XIX IMEKO World Congress
Fundamental and Applied Metrology
September 6–11, 2009, Lisbon, Portugal

COMPARISON OF STATISTICAL CONSISTENCY AND METROLOGICAL CONSISTENCY

Raghu N Kacker[1], *Ruediger Kessel*[1], *Klaus-Dieter Sommer*[2], *Xin Bian*[3]

[1]National Institute of Standards and Technology, Gaithersburg, USA, raghu.kacker@nist.gov, ruediger.kessel@nist.gov
[2] Physikalisch-Technische Bundesanstalt, Braunschweig, Germany, klaus-dieter.sommer@ptb.de
[3]National Institute of Metrology, Beijing, China, bianx@nim.ac.cn

Abstract – The conventional concept of consistency in multiple evaluations of the same measurand is based on statistical error analysis. This concept is based on regarding the evaluations as realizations from sampling probability distributions of potential evaluations which might be obtained in contemplated replications. The expected values of the sampling distributions are regarded as unknown but the standard deviations are assumed to be known. The multiple evaluations are said to be statistically consistent if their dispersion agrees with the hypothesis that the sampling distributions of potential evaluations have equal expected values. As the science and technology of measurement advanced, the limitations of the statistical error analysis view of uncertainty in measurement became a hindrance to communication of scientific and technical measurements. Therefore, a new concept of uncertainty in measurement was established by the Guide to the Expression of Uncertainty in Measurement (GUM). In the GUM view, an evaluation and uncertainty are, respectively, measures of centrality and dispersion of a state-of-knowledge probability distribution for the measurand. Statistical consistency is not compatible with the GUM concept of uncertainty in measurement; however, metrologists continue to use it as an approximate rule of thumb because no suitable alternative has been available until recently. The concept of metrological consistency is compatible with the GUM concept of uncertainty in measurement. It is a pair-wise concept. A pair of state-of-knowledge distributions are said to be metrologically consistent if the ratio of the absolute difference between evaluations and the standard uncertainty of the difference is less than some chosen benchmark. As the concept of metrological consistency becomes more widely known and its benefits realized, it should become the dominant approach to test consistency of multiple evaluations of the same measurand.

Keywords: Birge test, Interlaboratory evaluations, metrological consistency, statistical consistency

1. STATISTICAL CONSISTENCY

The Birge test is the most popular method to assess consistency of multiple measured values for the same measurand [1]. It is based on statistical error analysis and it led to the concept of statistical consistency of multiple measured values for the same measurand. Suppose n different results of measurement $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$ for a common reference are available, where $x_1, \ldots, x_n$ are the measured values and $u(x_1), \ldots, u(x_n)$ are the associated standard uncertainties. In the Birge test the measured values $x_1, \ldots, x_n$ are regarded as realizations (random draws) from sampling probability density functions (pdfs) which are assumed to be normal with known variances. To apply the Birge test, the squared standard uncertainties $u^2(x_1), \ldots, u^2(x_n)$ are (wrongly) regarded as the known variances of the sampling pdfs of $x_1, \ldots, x_n$. The Birge test is applicable when the measured values $x_1, \ldots, x_n$ are uncorrelated random variables. Birge [1] proposed that to check the statistical consistency of $x_1, \ldots, x_n$, calculate the test statistic

$$R^2 = \sum_{i=1}^{n} w_i (x_i - x_{\mathrm{W}})^2 / (n-1), \qquad (1)$$

where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$, and $x_{\mathrm{W}} = \sum_i w_i x_i / \sum_i w_i$ is the weighted mean of $x_1, \ldots, x_n$. If the calculated value of R^2 is substantially larger than one, then declare the measured values $x_1, \ldots, x_n$ to be inconsistent.

The Birge test of consistency can be interpreted as a classical test of the null hypothesis H_0 that the variances of the presumed normal (Gaussian) sampling pdfs of the results $x_1, \ldots, x_n$ are less than or equal to $u^2(x_1), \ldots, u^2(x_n)$ against the alternative hypothesis H_1 that the variances of the normal sampling pdfs of $x_1, \ldots, x_n$ are greater than $u^2(x_1), \ldots, u^2(x_n)$. 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 value. The classical p-value of a realization of $(n-1)\,R^2$ is

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

where $\chi^2_{(n-1)}$ denotes a variable with the chi-square probability distribution with degrees of freedom $(n-1)$. If the classical p-value is too small, say less than 0.05, then the null hypothesis is rejected and the measured values $x_1, \ldots, x_n$ are declared to be inconsistent.

The Birge test can be generalized to test the consistency of measured values $x_1, \ldots, x_n$ whose covariances $u(x_1, x_2), \ldots, u(x_{n-1}, x_n)$ are known. The Birge test led to the following view of statistical consistency [2] 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 expected value $\mathbf{1}\mu$ and variance-covariance matrix $\boldsymbol{D} = [u(x_i, x_j)]$, where $\mathbf{1} = (1, \ldots, 1)^t$ and $u(x_i, x_i) = u^2(x_i)$ for $i = 1, 2, \ldots, n$.

A review of the Birge test in [3] notes that if the realized value of the Birge test statistic is substantially less than one, then the stated variances $u^2(x_1), \ldots, u^2(x_n)$ may well be too large. To alert against pronouncements of statistically consistency arising from excessively overstating the variances, the following definition of statistical consistency was proposed in [4].

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 expected value $\mathbf{1}\mu$ and variance-covariance matrix $\boldsymbol{D} = [u(x_i, x_j)]$.

2. METROLOGICAL CONSISTENCY

The world's leading metrologists developed the concept of uncertainty in measurement described in the Guide to the Expression of Uncertainty in Measurement (GUM) [5]. The third edition of the International Vocabulary of Metrology (VIM3) [6] further elaborates the GUM concept of uncertainty in measurement. According to the GUM and VIM3, a result of measurement consists of a measured value and its associated standard uncertainty. The measured value is regarded as the expected value and the standard uncertainty is regarded as the standard deviation of a state-of-knowledge probability density function (pdf) attributed to the unknown measurand. Generally, the pdf attributed to the measurand is incompletely determined. The statistical view of consistency does not match the GUM view of uncertainty in measurement and it does not apply to the results of measurement expressed as measured values with standard uncertainties. Therefore the VIM3 introduced the concept of metrological compatibility of multiple results of measurement for the same measurand. We use the term metrological consistency for the VIM3 concept of metrological compatibility. Two or more results of measurement are metrologically comparable if they are traceable to the same reference [6] The concept of metrological consistency (compatibility) applies to only those results which are metrologically comparable. Metrological consistency is a pair-wise concept; that is, it applies to only two results at a time.

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

$$\zeta(x_1 - x_2) = \frac{|x_1 - x_2|}{u(x_1 - x_2)} \le k, \qquad (3)$$

for a chosen value of k, where $u(x_1 - x_2) = (u^2(x_1) + u^2(x_2) - 2r(x_1, x_2)u(x_1)u(x_2))^{1/2}$ and $r(x_1, x_2)$ is the correlation coefficient between the random variables represented by the results [6]. The value used for k is often set as two. When the results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$ are metrologically consistent, we can say that the measured values x_1 and x_2 agree with each other in view of the stated standard uncertainties $u(x_1)$ and $u(x_2)$. That is, the difference between x_1 and x_2 is not significant. If the measurement procedures are credible and the uncertainties are properly determined then two results for the same measurand should be consistent. When more than two results for the same measurand are available, one compares them one pair at a time.

3. COMPARISON

The major differences between statistical consistency and metrological consistency are as follows: (i) Concept of statistical consistency does not match the GUM concept of uncertainty. (ii) Statistical consistency does not apply to the results of measurement expressed as measured values with associated standard uncertainties. (iii) Statistical consistency does not require that the measured values be evaluations for the same measurand. Metrological consistency applies only to evaluations for the same measurand which are traceable to the same reference. (iv) The default assumption in statistical consistency is that the measured values are inconsistent. Credible results for the same measurand should be metrologically consistent unless something is wrong. (v) Metrological consistency is a pair-wise concept, while statistical consistency applies to any number of results. (vi) The theory of statistical consistency allows for some measured values to be outliers. In metrological consistency, outliers indicate problems with the measurement procedures or stated uncertainties.

4. CONCLUSION

The traditional view of consistency as used by metrologists is statistical. However, the statistical view of consistency does not match the concept of uncertainty in measurement established by the GUM. In particular, the statistical view of consistency does not apply to the results of measurement expressed as measured values with standard uncertainties. Therefore VIM3 introduced the concept of Metrological compatibility. We prefer and use the term metrological consistency for the VIM3 concept of metrological compatibility. The concept of metrological consistency matches the GUM view of uncertainty in measurement. The concept of metrological consistency is new and not yet very widely known. Therefore, many metrologists continue to use statistical consistency as a rule of thumb by treating the squared standard uncertainties $u^2(x_1), \ldots, u^2(x_n)$ as if they were the known variances of the sampling pdfs of $x_1, \ldots, x_n$. This is inappropriate use of the standard uncertainties.

ACKNOWLEDGMENTS

For presentation at IMEKO-2009, this short paper is extracted from a larger paper submitted for publication. The first two authors are employees of the US Government and this work was done as part of their duties; therefore, this paper is not subject to copyright. The following provided useful comments on earlier draft of this paper: Javier Bernal and Tyler Estler.

REFERENCES

[1] Birge, Raymond T. 1932 The calculation of errors by the method of least squares, *Physical Review*, **40**, pp 207-227

[2] 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

[3] 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**, pp 375-496

[4] 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

[5] GUM 1995 *Guide to the Expression of Uncertainty in Measurement* 2nd ed (Geneva: International Organization for Standardization) ISBN 92-67-10188-9

[6] 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)