XXI IMEKO World Congress "Measurement in Research and Industry"
August 30 - September 4, 2015, Prague, Czech Republic

ALTERNATIVE STATISTICAL ANALYSIS OF INTERLABORATORY COMPARISON MEASUREMENT RESULTS

Rossella Berni [1], *Carlo Carobbi* [2]

[1] Department of Statistics, Computer Science and Applications "G.Parenti", University of Florence, Italy
E-mail: berni@disia.unifi.it
[2] Department of Information Engineering, University of Florence, Italy
E-mail: carlo.carobbi@unifi.it

Abstract - Measurement results provided by laboratories involved in interlaboratory comparisons are processed in order to obtain the reference value of the quantity under measurement and its uncertainty. Measurement results are also compared with the reference value and its uncertainty in order to assess participants performance. Standards procedures are available in order to statistically process the measured values resulting from interlaboratory comparisons, such as those described in [1]. Unfortunately these procedures do not take into account the measurement uncertainty that each participant is required to declare. Well established statistical methods, such as errors measurement models, are however available that permit to statistically process both the measured values and their uncertainty, and therefore to fully evaluate the dispersion effects involved in the interlaboratory comparison activity.

Keywords: proficiency test, interlaboratory comparisons, key-comparisons, robust statistical analysis, error measurement model

1. INTRODUCTION

Test laboratories accredited to ISO/IEC 17025:2005 [2] have to comply with the requirements of this standard. In particular they have to calculate measurement uncertainty (always) and declare the calculated value in the test reports that they issue (when required by the relevant test method). Further, they have to assure the quality of test results. To this purpose participation in interlaboratory comparisons (more specifically proficiency tests) is mandatory when they are available for the specific test method in the scope of accreditation. Identical requirements apply to calibration laboratories [2], that have to participate in interlaboratory comparisons in order to confirm and provide experimental evidence of their calibration and measurement capabilities (CMCs). Finally, participation of national metrological institutes in key comparisons is essential in order to provide support to the Mutual Recognition Arrangement of the International Committee for Weights and Measures (CIPM MRA) of national measurement standards and of calibration and measurement certificates issued by national metrology institutes and to test the very fundamental measuring techniques in the field.

The most common scheme of participation in interlaboratory comparisons is the one where a sample that realizes the quantity to be measured (a stable source of electromagnetic field in the application described here) is circulated among participants. Each participant measures, according to a standard or specific procedure, the quantity (electromagnetic field) and reports to the coordinator of the comparison a measurement result consisting of a measured value e_i and its standard uncertainty u_{ei}, where $i = 1, 2, ..., p$ and p is the number of participants. It is assumed, and reasonable, that the measurement results provided by the participants are independent each other.

Whatever is the type of interlaboratory comparison (proficiency test involving test laboratories or experimental assessment of calibration laboratories or key comparison among national metrological institutes) a reference value and the corresponding uncertainty for the quantity realized by the sample under measurement must be established. This may be done before the start of the interlaboratory comparison, through accurate calibration of the sample (with greater accuracy than that expected to be achieved by the average participant), or after the completion of the comparison, through the statistical analysis of the measurement results provided by the participants. As a matter of fact the first option is not always viable. This is the case, for example, of electromagnetic compatibility measurements. Indeed in this case the electromagnetic field measurement accuracy reached by a well equipped test laboratory is comparable with (if not better than) the measurement capability of the national metrological institute of the country where the test laboratory operates. A highly accurate calibration of the sample then requires particular skills, e.g. the capability of combining numerical electromagnetic simulations of the sample with ad hoc (non standard) measurement techniques, see [7], [8] and [9].

A standard procedure, described in [1], is available for processing the measurement results of an interlaboratory comparison and obtaining a reference value and its uncertainty. Such procedure, adopted for the proficiency assessment of test laboratories [3], is elegantly explained in [5] and [6], and it only processes the measured values e_i without taking into account the measurement uncertainty u_{ei} that each participant is required to declare. Therefore an important piece of information about the variability (reliability) of the various measurement results is neglected. In the field of calibration and of key comparisons among national metrological institutes a different approach

is adopted and based on the weighted average of the measurement results e_i, the weight w_i being the reciprocal of the square of the standard uncertainty, i.e. $w_i = 1/u_{ei}^2$. Unfortunately the weighted average, although taking into account the measurement uncertainty declared by each participant, does not provide a robust estimate of the reference value.

Recently, a robust analysis has been proposed for processing the measurement results originating from key comparisons which takes into account measurement uncertainty [4], however those results that are identified as outliers are discarded by the process of derivation of the reference value. It is important to observe that measurement results may be discarded because, although the measured value of the quantity is correct, i.e. within the range of the e_i values provided by all the participants, its uncertainty is quite small, i.e. much smaller than the range the e_i values (for example because some significant contribution to measurement uncertainty is neglected). In our opinion this is not an acceptable solution: all measurement results must be taken into account for the purpose of deriving the reference value and its uncertainty.

A novel approach is here suggested, based on a well established statistical method, namely error measurement model analysis, in order to consider both the measured values as well as their uncertainty in the full evaluation of the dispersion effects involved in the interlaboratory comparison activity.

The organization of the paper is the following: the next section contains the theory related the error measurement model applied in this context; the section after describes data and model results; discussion final remarks follow.

2. THE OUTLINED THEORY

We assume to observe $(Y_i, X_i), i = 1, .., n$ random variables, with expected values $E(X_i) = \xi_i$ and $E(Y_i) = \eta_i$. At first we consider a linear relationship model, where the functional structure (formula (1)) is extended by including the random variables ξ_i.

$$\eta_i = \alpha + \beta\xi_i \tag{1}$$

Further, it is introduced the following statistical model (formula (2)), that implicitly includes a hierarchical structure which involves the X_i through the random variables ξ_i.

$$\begin{aligned} Y_i &= \alpha + \beta\xi_i + \epsilon_i \\ X_i &= \xi_i + \delta_i \end{aligned} \tag{2}$$

Model (2) is an error measurement model[10], where $\xi_i \sim IIDN(\xi, \sigma_\xi^2)$. The two error components ϵ_i and δ_i are supposedly Normally distributed with mean equal to zero and variances: σ_ϵ^2, σ_δ^2, respectively; therefore, the variance Var(X_i) is equal to $\sigma_\delta^2 + \sigma_\xi^2$.

2.1. The measurement error model for interlaboratory comparison measurement data

Let's start by considering $i = 1, ..., p$ laboratories; e_i is the i-th measurement within the i-th laboratory (Section 1). Therefore, we assume to measure e_i as the i-th realization of the random variable e within the i-th laboratory.

We name as $e^\star$ the robust average [1] of all the measurements performed by the p laboratories. We also consider u_{e_i} as the uncertainty related to the e_i measurement and declared by the i-th participant.

Therefore, when considering model (2) for interlaboratory comparison measurement results, we observe the measure $Y_i = e_i$ and the corresponding uncertainty $X_i = u_{e_i}$ for each i-th participant. In the model structure, the coefficient α is defined as the mean of the measurements performed by the p laboratories and is denoted by $\bar{e}$, thus it is assumed as an estimator of $e^\star$. The random variable ξ_i is defined as the non observed error for each measurement within each laboratory, so that $e_i = \bar{e} + u_{e_i}$.

Furthermore, we assume that the observed measurement uncertainty u_{e_i} is the sum of a non observed random component ξ_i and an additional error component δ_i; in addition, $\hat{\xi} = \bar{u}_e$ where $\bar{u}_e$ is the average of the measurement uncertainty calculated through the p laboratories.

Model (2) is expounded in order to consider: i) the two manifest variables e_i and u_{e_i}; ii) two latent variables related to the components ζ_i and ξ_i and the corresponding error components. More specifically, we define the following model:

$$e_i = \zeta_i + \epsilon_i \tag{3}$$

$$u_{e_i} = \xi_i + \delta_i$$

$$\zeta_i = \alpha + \beta\xi_i + r_i \tag{4}$$

and:

$$e_i = \hat{\alpha} + \hat{\beta}\xi_i = \bar{e} + \hat{\beta}u_{e_i}$$

$$e_i \sim N(\bar{e} + \hat{\beta}\bar{u}_e, \sigma_\epsilon^2)$$

$$u_{e_i} \sim N(\xi, \sigma_\delta^2)$$

.

Therefore, we observe (e_i, u_{e_i}) in place of the unobserved (but true) variables (ζ_i, ξ_i) with corresponding additive errors (ϵ_i, δ_i); in addition, the two latent variables ζ_i and ξ_i are assumed as linearly related, formula (4). We also add the r_i term to check the hypothesized linear relationship between the unobserved true variables. This term, also called equation error, [11, 12], is assumed $IID \sim (0, \sigma_r^2)$. It must be noted that the r term is null or negligible if the linear model assumption is valid.

Moreover, we aim at improving the reliability of the estimates, particularly the estimation of $e^\star$ through the α coefficient, and the estimation of the uncertainties (u_{e_i}) by means of β, across p laboratories. To this

end, we consider several estimation methods: Maximum Likelihood (ML), Un-weighted Least Squares (ULS), and we also evaluate weighted estimation methods to account for heteroscedasticity across measurements; nevertheless, estimates calculated through the ML method give the best results for the radiated emission measurements evaluated for each sampling frequency.

Statistical results are also analyzed through several criteria: i) indexes and hypothesis tests at evaluating the model significance and the goodness-of-fit; ii) model residuals; iii) indexes at evaluating the parsimony of the applied model (3); iv) convergency and optimization measurements. In the following Section, data and model results are illustrated.

3. DATA AND MODEL RESULTS

Data are related to a proficiency test performed through $p = 19$ laboratories (participants). The interlaboratory comparison of radiated emission measurements are carried out in the period May 2012-May 2013; measurements are performed in anechoic chambers and in the frequency range comprised between 200 and 3000 MHz; the sampling frequencies are: 260MHz, 560MHz, 1100MHz, 2200MHz, 2900MHz. A summary description of the measurement data is reported in Table 1, where values correspond to the electromagnetic field obtained from calibration of the travelling sample (E) and those obtained from the ISO 13528 robust analysis ($e^\star$). The corresponding standard uncertainties (u_e and $s^\star$) are also reported; for more details see [8]. In this application, we consider as Y_i the mean value

Table 1. Data: reference values by frequency

	260MHz	560MHz	1100MHz	2200MHz	2900MHz
E	72.4	72.8	73.8	64.3	57.0
u_e	0.45	0.45	0.45	0.45	0.45
$e^\star$	72.2	72.9	73.7	64.6	57.6
$s^\star$	1.5	0.9	1.0	1.7	2.3

e_i calculated on the repeated measurements performed by the i-th participant; the same for the uncertainty $X_i = u_{e_i}$, for each i-th participant. Therefore, the dataset consists of 19 measurements for each frequency.

The linear measurement error model described in Subsection 2.1, formula (3), is separately applied for each frequency by means of the TCALIS procedure (SAS software; Windows Platform vs.9.2). In Table 2 estimates of the α and β coefficients are reported with related standard errors; in the same Table (Table 2), we also include the estimated variances of the error components: σ^2_ϵ, σ^2_δ. The statistical results reported in Table 2 are satisfactory, both

Table 2. Estimates of coefficients and error variances, with related standard errors (s.e), by frequency
$(***= 1\%; ** = 5\%; * = 10\%)$

Par.	260MHz	560MHz	1100MHz	2200MHz	2900MHz
α	72.86^{***}	72.74^{***}	74.60^{***}	65.62^{***}	58.66^{***}
s.e.	0.43	0.27	1.27	0.57	1.11
β	-0.67^{*}	0.07	-2.27^{*}	-1.24^{**}	-0.56
s.e.	0.39	0.26	1.19	0.54	1.05
σ^2_ϵ	2.75^{**}	1.36^{***}	21.54^{*}	3.88	19.24^{***}
s.e.	1.27	0.46	11.36	2.52	7.05
σ^2_δ	2.40^{***}	2.26^{***}	2.24^{***}	2.26^{***}	2.23^{***}
s.e.	0.80	0.75	0.77	0.77	0.79

considering the achieved significant values and also when compared with reference values (Table 1). The estimates of $e^\star$, obtained through the coefficient α, are always very close to the reference value, without considering the uncertainty term, related to the estimate of the β coefficient. Moreover, for each frequency, the estimate of β, which is the slope of the latent variable for uncertainty, has always the right sign for a correct adjustment to the desired value. The two latent variables, ζ_i and ξ_i, which represent the true variables for the measurement e_i and the uncertainties u_{e_i} respectively, always show high scores and predicted covariances with the corresponding manifest variable. On the contrary, scores and covariances have always minus sign and/or very low values when considering each manifest variable with the non corresponding latent variable.

In Table (3) are reported some diagnostic measures, by frequency. As previously outlined, we consider four issues: the goodness-of-fit of the model; the residuals; the parsimony of the model; optimization and convergency. This last issue is evaluated through the stability coefficient (always < 1), the computational choices and the diagnostic results for the maximization of the ML function (the max-step during the search, optimization algorithm, max absolute gradient element, objective function value). More precisely, in Table 3, we include:

- Chi-square test and p-value: a significant p-value means that the the applied model is valid;
- Fit function: the vector of coefficients is estimated iteratively through a nonlinear optimization algorithm that minimizes a discrepancy function, which is also known as the fit function; therefore, a low value is desired;

- Root of the Mean of the Squared Residuals (RMSR);
- The Goodness-of-Fit (GFI) index is equal to one minus the ratio of the minimum ML function value and the function value before any model has been fitted. The GFI should be between 0 and 1; when GFI is negative or much larger than 1, the model does not fit well. The AGFI is the Adjusted GFI, i.e adjusted for the degrees of freedom of the model;
- Root Mean Square Error Approximation (RMSEA) coefficient, with Lower (L-CL) and Upper (U-CL) Confidence Limits: this index allows us to evaluate the parsimony of the model with respect to the degrees of freedom and the minimum ML function value;
- Akaike's information criterion (AIC): this is a criterion for selecting the best model among a number of candidate models; the model that yields the smallest value of AIC is considered as the best candidate model.

Table 3. Diagnostic measures for each estimated model (by frequency[MHz])

	260	560	1100	2200	2900
χ^2_1	42.26	36.62	33.90	34.55	32.93
p-value	0.0001	0.0001	0.0001	0.0001	0.0001
Fit function	2.35	2.03	1.99	2.03	2.06
RMRS	2.28	1.97	3.70	2.03	2.03
GFI	0.998	0.999	0.992	0.997	0.990
AGFI	0.995	0.997	0.961	0.989	0.954
RSMEA	1.51	1.41	1.39	1.40	1.41
RSMEA L-CL	1.45	1.04	1.01	1.03	1.02
RSMEA U-CL	1.92	1.81	1.81	1.82	1.84
AIC	40.26	34.62	31.90	32.56	30.93

When observing the obtained diagnostic measures, we may consider the good results related to the significance for each model (Chi-square test) and the low values achieved for the AIC indexes. The goodness-of-fit is always gained, with GFI and AGFI always lower then one. Residuals (RMRS index) presents a very low value for the 560MHz model, but the highest value for the 1100MHz model; if we consider the measurement data for this frequency, we observe two abnormal residual values. Nevertheless, diagnostic results are in general very similar, with very small differences in magnitude, for each applied model. Finally, for each frequency, the equation error r in model (3) does not present significant values, and thus the linear relationship between ζ_i and ξ_i can be accepted.

4. DISCUSSION AND FINAL REMARKS

This research aims at building alternative statistics in order to improve the analysis and the evaluation of interlaboratory comparison measurement results. In particular, the improvement of the statistical measures is suggested by considering two specific issues: i) robustness and ii) reliability. The concept of robustness means that we take into account each measurement and its uncertainty, also including outliers and their influence on proficiency test results; the concept of reliability is related to the estimation of $e^\star$ through α, and the estimation of the uncertainty, through β, obtained by applying the suggested error measurement model. The statistical results, illustrated in Section 3, show that the applied methods, e.g. the proposal of a structural model with latent variables, could be an useful tool to evaluate the reliability of interlaboratory comparison measurements. To this end, this study could be considered as a first attempt that should be expounded by further investigations directed to improve the accuracy of the estimate for the β coefficient, by considering the proposals existing in literature for building confidence intervals for the slope of an error measurement model.

In fact, in literature, a particular attention is devoted to the accuracy of a confidence set for the slope β. In [10], a consistent estimator for the variance of β is then used to build an approximate confidence set for the slope. To this end, a parameter $\tau^2 = \sigma^2_\xi/\sigma^2_\delta$ is defined as the amount of information useful at evaluating the reliability of the confidence set, given also the sample size (p). Also in [13], the suggested confidence set takes into account the coverage probability $(1-\alpha)$, which depends on τ^2; in this case, the proposal is valid for $p \geq 10$ or $\tau^2 \geq 0.25$, with a decreasing coverage (lower than 80%) if τ diminishes. More recently, in [12], a generalized confidence interval for the slope β is suggested and it is compared with the traditional asymptotic one. Nevertheless, the building of a confidence interval for this specific kind of data (proficiency tests) should be performed given particular attention to the coverage probability, the expected length, but also to the estimates of the error components.

REFERENCES

[1] *Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparison*, ISO 13528:2005.

[2] *Conformity Assessment – General Requirements for the Competence of Testing and Calibration Laboratories*, ISO/IEC 17025:2005.

[3] *Conformity Assessment – General Requirements for Proficiency Testing*, ISO/IEC 17043:2010.

[4] Maurice G Cox, "The evaluation of key comparison data: determining the largest consistent subset," Metrologia 44

(2007), 187-200.

[5] Analytical Methods Committee, "Robust Statistics – How Not to Reject Outliers – Part 1. Basic Concepts," Analyst, December 1989, vol. 114.

[6] Analytical Methods Committee, "Robust Statistics – How Not to Reject Outliers – Part 2. Inter-laboratory Trials," Analyst, December 1989, vol. 114.

[7] C. F. M. Carobbi,M. Cati, and C. Panconi, "Generation and measurement of a reference field for round-robin comparison purposes," presented at the IEEE Int. Symp. Electromagn. Compat., Detroit, MI, USA, Aug. 18-22, 2008.

[8] C. F. M. Carobbi, A. Bonci, M. Cati, C. Panconi, M. Borsero, and G. Vizio, "Validation of far-field numerical predictions through near-field measurements," in Proc. Int. Conf. Electromagn. Adv. Appl., Torino, Italy, Sep. 9-13, 2013, pp. 551-554.

[9] Carlo F. M. Carobbi, Alessio Bonci, Marco Cati, Carlo Panconi, Michele Borsero and Giuseppe Vizio, "Design, Preparation, Conduct and Result of a Proficiency Test of Radiated Emission Measurements," *IEEE Transactions on Electromagnetic Compatibility*, vol. 56, no. 6, pp. 1251-1261, Dec. 2014.

[10] G. Casella, R.L. Berger, *Statistical Inference*, Wadsworth & Brooks, Pacific Grove, California U.S.A., 1990.

[11] C.-L.Cheng, J.W. Van Ness, *Statistical Regression with Measurement Error*, Arnold, London, UK, 1999.

[12] J.-R. Tsai, "Generalized confidence interval for the slope in linear measurement error model", *Journal of Statistical Computation and Simulation*, 80 (2010), 927-936.

[13] L. Gleser, J. Hwang, "The Non-existence of $100(1-\alpha)\%$ confidence sets of finite expected diameter in error-in-variables and related models," *Annals of Statistics* 15 (1987), 1351-62.