comparison:

$$C = \frac{s_{max}^2}{\sum_{i=1}^{p} s_i^2}$$

Under the assumption H_0 of an equality of the variances, C is supposed to be distributed as a Cochran's distribution. As a result, the observed value C is compared with the critical value in the Cochran's table for p participants and n repeated measurements.
This test is commonly used in the analysis of interlaboratory comparisons. However, its conclusions are valid under the assumption that the measurements are distributed as a Gaussian distribution.

2.2.3.Grubbs' homogeneity test for the mean values

The Grubb's test aims at the identification of an outlier, either among the mean values of the different laboratories or among the repeated measurements of a single laboratory.

If $\bar{x}$ denotes the mean values of the set of observations and s their standard deviation, the test statistic may be either $G_p = \frac{x_p - \bar{x}}{s}$, if one wants to test whether the maximal value is an

 DOI: 10.1051/metrology/201712003

outlier, or $G_1 = \frac{\bar{x}-x_1}{s}$, if one is interested in the minimal value. Then the considered quantity is compared with the critical value in the Grubbs' table. However, the validity of the Grubbs' test is also conditioned to the Gaussian behaviour of the observations.

In Fig.4 & Fig5, we provide a representation of all results expressed in both units. The red dots (laboratory 8: figure 4 and Laboratory 4 figure 5) correspond to observations which were discarded after the homogeneity tests (Cochran and Grubbs).

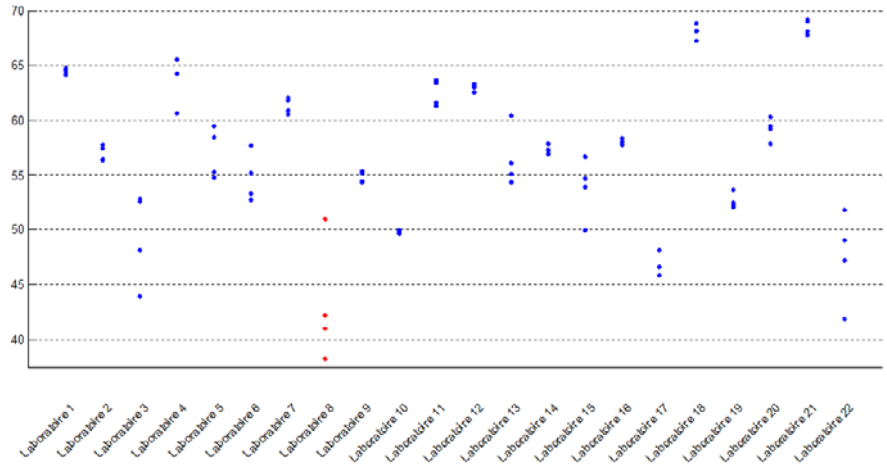

Fig. 4. Graph of the results from each laboratory with analysis in dBµV/m

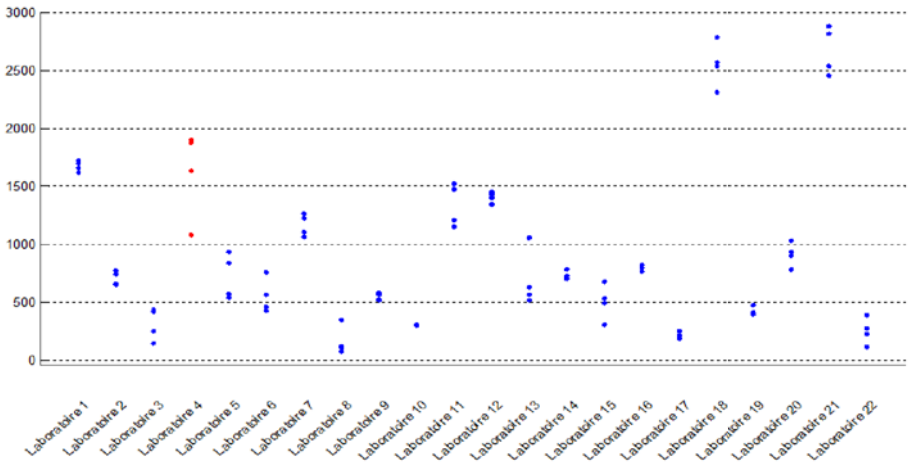

Fig. 5. Graph of the results from each laboratory with analysis in µV/m

In both cases, the outlier laboratory has a too large variance and has been discarded through the Cochran's test. However, it is not the same laboratory in both cases. Indeed, as the transformation is nonlinear, it has an effect on the variances. It can be observed also that a transformation in µV/m results in a higher spread of the measurement results.

2.2.4.Statistical parameters of performance of measurement method

Table 3: Results of the participants

Laboratory	Results
Labo1	y11, y12, y13, y14,……..,y1n
Labo 2	y21, y22, y23, y24,……..,y2n
Labo 3	y31, y32, y33, y34,……..,y3n
….	……..
Labo p	yp1, yp2, yp3, yp4,……..,ypn

After elimination with the tests of homogeneity, we obtain p ' participating laboratories with p ' <= p.

Table 4. Results of the participants after homogeneity tests

Laboratory	Results after homogeneity tests (Cochran and Grubbs)
Labo1	y11, y12, y13, y14,……..,y1n
Labo 2	y21, y22, y23, y24,……..,y2n
Labo 3	y31, y32, y33, y34,……..,y3n
….	……..
Labo p'	yp'1, yp'2, yp'3, yp'4,……..,yp'n

With the following descriptive statistics, table 5.

Table 5. Results of the participants after homogeneity tests

mean	standard-deviation
y1	**S1**
y2	**S2**
y3	**S3**
…	**…**
yp'	**Sp'**

The Evaluation of the parameters of precision (standard deviation of repeatability S_r and reproducibility S_R) also the parameter of position (the overall average) on the results on table 4 using formulas (1), (2) and (3) below.

Overall Average (1)

$$\bar{\bar{y}} = \frac{\sum_{p'} \bar{y}_i}{p'}$$

Repeatability standard deviation (2)

$$S_r = \sqrt{\frac{\sum_{p'} s_i^2}{p'}}$$

Reproducibility standard deviation (3)

$$s_R = \sqrt{\frac{1}{p'-1}\sum_{p'}(\bar{y}_i - \bar{\bar{y}})^2 + \frac{n-1}{n}s_r^2}$$

 DOI: 10.1051/metrology/201712003

The corresponding evaluated parameters are presented in Table 6.

Table 6. The evaluation of the performance of the measurement method in dBµV/m and in µV/m.

Statistical parameters	dBµV/m	µV/m
Overall average	57.5	895.32
Repeatability standard deviation s_r	1.86	126.59
Reproducibility standard deviation s_R	6.48	710.48

2.3. Evaluation of uncertainty of measurement

Alternately to the GUM [3] ,Guide for the expression of the uncertainty of measure (reference method for the evaluation of the uncertainty of measure), we can use the standard deviation of reproducibility obtained in a study of interlaboratory comparison using the standard ISO 5725-2 as an estimation of the standard uncertainty.
So, for every studied frequency, we have:

$$u(y) = S_R$$

The direct calculation in µV/m seems unrealistic in the sense that the standard deviation is of the order of magnitude of the overall mean. However, when data are expressed in dBµV/m, it is possible to compute an approximately 95% coverage interval using a coverage factor $k = 2$, which is adequate because of the Gaussian behaviour of the results.

In case it is required to express the measurement result in µV/m, care should be taken while applying the transformation. Indeed, Measurements expressed in dBVµ/m can be transformed in µV/m, but such transformation is not allowed for the variance (and thus for the standard uncertainty), as it is a nonlinear transformation. The corresponding results are represented in Fig.6, with the individual mean value for each participant.

 DOI: 10.1051/metrology/201712003

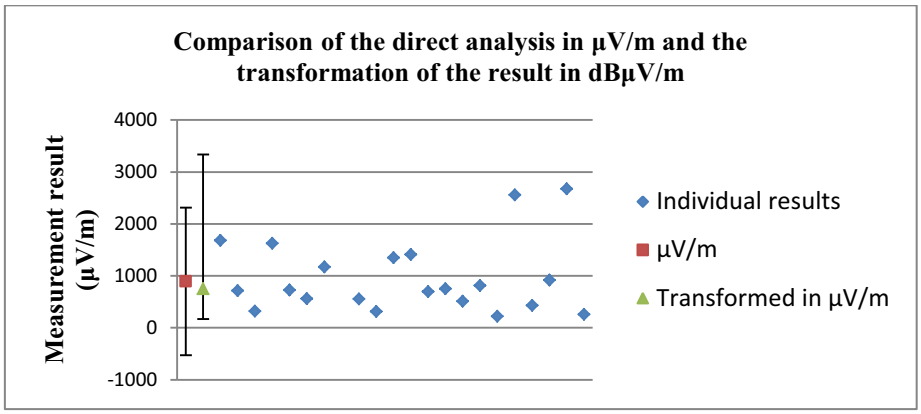

Fig. 6. Comparison of the coverage intervals obtained with a direct analysis in µV/m or with an analysis in dBµV/m followed by a transformation of the coverage interval

First, the coverage interval obtained for a direct analysis in µV/m overlaps 0, whereas the intensity of the electrical field is supposed to have a positive value. Of course, such a coverage interval was obtained with a "naïve" assumption of a Gaussian behaviour, which is false as explained above. On the other hand, the transformation of the coverage interval obtained in dBµV/m encloses only positive values, which makes it already more reliable. But it also has an asymmetric shape: the lower bound is closer to the overall mean than the upper bound. This is a consequence of the nonlinear transformation used. Moreover, Fig 6 shows that this last coverage interval is consistent with all individual data when expressed in µV/m.

Table 6. The evaluation of the performance of the measurement method (dBµV/m andvµV/m)

Frequency (GHz)	Horizontal Polarization				
	Mean (dBµV/m)	Expanded Uncertainty k=2 (dBµV/m)	Uncertainty interval (dBµV/m)	Mean (µV/m)	Uncertainty interval (µV/m)
1,15	58,0	10,0	[47.9; 68.1]	792.5	[248.3; 2529.3]
1,5	52,4	14,7	[37.3; 66.2]	386.8	[73.2; 2044.1]
1,8	53,1	9,5	[43.4; 62.0]	432.0	[148.4; 1257.5]
2,25	58,5	12,3	[44.5; 70.5]	749.9	[168.7; 3334.3]
2,6	56,7	10,3	[45.2; 65.7]	592.9	[182.0; 1932.0]
3,4	57,6	7,6	[49.8; 65.1]	746.4	[309.0; 1803.0]
3,95	54,7	12,0	[42.6; 65.9]	515.8	[134.4; 1979.2]
4,55	56,4	8,3	[48.0; 64.4]	644.2	[250.6; 1655.8]
5,35	56,4	13,7	[42.5; 69.4]	625.9	[132.6; 2954.6]

2.4. Results of the proficiency testing (ISO 13528)

The evaluation of the proficiency of laboratories bases on:

-An assigned value *Xpt* who can be calculated by several methods. For this study the assigned value will be taken equal to the overall average from the exploitation of the results above (Table 6). $X_{pt} = \bar{\bar{y}}$

-A proficiency standard deviation can be fixed or calculated. In our case, we have used the standard deviation of reproducibility SR estimated below in table 6. $\hat{\sigma}_{pt} = S_R$

The statistic of performance estimates the proficiency of the participant to realize the testing measurement. There are various statistics of performance (Zscore, Difference). In the configuration of this interlaboratory comparison, we used Zscore as formula (4) below.

$$Z_{score} = \frac{X_{lab} - X_{pt}}{\hat{\sigma}_{pt}}$$

The interpretation of Z-score :

- If |z| <= 2: the performance of the laboratory is satisfactory.
- If 2<|z ¦<= 3: then the performance of the laboratory is debatable, we generate a signal of warning;
- If |z |> 3, then the performance of the laboratory is "unsatisfactory", and we generate a signal of action

The table 7 below is an example represents Z-scores of every laboratory for one frequency and horizontal polarization by using results on both units.

 DOI: 10.1051/metrology/201712003

Table 7. The evaluation of the performance of the measurement method (dBµV/m andvµV/m) at frequency 2,25 GHz

Frequency : 2.25GHz				
Reference value dBµV/m : 57,50			**Reference value dBµV/m : 895,32**	
proficiency testing standard deviation dBµV/m : 6,48			**proficiency testing standard deviation µV/m : 710,48**	
Laboratory	**Result**	**Zscore**	**Result**	**Zscore**
labo 1	**64,51**	**1,1**	**1680,25**	**1,1**
labo 2	**57,05**	**-0,1**	**714,05**	**-0,3**
labo 3	**49,43**	**-1,2**	**320,92**	**-0,8**
labo 4	**64,03**	**1**		
labo 5	**57,03**	**-0,1**	**728,92**	**-0,2**
labo 6	**54,75**	**-0,4**	**560,5**	**-0,5**
labo 7	**61,35**	**0,6**	**1171,12**	**0,4**
labo 8	**43,15**	**-2,2**	**169,82**	**-1,0**
labo 9	**54,88**	**-0,4**	**554,98**	**-0,5**
labo 10	**49,85**	**-1,2**	**310,83**	**-0,8**
labo 11	**62,53**	**0,8**	**1346,95**	**0,6**
labo 12	**63**	**0,8**	**1412,35**	**0,7**
labo 13	**56,53**	**-0,2**	**697,82**	**-0,3**
labo 14	**57,53**	**0**	**752,78**	**-0,2**
labo 15	**53,85**	**-0,6**	**510,88**	**-0,5**
labo 16	**58,18**	**0,1**	**810,82**	**-0,1**
labo 17	**46,86**	**-1,6**	**221,28**	**-0,9**
labo 18	**68,13**	**1,6**	**2553,73**	**2,3**
labo 19	**52,68**	**-0,7**	**431,38**	**-0,7**
labo 20	**59,23**	**0,3**	**919,1**	**0,0**
labo 21	**68,53**	**1,7**	**2674,52**	**2,5**
labo 22	**47,55**	**-1,5**	**258,73**	**-0,9**

Conclusion

As a conclusion, our article aims at pointing out the importance of the verification of the assumptions underlying the use of statistical methods. In metrology, a wide majority of the statistical methods commonly used implicitly assume that the data are normally distributed. This is the case when applying the GUM [2] with the common usage of a coverage factor $k = 2$, and this is also the case in the analysis of interlaboratory comparisons, whether the objective is to characterize the measurement method or to evaluate the proficiency of a laboratory.

In the first case, Cochran's and Grubbs' test have in common to be accurate for Gaussian data. In the second case, the comparison of a Z-score with the values 2 or 3 also relies on a Gaussian assumption as they correspond to a 95% or 99% confidence level (the true values for a Gaussian distribution are then 1.96 and 2.58).

Another warning of this paper is to be careful when applying nonlinear transformations to data. In particular, such transformations cannot be applied to variance or standard deviation calculations.

Acknowledgements: The authors wish to thank the members of the workgroup CEM of Eurolab France, for the authorization to use the results of their interlaboratory comparison Eurolab

References

1. ISO 5725-2:1994, Application of statistics – Accuracy (and trueness) of measurement methods and results – Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method.

2. ISO/CEI 17025 2005 : General requirements for the competence of testing and calibration laboratories

3. JCGM 100 (2008) ,Evaluation of measurement data — Guide to the expression of uncertainty in measurement (GUM)

4. ISO/IEC 17043:2010, Conformity assessment - General requirements for proficiency testing

5. ISO 13528:2015 Statistical methods for use in proficiency testing by interlaboratory comparison

6. ISO 21748:2017 Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation

7. H. Lilliefors, "On the Kolmogorov–Smirnov test for normality with mean and variance unknown", J. Am. Stat. Ass., **62**, pp. 399–402 (1967).

8. T.W. Anderson, D.A. Darling, "A Test of Goodness-of-Fit", J. Am. Stat. Ass., **49**, 765–769 (1954).

9. EN 55016-2-3 2010 : Specification for radio disturbance and immunity measuring apparatus and methods. Methods of measurement of disturbances and immunity. Radiated disturbance measurements.