XVIII IMEKO WORLD CONGRESS
Metrology for a Sustainable Development
September, 17 – 22, 2006, Rio de Janeiro, Brazil

UNCERTAINTY PROPAGATION IN MULTI-STAGE MEASUREMENTS USING LINEAR REGRESSION ANALYSIS AND MONTE CARLO SIMULATION

A. Silva Ribeiro [1], *J. Alves e Sousa* [2], *C. Oliveira Costa* [1], *M. Pimenta de Castro* [1]

[1] National Laboratory of Civil Engineering, Lisbon, Portugal, asribeiro@lnec.pt, ocosta@lnec.pt, mpcastro@lnec.pt
[2] Madeira Regional Laboratory of Civil Engineering, Funchal, Portugal, jasousa@lrec.pt

Abstract: Linear Regression Analysis (LRA) is one of the statistical tools most intensively used in all branches of science, with many applications in the study of measurement processes and is therefore important in metrology. The implementation of metrology in quality systems, led to a widespread evaluation of measurement uncertainties based on the GUM uncertainty framework (Guide to the Expression of Uncertainty in Measurement, [1]). This methodology, however, has its own restrictions among which one could include the use of LRA in multi-stage measurement. To overcome these restrictions, an alternative approach considers the use of the Monte Carlo method to evaluate LRA uncertainties and, subsequently, to use it further in the evaluation of uncertainties in multi-stage measurement processes.

Keywords: measurement uncertainty, linear regression analysis, Monte Carlo method, multi-stage measurement.

1. INTRODUCTION

Linear Regression Analysis (LRA) [2,3] is one of the statistical techniques most largely used in science and in particular in Metrology. The special interest in LRA is due to its application in the study of measurement processes, e. g., those concerning the development of mathematical models or the correction of calibrated instruments.

Since LRA is based on the use of data that is regarded as estimates of random variables, the LRA coefficients that are determined for the data are also estimates of random variables. Hence, it is of interest to evaluate the uncertainties associated with these coefficients.

Today, measurement uncertainty is a crucial aspect concerning the measurement problem, ideally treated by analytical methods (often too complex real metrological problems) or by the GUM uncertainty framework. However, the latter has important restrictions that make it difficult to implement in circumstances such as those found in LRA multi-stage measurement problems, although sound solutions can be obtained for many classes of problem.

It is, therefore, useful to apply methodologies that are able to remove these restrictions. A Monte Carlo method (MCM) can be applied to this type of stochastic problem, using numerical computational simulation to evaluate measurement uncertainties. In this context, the purpose of this paper is to study the application of such a procedure in the evaluation of LRA coefficients and its further use in multi-stage measurement. This can be considered an important application, e.g., the correction of readings of calibrated instruments knowing the uncertainty of LRA, which is relevant to the needs and requirements of industry and laboratories.

The discussion of the restrictions and advantages of using MCM for these problems is another aim of this study, which will be illustrated with a practical example taken from the field of dimensional metrology.

2. THE ANALYTICAL APPROACH

In metrology, the expression of a measurement result should include the related uncertainty. The evaluation of the uncertainty is a metrological task that should involve all input quantities that affect the measurement. Therefore, if the result is influenced by the use of LRA corrections, the uncertainty related to their coefficients should be taken into account in the uncertainty budget.

This problem has at least two stages: the first requires the evaluation of LRA coefficients and the second involves the use of LRA in the correction of measurements. The starting point is the use of data to obtain the LRA coefficients.

In the first stage, consider data (x_i,y_i), i = 1,…,n, where the x_i denote stimulus values and the y_i corresponding response values. Associated with each y_i is a standard uncertainty (the same for all i) that is unstated *a priori*. The uncertainties associated with the x_i are regarded as negligible. The LRA equation is given by the model equation:

$$Y = B_0 + B_1 X \, , \tag{1}$$

fitted to data and the two coefficient estimates (b_0 and b_1) of B_0 and B_1 are obtained [2] using equations

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \, , \tag{2}$$

$$b_0 = \bar{y} - b_1 \bar{x}. \tag{3}$$

where $\bar{x} = \sum x_i / n$ and $\bar{y} = \sum y_i / n$.

In the context of LRA, B_0 and B_1 can be considered as random variables with their own variance. The equations for the variance and covariance associated with b_0 and b_1, and σ^2, the variance of the LRA residual errors, are [2,4]:

$$\mathrm{var}[b_0] = \frac{\sigma^2 \sum x_i^2}{n \cdot \sum (x_i - \bar{x})^2}, \tag{4}$$

$$\mathrm{var}[b_1] = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}, \tag{5}$$

$$\mathrm{cov}[b_0, b_1] = \frac{-\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}. \tag{6}$$

The variances of these coefficients (b_0 and b_1), result from the mentioned standard uncertainties associated with the y_i. Each y_i is typically determined as the average of a number of repeated indications corresponding to the stimulus x_i. The average points are used in the LRA leading to the main LRA curve (as illustrated in Fig. 1). However, a large number of other LRA curves could be obtained using other points drawn at random from probability density functions centered on the y_i with standard deviations equal to σ, creating an uncertainty region bounded as presented below.

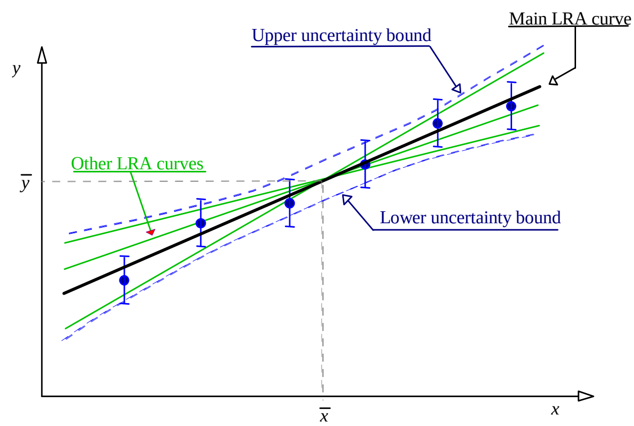

Figure 1: LRA curves and uncertainty bounds.

The second stage of the problem is related to the use of the LRA equation (1) in the context of the correction of measurement. The corrected value, x_i*, of a new indication, y_i*, follows equation (7), representing the expectation $\hat{E}[x_i * | y_i *]$.

$$x_i * = \frac{(y_i * - b_0)}{b_1}. \tag{7}$$

The introduction of the variance operator in (7) in order to quantify the variance of the corrected value gives

$$\mathrm{var}\left[x_i^*\right] = \mathrm{var}\left[\frac{y_i^* - b_0}{b_1}\right], \tag{8}$$

obtained using the following relations [2,4],

$$\mathrm{var}[X_1 \pm X_2] = \mathrm{var}[X_1] + \mathrm{var}[X_2] \pm 2\,\mathrm{cov}[X_1, X_2], \tag{9}$$

$$\mathrm{var}\left[\frac{X}{Y}\right] \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \cdot \left(\frac{\mathrm{var}[X]}{\mu_X^2} + \frac{\mathrm{var}[Y]}{\mu_Y^2} - \frac{2\,\mathrm{cov}[X,Y]}{\mu_X \mu_Y}\right). \tag{10}$$

Considering that all pairs of quantities on the right-hand side are uncorrelated, equation (8) becomes:

$$\mathrm{var}[x_i *] \approx \left(\frac{y_i * - b_0}{b_1}\right)^2 \cdot \left(\frac{\mathrm{var}[y_i *]}{(y_i * - b_0)^2} + \frac{\mathrm{var}[b_0]}{(y_i * - b_0)^2} + \frac{\mathrm{var}[b_1]}{b_1^2}\right), \tag{11}$$

which is an approximation of the exact solution due to the linearization implicit in deriving result (10).

3. EXPERIMENTAL DATA

To understand the treatment involved in the evaluation process, a set of experimental data related to a calibration in the field of dimensional metrology is included enabling the type of results that can be obtained to be illustrated.

The particular example concerns the calibration of a 1-D measuring machine, with resolution of 0.1 µm using standard block gauges. The results obtained from experimental testing are briefly presented in the following Table.

Table 1: Calibration results and associated standard uncertainties.

Conventional true values x_i /mm	Measured average values y_i /mm	Average standard deviation (n=5) $s(y_i)$ /10^{-5} mm
0.99994	1.0000	2.7
2.00005	2.0001	2.2
5.00006	5.0002	2.7
10.00006	10.0002	2.7
25.00004	25.0004	5.7
49.99993	50.0005	3.5
74.99991	75.0008	4.2
99.99990	100.0010	4.2

The standard deviation associated with each mean is indicated in Table 1. Because such a dispersion of values could arise from random sampling alone (e.g., by taking standard deviations of the means of groups of five values sampled from a normal distribution), a pooled value of 3.7×10^{-5} for the standard deviation was used and taken as the standard uncertainty associated with the y_i. A non-parametric statistical test was carried out, which gave no reason to doubt the hypothesis of homogeneity of variance. Such a pooled value can be expected to be more reliable than the individual values in such circumstances [5].

With this set of data, it is possible to evaluate analytical solutions of LRA coefficients b_0, b_1, their associated variances, standard deviation and covariance according to equations (2-6) given in table 2. The results obtained are

Table 2: Statistical parameters based on analytical evaluation.

Parameter	Analytical solution
B_0	$5.64039 \cdot 10^{-5}$ mm
B_1	≈ 1.000010667
σ	$\approx 3.077 \cdot 10^{-5}$ mm
Var (b_0)	$2.25692 \cdot 10^{-10}$ mm^2
s (b_0)	$\approx 1.50 \cdot 10^{-5}$ mm
Var (b_1)	$9.56322 \cdot 10^{-14}$
s (b_1)	$\approx 3.1 \cdot 10^{-7}$
Cov (b_0,b_1)	$-3.20 \cdot 10^{-12}$

These results are taken as reference values in order to test the numerical simulation results presented in the next sections.

3. MCM APPROACH TO THE FIRST STAGE

The use of a Monte Carlo method in the evaluation of uncertainty was developed using a validated pseudo-random number generator, which was then converted from uniform into normal distribution using a Box-Muller transformation [6]. Finally, the values of the output sequence thus obtained were sorted into non-descending order.

Input pseudo-random sequences with 100 000 trials were generated, their values converted in order to obtain 8 normal probability distribution sequences having average values identical to those presented in Table 1 and using the pooled standard deviation 3.7×10^{-5}. The combination of these values according to the mathematical model where used to determine b_0 and b_1, generating an output sequence for each parameter.

An approximation to the probability density function (PDF) for these output sequences was obtained by scaling a histogram of the values provided by the MCM. They both show a Gaussian distribution, as illustrated in Figure 2 and Figure 3, and the main results are presented in Table 3.

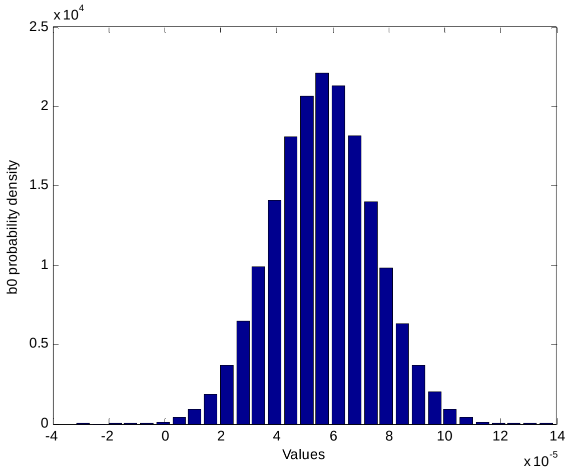

Figure 2: Approximation to the PDF of b_0 output sequence.

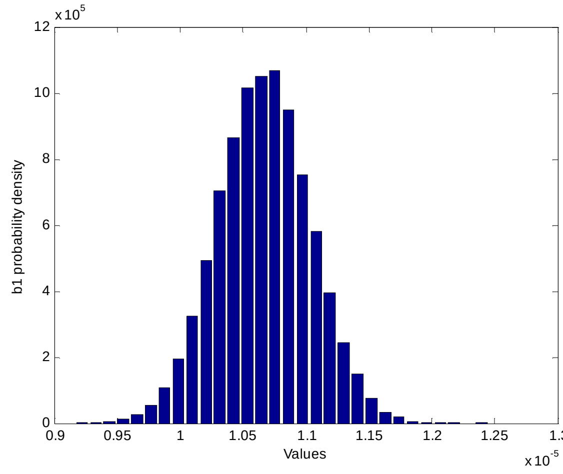

Figure 3: Approximation to the PDF of $(b_1 - 1)$ output sequence.

Table 3: Statistical parameters based on MCM (using the pooled standard deviation)

Parameter	MCM results
B_0	$5.64 \cdot 10^{-5}$ mm
B_1	1.000010666
s (b_0)	$1.78 \cdot 10^{-5}$ mm
s (b_1)	$3.62 \cdot 10^{-7}$
Cov (b_0,b_1)	$-4.5 \cdot 10^{-12}$

Because both have a Gaussian distribution, the pairs probability distribution is a bivariate gaussian distribution.

The variance-covariance matrix can be presented as:

$$\begin{bmatrix} \left(1.78 \cdot 10^{-5}\right)^2 & -4.5 \cdot 10^{-12} \\ -4.5 \cdot 10^{-12} & \left(3.62 \cdot 10^{-7}\right)^2 \end{bmatrix}. \tag{12}$$

The comparison between the analytical results and MCM results shows differences of about 15% in the standard deviations of the coefficients b_0 and b_1.

This difference was studied in order to find its origin. The study points out that, if instead of applying the analytical approach presented above, where no standard deviation is provided, the evaluation of the LRA coefficients was performed with the referred pooled variance [7], the results obtained were closer to the MCM results (within 1% of difference).

Sensitivity tests were carried out using 2 independent sets of generated pseudo-random sequences, and the results obtained, presented in Table 4, shows a very consistent behavior of the parameters found, with differences lower than 1 % (considering all cases).

Table 4: Sensitivity test results

Parameter	Data: 1st set	Data: 2nd set	Data: 3rd set
s (b_0)	$1.78 \cdot 10^{-5}$ mm	$1.79 \cdot 10^{-5}$ mm	$1.78 \cdot 10^{-5}$ mm
s (b_1)	$3.62 \cdot 10^{-7}$	$3.63 \cdot 10^{-7}$	$3.62 \cdot 10^{-7}$

The presentation of the data (b_0,b_1) in a 2-D map (Fig. 4) shows the correlation between the two random variables (coefficients). The slope evaluated from the sample (equal to – 0.014, and presented as the line over the data in Fig. 4) is a parameter influenced by that correlation that will be applied in further generation of the random variables (B_0,B_1).

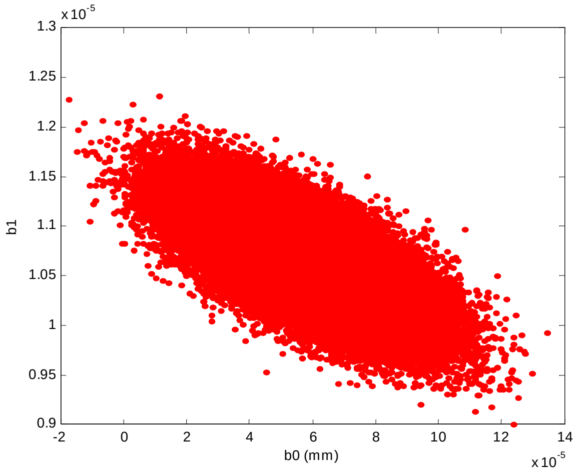

Figure 4: 100 000 pairs of coefficients $(b_0, b_1 - 1)$ **obtained using MCM**

It is important to point out that MCM can be applied using the original standard deviations presented in Table 1 instead of the pooled standard deviation, giving the results presented in Table 5.

Table 5: Statistical parameters based on MCM (using the original standard deviations)

Parameter	MCM results
B_0	$5.633{\cdot}10^{-5}$ mm
B_1	1.000010668
$s\ (b_0)$	$1.54{\cdot}10^{-5}$ mm
$s\ (b_1)$	$3.65{\cdot}10^{-7}$
Cov (b_0,b_1)	$-3.34{\cdot}10^{-12}$

This alternative simulation enhances how MCM can process directly the information of the experimental sample, namely, when input quantities have different standard deviations. The analytical method based on equations (2) to (6), however, assumes the same standard deviation value assigned to all points used in the regression.

4. MCM APPROACH TO THE FOLLOWING STAGES

The general purpose of the evaluation of the LRA coefficients is to use them in further measurement corrections using equation (7):

$$x_i{*} = \frac{\left(y_i{}^{*} - \mathrm{b}_0\right)}{\mathrm{b}_1}$$

At this stage, it is required to express the results with their associated uncertainties. Equation (11) can lead to a solution but it is an approximate solution and has some degree of complexity.

Another way to achieve the required uncertainty is to apply Monte Carlo to the mathematical model of equation (7), which only requires the uncertainty distribution of y_i^* and the bivariate distribution of the pairs (b_0,b_1).

The first task, which is to obtain the sequence related with the measurand y_i^*, requires the measurement sample average value and the associated standard deviation. For this purpose, an experimental sample gave an average value of 30.0004 mm and an average associated standard deviation not significantly different from pooled value of $3.7{\times}10^{-5}$ mm. These values were used to obtain a MCM gaussian distribution.

The second task is to obtain correlated estimates of the coefficients according to the Gaussian bivariate distribution, like the one presented in Figure 3, but independently from the data. One procedure to do so includes the following steps:

A1. Generate 2 uniform pseudo-random number sequences using the Hill-Wichmann algorithm [8];

A2. Apply Box-Muller Algorithm [6] to the A1. sequences, in order to obtain a new sequence with pairs of coordinates of Gaussian bivariate distribution ($n_0{}^*$,$n_1{}^*$);

A3. Perform the scaling of the original pairs to the values of s(b_0) and s(b_1);

A3’. Perform the translation of the data centered in the origin to be centered in the average pair $\left(\bar{b}_0, \bar{b}_1\right)$;

A3’’ Perform the rotation of the data according to the slope angle found previously, presented in Figure 3, as measure related with the correlation, obtaining the new sequence of pairs $\left(b_0^*, b_1^*\right)$.

The steps A3 to A3’’ can be performed in a single matrix operation, having the following description:

$$\mathbf{B} = \left[\mathbf{C}^{\mathrm{T}} \cdot \mathbf{Z} + \mathbf{T} \cdot \mathbf{1}\right] \tag{12}$$

where **C** is the Cholesky matrix of the variance-covariance matrix (12):

$$\mathbf{C} = \begin{bmatrix} 1.78{\cdot}10^{-5} & -2.528{\cdot}10^{-7} \\ 0 & 2.591{\cdot}10^{-7} \end{bmatrix} \tag{13}$$

Z is the generated matrix with the two Gaussian sequences, each in a row, **1** represents the unit matrix with dimension 1×100000; and **T** is the translation matrix required to perform the operation referred in (A3’):

$$\mathbf{T} = \begin{bmatrix} 5.64{\cdot}10^{-5} \\ 1.000010666 \end{bmatrix} \tag{14}$$

With this procedure a new set of pairs $\left(b_0^*, b_1^*\right)$ can be generated, independently from the first stage sequences (not biased by the first set), but with similar correlation. Fig. 5 exhibits the new set of generated pairs, presented with the evaluated slope value.

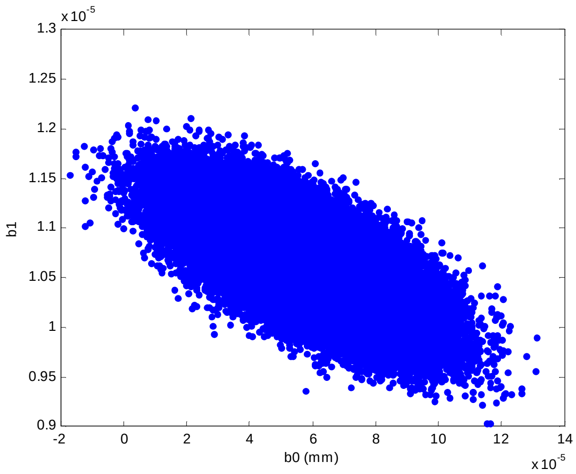

Figure 5: Data pairs $(b_0^*, (b_1^* - 1))$ generated for the 2nd stage.

The generated sequences of y^*, b_0^* and b_1^* can be applied in the mathematical model (7) generating an output sequence of values of x^*. From this output sequence, having the Gaussian configuration presented in Figure 6, the average value and the associated standard uncertainty of the variable x^*, can then be obtained:

Table 6: Output statistical parameters from MCM

Parameter	MCM results
x^*	30.000024 mm
$s\,(x^*)$	$3.9{\cdot}10^{-5}$ mm

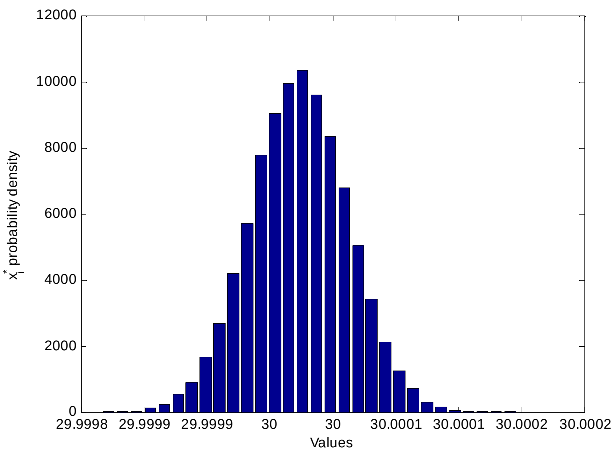

Figure 6: Probability distribution of x_i* at 30 mm.

These values can be compared with the results obtained using the analytical approximation (11) where the values of MCM presented in Table 3 are now the input values, which yields:

Table 7: Analytical approximation results

Parameter	MCM results
x^*	30.000024 mm
$s\,(x^*)$	$4.0{\cdot}10^{-5}$ mm

The percentage of difference obtained between both estimates of the standard deviation is of about 2% (lower in the MCM estimate). The percentiles accuracy for the evaluation of **s(*x**)**, based on [9], is of $2\text{x}10^{-7}$ mm.

A similar Monte Carlo was carried out in order to evaluate the uncertainty in other points of the measuring range. The standard deviation was assumed unchanged from the previous simulation process (3.7×10^{-5} mm) and the results are presented in Fig. 6.

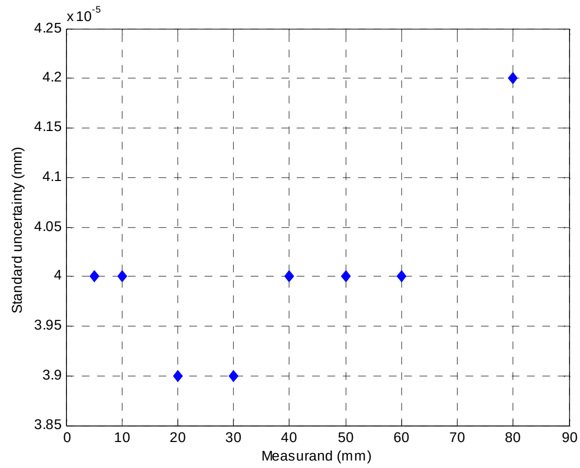

Figure 7: Standard uncertainties obtained using MCM.

The differences presented in Fig. 7 shows that the behavior of the evaluated uncertainty is near the minimum at the average x value and maximum in both ending points of the LRA range. This also agrees with the analytical predicted behavior for this type of LRA problems [4,10].

5. CONCLUDING REMARKS

Monte Carlo performed the evaluation of measurement uncertainties directly from the original model, overcoming the intricacies of analytical approximate solutions. Furthermore, a Monte Carlo method can easily handle individual standard deviations of LRA input quantities, using the same numerical algorithm.

The use of a matrix procedure presented in expression (12) has the additional advantage of portability, allowing its use with data obtained from different tools and applications.

MCM can be easily combined with analytical procedures in order to perform specific tasks, especially those that increase its efficiency, and can also be combined with bootstrap applications to provide solutions for more complicated regression models.

The approach presented in this study, provides a robust and flexible method to determine the probability distributions for the LRA coefficients, from which confidence intervals are readily obtained, thus enabling the use of linear regression with assurance.

ACKNOWLEDGMENTS

The authors are indebted to Professor Maurice Cox and Dr. Peter Harris for their continued interest in this work and for many useful discussions.

REFERENCES

[1] Guide for the Expression of Uncertainty in Measurement (ISO-GUM*).* ISO, BIPM, CEI, IFCC, IUPAC, IUPAP, OIML, 1995.

[2] Mood, A., Graybill, F. and Boes, D., "Introduction to the theory of statistics", 3rd Ed., McGraw-Hill International Editions, Singapore, 1974.

[3] Sen, A. and Srivastava, M., "Regression analysis. Theory, methods and applications", Springer-Verlag, New York, USA, 1990.

[4] Johnson, N. and Leone, F. "Statistics and experimental design in engineering and the physical sciences. Vol. I", 2nd Ed., John Wiley & Sons, USA, 1977.

[5] Burns, M., Cox, M.G., Ellison, S.L.R., Griffiths, M., Kamvissis, S., Keer, Jacquie, "Uncertainty evaluation for real-time PCR quantitation of nucleic acid content", In Preparation.

[6] Knuth, D., "The Art of Computer Programming. Vol.2: Seminumerical Algorithms", 3rd Edition, Addison–Wesley Longman Inc., 1998.

[7] Press, W., Flannery, B., Teukolsky, S. and Vetterling, W., "Numerical Recipes. The Art of Scientific Programming". Cambridge University Press, 1986

[8] Hill, I. And Wichmann, B., "An efficient and portable pseudo-random number generator", In *Applied Stat.* **31**, 188-190, 1982.

[9] Cox, M. G, Dainton, M.P. and Harris, P. M., "*Software Support for Metrology Best Practice Guide No. 6: Uncertainty and Statistical Modelling*". Crown, National Physical Laboratory. Teddington, UK, 2001.

[10] Ribeiro, A. S., Mimoso, J. O., Sousa, J. A. and Castro, M. P., "Uncertainty Related with the Use of Linear Regression Analysis for the Correction of Calibrated Instruments". In Proc. of IMEKO XVII World Congress, Cavtat-Dubrovnik, Croatia, Jun. 2003.

Authors:
National Laboratory of Civil Engineering, Av. do Brasil, 101, 1700-066 Lisboa, Portugal.
Tel. ++ 351 21 844 3000, Fax ++ 351 21 844 3041,
E-mail: asribeiro@lnec.pt, ocosta@lnec.pt, mpcastro@lnec.pt.

Madeira Regional Laboratory of Civil Engineering, Rua Agostinho Pereira de Oliveira, S.Martinho, 9000-264 Funchal, Portugal.
Tel. ++ 351.291 76 5223 /4/5, Fax ++ 351.291 76 5385,
E-mail: jasousa@lrec.pt.