ACTA IMEKO
ISSN: 2221-870X
February 20YY, Volume V, Number N, p1 – p2

An informed type A evaluation of standard uncertainty valid for any sample size $n \geq 1$

Carlo Carobbi

Departement of Information Engineering, Università degli Studi di Firenze, Via Santa Marta 3, 50139 Firenze, Italy

ABSTRACT
An informed type A evaluation of standard uncertainty is here derived based on Bayesian analysis. The result is mathematically simple, easily interpretable, applicable both in the theoretical framework of the Guide to the Expression of Uncertainty in Measurement (propagation of standard uncertainties) and in that of the Supplement 1 of the Guide (propagation of distributions), valid for any size $n \geq 1$ of the sample of present observations. The evaluation consistently addresses prior information in the form of the sample variance of a series of recorded experimental observations and in the form of an educated guess based on expert's experience. It turns out that distinction between type A and type B evaluation is, in this context, contrived.

Section: RESEARCH PAPER

Keywords: Measurement uncertainty; Type A evaluation; Pooled variance; Bayesian inference; Informative prior.

Citation: Carlo Carobbi, On the type A evaluation of standard uncertainty when a sample of small size is available, Acta IMEKO, vol. XX, no. YY, article ZZ, Month WWWW, identifier: IMEKO-ACTA-UU (YYYY)-VV-ZZ

Editor: Paolo Carbone, University of Perugia, Italy

Received month day, year; **In final form** month day, year; **Published** Month WWWW

Funding: None

Corresponding author: Carlo Carobbi, e-mail: carlo.carobbi@unifi.it

1. INTRODUCTION

The quantification of the type A uncertainty contribution in the case of a small sample ($n = 1, 2, 3$) is a subject of research and passionate debate in the Working Group 1 of the Joint Committee for Guides in Metrology (JCGM WG1), the standards working group involved in the maintenance and development of the Guide to the Expression of Uncertainty in Measurement (GUM, [1]) and its supplements. The topic is so felt that, at the end of 2019, the "JCGM WG1 Workshop on Type A evaluation of measurement uncertainty for a small set of observations" was held at the Bureau International des Poids et Mesures (BIPM, Sèvres, Paris). The problem arose following the negative reaction to the Committee Draft (CD) of the review of the GUM, circulated at the end of 2014.

One of the most criticized issues of the draft of the "new GUM" is the type A evaluation of uncertainty based on the use of a Student's t probability density function having $n-1$ degrees of freedom, shifted by the mean $\overline{y}$ of the n observations y_i, $i = 1, 2, ..., n$, and scaled by the standard deviation of the mean $s/\sqrt{n}$, where

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1}$$

and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 . \tag{2}$$

By following this approach, the type A evaluation of standard uncertainty is

$$u\left(\overline{y}\right) = \sqrt{\frac{n-1}{n-3}}\,\frac{s}{\sqrt{n}}, \tag{3}$$

which is not valid for a sample having a size of less than $n = 4$. Such solution originates from a Bayesian approach to inference, where improper priors (*Jeffreys prior*) are adopted for the mean μ and variance σ^2 parameters of the parent normal probability density function (PDF), i.e.

$$\mu \sim \text{const.} \tag{4}$$

and

$$\sigma^2 \sim \frac{1}{\sigma^2} . \tag{5}$$

The Bayesian approach is the one followed by the Supplement 1 of the GUM (GUMS1, [2]) and the intent of JCGM WG1 was precisely to align the GUM to GUMS1 by attributing the same Student's t probability density to a sample of repeated observations. The problem is that, by doing so, it is possible to propagate the distributions (as foreseen by GUMS1) but it is not possible to propagate the standard uncertainties (as foreseen by the GUM) if the sample size is less than $n = 4$. This is generally not acceptable (e.g., in destructive testing), particularly if implemented as a standard (mandatory) method.

The GUM and the GUMS1 approaches are therefore inconsistent. They produce substantially different results when random variability is a significant contribution to measurement uncertainty and the number of measurements used for its estimate is low [3]. The JCGM WG 1 did not seemingly yet identify a way out of the inconsistence between the GUM and the GUMS1.

Both frequentists and Bayesians can agree on the fact that the estimate of the average value obtainable from such a small sample is not very reliable. In favor of the Bayesian approach to inference, one can observe that no other way to enrich the estimate is available than the use of prior information on the variability of the measurement process that integrates the meagre experimental observation. In this sense, a Bayesian approach is useful because, differently from the frequentist approach, it provides us with a method for combining prior information with experimental observation.

From the applicative point of view these concepts have relevance to the evaluation of measurement repeatability. Measurement repeatability quantifies the variability of measurement results obtained under specified repeatability conditions. Measurement repeatability is an essential contribution to measurement uncertainty in every field of experimental activity.

In the context of testing and calibration if a stable item is re-tested or re-calibrated, the new measurement results are expected to be compatible with the old ones. Two distinct operators should provide compatible measurement results when testing or calibrating the same item. Measurement repeatability is then a reference for qualification of personnel. Monitoring measurement repeatability contributes to assuring the validity of test and calibration results. In an accreditation regime [4], measurement repeatability must be kept under statistical control. Periodic assessments are carried out by the accreditation body aimed at verifying, through an appropriate experimental check, the robustness of the estimate of measurement repeatability [5].

The GUM provides type A evaluation of standard uncertainty as the tool to quantify measurement repeatability. Type A evaluation is based on a frequentist approach, thus implying that information on the quality of the estimate of measurement uncertainty must be conveyed to the user. This is done in terms of effective degrees of freedom. The GUMS1 adopts a knowledge based (in contrast to frequentist) approach to model measurement repeatability. The quality of the estimate of measurement uncertainty is accounted for by the available prior knowledge, which eventually determines the width of the coverage interval.

The use of numerical methods for professional (accredited) evaluation of measurement uncertainty is expected to increase in the future. Indeed, the GUMS1 numerical method, which is based on the propagation of probability distributions, accounts for possible non-linearity of the measurement model, is simple, less prone to mistakes (partial derivatives are not required), provides all the available information about the measurand in terms of its probability distribution. Further, the use of numerical methods is practically unavoidable when the measurement model is complex and/or the measurand is an ensemble of scalar quantities (vector). On the other extreme, the analytical method (based on the law of propagation of uncertainty) is consolidated and the one predominantly adopted nowadays. A further point of strength of the analytical method is its great pedagogical value. Achieving consistence between the analytical and numerical approaches to measurement uncertainty quantification is therefore desirable since both have arguments of strength and are expected to coexist in the future.

What is proposed here is a knowledge-based approach to the type A evaluation of measurement uncertainty and, specifically, measurement repeatability. An estimate of the repeatability of a measurement system may be available, representative of its performance in testing. This knowledge may be derived from:

- Systematic recording of periodic verifications of the measurement system;
- Analysis and quantification of the individual sources of variability in the measurement chain;
- Normative reference (for standard measurement systems used in testing);
- Information from manufacturers of measuring instruments;
- Experience with the specific measurement chain or similar ones.

As in the GUMS1, use is made here of Bayesian inference since it provides a straightforward method to incorporate prior knowledge. Differently from the GUMS1 Bayesian approach, here an informative prior PDF is assigned to σ^2. To obtain analytical results, useful in the framework of the law of propagation of uncertainty, a normal probability model is assumed with a non-informative prior PDF for the mean and a conjugate prior PDF for the variance.

In section 2 the theoretical approach is described and in subsection 2.1 is compared with another one [6] previously presented in the scientific literature and proposed by a member of JCGM WG1. In section 3 theoretical results are applied to a practical case, based on the experience of the author, as an assessor of accredited testing laboratories. Conclusions follow in section 4. Finally, an appendix is devoted to the mathematical derivations supporting the results presented in section 2.

2. TYPE A EVALUATION WHEN PRIOR INFORMATION IS AVAILABLE

By prior information we mean here information on the variability of the measurement process obtained before that a certain test (or calibration) is carried out. Let us consider the case in which the a priori information consists of a relatively long series of experimental observations. The important hypothesis that must be verified is that the previous experimental observations are obtained under repeatability conditions that are representative of those that occur during the test, both as regards the measurement system and the measurand. If this is not verified, the a priori information is not valid to represent the variability observed during the test. This hypothesis is necessarily realized following an experimental procedure based on a physical modeling aimed at identifying the causes of the variability and at limiting its effects. It is the experimenter's task to ensure that the hypothesis is verified in practice.

In mathematical terms, the Bayesian inference is made on the mean value μ and the variance σ^2 of a Gaussian PDF assuming an improper uniform PDF for μ and an inverse χ^2 PDF for σ^2. The choice of the improper uniform PDF for μ is justified by the desire to avoid introducing an a-priori bias on the best estimate of the measurand value, which in this way depends solely on the experimental observation obtained during the test. The choice of the inverse χ^2 PDF for σ^2 is justified by the desire to incorporate prior information while retaining the well-known Student's *t* as the posterior PDF of μ. The parameters of the inverse χ^2 PDF are the prior variance σ_0^2 and the associated degrees of freedom ν_0. In this way (see the appendix for the derivation) we obtain, for the posterior marginal PDF of μ a Student's *t* PDF with degrees of freedom

$$\nu_n = (n-1)+\nu_0 , \tag{6}$$

shifted in

$$\mu_n = \bar{y} \tag{7}$$

and with scaling factor σ_n^2/n, where

$$\sigma_n^2 = \frac{(n-1)s^2+\nu_0\sigma_0^2}{(n-1)+\nu_0} . \tag{8}$$

According to this approach, the type A evaluation of standard uncertainty will be

$$\sigma_\mu = \sqrt{\frac{\nu_n}{\nu_n-2}}\frac{\sigma_n}{\sqrt{n}} . \tag{9}$$

We observe from (6) that the number of degrees of freedom ν_0 of the prior evaluation of variability, σ_0, add up to the number of degrees of freedom $n-1$ with which the variability s is evaluated during testing. The result is valid if the assumption that repeatability conditions are kept the same both in the prior investigation and testing is verified.

The estimate (7) is determined by repeated observations obtained during the testing phase because a constant and improper prior PDF for μ has been chosen.

The result (8) is particularly simple and convincing: the variance σ_n^2 which quantifies the variability of the measurement process is the result of the *pooling* of the prior variance σ_0^2 and the sample variance observed in testing s^2 through a weighted average, the weights being the corresponding degrees of freedom. The type A evaluation of standard uncertainty passes from (3), in absence of prior information, to (9), which is valid also for $n=1$ provided that $\nu_0 \geq 3$.

The following consideration is also of interest. The prior information about the variability of the measurement process may derived, for example, from the assessment of an expert. A simple form of this prior information is a best estimate σ_0 and a quantile σ_α that the expert judges to be exceeded with a small probability α. A link can be established among σ_α, α and ν_0 for a given σ_0. This can be done through the cumulative distribution function of the inverse χ^2 prior of σ^2 evaluated at σ_α^2. The result is

$$\frac{\Gamma\left(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2\sigma_\alpha^2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)} = 1-\alpha , \tag{10}$$

where

$$\Gamma(\nu,\tilde{z}) = \int_{\tilde{z}}^{\infty} t^{\nu-1}\exp(-t)\,dt$$

is the upper incomplete gamma function with parameters ν and $\tilde{z}$, and $\Gamma(\tilde{z})$ is the gamma function. If σ_α/σ_0 is known then (10), for any given α, implicitly provides a value for ν_0. This relationship can be represented through a plot such as the one in Figure 1. Note from Figure 1 that the larger is σ_α/σ_0 the smaller is ν_0 for a given α. The smaller is α for a given σ_α/σ_0 the larger is ν_0.

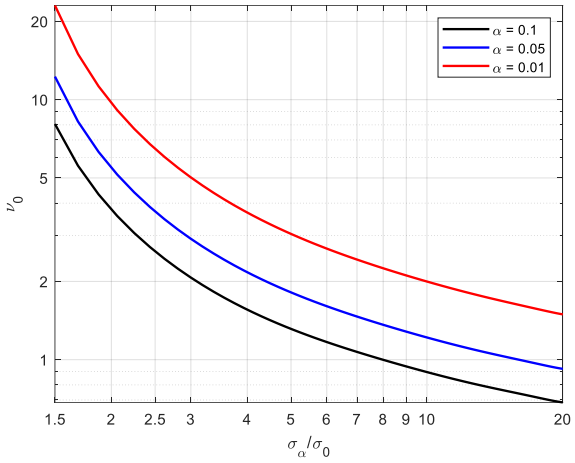

Figure 1: plots of the degrees of freedom ν_0 as a function of the ratio σ_α/σ_0 obtained by solving the implicit equation (10) for three values of probability α (see the legend).

The idea of pooling prior and observed variability is not new, even in the context of the GUM. The described solution, here traced back to a problem of inference solved through Bayesian analysis, is indeed briefly mentioned in clause 6.4.9.6 of the GUMS1. It is admittedly surprising that in the debate within JCGM WG1 this standard available option is not even mentioned.

2.1. Comparison with the type A evaluation obtained truncating the improper prior for σ^2

In a recent paper [6] Cox and Shirono propose a solution to the problem of the type A evaluation in case of small sample where σ_T is an upper bound (truncation) value for the improper prior of σ^2, i.e.

$$\sigma^2 \sim \begin{cases} \dfrac{1}{\sigma^2} & 0<\sigma<\sigma_T \\ 0 & \sigma \geq \sigma_T \end{cases} \tag{11}$$

The prior PDF of μ is in [6], as in this work, a constant improper prior. By following [6], the type A evaluation of standard uncertainty can be expressed as $\phi \cdot s/\sqrt{n}$, where

$$\phi = \left[\frac{n-1}{2} \frac{\Gamma\left(\frac{n-3}{2}, 2\frac{\sigma_T^2}{(n-1)s^2} \right)}{\Gamma\left(\frac{n-1}{2}, 2\frac{\sigma_T^2}{(n-1)s^2} \right)} \right]^{\frac{1}{2}}. \quad (12)$$

$\phi > 0$ if $n \geq 2$ is a function of s, n and σ_T.

As shown in Figure 2, it results that $\phi \cdot s < \sigma_T$ also when $s > \sigma_T$ and n is arbitrarily large. This is problematic because, when observed variability is more credible (larger number of degrees of freedom) than prior knowledge of variability, then the observed variability, not its prior estimate, should dominate the type A evaluation. In other words, setting an upper bound on σ^2 is acceptable provided that irrefutable evidence is available of an upper truncation value. Otherwise, setting a large value with an associated small probability of being exceeded is a more cautionary approach.

Another limitation of the approach in [6] is that necessarily $n \geq 2$ (see (12), $\phi = 0$ if $n = 1$) while, according to the solution here proposed, also the case $n = 1$ is tractable.

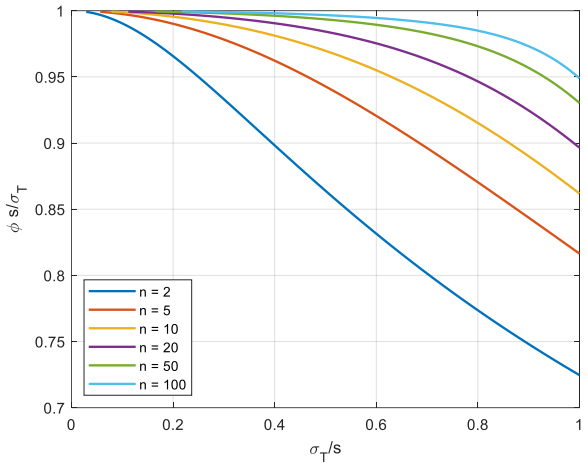

Figure 2: plots of $\phi \cdot \sigma / \sigma_T$ as a function of σ_T / s and for selected values of n (see the legend). Note that $\phi \cdot s < \sigma_T$ for any value of $s > \sigma_T$ and for any value of n.

3. APPLICATION IN THE CONTEXT OF ACCREDITATION TO ISO/IEC 17025

National accreditation bodies require evaluation of measurement repeatability of the test methods in the scope of accreditation. Such evaluation is carried out by testing laboratories through periodic recording of measurement results obtained in representative conditions of actual testing. An estimate σ_0 with corresponding degrees of freedom ν_0 is thus obtained. How to incorporate this prior knowledge into test outcome? We here provide a numerical example in the context of electromagnetic compatibility (EMC) testing. Suppose that the estimate of the non-repeatability of the radiated emission measurement chain is $\sigma_0 = 0.8$ dB and $\nu_0 = 9$. Testing two times ($n = 2$) an absolute deviation between measured values of 1.5 dB is obtained, then $s = 1.5/\sqrt{2}$ dB $= 1.06$ dB. By pooling standard deviations σ_0 and s we have

$$\nu_n = (n-1) + \nu_0 = 1 + 9 = 10,$$

$$\sigma_n = \sqrt{\frac{(n-1)s^2 + \nu_0 \sigma_0^2}{(n-1) + \nu_0}} = \sqrt{\frac{1 \cdot 1.06^2 + 9 \cdot 0.8^2}{1+9}} \text{ dB} = 0.76 \text{ dB}$$

and

$$\sigma_\mu = \sqrt{\frac{\nu_n}{\nu_n - 2}} \frac{\sigma_n}{\sqrt{n}} = \sqrt{\frac{10}{10-2}} \frac{0.76}{\sqrt{2}} = 0.60 \text{ dB}$$

As a second example consider the case where an expert of the specific test method provides a guess $\sigma_0 = 1$ dB, based on experience with similar test systems. The expert is also confident that, with a low probability $\alpha = 5\,\%$, σ exceeds $\sigma_\alpha = 2.5$ dB. This state of knowledge corresponds to approximately (see Figure 1) $\nu_0 = 4$, from which $\nu_n = 5$ (instead of 10, as in the previous example), $\sigma_n = 0.86$ dB (instead of 0.76 dB), and $\sigma_\mu = 0.78$ dB (instead of 0.60 dB).

4. CONCLUSIONS

Reliable statistical techniques to incorporate prior knowledge into the so-called "type A" evaluation of standard uncertainty should be identified to make evaluation more robust in case of small sample. The use of these statistical techniques should be promoted and confidently accepted in accredited testing if competence requirements are fulfilled. GUMS1 already provides such a tool by pooling prior variance and sample variance. A Bayesian derivation of the GUMS1 pooled variance is here illustrated along with and a more flexible interpretation aimed at addressing expert's knowledge as a useful source of reliable information.

According to the results described in this work there is no need to distinguish between type A and type B evaluations since a homogeneous mathematical treatment is used to address prior information about variability (notwithstanding is originated from experimental evidence or expert's experience) and its pooling with present observation.

APPENDIX

We here derive the marginal posterior PDF of μ given prior information in terms of the prior PDFs of μ and σ^2 and the set of observations y_i, where $i = 1, 2, ..., n$.

A uniform prior PDF is assigned to μ as

$$\mu | \sigma^2 \sim \text{const.}, \quad (13)$$

while the prior of σ^2 is an inverse χ^2 PDF with prior variance σ_0^2 and associated degrees of freedom ν_0

$$\sigma^2 \sim \text{Inv} - \chi^2 (\nu_0, \sigma_0^2). \quad (14)$$

μ and σ^2 are a-priori independent, then the joint prior PDF of μ and σ^2 is, from (13) and (14),

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2} \right). \quad (15)$$

The likelihood of the observations is easily obtained as

$$l(\mathbf{y}; \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right), \quad (16)$$

where $\mathbf{y}$ is a vector representing the set of observations y_i, $i = 1, 2, ..., n$. Due to Bayes theorem the joint posterior PDF of μ and σ^2 is given by

$$p\left(\mu, \sigma^2 \mid \mathbf{y}\right) \propto l\left(\mathbf{y}; \mu, \sigma^2\right) \times p\left(\mu, \sigma^2\right). \tag{17}$$

Substituting (15) and (16) into (17) and marginalizing with respect to σ^2 it is readily obtained

$$p\left(\mu \mid y\right) \propto \left(1 + \frac{n\left(\bar{y} - \mu\right)^2}{\left(n-1\right)s^2 + \nu_0 \sigma_0^2}\right)^{-\frac{n+\nu_0}{2}} \tag{18}$$

where $p\left(\mu \mid y\right)$ represents the marginal posterior PDF of μ. It is evident from (18) that $p\left(\mu \mid y\right)$ is a Student's *t* PDF shifted in $\bar{y}$ and scaled by σ_n^2 / n, where σ_n^2 is given by (8).

REFERENCES

[1] GUM: BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, 2008 Guide to the Expression of Uncertainty in Measurement, JCGM 100:2008, GUM 1995 with minor corrections.

[2] GUMS1: BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, 2008 Supplement 1 to the 'Guide to the Expression of Uncertainty in Measurement' – Propagation of distributions using a Monte Carlo method JCGM 101:2008.

[3] Walter Bich et al Metrologia 49 (2012) 702–705.

[4] ISO/IEC 17025, Conformity Assessment–General Requirements for the Competence of Testing and Calibration Laboratories, Int. Org. Standardization, Geneva, Switzerland (2017).

[5] SINAL DT-0002/6, GUIDA AL CALCOLO DELLA RIPETIBILITÀ DI UN METODO DI PROVA ED ALLA SUA VERIFICA NEL TEMPO, Rev. 0, Dicembre 2007.

[6] M Cox and T Shirono Metrologia 54 (2017) 642–652.