UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2577

Statistical Analysis of Proficiency Testing Results

Dorival Leão P. Jr.
Reiko Aoki
Gecirlei F. da Silva

Nº 73

NOTAS

Série Estatística

São Carlos – SP
Nov./2002

Statistical Analysis of Proficiency Testing Results

Dorival Leão P. Jr., Reiko Aoki, Gecirlei F. da Silva

Departamento de Ciência da Computação e Estatística,
ICMC/USP-São Carlos
Caixa Postal 668
13560-970 São Carlos, SP

RESUMO. O uso de ensaios de comparação interlaboratorial para acompanhar a performance de laboratórios na realização de testes específicos (ou calibrações) é denominado Ensaios de Proficiência (PT). Neste trabalho propomos um teste da razão de verossimilhança generalizada para comparar a performance de um grupo de laboratórios na realização de testes específicos e ilustrar o procedimento considerando dados do programa de PT organizado pela REMESP na área de volume. Este programa foi conduzido como um esquema de comparação de medidas, onde um único artefato (um balão de 50 ml em nossa ilustração) é distribuído sequencialmente entre os participantes, enviando o artefato de um laboratório para outro via correio. Em nosso estudo o balão foi medido por seis laboratórios, todos membros da REMESP, sendo que cada laboratório realizou dez medições do artefato.

SUMMARY. The use of interlaboratory test comparisons to determine the performance of individual laboratories for specific tests (or calibration) [Guide 43] is called Proficiency Testing (PT). In this paper we propose the use of the generalized likelihood ratio test to compare the performance of the group of laboratories for specific tests relative to the assigned value and illustrate the procedure considering an actual data from the PT program organized by REMESP in the area of volume. This program was conducted as the measurement comparison scheme, where a single test item (a balloon of 50 ml in our illustration) is distributed sequentially among participants, by mailing the item from one laboratory to the next laboratory. In that study, the balloon was mailed to 6 laboratories, which are members of the REMESP and the item was measured by each laboratory 10 times.

KEY WORDS: Proficiency Testing; calibration; Generalized Likelihood Ratio Test; Maximum Likelihood Estimator; Assigned Value.

1 Introduction

A key component of industrial quality control is the gathering of the reliable data sets by means of measurement systems. A measurement system is a process which combine instruments, operators and methods for obtaining the measurand, i.e. the value of a quantity to be measured. Usually, the value obtained as a result of a measurement is only an approximation or estimate of the measurand and thus it must be accompanied by a statement of the standard deviation (uncertainty) of the estimate. See ISO GUM (1995) for more details. As a consequence, it is essential to monitor the consistency and the competence of the laboratories that realize this specific measurements. In the state of São Paulo, Brazil, it was created an organization with the responsibility for coordinating activities involved with the monitoration of the consistency and competency of the laboratories, which is called REMESP (Rede de Metrologia do Estado de São Paulo) . This organization is composed by companies and laboratories interested in the improvement of their measurement systems.

Proficiency Testing (PT) is the use of interlaboratory test comparisons to determine the performance of individual laboratories for specific tests (or calibration) [Guide 43]. It monitors the consistency and comparability of laboratory's test data. PT techniques depends on the nature of the item or material under test, as well as the test methods in use and also the number of laboratories which are participating. There are six distinct types of PT discussed in Guide 43: Measurement Comparison Schemes, Interlaboratory Testing Schemes (Bulk Material), Split-Sample Schemes, Qualitative Schemes, Known-Value Schemes and Partial-Process Schemes. Besides monitoring the consistency and comparability of Laboratory's test data, PT program may improve the Laboratory's test data (see Richardson et al. (1996), for example).

The statistical techniques used to analyze PT results has three basic steps, which are common in all PT tests. After participants results are evaluated, the next three steps are applied: Determination of the Assigned Value, Performance Statistics and Comparison Results. The participants results should be compared to the value or answer (assigned value) that best demonstrates competence with the method. There are a variety of common methods to determine the assigned value. Different situations require different procedures, however the primary concern is on the uncertainty (primarily traceability) versus the convenience [Guide 43]. To judge the suitability of the assigned value and determine performance statistics of each laboratory, it is necessary to estimate the uncertainty of the assigned value. There are a number of methods to determine the assigned value as described in Guide 43. After the determination of the assigned value, the next step consists in transforming the measurement of the participants into performance statistics. The objective is to measure the difference from the assigned value in a way that allows the comparison with a defined criteria. Nowadays, the analysis are made via statistics that provide a direct comparison between the measurements of the participants and the assigned value. A common performance statistics in measurement comparisons programs, proposed in Guide 43, is the E_n score, which is given by

$$E_n = \frac{y - \mu_x}{\sqrt{U_{lab}^2 + U_{ref}^2}}$$

where , y is the laboratory measurement, μ_x is the estimated mean of the assigned value, U_{lab} and U_{ref} corresponds, respectively, to the expanded uncertainty of the laboratory value and to the expanded uncertainty of the assigned value. This score describes the difference between the laboratory measurement and the assigned value relative to the uncertainties of the values involved in the difference. In this case, each laboratory measurement is compared with the assigned value via individual testing hypothesis. However, in many situations we are also interested in establishing multiple comparisons to evaluate the performance of the group of participants laboratories relative to the assigned value.

In this paper, we propose a generalized likelihood ratio procedure to test the competence of the group of laboratories and we apply to an actual data from the PT program organized by REMESP in the area of volume. The program was conducted as the first PT type (Measurement Comparison Schemes) which consists in distributing a single test item (a balloon of 50 ml in our illustration) sequentially among participants, by mailing the item from one laboratory to the next laboratory. In that study, the balloon was mailed to 6 laboratories, which are members of the REMESP. Each laboratory measured the item 10 times and the corresponding measurements together with the related uncertainty was obtained. The assigned value is determined by considering two methodologies, namely 'consensus value from participants' and 'Reference

Value'. In the reference value procedure the assigned value is determined by analysis, measurement or comparison of the test item with a reference material or standard that is traceable to a national or international standard. This procedure is common in measurement calibration schemes. Consensus value from participants are very sensitive to outliers. Outliers can have profound influence on means and standard deviations. When consensus mean and consensus standard deviation are used to estimate the mean and the standard deviation of the distribution of the assigned value, there must be procedures in order to control the effects of outliers. Different methods could be applied to identify outliers, see Dixon (1953), Grubbs (1969) and Healy (1979) for example. Another way for dealing with outliers is the use of robust statistical procedures. These procedures do not eliminate outliers from the calculations, rather, they minimize the weight assigned to values far from the center, see Rocke (1983) and Thompson (1989).

Jaech (1985) proposed a measurement error model, establishing and developing methods for estimating and testing variance components in errors of measurements (or in the 'imprecision' of the measurements). Another way commonly used in the literature is the errors-in-variables regression models [see Fuller (1987), Tan and Iglewicz (1999) and Chang and van Ness (1999), for example]. In this paper, we consider the measurement error model proposed by Jaech (1985), to analyze the PT results, as in our study the covariate (assigned value) is evaluated without error. This procedure is simpler than the use of errors-in-variables models and the covariate is assumed to be a random variable.

In Section 2, we describe the Proficiency Testing Results. In Section 3 we define the model. The explicit maximum likelihood estimator (MLE) for the bias of the measurements of each laboratories with respect to the assigned value are evaluated in Section 4, as well as some properties of the MLE and the development of the generalized likelihood ratio test to compare the performance of the group of participants laboratories relative to the assigned value. Finally, in Section 5, we apply the methodology described in earlier sections to the PT developed by REMESP to measure the volume of a balloon and in Section 6 we discuss the results.

2 Proficiency Testing Results

In this section we describe the data and the methodology to evaluate the uncertainty of the measurements. In most cases a measurand Y (in our application, it represents the volume of the balloon) is not measured directly, but it is determined from n other quantities $X_1, X_2, \cdots, X_n$ through a functional relationship f:

$$Y = f(X_1, X_2, \cdots, X_n).$$

In order to evaluate the volume of the balloon, the laboratories filled it up with water and then they considered the following relationship between volume, density and mass:

$$V = \left\{ \frac{m_{water}}{\rho_{water} - \rho_{air}} \right\} \left\{ 1 - \frac{\rho_{air}}{\rho_{mass}} \right\} [1 - \gamma (t_{water} - t_{ref})], \qquad (1)$$

with

- t_{ref}: Reference temperature of the water - 20,0°C;
- t_{water}: Temperature of the water during the calibration;
- m_{water}: Mass of the water;
- ρ_{water} : Specific mass of the water;
- ρ_{air} : Specific mass of the air;

- ρ_{mass} : Specific mass of the master mass;
- γ : Correction factor.

The variables m_{water}, ρ_{water}, ρ_{air}, ρ_{mass} and t_{water} are the input quantities and V is the response variable (Volume). We assume that the input quantities are random variables with appropriate means and variances. The estimated standard deviation associated with the measurement result V, termed *combined standard uncertainty* is denoted by $u_c(V)$. It is determined from the estimated standard deviation associated with each input estimate x_i, which is termed standard uncertainty and denoted by $u(x_i)$. Each input estimate x_i and its associated standard uncertainty (or standard deviation) $u(x_i)$ are obtained from the distributions of the input variables X_i, $i = 1, \cdots, n$. These probability distributions may be based on a series of observations $X_{i,k}$ of X_i, or it may be an *a priori* distribution. It must be recognized that in both cases the distributions of the input quantities X_i are models that are used to represent the state of our knowledge. It follows from Kendal and Stuart (1947, pp 231) and ISO GUM (1995), that the combined standard uncertainty could be approximated by

$$u_c^2(V) = \sum_{i=1}^{n}\left(\frac{\partial f}{\partial X_i}\right)^2 Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{\partial f}{\partial X_i}\frac{\partial f}{\partial X_j}Cov[X_i, X_j] \quad (2)$$

where $Cov[X_i, X_j]$ corresponds to the covariance between X_i and X_j for all $i, j = 1, \cdots, n$.

In our case, we assume that the input quantities are independent random variables. Thus, it follows from Equations (1) and (2), that

$$u_c^2(V) = \left(\frac{\partial f}{\partial m_{water}}\right)^2 Var(m_{water}) + \left(\frac{\partial f}{\partial \rho_{water}}\right)^2 Var(\rho_{water}) + \left(\frac{\partial f}{\partial t_{water}}\right)^2 Var(t_{water}) + \left(\frac{\partial f}{\partial \rho_{air}}\right)^2 Var(\rho_{air}) + \left(\frac{\partial f}{\partial \rho_{mass}}\right)^2 Var(\rho_{mass}). \quad (3)$$

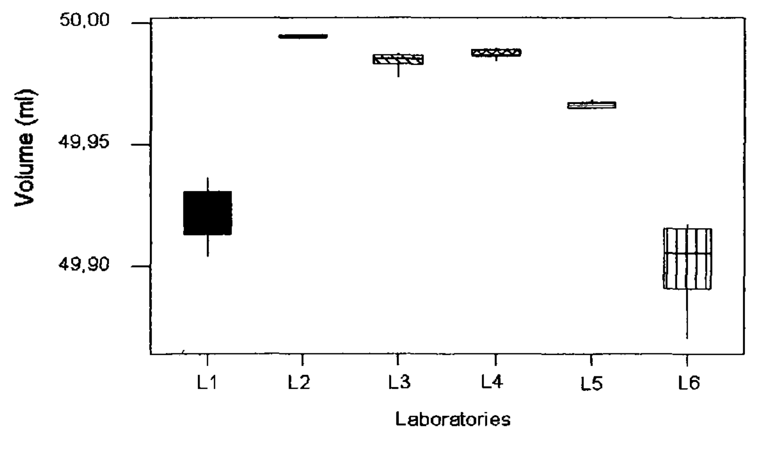
In most practical measurement situations, the calculation of intervals with a specified confidence level is very important. Let us define the interval for the measurand Y by $y-U_E \leq Y \leq y+U_E$, where y is the estimate of Y and $U_E = K_p u_c(y)$ is an expanded uncertainty, with K_p the coverage factor corresponding to a specified confidence level p. In general, the distribution of the random variable $(y-Y)/u_c(y)$ can be approximated by a t_Student distribution with an effective degrees of freedom ν_{eff} obtained from the Welch-Satterthwaite formula (see, ISO GUM (1995)). The 6 laboratories members of the REMESP considered the Welch-Satterthwaite formula to determine the effective degrees of freedom and the correspondent coverage factor K_p, with a 95% confidence level.

Table 1: Measurements of the Laboratories

	Lab1	**Lab2**	**Lab3**	**Lab4**	**Lab5**	**Lab6**
	49.9273	49.9936	49.9844	49.9866	49.966	49.8798
	49.9365	49.9940	49.9844	49.9868	49.965	49.9039
	49.9035	49.9938	49.9790	49.9901	49.966	49.9169
	49.9134	49.9940	49.9772	49.9869	49.967	49.9149
	49.9237	49.9949	49.9853	49.9893	49.966	49.9139
	49.9279	49.9946	49.9868	49.9869	49.969	49.9069
	49.9304	49.9947	49.9876	49.9891	49.965	49.8708
	49.9308	49.9945	49.9868	49.9853	49.965	49.8969
	49.9248	49.9952	49.9874	49.9890	49.967	49.8948
	49.9118	49.9950	49.9855	49.9841	49.968	49.9179
Mean	49.9230	49.9944	49.9844	49.9874	49.9664	49.9017
Combined Uncertainty	0.0036	0.0138	0.008	0.003	0.009	0.01
Expanded Uncertainty	0.016	0.028	0.016	0.006	0.02	0.02

In Table 1, we introduce the data obtained from the PT described in Section 1, where six laboratories measured the same item (a balloon of 50 ml) ten times, together with the mean value, the respective combined standard uncertainty given in (3) and the expanded uncertainty. Figure 1 presents the Multiple Box-Plot of the data set. We observe that the laboratories L1 and L6 obtained measurements far from the group of other laboratories. As discussed in the Introduction the detection of outliers is very important, when we consider consensus values from participants to obtain the assigned value. Here, we applied Grubbs and Dixon's test to detect outliers. These tests were applied to detect outliers in the univariate data set formed by the mean of the measurements of each laboratory. In both cases, the mean value of the measurements of the laboratories are not considered as outlier.

Figure 1: Box Plot



3 Statistical Model

Consider a PT program, where a group of k laboratories are participating, with each laboratory obtaining n_i measurements of the same item , $i = 1, \cdots, k$. As the same item is measured by each laboratory n_i times, the model should take into account the possible dependence on the outcome measurements for $i = 1, \cdots, k$ and $j = 1, \cdots, n_i$. A possible model to describe these measurements may be given by

$$Y_{ij} = \alpha_i + X + \epsilon_{ij}, \quad j = 1, \cdots, n_i \; ; \; i = 1, \cdots, k, \tag{4}$$

where α_i and ϵ_{ij}, respectively, correspond to the bias with respect to the assigned value and the measurement error associated with the laboratory i, for the jth measurement. We assume that the assigned value X has normal distribution with mean μ_x and standard deviation σ_x. Furthermore, the measurement error ϵ_{ij} has normal distribution with zero mean and standard deviation σ_i. The parameters of the distribution of the assigned value may be estimated by

i. Consensus mean and consensus uncertainty from the laboratories which measurements are not considered as an outlier; or

ii. Reference Value;

among other forms described in Guide 43. The parameter of the measurement error ϵ_{ij} are estimated by the combined standard uncertainty of the laboratory i. In that way, the parameters of the distribution of the assigned value and the parameters of the distribution of the measurement errors are assumed to be known. Since each laboratory has different instruments, environment conditions and operators, we consider that ϵ_{ij} and ϵ_{mj} are independent and also, as the replicas of the measurements of each laboratories are obtained from repetitivity condition, we assume that ϵ_{ij} and ϵ_{il} are independent for all $j, l = 1, \cdots, n_i$ and $i, m = 1, \cdots, k$. Furthermore, we consider that ϵ_{ij} and X are independently distributed. Observe that the competence of each laboratory is determined by the respective bias (α_i) with respect to the assigned value, and the respective combined standard uncertainty (σ_i).

Let us denote by $\mathbf{1}_n$ a vector composed by n ones; $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_k^T\right)^T$, with $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{in_i})^T$ the measurements of the laboratory i; $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k)^T$; $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_k^T)^T$, with $\boldsymbol{\alpha}_i = \alpha_i \, \mathbf{1}_{n_i}$; $\boldsymbol{\mu}_x^* = (\boldsymbol{\mu}_x(1)^T, \cdots, \boldsymbol{\mu}_x(k)^T)^T$, with $\boldsymbol{\mu}_x(i) = \mu_x \, \mathbf{1}_{n_i}$, $i = 1, \cdots, k$ and the superscript T denotes the transpose operation. We assume that $\boldsymbol{Y}$ has p-multivariate normal distribution with $p = n_1 + \cdots + n_k$. Then, $\boldsymbol{Y} \sim N_p(\boldsymbol{\xi}, \boldsymbol{\Sigma})$ where $\boldsymbol{\xi} = \boldsymbol{\alpha}^* + \boldsymbol{\mu}_x^*$ and $\boldsymbol{\Sigma} = A + \sigma_x^2 \mathbf{1}_p \mathbf{1}_p^T$, with $A = D(\boldsymbol{\sigma}_1^T, \cdots, \boldsymbol{\sigma}_k^T)$ and $\boldsymbol{\sigma}_i = \sigma_i \, \mathbf{1}_{n_i}$, with D denoting a diagonal matrix.

4 Maximum Likelihood Estimator and Generalized Likelihood Ratio Test

In this section, we derive the MLE of the parameters and prove that the covariance matrix of these estimators is the inverse of the Fisher Information matrix. Next, we develop the generalized likelihood ratio test to determine the competence of laboratories with respect to the assigned value. By using general properties of the multivariate normal distribution and after some algebraic manipulations, it follows that the log-likelihood function for the vector of parameters $\boldsymbol{\alpha}$ (all other parameters are supposed known, as described in the previous sections) can be written as

$$l(\boldsymbol{y};\boldsymbol{\alpha}) = lnf(\boldsymbol{y};\boldsymbol{\alpha}) = const \; - \; \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_x^*)^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_x^*) \; + \; (\boldsymbol{y} - \boldsymbol{\mu}_x^*)^T \Sigma^{-1} J\boldsymbol{\alpha} \; - \; \frac{1}{2}(J\boldsymbol{\alpha})^T \Sigma^{-1} J\boldsymbol{\alpha}, \quad (5)$$

where f corresponds with the likelihood function obtained from the distribution of $\boldsymbol{Y}$,

$$\Sigma^{-1} = A^{-1} + \gamma \boldsymbol{b}^* \, \boldsymbol{b}^{*T},$$

with

$$b^* = (\frac{1}{\sigma_1}\mathbf{1}_{n_1}^T, \frac{1}{\sigma_2}\mathbf{1}_{n_2}^T, \cdots, \frac{1}{\sigma_k}\mathbf{1}_{n_k}^T)^T \quad \text{and} \quad \gamma = -\sigma_x^2 \left(1 + \sigma_x^2 \sum_{i=1}^{k} \frac{n_i}{\sigma_i}\right)^{-1};$$

and J corresponds to the following matrix

$$\begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \dots & \mathbf{0} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{n_k} \end{pmatrix}.$$

From the likelihood equations we obtain

$$\hat{\boldsymbol{\alpha}} = \left(J^T \Sigma^{-1} J\right)^{-1} J^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_x^*),$$

and after some algebraic manipulations, it follows that

$$\hat{\boldsymbol{\alpha}} \;=\; \Gamma(\boldsymbol{y} - \boldsymbol{\mu}_x^*) \;=\; \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ . \\ . \\ . \\ \hat{\alpha}_k \end{pmatrix} \;=\; \begin{pmatrix} \bar{y_1}. - \mu_x \\ \bar{y_2}. - \mu_x \\ . \\ . \\ . \\ \bar{y_k}. - \mu_x \end{pmatrix}, \tag{6}$$

with

$$\Gamma \;=\; \begin{pmatrix} \frac{1}{n_1}\mathbf{1}_{n_1}^T & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\mathbf{1}_{n_2}^T & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \frac{1}{n_k}\mathbf{1}_{n_k}^T \end{pmatrix}$$

and $\bar{y_i}. = \sum_j \frac{y_{ij}}{n_i}$. Thus we obtained the MLE of the vector of parameters $\boldsymbol{\alpha}$, where α_i corresponds to the bias of the *ith* laboratory with respect to the assigned value.

Since $\boldsymbol{Y}$ has p-multivariate normal distribution with mean vector $\boldsymbol{\alpha}^*$ and covariance matrix $\boldsymbol{\Sigma}$, it follows that $\hat{\boldsymbol{\alpha}} = \Gamma(\boldsymbol{Y} - \boldsymbol{\mu}_x{}^*)$ is distributed as a k-multivariate normal distribution with mean vector $\boldsymbol{\alpha}$ and covariance matrix $\Gamma\Sigma\Gamma^T$. On the other hand, after some algebraic manipulations, the Fisher information matrix can be obtained and it is given by

$$\mathbf{I}(\boldsymbol{\alpha}) = E\left\{\left[\frac{\partial lnf(\boldsymbol{Y};\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right]\left[\frac{\partial lnf(\boldsymbol{Y};\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right]^T\right\} = J^T\Sigma^{-1}J,$$

and then, after algebraic manipulations, it follows that

$$[\mathbf{I}(\boldsymbol{\alpha})]^{-1} = (J^T\Sigma^{-1}J)^{-1} = \Gamma\Sigma\Gamma^T,$$

that is, the covariance matrix of $\hat{\boldsymbol{\alpha}}$ is the inverse of the Fisher information matrix.

Consider the general linear hypothesis given by

$$\begin{cases} H_0 : C\boldsymbol{\alpha} = \boldsymbol{a} \\ H_1 : C\boldsymbol{\alpha} \neq \boldsymbol{a} \end{cases}$$

with C a $m \times k$ matrix of constants, $\boldsymbol{a}$ a $m \times 1$ vector of constants and λ the generalized likelihood ratio defined as

$$\lambda = \frac{\sup_{\boldsymbol{\alpha}\in Q_0} f(\boldsymbol{y}\ ,\ \boldsymbol{\alpha})}{\sup_{\boldsymbol{\alpha}\in\mathbb{R}^k} f(\boldsymbol{y}\ ,\ \boldsymbol{\alpha})} \tag{7}$$

where $Q_0 = \{\boldsymbol{\alpha} \in \mathbb{R}^k \mid C\boldsymbol{\alpha} = \boldsymbol{a}\}$. Then, the generalized likelihood ratio test to determine the competence of laboratories with respect to the assigned value can be summarized in the following theorem.

Theorem 4.1 *Under the model defined by (4) with the log-likelihood function given by (5) the generalized likelihood ratio is given by*

$$\lambda = exp\left\{-\frac{1}{2}\left[(C\hat{\boldsymbol{\alpha}} - \boldsymbol{a})^T(B^TJ^T\Sigma^{-1}JB)(C\hat{\boldsymbol{\alpha}} - \boldsymbol{a})\right]\right\},$$

and denoting

$$W = -2ln(\lambda),$$

W has noncentral chi-square distribution with m degrees of freedom and noncentrality parameter s, where

$$s = \frac{1}{2}(C\boldsymbol{\alpha} - \boldsymbol{a})^T \left[B^T J^T \Sigma^{-1} J B\right](C\boldsymbol{\alpha} - \boldsymbol{a}),$$

with $B = (J^T\Sigma^{-1}J)^{-1}C^T\left(C(J^T\Sigma^{-1}J)^{-1}C^T\right)^{-1}$.

Proof: The proof of the theorem is given in Appendix A.

5 Applications

In this section, we analyze the actual data obtained from the PT program organized by REMESP (see, Sections 1 and 2). The parameters of the assigned value was estimated considering two different procedures: consensus value from participants and Reference Value. Independent of the choice of the procedure, to test the consistency of the group of laboratories the following joint hypothesis testing can be applied,

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0 \\ H_1 : \text{ at least one different from zero.} \end{cases}$$

If we reject H_0, we say that the group of laboratories is not consistent. Next, we propose the following hypothesis testing to evaluate the performance of individual laboratories with respect to the assigned value

$$\begin{cases} H_0 : \alpha_i = 0 \\ H_1 : \alpha_i \neq 0, \end{cases}$$

for $i = 1, \cdots, k$. If we do not reject H_0, we conclude that the measurement of the correspondent laboratory is GLR-satisfactory.

Finally, we evaluate the normalized error E_n, proposed in Guide 43, which is given by

$$E_n(i) = \frac{\bar{y_{i.}} - \mu_x}{\sqrt{U_{lab}^2 + U_{ref}^2}}$$

where, $\bar{y_{i.}}$ is the mean of i-th laboratory measurements ($i = 1, \cdots, k$), μ_x is the estimated mean of the assigned value and U_{ref} (U_{lab}) corresponds to the expanded uncertainty of the assigned value (expanded uncertainty of the laboratory value). If the normalized error E_n is less than 1, we say that the measurement of the correspondent laboratory is satisfactory, see Guide 43.

5.1 Reference Value

In this procedure the parameters of the distribution of the assigned value is determined by a calibration process. In our application the Laboratory 5 are recognized as having competence that exceeds the competence of the others laboratories to measure (or calibrate) the item of interest. Then, we consider its mean and combined standard uncertainty to estimate the parameters of the assigned value, that is, $\mu_x = 49.9664$, $\sigma_x^2 = 0.000081$ and $U_{ref} = 0.02$. In table 2, we show the bias and test for the competence of the laboratories with respect to the assigned value.

Table 2: Bias and Test for Competence of the 5 Laboratories

Reference Value

Test	Estimative of the Bias	Statistic W	P-value	E_n	Expanded Uncertainty
All X L5		2.384.7	0.0001		
L1 X L5	-0.0434	22.8876	0.0001	1.7	0.016
L2 X L5	0.028	7.8366	0.0051	0.81	0.028
L3 X L5	0.018	3.7071	0.0542	0.7	0.016
L4 X L5	0.021	5.3846	0.0203	1.01	0.006
L6 X L5	-0.0647	46.0010	0.0001	2.28	0.02

In this section, we propose the size of the tests $\eta = 0.05$ to decide between the hypothesis. Since the P-value for the first test is less than 0.05, we reject the hypothesis that the bias of the laboratories are all zero, so we conclude that the group of laboratories is not consistent. Furthermore, it follows from the multiple comparisons tests that only the measurements of the laboratories L3 is GLR-satisfactory. However, it follows from the E_n criteria that the measurements of the laboratories L2 and L3 are considered satisfactory. Observe that the measurements of the laboratory L2 is satisfactory, but it is not GLR-satisfactory. This is a consequence of the magnitude of its expanded uncertainty ($u_c = 0.028$), which is the greatest uncertainty among the participants. Moreover, the laboratory L2 has the greatest bias among the laboratories considered as satisfactory (L2, L3 and L4) .

5.2 Consensus Value from Participants

Another way to estimate the parameters of the distribution of the assigned value is by means of the consensus value from participants. This is one of the most commonly used procedure in accreditation programs with routine PT. The experience of the College of American Pathologists concluded that the consensus mean of the participants laboratories is consistently very close to the reference value and is more reliable than the consensus of experts laboratories [Tholen (1993)]. In this case we consider consensus mean and consensus uncertainty to estimate the parameters of the assigned value, that is

$$\mu_x = \bar{y}_{..} = \sum_{i=1}^{k} \frac{\bar{y}_{i.}}{k} = 49.9596 \quad \text{and} \quad \sigma_x^2 = \frac{\sum_{i=1}^{k} \sigma_i^2}{k} = 0.000076233$$

and determine the expanded uncertainty for the assigned value from the normal distribution. Thus, we obtain that $U_{ref} = 2\sigma_x = 0.017$. In table 3, we show the results of the tests and the normalized error. We observe that the group of laboratories is not consistent with respect to the assigned value. Moreover, only the laboratory L5 are GLR-satisfactory (and satisfactory).

6 Discussion

In general, the PT results are analyzed via E_n score described in Section 5. Here, we proposed a methodology to analyze the PT results organized as Measurement Calibration Scheme. We developed a GLR test to determine the consistency of the measurements of the laboratories and also the competence of individual laboratory with respect to the assigned value. In order to construct the GLR test, we proposed a model similar to Jaech (1985), where the covariate is a random variable associated with the assigned value. Considering this model, we evaluated the MLE's for the bias of the measurements of the laboratories with respect to the assigned value. Note that, in this case, the moment estimators coincide with the MLE's.

Table 3: Bias and Test for Competence of the 6 Laboratories

Consensus Value from Participants

Test	Estimative of the Bias	Statistic W	P-value	E_n	Expanded Uncertainty
All X Ref.		2388.0	0.0001		
L1 X Ref.	-0.0366	17.2782	0.0001	1.54	0.016
L2 X Ref.	0.0348	12.7107	0.0004	1.05	0.028
L3 X Ref.	0.0248	7.4430	0.0064	1.04	0.016
L4 X Ref.	0.0278	10.0196	0.0015	1.51	0.003
L5 X Ref.	0.0068	0.5483	0.4590	0.26	0.020
L6 X Ref.	-0.0579	38.8762	0.0001	2.18	0.020

In the example described in the previous sections, the six laboratories, members of REMESP, realized a series of ten measurements of the volume of the same balloon. Considering Y_{ij}, where Y_{ij} represents the j-th measurement of the i-th laboratory for

$j = 1, \cdots 10$ and $i = 1, \cdots, 6$, the variance of $\bar{Y_{ij}}$ is given by $\frac{\sigma_x^2+\sigma_i^2}{n_i}$. On the other hand, if we consider $n_i = 1$ and Y_{ij} as the mean of the measurements of the i-th laboratory, the variance of $\bar{Y_{ij}}$ is $\sigma_x^2 + \sigma_i^2$. Observe that the variance of $\bar{Y_{ij}}$ is maximum in the second case. Hence, if we would like to protect the hypothesis of the consistency (or the GLR-satisfactory) of the laboratories with respect to the assigned value, we can consider the second case.

Acknowledgements

This research was supported by grants from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We would like to thank the REMESP and the six laboratories for kindly providing the actual data set from a PT program and also information about the PT program.

Appendix A

Proof of the *Theorem 4.1*:

First we show that

$$\lambda = exp\left\{-\frac{1}{2}\left[(C\hat{\boldsymbol{\alpha}} - \mathbf{a})^T(B^TJ^T\Sigma^{-1}JB)(C\hat{\boldsymbol{\alpha}} - \mathbf{a})\right]\right\},$$

with $B = (J^T\Sigma^{-1}J)^{-1}C^T\left(C(J^T\Sigma^{-1}J)^{-1}C^T\right)^{-1}$.

Considering the denominator of the equation (7), it is easy to see that

$$\sup_{\boldsymbol{\alpha}\in\mathbb{R}^k} f(\boldsymbol{y}\ ,\ \boldsymbol{\alpha}) = f(\boldsymbol{y}\ ,\ \hat{\boldsymbol{\alpha}}) = \frac{1}{(2\pi)^{\frac{p}{2}}}\frac{1}{\sqrt{|\Sigma|}}\ exp\left\{-\frac{1}{2}(\boldsymbol{y}\ -\ J\bar{\boldsymbol{y}})^T\Sigma^{-1}(\boldsymbol{y}\ -\ J\bar{\boldsymbol{y}})\right\}.$$

In order to evaluate the numerator of the likelihood ratio, we consider the Lagrange technique to solve the corresponding maximization problem, that is

$$\sup_{\boldsymbol{\alpha} \in Q_0} f(\boldsymbol{y}\ ,\ \boldsymbol{\alpha}).$$

Defining,

$$G = f(\boldsymbol{y}\ ,\ \boldsymbol{\alpha}) - \boldsymbol{\psi}^T (C\boldsymbol{\alpha} - \mathbf{a}),$$

where $\boldsymbol{\psi}$ is a vector $k \times 1$ which corresponds to the Lagrange multipliers, then

$$\frac{\partial G}{\partial \boldsymbol{\alpha}} = f(\boldsymbol{y}\ ,\ \boldsymbol{\alpha}) \left(J^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_x^*) - J^T \Sigma^{-1} J \boldsymbol{\alpha} \right) - C^T \boldsymbol{\psi} \tag{8}$$

and

$$\frac{\partial G}{\partial \boldsymbol{\psi}} = -C\boldsymbol{\alpha} + \mathbf{a}. \tag{9}$$

Equating (8) and (9) to zero we obtain

$$J^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_x^*) - J^T \Sigma^{-1} J \boldsymbol{\alpha}_0 = \frac{C^T \boldsymbol{\psi}_0}{f(\boldsymbol{y}\ ,\ \boldsymbol{\alpha}_0)}$$

and

$$C\boldsymbol{\alpha}_0 = \mathbf{a},$$

where $\boldsymbol{\alpha}_0$ and $\boldsymbol{\psi}_0$ correspond to the maximum of $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$, respectively.

If we take $\boldsymbol{\psi}^* = \frac{\boldsymbol{\psi}_0}{f(\boldsymbol{y}\ ,\ \boldsymbol{\alpha}_0)}$, it follows that

$$-J^T \Sigma^{-1} J \boldsymbol{\alpha}_0 = C^T \boldsymbol{\psi}^* - J^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_x^*)$$

and

$$\boldsymbol{\alpha}_0 = \Gamma(\boldsymbol{y} - \boldsymbol{\mu}_x^*) - (J^T \Sigma^{-1} J)^{-1} C^T \boldsymbol{\psi}^* = \hat{\boldsymbol{\alpha}} - (J^T \Sigma^{-1} J)^{-1} C^T \boldsymbol{\psi}^*. \tag{10}$$

Hence, we conclude that

$$C\boldsymbol{\alpha}_0 = C\left(\hat{\boldsymbol{\alpha}} - (J^T \Sigma^{-1} J)^{-1} C^T \boldsymbol{\psi}^* \right) = C\hat{\boldsymbol{\alpha}} - C(J^T \Sigma^{-1} J)^{-1} C^T \boldsymbol{\psi}^* = \mathbf{a}$$

and

$$\boldsymbol{\psi}^* = \left[C(J^T\Sigma^{-1}J)^{-1}C^T\right]^{-1}(C\hat{\boldsymbol{\alpha}} - \mathbf{a}). \tag{11}$$

Thus, it follows from Equations (10) and (11), that

$$\boldsymbol{\alpha}_0 = \hat{\boldsymbol{\alpha}} - (J^T\Sigma^{-1}J)^{-1}C^T\left(C(J^T\Sigma^{-1}J)^{-1}C^T\right)^{-1}(C\hat{\boldsymbol{\alpha}} - \mathbf{a}) \;=\; \hat{\boldsymbol{\alpha}} - B\,(C\hat{\boldsymbol{\alpha}} - \mathbf{a}),$$

where $B = (J^T\Sigma^{-1}J)^{-1}C^T\left(C(J^T\Sigma^{-1}J)^{-1}C^T\right)^{-1}$.

Then, we obtain that

$$f(\boldsymbol{y};\boldsymbol{\alpha}_0) \;=\; \frac{1}{(2\pi)^{\frac{p}{2}}}\frac{1}{\sqrt{|\Sigma|}}\,exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - J\bar{\boldsymbol{y}} + JB\,(C\hat{\boldsymbol{\alpha}} - \mathbf{a})\right)^T\Sigma^{-1}\left(\boldsymbol{y} - J\bar{\boldsymbol{y}} + JB\,(C\hat{\boldsymbol{\alpha}} - \mathbf{a})\right)\right\}.$$

Finally, the generalized likelihood ratio is given by

$$\lambda = \frac{f(\boldsymbol{y};\ \hat{\boldsymbol{\alpha}})}{f(\boldsymbol{y};\ \boldsymbol{\alpha}_0)} \;=$$

$$exp\left\{-\frac{1}{2}\left[2(\boldsymbol{y} - J\bar{\boldsymbol{y}})^T\Sigma^{-1}JB\,(C\hat{\boldsymbol{\alpha}} - \mathbf{a}) + (C\hat{\boldsymbol{\alpha}} - \mathbf{a})^T(B^TJ^T\Sigma^{-1}JB)(C\hat{\boldsymbol{\alpha}} - \mathbf{a})\right]\right\}$$

$$= exp\left\{-\frac{1}{2}\left[(C\hat{\boldsymbol{\alpha}} - \mathbf{a})^T(B^TJ^T\Sigma^{-1}JB)(C\hat{\boldsymbol{\alpha}} - \mathbf{a})\right]\right\},$$

as $(\boldsymbol{y} - J\bar{\boldsymbol{y}})^T\Sigma^{-1}JB\,(C\hat{\boldsymbol{\alpha}} - \mathbf{a}) = 0$. Notice that $\boldsymbol{\alpha}_0$ is the MLE under the reduced model ($C\boldsymbol{\alpha} = 0$).

Next, we determine the exact distribution of $W = -2ln(\lambda)$.

Since $\hat{\boldsymbol{\alpha}}$ has k-multivariate normal distribution with mean vector $\boldsymbol{\alpha}$ and covariance matrix $\Gamma\Sigma\Gamma^T$, it follows that $C\hat{\boldsymbol{\alpha}}$ has m-multivariate normal distribution with mean vector $C\boldsymbol{\alpha}$ and covariance matrix $C\Gamma\Sigma\Gamma^TC^T$.

Thus, it follows from Graybill (1976, Theorem 4.4.2, pp. 135) that W has noncentral chi-square distribution with m degrees of freedom and noncentrality parameter s, if and only, if

$$H = \left[B^T J^T \Sigma^{-1} J B\right] \left[C \Gamma \Sigma \Gamma^T C^T\right]$$

is idempotent. After some algebraic manipulations, we obtain $H = I$, where I is the identity matrix. Hence, H is idempotent.

Notice that under H_0, $C\boldsymbol{\alpha} = \mathbf{a}$, W has central chi-square distribution with τ degrees of freedom. □

References

[1] ISO/IEC Guide 43-1 (1997) - Proficiency testing by interlaboratory comparisons - Part I: Development and operation of proficiency testing schemes.

[2] ISO GUM (1995), *Guide to the Expression of Uncertainty in Measurement*, BIPM, IEC, IFCC, IUPAC, IUPAP, OIML, First Edition.

[3] C. L. Cheng and J. W. Van Ness (1999), *Statistical Regression With Measurement Error*, Kendall's Library of Statistic 6, Arnold.

[4] W. J. Dixon (1953) - Processing Data for Outliers, Biometrics, BIOMA, March, No. 1, pp. 74-89.

[5] W. A. Fuller (1987) - it Measurement Error Model, Wiley, New York.

[6] F. E. Grubbs (1969) - Procedures for Detecting Outlying Observations in Samples, Technometrics, TCTMTA, **11**, No. 4, pp. 1-21.

[7] M. J. R. Healy (1979) - Outliers in Clinical Chemistry Quality Control Schemes, Clinical Chemistry, **25**, No. 5, pp. 675-677.

[8] J. L. Jaech (1985) *Statistical Analysis of Measurement Error*, Exxon Monograph, John Wiley and Sons.

[9] M. G. Kendall and A.Stuart (1947), *The Advanced Theory of Statistics*, Charles Griffin & Company Limited, London, 3 edition, vol. 1.

[10] H. Richardson, D. Wood, J. Whitby, R. Lannigan, C. Fleming (1996) - Quality Improvment of Diagnostic Microbiology Through a Peer-group Proficiency Assessment Program: a 20 year Experience in Ontario, Archives of Pathology and Lasboratory Medicine, **120**, pp. 445-455.

[11] D. M. Rocke (1983) - Robust Statistical Analysis of interlaboratory Studies, Biometrika, **70**, No. 2, pp. 421-431.

[12] C. Y. Tan and B. Iglewics (1999) - Measurement-Methods Comparisons and Linear Statistical Relationship, Technometrics, **41**, No. 3, pp. 192-201.

[13] D. W. Tholen (1993) - Reference Values and Participants Means as Target in Proficiency Testing, Archives of Pathology and Lasboratory Medicine, **117**, pp. 885-889.

[14] M. Thompson (1989) - Robust Statistics and Functional Relationship Estimation for Comparing the Bias of Analytical over Extending Concentration Ranges, Analytical Chemistry, **61**, pp. 1942-1945.

NOTAS DO ICMC

SÉRIE ESTATÍSTICA

072/2002 ANDRADE, MARINHO G.; FERREIRA, VALÉRIA A. MARTINS; OLIVEIRA, SANDRA C. – Abordagem bayesiana com MCMC versus máxima verossimilhança para modelos ARCH(p)

071/2002 LOUZADA-NETO, F.; TOMAZELLA, V.L.D.; ANDRADE, M.G. – Bayesian analysis for reccurent lifetime data with a non homogeneous Poisson process with a frailty term.

070/2002 LEÃO JR., D.; FRAGOS, M. D.; VAL, J. B. R. DO; ANDRADE, M. G. – Separable Hansdorff measurable Radon spaces

069/2002 AOKI, R.; ACHCAR, J.A.; BOLFARINR, H.; SINGER, J.M. – Bayesian analysis of paired data in mixture of normal distributions.

068/2001 LOIBEL, S.; VAL, J.B.R.; ANDRADE, M.G. – Comparação entre as abordagens
Bayesiana e clássica do problema de inferência com modelo de crescimento populacional de Richards.

067/2001 ACHCAR, J.A; BARBOSA, V.F. – Análise bayesiana para dados de sobrevivência com funções de risco em forma de U.

066/2001 MAZUCHELI, J.; LOUZADA-NETO, F.; ACHCAR, J.A. – A survival model with a surviving fraction.

065/2001 MAZUCHELI, J.; LOUZADA-NETO, F.; ACHCAR, J.A. – Infant mortality model for lifetime data.

064/2001 SOUZA, C.N.; ACHCAR, J.A.; MAZUCHELI, J. – Uso de métodos MCMC para análise bayesiana de dados de sobrevivência na presença de covariáveis.

063/2001 ACHCAR J.A.; JUNQUEIRA, J.J.G. – Extra-biomial variability: a bayesian approach.