October 4, 2010

Assessing differences between results determined according to the Guide to the Expression of Uncertainty in Measurement

Raghu N Kacker[1], Rüdiger Kessel[1], Klaus-Dieter Sommer[2]
[1]National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA
[2] Physikalisch-Technische Bundesanstalt, D-38116 Braunschweig, Germany

Abstract

When the data consist of multiple results of measurement for a common measurand, often one needs to determine whether the results agree with each other. A result of measurement based on the Guide to the Expression of Uncertainty in Measurement (GUM) consists of a measured value together with its associated standard uncertainty. In the GUM, the measured value is regarded as the expected value and the standard uncertainty is regarded as the standard deviation, both known values, of a state-of-knowledge probability distribution. A state-of-knowledge distribution represented by a result is not required to be completely known. Then how can one assess the differences between the results based on the GUM? Metrologists have for many years used the Birge chi-square test as 'a rule of thumb' to assess the differences between two or more measured values for the same measurand by pretending that the standard uncertainties were the standard deviations of the presumed sampling probability distributions from random variation of the measured values. We point out that this is misuse of the standard uncertainties; the Birge test and the concept of statistical consistency motivated by it do not apply to the results of measurement based on the GUM. In 2008, the International Vocabulary of Metrology, third edition (VIM3) introduced the concept of metrological compatibility. We show that the concept of metrological compatibility can be used to assess the differences between results based on the GUM for the same measurand. A pairwise Birge test of statistical consistency and a test of metrological compatibility do not conflict.

Key words: Birge test; Interlaboratory evaluations; Predictive *p*-value; Uncertainty

1. Introduction

To test the proficiency of individual laboratories in conducting specific tasks, interlaboratory comparisons (ILC) are often used. In ILC between measurement laboratories, the task is generally the measurement of a common artifact or fractions of the same sample of material. To develop a certified reference material, a well characterized material is measured by two or more methods in one or more laboratories. In both cases the data consist of multiple evaluations of a common measurand. To assess the differences between two or more measured values for the same measurand, metrologists have for many years used a test proposed by physicist Raymond T. Birge in 1932 [1]. Birge introduced the term consistency for lack of significant differences between measured values. The Birge test is also referred to as the Birge ratio test or Birge chi-square test or simply chi-square test. The Birge test is based on treating the measured values as realizations of

random draws from sampling probability density functions (pdfs). A sampling pdf models possible outcomes for measured values in contemplated replications of the measurement procedure in the same conditions. Therefore, the consistency of measured values assessed by the Birge test is statistical consistency. The Birge test applies to uncorrelated measured values only. In section 2, we review a concept of statistical consistency motivated by the Birge test. The idea of statistical consistency belongs to the period when the error analysis view of measurements was prevalent. The error analysis view of measurements was a hindrance to communicating the results of measurement and in advancing the science and technology of measurement. Therefore leading authorities in the field of metrology developed the Guide to the Expression of Uncertainty in Measurement (GUM) [2]. According to the GUM, a result of measurement consists of a measured value together with its associated standard uncertainty. In the GUM, the measured value is regarded as the expected value and the standard uncertainty is regarded as the standard deviation, both known values, of a state-of-knowledge probability distribution which is not required to be completely determined. We note in section 3 that the Birge test and the concept of statistical consistency are not applicable to the results of measurement based on the GUM. Then how can one assess the differences between the results based on the GUM for the same measurand? In 2008, the International Vocabulary of Metrology, third edition (VIM3) [3] introduced the concept of metrological compatibility of two or more results of measurement determined according to the (GUM). In section 4, we review the VIM3 concept of metrological compatibility and show that this concept can be used to assess the differences between multiple results based on the GUM for the same measurand. In section 5, we show that a pairwise Birge test of statistical consistency and a test of metrological compatibility do not conflict.

2. The Birge test and concept of statistical consistency

Suppose $x_1, \ldots, x_n$ are n measured values for a common measurand which is believed to be sufficiently stable. The Birge test is based on regarding the measured values $x_1, \ldots, x_n$ as realizations of random draws from their presumed sampling probability density functions (pdfs). A sampling pdf models possible outcomes in contemplated replications of a measurement procedure subject to random effects in the same conditions. Therefore, the consistency (lack of significant differences between measured values) assessed by the Birge test is statistical consistency. The Birge test is applicable when the sampling pdfs of the measured values $x_1, \ldots, x_n$ are uncorrelated. The Birge test requires knowledge of the variances $\sigma_1^2, \ldots, \sigma_n^2$ of the sampling pdfs of $x_1, \ldots, x_n$, respectively. Statistical consistency of the measured values $x_1, \ldots, x_n$ means that their expected values are indistinguishable[1] in view of the corresponding variances. Specifically, the Birge test checks whether the measured values $x_1, \ldots, x_n$ may be modeled as realizations from normal (Gaussian) sampling pdfs with unknown but equal expected values and known variances $\sigma_1^2, \ldots, \sigma_n^2$. Birge proposed that to check the consistency of the measured values $x_1, \ldots, x_n$, one can calculate the test statistic

[1] In statistical literature the term consistency is applied to a statistical estimator. A point statistical estimator is said to be consistent if it approaches the parameter being estimated as the sample size increases.

$$R^2 = \sum_{i=1}^{n} w_i (x_i - x_W)^2 / (n-1) , \quad (1)$$

where $w_i = 1/\sigma_i^2$, for $i = 1, 2, \ldots, n$, and $x_W = \sum_i w_i x_i / \sum_i w_i$ is the weighted mean of $x_1, \ldots, x_n$. If the calculated value of R^2 is substantially larger than one, then the dispersion of $x_1, \ldots, x_n$ is larger than what can be expected from the normal pdfs with equal expected values and variances $\sigma_1^2, \ldots, \sigma_n^2$. In that case the measured values $x_1, \ldots, x_n$ can be declared to be statistically inconsistent.

Statistical interpretation of the Birge test: Birge [1] was a physicist and he proposed his test independently of and before much of the statistical theory as it is known today was established. However, the Birge test of consistency can now be interpreted as a classical (sampling theory) statistical test of hypothesis. The measured values $x_1, \ldots, x_n$ are presumed to have normal sampling pdfs with unknown but equal expected values and variance-covariance matrix $\tau^2 \times \text{Diag}[\sigma_1^2, \ldots, \sigma_n^2]$, where τ^2 is an unknown parameter and $\sigma_1^2, \ldots, \sigma_n^2$ are known. The null hypothesis H_0 is that $\tau^2 \leq 1$ and the alternative hypothesis H_1 is that $\tau^2 > 1$. The null hypothesis H_0 means that the variances of $x_1, \ldots, x_n$ are not greater than $\sigma_1^2, \ldots, \sigma_n^2$, respectively. The alternative hypothesis H_1 means that the variances of $x_1, \ldots, x_n$ are greater than $\sigma_1^2, \ldots, \sigma_n^2$ [4]. The classical *p*-value p_C is the maximum probability under the null hypothesis of realizing in contemplated replications of the *n* measurements a value of the test statistic more extreme than its realized (calculated) value. The classical *p*-value of a realization of $(n-1)\,R^2$ is

$$p_C = \Pr\{\chi^2_{(n-1)} \geq (n-1)R^2\} , \quad (2)$$

where $\chi^2_{(n-1)}$ denotes a variable with the chi-square probability distribution with degrees of freedom $(n-1)$ [4]. If the classical *p*-value p_C is too small, say, less than 0.05, then the null hypothesis is rejected with level of significance 0.05 or less. A rejection of the null hypothesis means that the dispersion of the measured values $x_1, \ldots, x_n$ is greater than what can be expected from normal distributions for $x_1, \ldots, x_n$ with equal expected values and variances $\sigma_1^2, \ldots, \sigma_n^2$, respectively. If the stated variances $\sigma_1^2, \ldots, \sigma_n^2$ are not doubtful then the assumption that the expected values of the pdfs of $x_1, \ldots, x_n$ are equal is questionable. In that case, the measured values $x_1, \ldots, x_n$ can be declared to be statistically inconsistent.

Limitations of the Birge test: A limitation of the Birge test is that it is applicable for uncorrelated measured values $x_1, \ldots, x_n$ only. However, it can be easily generalized to correlated measured values $x_1, \ldots, x_n$ whose covariances denoted by $\sigma_{12}, \ldots, \sigma_{(n-1)n}$ are known [4] . The Birge test suggests the following notion of the statistical consistency of the measured values $x_1, \ldots, x_n$ [4]: The measured values $\boldsymbol{x} = (x_1, \ldots, x_n)^t$ are said to be statistically consistent if their dispersion *is not greater than* what can be expected from the *normal consistency model* which postulates that the joint *n*-variate sampling pdf of $\boldsymbol{x}$ is normal $N(\mathbf{1}\mu, \boldsymbol{D})$ with unknown expected value $\mathbf{1}\mu$ and variance-covariance matrix $\boldsymbol{D} = [\sigma_{ij}]$, where $\mathbf{1} = (1, \ldots, 1)^t$, σ_{ij} is the covariance between x_i and x_j, and $\sigma_{ii} = \sigma_i^2$ for $i, j = 1, 2, \ldots, n$.

The Birge test and its generalized version for correlated measured values is a one sided test of hypothesis which checks whether the dispersion of $x_1, \ldots, x_n$ is more than what can be

expected from a normal consistency model. A review of the Birge test in [5] notes that if the realized value of the Birge test statistic R^2 is substantially less than one, then the stated variances $\sigma_1^2, \ldots, \sigma_n^2$ may well be too large. To avoid declarations of statistical consistency from overstated variances, the following definition of statistical consistency was proposed in [6].

Definition of statistical consistency: The measured values $\boldsymbol{x} = (x_1, \ldots, x_n)^t$ are said to be statistically consistent if they *reasonably fit* the normal consistency model which postulates that the joint n-variate sampling pdf of $\boldsymbol{x}$ is normal N($\mathbf{1}\mu$, $\boldsymbol{D}$) with unknown expected value $\mathbf{1}\mu$ and variance-covariance matrix $\boldsymbol{D} = [\sigma_{ij}]$.

This definition requires a different approach for testing statistical consistency than the Birge test and its generalized version for correlated values. A modern method to assess the fit of a statistical model to the data is Bayesian posterior predictive checking [6]. Posterior predictive checking is a Bayesian adaptation of the classical (sampling theory) statistical hypothesis testing. A function of the data (and possibly unknown parameters) called 'discrepancy measure' is defined to characterize a potential discrepancy between the statistical model and the data. The posterior predictive p-value p_P of a discrepancy measure $T(\boldsymbol{x})$ is the probability of realizing in contemplated replications a value of the discrepancy measure more extreme than its realized value. If the posterior predictive p-value is close to zero (or to one) then the fit of the statistical model to data is suspect.

If the measured values $x_1, \ldots, x_n$ were uncorrelated, then the statistic $T_c(\boldsymbol{x}) = (n-1)\, R^2 = \sum_i w_i (x_i - x_W)^2$ is a useful discrepancy measure to check the overall fit of the normal consistency model N($\mathbf{1}\mu$, $\boldsymbol{D}$) to the measured values $x_1, \ldots, x_n$. The posterior predictive p-value of the realized discrepancy measure $T_c(\boldsymbol{x}) = (n-1)\, R^2$ is

$$p_P = \Pr\{\chi^2_{(n-1)} \geq (n-1)R^2\}\,. \tag{3}$$

We note that (3) is identical to the classical p-value p_C given in (2). Thus Bayesian posterior predictive checking of the discrepancy measure $T_c(\boldsymbol{x}) = (n-1)\, R^2$ is equivalent to the Birge test of statistical consistency.

In Bayesian posterior predictive checking, one can investigate any number of potential discrepancies between the statistical model and the data. To assess the difference between two particular measured values x_i and x_j, the statistic $T_{i-j}(\boldsymbol{x}) = |\, x_i - x_j\,|$ is a useful discrepancy measure, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. The Bayesian posterior predictive p-value of the realized discrepancy measure $|\, x_i - x_j\,|$ is

$$p_P = \Pr\{Z \geq \frac{|\, x_i - x_j\,|}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}}\}\,, \tag{4}$$

where ρ_{ij} is the correlation coefficient between the presumed normal sampling pdfs of x_i and x_j; the covariance between x_i and x_j is $\sigma_{ij} = \rho_{ij}\, \sigma_i\, \sigma_j$, and Z denotes a variable with standard normal distribution $N(0, 1)$ [6, section 3.2]. A posterior predictive p-value p_P close to zero suggests that the difference between x_i and x_j is larger than what can be expected from the normal statistical consistency model N($\mathbf{1}\mu$, $\boldsymbol{D}$). That is, the measured values x_i and x_j do not

seem to have the same expected value and hence they are not mutually statistically consistent.

3. Concept of statistical consistency does not apply to results based on the GUM

A result of measurement determined according to the GUM consists of a measured value together with its associated standard uncertainty. Suppose $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$ are n results of measurement for a common measurand, where x_1, …, x_n are the measured values and $u(x_1)$, …, $u(x_n)$ are the corresponding standard uncertainties. According to the GUM, a measured value x_i and its associated standard uncertainty $u(x_i)$ represent a state-of-knowledge pdf attributed to the measurand, for $i = 1, 2, \ldots, n$. Following the GUM, we use the symbol X_i for a quantity as well as for a variable with a state-of-knowledge pdf represented by the result $[x_i, u(x_i)]$, for $i = 1, 2, \ldots, n$. The measured value x_i is regarded as the expected value $\mathrm{E}(X_i)$ and the standard uncertainty $u(x_i)$ is regarded as the standard deviation $\mathrm{S}(X_i)$ of the pdf of X_i, for $i = 1, 2, \ldots, n$. The mainstream GUM requires knowledge of only the expected value $\mathrm{E}(X_i)$ and the standard deviation $\mathrm{S}(X_i)$ of a state-of-knowledge pdf of X_i. The GUM does not require that the state-of-knowledge pdf of X_i be completely determined. When the state-of-knowledge pdfs of X_1, …, X_n are correlated, the correlation coefficients are assumed to be known. Following the GUM we denote the correlation coefficient $\mathrm{R}(X_i, X_j)$ between the state-of-knowledge pdfs of X_i and X_j by the symbol $r(x_i, x_j)$. Note that $\{x_1, \ldots, x_n\}$, $\{u(x_1), \ldots, u(x_n)\}$, and $\{r(x_1, x_2), \ldots, r(x_{(n-1)}, x_n)\}$ are symbols for known values.

For many years, metrologists have used the Birge test as ‘a rule of thumb’ to assess the consistency of the measured values by treating the squared standard uncertainties $u^2(x_1)$, …, $u^2(x_n)$ as the known variances σ_1^2, …, σ_n^2 of the presumed normal (Gaussian) sampling pdfs of the measured values x_1, …, x_n; see, for example [8]. The guideline for the analysis of key comparisons developed by the BIPM Director’s Advisory Group on Uncertainties recommends the use of Birge chi-square test to assess the consistency of measured values by treating the squared standard uncertainties as the known variances of the presumed sampling pdfs of the measured values [9]. The consistency of the measured values from CIPM key comparisons and supplementary comparisons is almost always assessed using the Birge test [10].

The squared standard uncertainties $u^2(x_1)$, …, $u^2(x_n)$ cannot in any logical sense be identified with the known variances σ_1^2, …, σ_n^2 of the presumed normal (Gaussian) sampling pdfs of the measured values x_1, …, x_n. The standard deviation of a sampling pdf represents possible dispersion from random variation in contemplated replications of the measurement procedures. A standard uncertainty expresses the dispersion of a state-of-knowledge pdf which could be attributed to the measurand based on all available statistical and non-statistical information. A standard uncertainty includes all significant components whether arising from random effects or from corrections applied for systematic effects. All available statistical and non-statistical information is used to evaluate a standard uncertainty. In measurements done in high echelon laboratories, the component of uncertainty arising from random effects is generally a very small part of the combined standard uncertainty. Treating the squared standard uncertainties $u^2(x_1)$, …, $u^2(x_n)$

determined according to the GUM as the known variances σ_1^2, …, σ_n^2 from random variation (in contemplated replications of the measurements) is a misuse of the standard uncertainties. Also, as noted earlier, the state-of-knowledge pdfs represented by the results $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$ may not be completely determined. Therefore the Birge test and the concept of statistical consistency do not apply to the results of measurement determined according to the GUM.

4. VIM3 concept of metrological compatibility

A measured quantity value [3, definitions 1.19 and 2.10] is a product of a numerical value and a measurement unit. The measurement unit implies that the measured value is traceable to a reference for that measurement unit. A result of measurement (measured value together with its associated standard uncertainty) is traceable to a reference only if the result can be related to a practical realization of the reference through a documented unbroken chain of calibrations each contributing to the measurement uncertainty [3, definition 2.41]. Two or more results of measurement are *metrologically comparable* only if they are traceable to the same reference [3, definition 2.46]. Metrological comparability does not imply that the measured values have similar magnitudes. Thus, for example, distance between my apartment and my office expressed in meters is metrologically comparable to the distance between my apartment and the moon also expressed in meters. The concept of metrological compatibility discussed in the next section applies only to those results of measurement for a common measurand which are metrologically comparable. That is, the results must be traceable to the same reference.

The concept of statistical consistency can be applied to any set of numerical values which have similar magnitudes. They do not have to be measured values. Thus, for example, one can test statistical consistency of deviations (or relative deviations expressed as percentage) from a benchmark value. Although a metrologist is expected to assess consistency of only those measured values which have the same measurement unit, it is not a requirement of statistical consistency.

All n results $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$ for a common measurand must be traceable to the same reference for them to be metrologically comparable [3, definition 2.46]. The VIM3 concept of metrological compatibility is defined for two results of measurement at a time. The following definition is derived from [3, definition 2.47].

Definition of metrological compatibility: Two metrologically comparable results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$ for the same measurand are said be metrologically compatible if

$$\zeta(x_1 - x_2) = \frac{|\, x_1 - x_2 \,|}{\sqrt{u^2(x_1) + u^2(x_2) - 2r(x_1, x_2)u(x_1)u(x_2)}} \le \kappa, \qquad (5)$$

for a specified threshold κ, where $r(x_1, x_2)$ denotes the correlation coefficient $\mathrm{R}(X_1, X_2)$ between the variables X_1 and X_2. The quantity in the denominator of (5) is the standard deviation of the state-of-knowledge pdf for $X_1 - X_2$, which may be incompletely determined. When the pdfs represented by $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$ are uncorrelated, (5) reduces to

$$\zeta(x_1 - x_2) = \frac{|x_1 - x_2|}{\sqrt{u^2(x_1) + u^2(x_2)}} \le \kappa. \quad (6)$$

A set of metrologically comparable results $[x_1, u(x_1)]$, $[x_2, u(x_2)]$, …, $[x_n, u(x_n)]$ for the same measurand is said be metrologically compatible if for every one of the $n(n-1)/2$ pairs of results $[x_i, u(x_i)]$ and $[x_j, u(x_j)]$ we have

$$\zeta(x_i - x_j) = \frac{|x_i - x_j|}{\sqrt{u^2(x_i) + u^2(x_i) - 2r(x_i, x_j)u(x_i)u(x_j)}} \le \kappa, \quad (7)$$

for a specified threshold κ [3, definition 2.47]. The VIM3 does not discuss how the threshold κ should be determined. A conventional value of κ is two.

The concept of metrological compatibility can be used to assess the differences between the results of measurement based on the GUM for the same measurand. The concepts of metrological comparability and compatibility do not require that the state-of-knowledge pdfs represented by the results $[x_1, u(x_1)]$, $[x_2, u(x_2)]$, …, $[x_n, u(x_n)]$ be completely known. Thus they fit the GUM. When the set of results $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$ is metrologically compatible, we can say that the differences between the measured values x_1, …, x_n are insignificant in view of the uncertainties $u(x_1)$, …, $u(x_n)$.

To assess metrological compatibility of results based on GUM using the criteria (5), (6), or (7), the threshold κ needs to be specified. A proper choice of κ is to a large extent a matter of agreement because it requires accepting the economic consequences of that choice. Although a conventional value of κ is two, depending on the application, the interested parties could agree on a different value for κ. Once the value of the threshold κ is set the conclusion of a test of metrological compatibility based on the VIM3 definition is dichotomous, either a set of results is metrologically compatible or incompatible. The concept of metrological compatibility is being used by metrologists who are familiar with it; see for example [11, 12].

The VIM3 definition of metrological compatibility can be easily extended to metrological compatibility of a set of results and a reference result $[x_R, u(x_R)]$, where x_R is the reference value with standard uncertainty $u(x_R)$. Suppose the pdfs represented by the measurement results are uncorrelated with the pdf represented by the reference result. A set of metrologically comparable results $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$ is metrologically compatible with a reference result $[x_R, u(x_R)]$ if

$$\zeta(x_i - x_R) = \frac{|x_i - x_R|}{\sqrt{u^2(x_i) + u^2(x_R)}} \le \kappa, \quad (8)$$

for $i = 1, 2, \ldots, n$ [13]. Similarly a set of metrologically comparable results $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$ is compatible with a combined result $[x_C, u(x_C)]$, where x_C is the combined value (such as arithmetic mean or a weighted mean) with standard uncertainty $u(x_C)$ if

$$\zeta(x_i - x_C) = \frac{|\, x_i - x_C \,|}{\sqrt{u^2(x_i) + u^2(x_C) - 2r(x_i, x_C)u(x_i)u(x_C)}} \le \kappa, \qquad (9)$$

where $r(x_i, x_C)$ denotes the correlation coefficient between the pdfs represented by $[x_i, u(x_i)]$ and $[x_C, u(x_C)]$, for $i = 1, 2, \ldots, n$ [13].

5. Concluding remarks

For many years, metrologists have used the Birge chi-square test as 'a rule of thumb' to assess the differences between two or more measured values for the same measurand by pretending that the squared standard uncertainties were the known variances of the presumed normal sampling pdfs of the measured values. This is misuse of the standard uncertainties based on the GUM. The Birge test and the concept of statistical consistency do not apply to the results of measurement based on the GUM. As discussed in this paper, the VIM3 concept of metrological compatibility can be used to assess the differences between the results of measurement determined according to the GUM. Thus metrologists can start using the VIM3 concept of metrological compatibility in place of the Birge test to assess the differences between multiple evaluations of the same measurand.

The following is a pertinent question. Could the conclusions (about mutual agreement of results) based on the VIM3 concept of metrological compatibility and the Birge test differ? It is difficult to directly compare the Birge test and a test of metrological compatibility because the former is defined for an arbitrary positive integer $n > 1$ and the latter is defined for only two results at a time. For pairwise comparisons ($n = 2$), the Birge test statistic $R^2 = \sum_i w_i (x_i - x_W)^2 / (n - 1)$ reduces to

$$R^2 = \frac{(x_1 - x_2)^2}{(\sigma_1^{\,2} + \sigma_2^{\,2})}, \qquad (10)$$

which is square of $(x_1 - x_2)/\sqrt{(\sigma_1^{\,2} + \sigma_2^{\,2})}$. Under the null hypothesis that the presumed normal sampling pdfs of x_1 and x_2 have the same expected value, the distribution of $(x_1 - x_2)/\sqrt{(\sigma_1^{\,2} + \sigma_2^{\,2})}$ is normal N(0, 1). Therefore when $n = 2$, the normal distribution can be used to assess the absolute difference $|\, x_1 - x_2 \,|$. The square of a normal N(0, 1) variable has a chi-square distribution $\chi^2_{(1)}$ with degrees of freedom 1. Therefore the square of the $(1 - \alpha/2)\times 100$-th percentile $z_{[1 - \alpha/2]}$ of normal N(0, 1) distribution is equal to the $(1 - \alpha)\times 100$-th percentile $\chi^2_{(1)}[1 - \alpha]$ of $\chi^2_{(1)}$ distribution. Thus the realized value of (10) being less than $\chi^2_{(1)}[1 - \alpha]$ is equivalent to the ratio $|\, x_1 - x_2 \,|/\sqrt{(\sigma_1^{\,2} + \sigma_2^{\,2})}$ being less than $z_{[1 - \alpha/2]}$. It follows that declaration of Birge statistical consistency when the classical p-value p_C of the Birge test (2) is less than 0.05 (for example) is equivalent to the realization that

$$\frac{|\, x_1 - x_2 \,|}{\sqrt{(\sigma_1^{\,2} + \sigma_2^{\,2})}} \le z_{[0.975]} = 1.96 \approx 2. \qquad (11)$$

We note from (6) and (11) that if the threshold κ for metrological compatibility is set as $\kappa = 2$ then the conclusion of a check of metrological compatibility between a pair of results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$ would be identical to the assessment of statistical consistency

between x_1 and x_2 based on the Birge test by (wrongly) treating $u^2(x_1)$ and $u^2(x_2)$ as σ_1^2 and σ_2^2, respectively (and treating the correlation coefficient $R(X_1, X_2)$ as ρ_{12} which is zero in the Birge test). Therefore a pairwise Birge test of statistical consistency and a test of metrological compatibility do not conflict.

Acknowledgment: We thank Javier Bernal and Tyler Estler for their comments on an earlier draft of this paper.

References

[1] Birge R T 1932 The calculation of errors by the method of least squares, *Physical Review* **40** 207-227
[2] GUM 1995 *Guide to the Expression of Uncertainty in Measurement* 2nd ed (Geneva: International Organization for Standardization) ISBN 92-67-10188-9 (2008 version available at http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf)
[3] BIPM/JCGM 2008 *International Vocabulary of Metrology* – Basic and general concepts and associated terms 3rd ed (Sèvres: Bureau International des Poids et Mesures, Joint Committee for Guides in Metrology) (available at http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf)
[4] Kacker R N, Forbes A B, Kessel R and Sommer K 2008 Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations *Metrologia* **45** 257-264
[5] Taylor B N, Parker W H and Langenberg D N 1969 Determination of *e*/*h*, Using Macroscopic Quantum Phase Coherence in Superconductors: Implications for Quantum Electrodynamics and the Fundamental Physical Constants, *Review of Modern Physics* **41** 375-496
[6] Kacker R N, Forbes A B, Kessel R and Sommer K 2008 Bayesian posterior predictive p-value of statistical consistency in interlaboratory evaluations *Metrologia* **45** 512-523
[7] Gelman A, Carlin J B, Stern H S and Rubin D B 2004 *Bayesian Data Analysis*, 2nd ed, Chapman & Hall
[8] Mohr P J and Taylor B N 2000 CODATA recommended values of the fundamental physical constants: 1998 *Reviews of Modern Physics* **72** 351-495 (current version available at http://physics.nist.gov/cuu/Constants/index.html)
[9] Cox M G 2002 The evaluation of key comparison data *Metrologia* **39** 589-595 (these guidelines were developed by the BIPM Director's Advisory Group on Uncertainties)
[10] The *BIPM key comparison database* 2010 http://kcdb.bipm.org/
[11] Wellum R, Verbruggen A and Kessel R 2009 A new evaluation of the half-life of ^{241}Pu *Journal of Analytical Atomic Spectrometry* **24** 801-807
[12] Datla R U, Kessel R, Smith A W, Kacker R N and Pollock D B 2010 Uncertainty analysis of remote sensing optical sensor data: guiding principles to achieve metrological consistency *International Journal of Remote Sensing* **31** 867-880
[13] Kessel R, Kacker R N, and Sommer K 2009 Proposal for combining results from multiple evaluations of the same measurand, submitted for publication

About the Authors:

Raghu N Kacker is a mathematical statistician in the Information Technology Laboratory of the National Institute of Standards and Technology, Gaithersburg, MD 20899, USA.

Rüdiger Kessel is a guest researcher in the Information Technology Laboratory of the National Institute of Standards and Technology, Gaithersburg, MD 20899, USA.

Klaus-Dieter Sommer is director of the Chemical Physics and Explosion Protection Division of the of the National Metrology Institute of Germany, Physikalisch-Technische Bundesanstalt, D-38116 Braunschweig, Germany.