Thursday, April 07, 2009

Comparison of statistical consistency and metrological consistency

Raghu N Kacker and Ruediger Kessel
National Institute of Standards and Technology
Gaithersburg, Maryland 20899, USA

Emails:
raghu.kacker@nist.gog
ruediger.kessel@nist.gov

Abstract

The traditional concept of consistency in multiple evaluations of the same measurand is statistical. The statistical view of consistency does not match the modern view of uncertainty in measurement; in particular, it does not apply to the results of measurement expressed as measured values with standard uncertainties. Therefore, the International Vocabulary of Metrology, 3rd ed (VIM3) introduced the concept of metrological compatibility of multiple results of measurement for the same measurand. We prefer the term metrological consistency for the VIM3 concept of metrological compatibility. This paper discusses the differences between the two concepts of consistency.

1. Introduction

The most widely used method to assess consistency of multiple measured values for the same measurand is the Birge test published by Raymond T. Birge, a physicist, in 1932 [1]. The Birge test is based on statistical error analysis. It led to the concept of statistical consistency of multiple measured values for the same measurand. As the science and technology of measurement advanced, the limitations of statistical error analysis view of measured values became hindrance to communication of scientific and technical measurements; therefore, the world's leading metrologists developed the modern concept of uncertainty in measurement. The modern view is described in the Guide to the Expression of Uncertainty in Measurement (GUM) [2] and extended in the third edition of the International Vocabulary of Metrology (VIM3) [3]. 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 value of the 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.

In section 2, we review the concept of statistical consistency. In section 3, we review the concept of metrological consistency (compatibility). In section 4, we discuss the differences between the two concepts. Conclusion is given in section 5.

2. Statistical consistency

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 standard uncertainties. The purpose of a test of consistency is to check whether the results agree with each other. In the Birge test of consistency 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 to the measured values $x_1, \ldots, x_n$, 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 pdfs of measured values $x_1, \ldots, x_n$ are uncorrelated. Birge [1] proposed that to check the statistical consistency of $x_1, \ldots, x_n$ calculate the test statistic

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

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

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

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 [4]: The measured values $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$ 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 $u(x_i, x_j) = u^2(x_i)$ for $i = 1, 2, \ldots, n$.

A review of the Birge test in [5] 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 [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 expected value $\mathbf{1}\mu$ and variance-covariance matrix $\boldsymbol{D} = [u(x_i, x_j)]$.

A modern method to assess the fit of data to a statistical model is a Bayesian adaptation of the classical statistical theory of hypothesis testing called posterior predictive checking [7]. A discrepancy measure is a function of the data used to characterize a discrepancy, which one wishes to investigate, between the statistical model and the data. A great advantage of the posterior predictive checking is that there is no limit on the number of potential discrepancies between the statistical model and the data which may be investigated. The Bayesian 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. The fit of the statistical model to the data is suspect if the posterior predictive p-value is close to zero (say, less than 0.05) or close to one (say, more than 0.95).

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

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

which is identical to the classical p-value p_C given in (2). Thus a comparison of the posterior predictive p-value p_P relative to 0.05 is equivalent to the Birge test of consistency. When $(n - 1)\, R^2$ is too small, the posterior predictive p-value p_P is close to one raising doubt about the overall fit of the normal consistency model to the measured values.

3. Metrological consistency

Metrological consistency (compatibility) is a pair-wise concept; that is, it applies to only two results at a time. The concept of metrological consistency applies to only those

results which are metrologically comparable; that is, the results are traceable to the same reference.

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

$$\zeta(x_1 - x_2) = \frac{|x_1 - x_2|}{u(x_1 - x_2)} \leq k, \tag{4}$$

for a chosen value of k, where $u(x_1 - x_2) = \sqrt{[u^2(x_1) + u^2(x_2) - 2r(x_1, x_2)u(x_1)u(x_2)]}$ and $r(x_1, x_2)$ is the correlation coefficient between the pdfs represented by the results [3]. 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. One of the two results may be a reference result $[x_R, u(x_R)]$, where x_R is the reference value with standard uncertainty $u(x_R)$, or a consensus result $[x_C, u(x_C)]$, where x_C is consensus value and $u(x_C)$ is the standard uncertainty associated with x_C [8].

4. Differences between statistical consistency and metrological consistency

The major differences between statistical consistency and metrological consistency are as follows

(i) *Statistical consistency does not match the modern concept of uncertainty*

In the modern view of uncertainty in measurement, based on the GUM, the measured values $x_1, \ldots, x_n$ are known and the uncertainty is about the unobservable value of the measurand. Specifically, a result of measurement consists of a measured value and its associated standard uncertainty (or its equivalent). The measured value is interpreted as the known expected value and the standard uncertainty is interpreted as the known standard deviation of a state-of-knowledge probability density function (pdf) that could reasonably be attributed to the measurand. In the modern view, there is no inherent difference between an uncertainty component arising from random variation and one stemming from systematic effects. A properly evaluated expression of uncertainty includes all significant components of uncertainty determined using all available information. The concept of uncertainty applies to every single measured value.

The concept of statistical consistency is based on interpreting the measured values $x_1, \ldots, x_n$ as realizations (random draws) from sampling pdfs which in turn is based on the statistical error analysis view of the uncertainty in measurement. In error analysis the

uncertainty is about the measured values $x_1, \ldots, x_n$; specifically, the uncertainty is an estimate of the likely limits of error in $x_1, \ldots, x_n$ expressed by two expressions an estimate of the imprecision (estimated standard deviation from random error) and an assessment of the bound on bias (systematic error) in $x_1, \ldots, x_n$ [9], [10]. The bound on bias is deliberately chosen as a value that is not likely to be exceeded, which may make a statement of uncertainty based on error analysis unrealistically large. When only one measured value is available, error analysis is of little use. In summary, the concept of statistical consistency does not match the modern view of uncertainty in measurement.

(ii) *Statistical consistency does not apply to the results of measurement expressed as measured values with associated standard uncertainties*

To assess statistical consistency of the measured values $x_1, \ldots, x_n$, the metrologists treat 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 $x_1, \ldots, x_n$. A standard uncertainty represents the uncertainty about the unknown value of the measurand from all significant sources including random effects and corrections applied for systematic effects. The variance of a sampling pdf characterizes the possible dispersion of a measured value from random effects in contemplated replications. Thus treating $u^2(x_1), \ldots, u^2(x_n)$ as the known variances of the sampling pdfs of $x_1, \ldots, x_n$ is misuse of the standard uncertainties. In the modern concept of uncertainty in measurement, the results of measurement $[x_1, u(x_1)], [x_2, u(x_2)], \ldots, [x_n, u(x_n)]$, where $n \geq 2$, represent the known expected values and the standard uncertainties of different state-of-knowledge pdfs for the same measurand. The concept of 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 evaluation for the same measurand which are traceable to the same reference*

A test of statistical consistency can be applied to any set of numbers of similar magnitude with stated variances. For example, if the results $x_1, \ldots, x_n$ are differences or relative differences[1] of the measurements by participating laboratories and the corresponding measurements by a reference laboratory, then one can check their statistical consistency. It does not make sense to speak of the metrological consistency of differences and relative differences.

(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*

In statistical consistency, the unknown expected values $E(x_1), \ldots, E(x_n)$ of the measured values are not regarded as equal a priori. A check of statistical consistency checks

[1] If a_i is the result from the laboratory labeled i and a_R is the corresponding result from the reference laboratory, then $x_i = (a_i - a_R)$ is the difference and $x_i = (a_i - a_R)/a_R$ is the relative difference, for $i = 1, \ldots, n$. The relative difference may be expressed as a unit less number or as a percent.

whether the differences between the unknown expected values $E(x_1)$, …, $E(x_n)$ appear to be sufficiently small in view of their variances and covariances in which case the results may be regarded as statistically consistent. If the uncertainties are properly determined according to the GUM then two credible measured values for the same measurand should be metrologically consistent. Metrological inconsistency suggests that either the uncertainties are not properly evaluated or something is wrong with the measurement procedures.

(v) *Metrological consistency is a pair-wise concept, while statistical consistency applies to any number of results.*

Statistical consistency is defined for any number of results; however, metrological consistency is a pair-wise concept. When more than two results of measurement for a common measurand are available, then one either checks consistency of all pairs or checks consistency relative to a consensus mean or a reference value.

(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.*

In the theory underlying statistical consistency, the measured values are regarded as random selections from normal sampling pdfs. In theory, if the number of measured values is large, then some of them are likely have extreme values. Thus the theory of statistical consistency admits outliers. Metrological inconsistency indicates that some thing has gone awry in the measurement or the uncertainties are not properly determined.

5. Conclusion

The world's commerce, trade, manufacturing, engineering, and scientific research all require that different measured values for the same measurand determined in various places, at various times, and by various measurement procedures should be mutually consistent. The traditional view of consistency as used by metrologists is statistical. However, the modern concept of uncertainty in measurement, established by the GUM, has rendered the statistical view of consistency obsolete and inapplicable to the results of measurement expressed as measured values with standard uncertainties. Therefore VIM3 introduced the concept Metrological compatibility. We prefer and use the term metrological consistency for the VIM3 concept of metrological compatibility. The concept of metrological consistency matches the modern view of uncertainty in measurement and it applies to the results of measurement expressed as measured values with standard uncertainties. The concept of metrological consistency is new and not yet very widely known; also, it applies to only two results at a time. Therefore, many metrologists continue to use statistical consistency as a rule of thumb by treating the squared standard uncertainties $u^2(x_1)$, …, $u^2(x_n)$ as if they were the know variances of the sampling pdfs of x_1, …, x_n. This is inappropriate use of the standard uncertainties. In our view the use of statistical consistency should be supplanted by metrological consistency in the field of metrology.

References

[1] Birge, Raymond T. 1932 The calculation of errors by the method of least squares, *Physical Review*, **40**, pp 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
[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)
[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**, pp 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] Kessel R, Kacker R N, and Sommer K 2009 Metrological consistency and consensus result of multiple evaluations, submitted for publication
[9] Eisenhart C 1963 Realistic evaluation of the precision and accuracy of instrument calibration systems, *J Res NBS*, 67C, pp161-187
[10] Eisenhart C 1968 Expression of the uncertainties of final results, *Science*, 160, pp 1201-1204