Int J Thermophys (2008) 29:871–880
DOI 10.1007/s10765-008-0396-y

Components of a Variance Model for the Analysis of Repeated Measurements in Fixed-Point Comparisons

Eduarda Filipe

Published online: 5 March 2008

Abstract Experimental design is a statistical tool concerned with the planning of experiments to obtain the maximum amount of information from the available resources. This tool may be applied to metrology, especially for the analysis of a large number of repeated measurements (replicates) of short-term repeatability and the medium-term and long-term reproducibilities, enabling the inclusion of these "time-dependent sources of variability" in the uncertainty budget. The realization of the International Temperature Scale of 1990 (ITS-90) scale requires that laboratories usually have more than one cell for each fixed point, for comparison on a regular basis. The calculation of the uncertainty of such comparisons is considered here, taking into account these time-dependent sources of variability. These components of the uncertainty evaluated by a Type A method are obtained by the statistical analysis of the experimental results using the components of a variance model for designs consisting of nested or hierarchical sequences of measurements, as foreseen by the mainstream GUM. An application example of a balanced nested structure in the comparison of two fixed-point cells is presented.

Keywords Experimental design · Nested structures · Uncertainty

1 Introduction

Thermometry laboratories equipped to realize the International Temperature Scale of 1990 (ITS-90) usually have more than one cell for each fixed point. The laboratory may consider one of the cells as the reference cell, and the other(s) considered as the

E. Filipe (✉)
Laboratório Central de Metrologia, Instituto Português da Qualidade, Rua António Gião, 2,
2829-513 Caparica, Portugal
e-mail: efilipe@mail.ipq.pt

working standard(s). Alternatively, the laboratory may consider its reference value to be the average of the cells. In both cases, the cells must be compared on a regular basis and the uncertainty of these comparisons calculated. A similar situation exists when the laboratory compares its own reference value with that of a traveling standard during an inter-laboratory comparison.

To obtain a value for the difference between the cells, measurements of repeatability are performed at the equilibrium plateau. In the case of triple-point-of-water cells, where the equilibrium plateau lasts for several days, it is appropriate to repeat the experiment for the same plateau on subsequent days. It is also appropriate to repeat the experiments with different equilibrium plateaux.

The uncertainty calculation will take into account these time-dependent sources of variability, arising from short-term repeatability, day-to-day reproducibility, and the long-term random variations in the results. These components of uncertainty are evaluated by a statistical analysis of the data obtained from the experiment using the components of a variance model (ANOVA) for designs consisting of nested or hierarchical [1] sequences of measurements.

The ANOVA is defined [5, see 2.6] as a "technique, which subdivides the total variation of a response variable into meaningful components, associated with specific sources of variation." This technique is described here for a nested balanced design and an application example with the estimation of the variances by a Type A method for a comparison of two triple-point-of-water cells.

2 General Principles and Concepts

Experimental design is a statistical tool concerned with the planning of experiments to obtain the maximum amount of information from the available resources [7]. This tool is used generally for the improvement and optimization of processes where the experimenter, controlling the changes in the inputs and observing the corresponding changes in the outputs, is able to make the inference by rejecting the *null hypothesis* (H_0) of the outputs as statistically different for a significance level α, also known as the "producer's risk."

In addition, it can be used to test the homogeneity of a sample(s) for the same significance level, to identify the results that can be considered as "outliers," or to evaluate the components of variance between the "controllable" factors.

This tool can be applied to metrology, especially for the analysis of a large number of repeated measurements (replicates) to estimate the short-term repeatability and the medium-term and long-term reproducibilities, permitting the inclusion of these "time-dependent sources of variability" in the uncertainty budget.

2.1 Nested or Hierarchical Design

The nested design is defined [5,6] as "the experimental design in which each level of a given factor appears in only a single level of any other factor." The objective of this model is to deduce the values of the variance components that cannot be measured directly. The factors are arranged in hierarchical form, like a "tree" (see Fig. 1), and

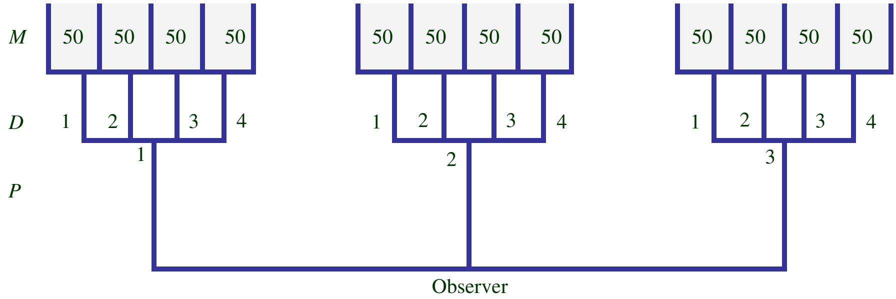


Fig. 1 A (3 × 4 × 50) nested or hierarchical design

any path from the "trunk" to the "extreme branches" will find the same number of nodes.

In the comparison experiment to be described, the factors are the daily measurements and the run (plateau) measurements. These factors are considered *random samples* of the *population* from which we are interested in drawing conclusions.

In this design, each factor is analyzed using the one-way analysis of the variance model, nested in the subsequent factor.

2.1.1 Model for One Factor

Considering first only one factor with a different levels taken *randomly* from a large population [4,7], any observation made at the ith factor-level with j observations will be denoted by y_{ij}.

The mathematical model that describes the set of data is

$$y_{ij} = M_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (i = 1, 2, \ldots, a \quad \text{and} \quad j = 1, 2, \ldots, n) \tag{1}$$

where M_i is the expected (random) value of the group of observations i, μ is the overall mean, τ_i is the parameter associated with the ith factor-level designated by the ith factor effect, and ε_{ij} is the random error component. This model with random factors is called the *random effects or components-of-variance model*.

For the hypothesis testing, the errors and the factor effects are assumed to be normally and independently distributed, respectively, with mean zero and variance σ^2 or $\varepsilon_{ij} \sim N(0, \sigma^2)$ and with mean zero and variance σ_τ^2 or $\tau_i \sim N(0, \sigma_\tau^2)$. The variance of any observation is composed of the sum of the variance components according to

$$\sigma_y^2 = \sigma_\tau^2 + \sigma^2 \tag{2}$$

The test is unilateral and the hypotheses are

$$\left| \begin{array}{l} H_0\colon \sigma_\tau^2 = 0 \\ H_1\colon \sigma_\tau^2 > 0 \end{array} \right. \tag{3}$$

That is, if the null hypothesis is true, all factor effects are "equal" and each observation is made up of the overall mean plus the random error $\varepsilon_{ij} \sim N(0, \sigma^2)$.

The total sum of squares, which is a measure of the total variability in the data, may be expressed by

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y})^2=\sum_{i=1}^{a}\sum_{j=1}^{n}\left[(\bar{y}_i-\bar{y})+(y_{ij}-\bar{y}_i)\right]^2$$
$$=n\sum_{i=1}^{a}(\bar{y}_i-\bar{y})^2+\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$$
$$+2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_i-\bar{y})(y_{ij}-\bar{y}_i) \tag{4}$$

As the cross-product is zero [8], the total variability of data (SS_T) can be separated into the sum of squares of differences between factor-level averages and the grand average (SS_{Factor}), a measure of the differences between factor-levels, and the sum of squares of the differences of observations within factor-levels from the factor-levels average (SS_{E}), due to the random error. Dividing each sum of squares by the respective degrees of freedom, we obtain the corresponding mean squares (*MS*):

$$MS_{\text{Factor}}=\frac{n}{a-1}\sum_{i=1}^{a}(\bar{y}_i-\bar{y})^2$$
$$MS_{\text{Error}}=\frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2}{a(n-1)}=\sigma^2 \tag{5}$$

The mean square between factor-levels (MS_{Factor}) [9] is an unbiased estimate of the variance σ^2, if H_0 is true, or a surestimate of σ^2 (see Eq. 7), if it is false. The mean square within factor (error) (MS_{Error}) is always the unbiased estimate of the variance σ^2.

To test the hypotheses, we use the statistic,

$$F_0=\frac{MS_{\text{Factor}}}{MS_{\text{Error}}}\sim F_{\alpha,a-1,a(n-1)} \tag{6}$$

where F[1] is the *Fisher-Snedcor* sampling distribution with a and $a\times(n-1)$ degrees of freedom.

If $F_0>F_{\alpha,a-1,a(n-1)}$, we reject the null hypothesis and conclude that the variance σ_τ^2 is significantly different from zero, for a significance level α.

The expected value of the MS_{Factor} is [8]

$$E(MS_{\text{Factor}})=E\left[\frac{n}{a-1}\sum_{i=1}^{a}(\bar{y}_i-\bar{y})^2\right]=\sigma^2+n\sigma_\tau^2 \tag{7}$$

[1] F distribution—Sampling distribution. If χ_u^2 and χ_v^2 are two independent chi-square random variables with u and v degrees of freedom, then its *ratio* $F_{u,v}$ is distributed as F with u numerator and v denominator degrees of freedom.

The variance component of the factor is then obtained by

$$\sigma_\tau^2 = \frac{E(MS_{\text{Factor}}) - \sigma^2}{n} \tag{8}$$

2.1.2 Model for Two Factors

Considering now a two "stage" nested design, the structure of the example to be described, the mathematical model is

$$y_{pdm} = \mu + \Pi_p + \Delta_d + \varepsilon_{pdm} \tag{9}$$

where y_{pdm} is the pdmth observation, μ is the overall mean, Π_p is the Pth random level effect, Δ_d is the dth random level effect, and ε_{pdm} is the random error component.

The errors and the level effects are assumed to be normally and independently distributed, respectively, with mean zero and variance σ^2 or $\varepsilon_{pdm} \sim N(0, \sigma^2)$ and with mean zero and variances σ_p^2 and σ_d^2. The variance of any observation is composed of the sum of the variance components, and the total number of measurements, N, is obtained from the product of the dimensions of the factors ($N = P \times D \times M$).

The total variability of the data [2,10] can be expressed by

$$\sum_p \sum_d \sum_m (y_{pdm} - \bar{y})^2 = \sum_p DM(\bar{y}_p - \bar{y})^2 + \sum_p \sum_d M(\bar{y}_{pd} - \bar{y}_p)^2 + \sum_p \sum_d \sum_m (\bar{y}_{pdm} - \bar{y}_{pd})^2$$

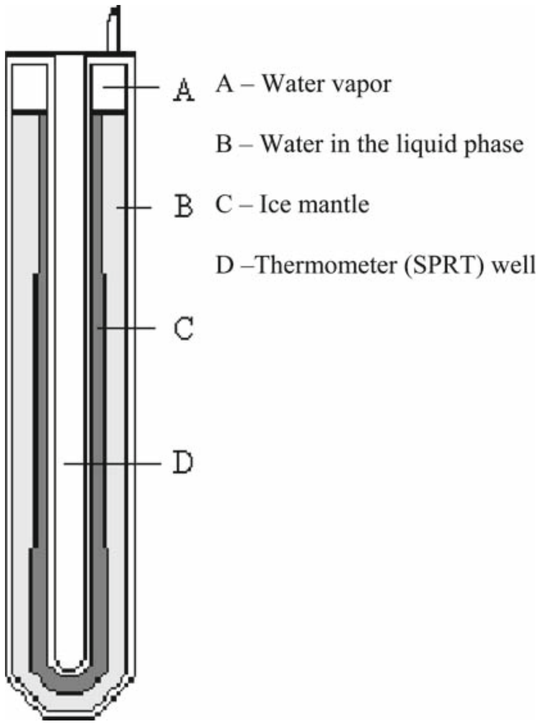
or

$$SS_T = SS_P + SS_{D|P} + SS_E \tag{10}$$

This total variability of the data is the sum of squares of factor $P(SS_P)$, the P-factor effect, plus the sum of squares of factor D for the same $P(SS_{D|P})$ and SS_E, the residual variation. Dividing by the respective degrees of freedom, $(P-1)$, $P \times (D-1)$, and $P \times D \times (M-1)$, we obtain the mean squares of the nested factors, which are estimates of σ^2, if there were no variation due to the factors. The estimates of the components of the variance are obtained by equating the mean squares to their expected values and solving the resulting equations:

$$\begin{aligned} E(MS_P) &= E\left[\frac{SS_P}{P-1}\right] = \sigma^2 + M\sigma_D^2 + DM\sigma_P^2 \\ E(MS_{D|P}) &= E\left[\frac{SS_{D|P}}{P(G-1)}\right] = \sigma^2 + M\sigma_D^2 \\ E(MS_E) &= E\left[\frac{SS_E}{PD(M-1)}\right] = \sigma^2 \end{aligned} \tag{11}$$

Fig. 2 Triple-point-of-water cell



3 Experimental Results

3.1 Short Description of the Laboratory Work

In the comparison of the two triple-point-of-water cells (Fig. 2), JA1 and JA3, we used one standard platinum resistance thermometer (SPRT). After the preparation of the ice mantles, the cells were maintained in a thermally regulated water bath at a temperature of 0.007°C. This bath can host two cells and is able to maintain them at the triple point of water ($t = 0.01$°C) for several weeks. The ice mantle in the cells was prepared according to the laboratory procedure, 48 h before starting the measurements.

Fifty independent measurement differences were obtained daily, and this set of measurements was repeated for four consecutive days. In the following weeks, two more ice mantles were prepared and two complete runs were performed (Run/Plateau 2 and 3). In total, six hundred measurements were analyzed.

3.2 Measurement Differences Analysis

Figures 3 and 4 represent the Boxplot diagrams of the daily measurements of cells JA1 and JA3, respectively, with the values being thermometer resistance (Ω). The Boxplot diagrams enable a quick "view" of the variability of the measurements and its

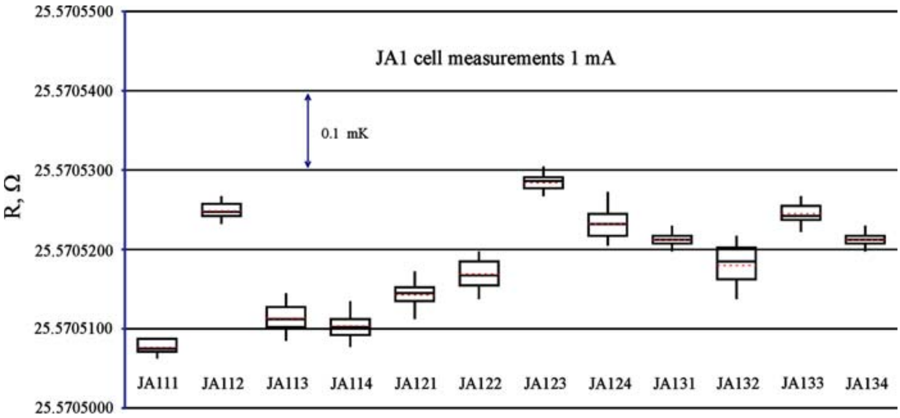


Fig. 3 Boxplot diagrams representing the JA1XY cell measurements, where X corresponds to the plateau and Y to the day

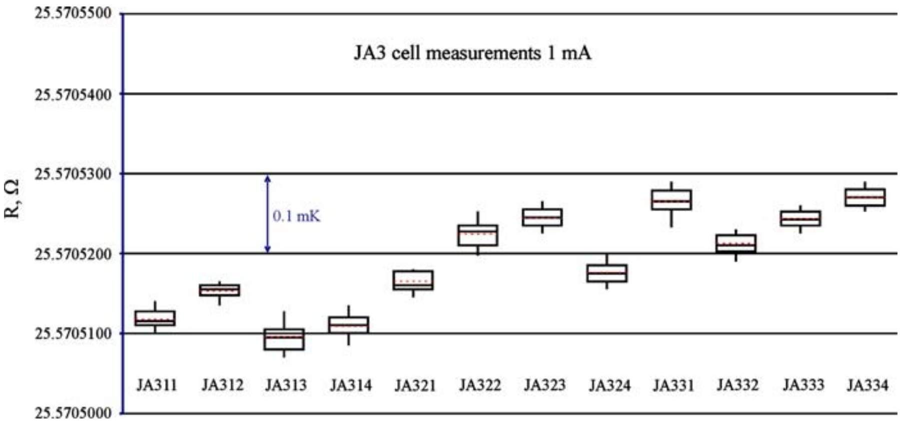


Fig. 4 Boxplot diagrams representing the JA3XY cell measurements, where X corresponds to the plateau and Y to the day

distribution symmetry. Indeed, in a single chart, they display the first quartile, median, third quartile, the interquartile range, the minimum and maximum values that are not considered as outliers (whiskers), and also plot the outliers ("Suspected"), if any, as considered by Tukey [3], i.e., the values exterior to $\pm 1.5\times$ interquartile range. All the values are included in a ± 0.1 mK interval in what was at first glance considered acceptable. The "zero" line in the boxes corresponds to the median of the measurements and the dotted red line to the average.

In this nested experiment, we considered the effects of Factor-P from the *Plateaux* ($P = 3$), the effects of the Factor-D from the *Days* ($D = 4$) for the same *Plateau*, and the variation between *Measurements* ($M = 50$) for the same *Day* and *Plateau* or the residual variation. Figure 5 shows the mean differences between the two cells, JA3-JA1, obtained daily for each plateau.

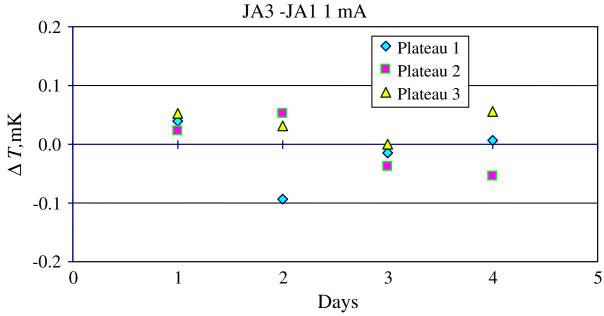


Fig. 5 Differences between the two cells JA3-JA1, average of the values: −0.005 mK (1 mA current)

Table 1 Analysis of variance table of the complete set of measurements ($\alpha = 5\%$)

Source of variation	Sum of squares	Degrees of freedom	Mean square	Expected value of mean square	F_0	Critical values $F_{\nu 1,\nu 2}$
Plateaus	0.2831	2	0.1415	$\sigma^2 + 50\sigma_\Gamma^2 + 200\sigma_P^2$	1.324	4.2565
Days	0.9619	9	0.1069	$\sigma^2 + 50\sigma_\Gamma^2$	299.679	1.8958
Measurements	0.2097	588	0.0004	σ^2		
Total	1.4546	599				

The variance analysis is usually presented as a “formatted” *ANOVA* Table, displaying the sums of squares, the degrees of freedom, the mean squares, the expected values of the mean squares, and the statistics F_0 obtained by calculating the ratios of subsequent levels mean squares.

From the ANOVA Table 1, we obtain F_0 values that will be compared with the critical values of the F distribution for $\alpha = 5\%$ and 2 and 9 degrees of freedom, $F_{0.05,2,9} = 4.2565$ for the *Plateau*/*Run* effect and 9 and 588 degrees of freedom $F_{0.05,9,588} = 1.8958$ for the Days effect.

We observe that the F_0 value is inferior to the F distribution for the *Plateau*/*Run* factor, so the null hypothesis is not rejected. For the *Days* factor, the null hypothesis is rejected for $\alpha = 5\%$; therefore, a significant difference exists among the daily measurements.

Equating the mean squares to their expected values, we can now calculate the variance components and include them in the uncertainty budget (see Table 2). These components of uncertainty, evaluated by a Type A method, reflect the random components of variance due to the factors effects.

3.3 Residual Analysis

It was mentioned (Sects. 2.1.1 and 2.1.2) that, in this model, the errors and the level effects are assumed to be normally and independently distributed. This assumption was checked by drawing a normality plot (Henry line) of the obtained residuals (see Fig. 6).

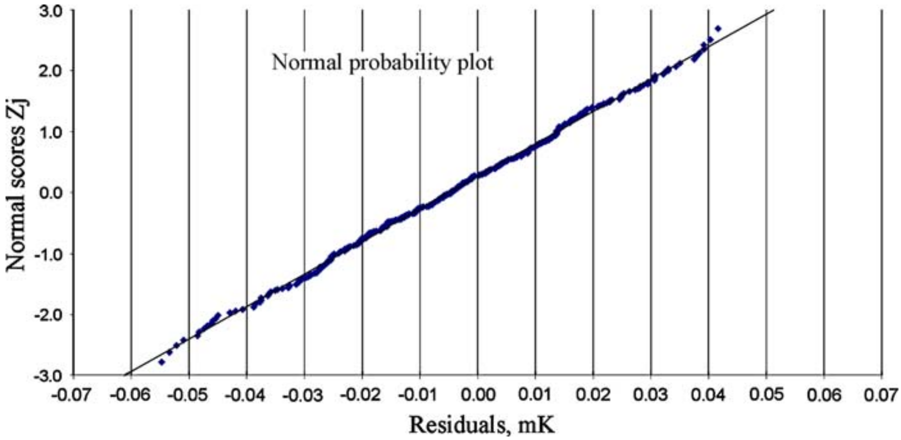


Fig. 6 Normal probability plot (Henry line) of the residuals of the comparison experiment. The residuals are plotted against the normal score Z_j and form an approximate straight line

Table 2 Components of uncertainty evaluated by a Type A method of the mean $\Delta T = 0.005\,\text{mK}$

Components of Type A uncertainty	Variance $(\text{mK})^2$	Standard deviation (mK)
Plateaus	0.0002	0.013
Days	0.0021	0.046
Measurements	0.0004	0.019
Total	0.0027	0.052

3.4 Remarks

(a) This model for "Type A" uncertainty evaluation that takes into account these time-dependent sources of variability is foreseen by the GUM. The obtained value using this nested design $u_A = 0.052\,\text{mK}$ (Table 2) is considerably larger than that obtained by calculating the standard deviation of the mean ($\Delta T = 0.005\,\text{mK}$) of the 600 measurements $u_A = 0.002\,\text{mK}$. This last approach is generally used and evidently underestimates this component of standard uncertainty.

(b) These two triple-point-of-water cells have the same geometry and contain "SMOW"[2] water. Cell JA1 was purchased in the year 2000 and cell JA3 during 2005. Only this last cell has a water isotopic composition certificate. Cell JA1 was used in CCT-K7 (Key Comparison of water-triple-point cells), and a difference of 0.1 mK was found with respect to older cells.

4 Summary

The nested-hierarchical design was described as a tool to identify and evaluate components of uncertainty arising from random effects. Applied to measurements, it is

[2] SMOW—Standard Mean Ocean Water.

suitable to estimate the components of standard uncertainty evaluated by a Type A method in time-dependent situations.

An application of the design has been presented to illustrate the variance components analysis in a three-factor nested model of a short, medium, and long-term comparison of two thermometric fixed points. The same model can be applied to other numbers of factors and is easily treated in an *Excel* spreadsheet.

References

1. BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, *Guide to the Expression of Uncertainty in Measurement (GUM)*, 2nd edn. (International Organization for Standardization, Genève, 1995), pp. 11, 83–87
2. G.E.P. Box, W.G. Hunter, J.S. Hunter, *Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building*, 1st edn. (John Wiley & Sons, New York, 1978), pp. 571–582
3. C.S. Davis, *Statistical Methods for the Analysis of Repeated Measurements*, 1st edn. (Springer, New York, 2002)
4. R.C. Guimarães, J.S. Cabral, *Estatística*, 1st edn. (Mc-Graw Hill de Portugal, Amadora, 1999), pp. 444–480
5. ISO 3534-3, *Statistics—Vocabulary and Symbols—Part 3: Design of Experiments*, 2nd edn. (International Organization for Standardization, Genève, 1999), pp. 31 (2.6) and 40–42 (3.4)
6. ISO TS 21749, *Measurement Uncertainty for Metrological Applications—Simple Replication and Nested Experiments* (International Organization for Standardization, Genève, 2005)
7. G.A. Milliken, D.E. Johnson, *Analysis of Messy Data*, vol. I: *Designed Experiments*, 1st edn. (Chapmann & Hall, London, 1997)
8. D. Montgomery, *Introduction to Statistical Quality Control*, 3rd edn. (John Wiley & Sons, New York, 1996), pp. 496–499
9. B. Murteira, *Probabilidades e Estatística*, vol. II, 2nd edn. (Mc-Graw Hill de Portugal, Amadora, 1990), pp. 361–364
10. J. Poirier, Analyse de la Variance et de la Régression, Plans d'Expérience, Techniques de l'Ingénieur, R1 (1993), pp. R260-1 to R260-23