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

UNCERTAINTY CALCULATION IN NESTED STRUCTURES

Eduarda Filipe

Instituto Português da Qualidade, Caparica, Portugal, efilipe@mail.ipq.pt

Abstract: The experimental design, active statistical tool generally used for the optimization of processes, can also be considered for the evaluation of sample(s) homogeneity. This tool may be applied to Metrology for the analysis of large amount of repeated measurements permitting the "mining" of the results and include this "time-dependent sources of variability" information at the uncertainty calculation.

Keywords: experimental design, uncertainty, nested.

1. INTRODUCTION

The experimental design is a statistical tool concerned with the planning of the experiments in order to obtain the maximum amount of information from the available resources [1]. 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 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 may be applied to Metrology especially for the analysis of large amount of repeated measurements [2] in short-term repeatability, the day-to-day and the long-term reproducibility, permitting the "mining" of the results and to include this "time-dependent sources of variability" information at the uncertainty calculation.

These components of uncertainty are evaluated by the statistical analysis of the results obtained from the experiments using the Analysis of Variance (ANOVA) for designs consisting of nested or hierarchical (ISO 3534-3, §2.6) [3] sequences of measurements. The Analysis of Variance as defined by the standard ISO 3534-3, §3.4, "is a technique that subdivides the total variation of a response variable into meaningful components, associated with specific sources of variation".

An application example in a calibration-nested structure with the corresponding estimation of the variances components is described.

2. PURPOSE

The Metrology Laboratories especially those concerned with the Scientific Metrology perform the calibration of the reference and working standards for the dissemination of the related unit.

This calibration work is designed in order to control or evaluate the different sources of variability. These laboratories usually have several standards and those are compared in a regular basis in order to have record of its regular behavior. When the laboratory performs the calibration of an external standard the acquired knowledge is included in this new calibration through the written procedures and experience allowing the validation of the obtained measurements.

3. THE NESTED OR HIERARCHICAL DESIGN. GENERAL MODEL

The nested design is defined [3,4] 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 hierarchized like a "tree" and any path from the "trunk" to the "extreme branches" will find the same number of nodes.

In the calibration experiment to be described, the calibration of a 25 ohm standard platinum resistance thermometer (SPRT) at the aluminum point, the factors are the repeated measurements obtained at regular intervals and the plateau/run measurements. These factors are considered as *random samples* of the *population* from which we are interested to draw conclusions.

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

3.1 Model for one factor

Considering firstly only one factor with a different levels taken randomly from a large population, [1, 5] any observation made at the i^{th} factor level with j observations, will be denoted by y_{ij}.

The mathematical model that describes the set of data is:

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

Where M_i is the expected (random) value of the group of observations i, μ the overall mean, τ_i the parameter associated with the i^{th} factor level designated by i^{th} factor effect and ε_{ij} 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 by 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:

$$\begin{cases} H_0 : \sigma_\tau^2 = 0 \\ H_1 : \sigma_\tau^2 > 0 \end{cases} \tag{3}$$

That is, if the null hypothesis is true, all factors 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 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 [6], the total variability of data (SS_T) can be separated into: the sum of squares of differences between factor-levels 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 a factor-levels from the factor-levels average (SS_E), due to the random error. Dividing each sum of squares by the respectively degrees of freedom, we obtain the corresponding mean squares (MS):

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

The mean square between factor-levels ($MSFactor$) [7] 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.

In order to test the hypotheses, we use the statistic:

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

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

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

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

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

The variance component of the factor is then obtained by:

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

3.2 Model for two factors

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

$$y_{pgm} = \mu + \Pi_p + \Gamma_g + \varepsilon_{pgm} \tag{9}$$

where y_{pgm} is the $(pgm)^{th}$ observation, μ the overall mean, Π_p the P^{th} random level effect, Γ_g the G^{th} random level effect, and ε_{pgm} 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 $\varepsilon_{pgm} \sim N(0, \sigma^2)$ and with mean zero and variances σ_P^2 and σ_G^2. The variance of any observation is composed by the sum of the variance components and the total number of measurements, N, is obtained by the product of the dimensions of the factors ($N = P \times G \times M$).

The total variability of the data [4, 9] can be expressed by:

$$\sum_p\sum_g\sum_m(y_{pdtm}-\bar{y})^2 = \sum_p GM(\bar{y}_p-\bar{y})^2 + \\ + \sum_p\sum_g M(\bar{y}_{pg}-\bar{y}_p)^2 + \sum_p\sum_d\sum_t(\bar{y}_{pgm}-\bar{y}_{pg})^2 \tag{10}$$

or

$$SS_T = SS_P + SS_{G|P} + SS_E$$

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 G for the same P ($SS_{G|P}$) and SS_E, the residual variation. Dividing by the respective degrees of freedom, $(P-1)$, $P \times (G\text{ - }1)$ and $P \times G \times (M\text{ - }1)$ 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:

---

[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.

$$E(MS_P) = E\left[\frac{SS_P}{P-1}\right] = \sigma^2 + M\sigma_G{}^2 + GM\sigma_P{}^2$$

$$E(MS_{G|P}) = E\left[\frac{SS_{G|P}}{P(G-1)}\right] = \sigma^2 + M\sigma_G{}^2$$

$$E(MS_E) = E\left[\frac{SS_E}{PG(M-1)}\right] = \sigma^2$$

4. CALIBRATION OF A SPRT AT THE ALUMINUM FREEZING POINT

Are presented the results/evaluation of two plateaus used for the calibration of a 25 ohm SPRT at the aluminum freezing point (660,323 ºC), performed in two different days. The measurements were taken in sequences of 30 values with a 5 minutes interval between each sequence and 6 groups of data were obtained, during approximately 1 hour. At each plateau/run the measurements started one hour after the stabilization of the freeze. These freeze plateaus usually last more than 10 hours.

The 360 values grouped by the 6 groups at each plateau were analyzed in order to study the stability of each plateau and include this information at the evaluation, by a type A method, of the uncertainty components.

From the measurements were obtained the values W_{Al} (ratio between the measured resistance at the aluminum point and the measured resistance at the triple point of water, t = 0,01 ºC), corrected by the hydrostatic pressure and extrapolated to a current of 0 mA.

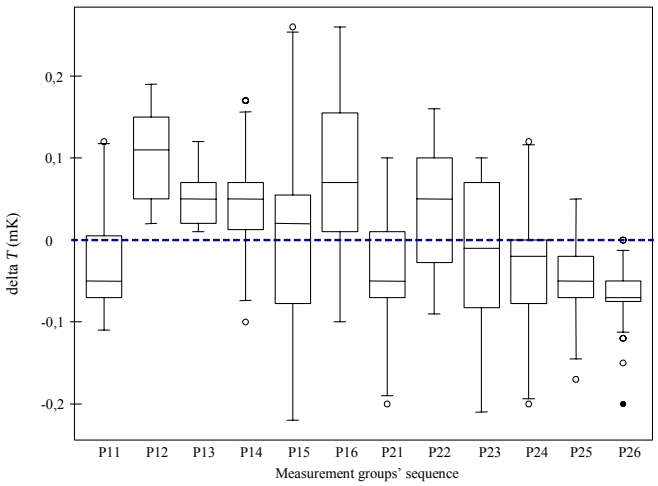

Figure 1 - Box plot diagrams of the groups of measurements from the first and second plateaus.

The boxplot diagrams of the groups are shown (see fig. 1) enabling a "view" of the variability of the measurements. We verify that the values are included in a ± 0,3 mK interval what in a first glance is very acceptable for a SPRT calibration at the aluminum point. "Suspected" outliers, as consider by J. Tukey, are the values exterior to a ± 1,5 × interquartile range (the difference between the upper and lower quartile), are marked. The "zero" line corresponds to the median of all measurements.

Analyzing this data with the analysis of variance (ANOVA) as described in [10] we see (Annex - Table 1) that there is no effect between the plateaus but an effect between the groups of measurements.

Equating the mean squares to their expected values, we can calculate the variance components and include them in the budget of the components of uncertainty drawn at Table 1.

Table 1 - Uncertainty budget for the components of uncertainty evaluated by a Type A method of the mean W_{Al} = 3,3757987

Expected value of mean square	Components of Type A uncertainty	Variance $(mK)^2$	Standard - deviation (mK)
$\sigma^2+30\sigma_I{}^2+180\sigma_P{}^2$	Plateaus	0,002	0,05
$\sigma^2+30\sigma_I{}^2$	Groups	0,002	0,04
σ^2	Measurements	0,005	0,07
	Total	0,009	0,09

For the purpose of the calibration of the SPRT these values are acceptable as they are consistent with the experimental equipment, the aluminum cell and the stability of the SPRTs at this temperature.

We can continue the example and withdraw the samples using one of the several methods, described in literature [1] that perform the multiple comparisons between each sample mean with all other sample means and suitable for the cases were the null hypothesis is rejected.

It was chosen for comparing the means μ_i and μ_j the LSD (Least Significance Difference) multiple comparison method α significance value:

$$LSD_\alpha = t_{\alpha/2,\nu} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

The pairs of means are different if

$$\left|\mu_i - \mu_j\right| > LSD_\alpha$$

This method is the easiest to apply but caution must be taken as the "experimentwise error rate" [1] increases with the number of treatments.

From the first plateau were withdrawn the groups P11, P13, P14 and from the second plateau the groups P21, P25 and P26. Another ANOVA table (see Annex – Table 2) was built and in this case the F_0 values are inferior to the critical values of the F distribution. The calculated uncertainty is very similar to the previous calculation (see Annex – Table 3).

5. REMARKS

At the aluminum point the SPRTs suffer stability problems that can be due to an insufficient previous annealing or to some "quenching" when withdrawing the SPRTs from the hot furnace. Repeating the plateaus allows us to verify the stability of the measuring instrument.

The plateaus values continuously taken where previously analyzed in terms of its normality. The size of the samples and the time interval were procedure options. This stability studies are important for the validation of the obtained values.

This model for "Type A" uncertainty evaluation that takes into account these time-dependent sources of

variability it is foreseen at the GUM [1]. The obtained value using this nested design u_A= 0,09 mK (Table 1) is considerably larger from that obtained by calculating the standard deviation of the mean of the 360 measurements u_A= 0,0046 mK. This last approach is generally used and evidently sub-estimates this component of uncertainty.

6. CONCLUSION

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

An application of the design has been drafted to illustrate the variance components analysis in a SPRT calibration at the aluminum point in a two stage nested model. This same model can be applied to other number of factors, easily treated in an *Excel* spreadsheet.

ANNEX

Table 1 - Analysis of variance table of the complete set of measurements. α = 5%

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0	Critical values $F_{v1,v2}$
Plateaus	0,4291	1	0,4291	8,1681	4,9646
Groups	0,5254	10	0,0525	10,3848	1,8579
Measurements	1,7605	348	0,0051		
Total	2,7150	359			

Table 2 - Analysis of variance table withdrawing the groups P11, P13, P14, P21, P25 and P26. α = 5%

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0	Critical values $F_{v1,v2}$
Plateaus	0,1922	1	0,1922	7,0062	18,5128
Time	0,0549	2	0,0274	4,3176	3,0744*
Measurements	0,7369	116	0,0064		
Total	0,9839	119			

* Critical value for α = 1% $F_{0,01,\,2\,,\,116}$=4,7929

Table 3 - Uncertainty budget for components of uncertainty evaluated by a Type A method of the mean W_{Al} = 3,3757989

Expected value of mean square	Components of Type A uncertainty	Variance $(mK)^2$	Standard - deviation (mK)
$\sigma^2 + 30\sigma_T^2 + 60\sigma_P^2$	Plateaus	0,003	0,05
$\sigma^2 + 30\sigma_T^2$	Time	0,001	0,03
σ^2	Measurements	0,006	0,08
	Total	0,010	0,10

7. REFERENCES

1. Milliken, G.A., Johnson D. E., Analysis of Messy Data. Vol. I: Designed Experiments. 1st ed., London, Chapmann & Hall, 1997.
2. BIPM et al, Guide to the Expression of Uncertainty in Measurement (GUM), 2nd ed., International Organization for Standardization, Genève, 1995, pp. 11, 83-87.
3. ISO 3534-3, Statistics – Vocabulary and Symbols – Part 3: Design of Experiments, 2nd ed., Genève, International Organization for Standardization, 1999, pp. 31 (2.6) and 40-42 (3.4).
4. ISO TS 21749 Measurement uncertainty for metrological applications — Simple replication and nested experiments Genève, International Organization for Standardization, 2005
5. Box, G.E.P., Hunter, W.G., Hunter J.S., Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building, 1st ed., New York, John Wiley & Sons, 1978, pp. 571-582
6. Guimarães R.C., Cabral J.S., Estatística, 1st ed., Amadora: Mc-Graw Hill de Portugal, 1999, pp. 444-480
7. Murteira, B., Probabilidades e Estatística. Vol. II, 2nd ed., Amadora, Mc-Graw Hill de Portugal, 1990, pp. 361-364.
8. Montgomery, D., Introduction to Statistical Quality Control, 3rd ed., New York, John Wiley & Sons, 1996, pp. 496-499.
9. Poirier J., "Analyse de la Variance et de la Régression. Plans d'Experience", *Techniques de l'Ingenieur,* **R1**, 1993, pp. R260-1 to R260-23.
10. Filipe E., "Evaluation of Standard Uncertainties in nested structures", Advanced Mathematical & Computational Tools in Metrology VII, Series on Advances in Mathematics for Applied Sciences – Vol. 72, World Scientific, 2006.