Metrologia 43 (2006) 1–11 doi:10.1088/0026-1394/43/1/001

Bayesian alternative to the ISO-GUM's use of the Welch–Satterthwaite formula

Raghu N Kacker

National Institute of Standards and Technology, Gaithersburg, MD 20899-8910, USA

E-mail: raghu.kacker@nist.gov

Received 2 September 2005
Published 18 November 2005
Online at stacks.iop.org/Met/43/1

Abstract
In certain disciplines, uncertainty is traditionally expressed as an interval about an estimate for the value of the measurand. Development of such uncertainty intervals with a stated coverage probability based on the International Organization for Standardization (ISO) *Guide to the Expression of Uncertainty in Measurement* (GUM) requires a description of the probability distribution for the value of the measurand. The ISO-GUM propagates the estimates and their associated standard uncertainties for various input quantities through a linear approximation of the measurement equation to determine an estimate and its associated standard uncertainty for the value of the measurand. This procedure does not yield a probability distribution for the value of the measurand. The ISO-GUM suggests that under certain conditions motivated by the central limit theorem the distribution for the value of the measurand may be approximated by a scaled-and-shifted t-distribution with effective degrees of freedom obtained from the Welch–Satterthwaite (W–S) formula. The approximate t-distribution may then be used to develop an uncertainty interval with a stated coverage probability for the value of the measurand. We propose an approximate normal distribution based on a Bayesian uncertainty as an alternative to the t-distribution based on the W–S formula. A benefit of the approximate normal distribution based on a Bayesian uncertainty is that it greatly simplifies the expression of uncertainty by eliminating altogether the need for calculating effective degrees of freedom from the W–S formula. In the special case where the measurand is the difference between two means, each evaluated from statistical analyses of independent normally distributed measurements with unknown and possibly unequal variances, the probability distribution for the value of the measurand is known to be a Behrens–Fisher distribution. We compare the performance of the approximate normal distribution based on a Bayesian uncertainty and the approximate t-distribution based on the W–S formula with respect to the Behrens–Fisher distribution. The approximate normal distribution is simpler and better in this case. A thorough investigation of the relative performance of the two approximate distributions would require comparison for a range of measurement equations by numerical methods.

1. Introduction

The International Organization for Standardization (ISO) *Guide to the Expression of Uncertainty in Measurement* (GUM) is being increasingly regarded as a *de facto* international standard for evaluating and expressing uncertainty in measurement. The ISO-GUM is intended for a broad spectrum of measurements including those for quality control, enforcing laws and regulations, research and development, calibrations, traceability and developing and maintaining international and national physical reference standards for measurement [1, section 0.4]. It is, therefore, reasonable for a user to expect

 © 2006 BIPM and IOP Publishing Ltd Printed in the UK

that the ISO-GUM has a single, unambiguous and consistent technical interpretation. Is it so or is it not so?

In my opinion the ISO-GUM is consistent; but certain sections, mostly in its annex G, are ambiguous. In the ISO-GUM, all Type A evaluations are estimates determined from sampling theory (frequentist statistics). However, the ISO-GUM (sections 4.1.6 and 6.2.2) interprets the Type A estimates determined from sampling theory as parameters of *state-of-knowledge probability distributions*. The Type B evaluations are, by definition, parameters of state-of-knowledge probability distributions. Thus both Type A and Type B evaluations have a common probabilistic and statistical interpretation in the ISO-GUM. This common interpretation makes the ISO-GUM consistent. The ISO-GUM's prescription that the sampling theory estimates be interpreted as parameters of state-of-knowledge probability distributions has no justification. It has previously been shown that the ISO-GUM's interpretation is justified when the Type A evaluations are either determined from Bayesian statistics or are regarded as approximations to Bayesian estimates determined from sampling theory [2]. In this paper, we propose a simpler Bayesian alternative to the ISO-GUM's approximate t-distribution with effective degrees of freedom obtained from the Welch–Satterthwaite (W–S) formula.

In section 2, we present a review of the ISO-GUM. The ISO-GUM propagates (through a linear approximation of the measurement equation) the estimates and their associated standard uncertainties for the input quantities rather than their probability distributions. This procedure does not yield a probability distribution for the value of the measurand. In certain disciplines, uncertainty is traditionally expressed as an interval about an estimate for the value of the measurand. Expression of uncertainty as an interval with a stated (supposed) coverage probability [1, section 6.2.2], requires a description of the probability distribution for the value of the measurand. The ISO-GUM suggests that when the measurement equation is a linear combination of independently distributed input variables and certain conditions, motivated by the central limit theorem, are met, the probability distribution for the value of the measurand may be approximated by a scaled-and-shifted t-distribution with effective degrees of freedom obtained from the W–S formula. In section 3, we describe the ISO-GUM's approximate t-distribution based on the W–S formula. The ISO-GUM does not state completely and explicitly the conditions required for approximating the probability distribution for the value of the measurand by a t-distribution. The ISO-GUM does not discuss the accuracy of uncertainty intervals so obtained. In section 4, we propose an approximate normal distribution based on a Bayesian uncertainty as an alternative to the approximate t-distribution based on the W–S formula. In section 5, we discuss the benefits of a Bayesian uncertainty and the approximate normal distribution. In particular, an approximate normal distribution based on a Bayesian uncertainty greatly simplifies the expression of uncertainty by eliminating altogether the need for calculating effective degrees of freedom from the W–S formula.

A probability distribution for the value of the measurand is not analytically tractable in most metrology applications even when the measurement equation is a linear combination of independent input variables. In the following non-trivial metrology application, a probability distribution for the value of the measurand is analytically tractable. The measurand is the difference between two means, each evaluated from statistical analyses of independent normally distributed measurements with unknown and possibly unequal variances. In this special application the probability distribution for the measurand is known to be a Behrens–Fisher distribution. In section 6, we compare the performance of the proposed approximate normal distribution based on a Bayesian uncertainty and the approximate t-distribution based on the W–S formula with respect to the Behrens–Fisher distribution. It is shown that the approximate normal distribution is not only simpler but also better generally. This illustration suggests that perhaps the simpler approximate normal distribution may be preferable to the approximate t-distribution for other measurement equations as well. A thorough investigation would require comparison for a range of measurement equations by numerical methods. In section 7, we discuss the parameters for comparing the two approximate distributions by numerical methods. The conclusion appears in section 8.

2. Review of the ISO-GUM

A *measurand*, denoted by Y, is a quantity subject to measurement or prediction. An *estimate* for Y, denoted by y, is a central value of the distribution of values that could reasonably be attributed to Y. The *uncertainty* is a parameter associated with the estimate y which characterizes the dispersion of the values that could be attributed to Y based upon all available information. The *standard uncertainty* is uncertainty expressed as a standard deviation, denoted by $u(y)$. The ISO-GUM is based on the concept of a *measurement equation*

$$Y = f(X_1, \ldots, X_N) \tag{1}$$

that mathematically represents the process (ingredients and recipe) for determining the estimate y and its associated standard uncertainty $u(y)$ from the estimates and their associated standard uncertainties for various input quantities $X_1, \ldots, X_N$. The measurement equation (1) should include an input quantity for every significant source of uncertainty in determining y and $u(y)$; otherwise, $u(y)$ would be a poor evaluation. Each input and output quantity of a measurement equation is regarded as a variable with a *state-of-knowledge probability distribution* having an expected value and a finite standard deviation. The measurement equation (1) could represent a system of equations where each input variable X_i may have its own measurement equation.

The expected values, standard deviations and correlation coefficients of the input variables may be evaluated from statistical methods or determined by other means. The input quantities evaluated from statistical analyses of the current data are referred to as Type A and the input quantities evaluated by other means, generally scientific judgement, are referred to as Type B [1, sections 2.3.2 and 2.3.3]. An input variable is Type A or Type B depending on whether its probability distribution or its parameters are specified by statistical methods or by other means.

A common Type A evaluation of an input quantity X_i is the arithmetic mean of a series $\{x_{i1}, x_{i2}, \ldots\}$ of m_i measurements that may reasonably be regarded as independent realizations from the same *sampling distribution* and that this distribution is normal with expected value[1] X_i and some unknown standard deviation σ_i. Suppose the arithmetic mean, the experimental (estimated) standard deviation and the estimated standard deviation of the mean of the m_i measurements are $x_i = \sum_j x_{ij}/m_i$, $s_i = \sqrt{[\sum_j (x_{ij} - x_i)^2/(m_i - 1)]}$ and $s(x_i) = s_i/\sqrt{m_i}$, respectively. Then x_i is an estimate of its expected value X_i and the standard uncertainty associated with x_i is $u_A(x_i) = s(x_i)$ [1, section 4.2]. The subscript A in $u_A(x_i)$ indicates that it is a Type A standard uncertainty. The estimate x_i and the uncertainty $u_A(x_i)$ are determined from sampling theory (frequentist statistics). However, the ISO-GUM regards x_i and $u_A(x_i)$ as the expected value and approximate standard deviation of a state-of-knowledge distribution for X_i, i.e. $E(X_i) = x_i$ and $S(X_i) \approx u_A(x_i) = s(x_i)$ [1, section 4.1.6], [2]. Depending on the number m_i of independent measurements, the Type A standard uncertainty $u_A(x_i)$ is uncertain because of the statistical reason of limited sampling [1, section E.4.3]. The uncertainty concerning $u_A(x_i)$ arising from a limited number of measurements is a *statistical-uncertainty*. Following the ISO-GUM [1, section E.4.3], the statistical-uncertainty concerning $u_A(x_i)$ is quantified by *degrees of freedom*. The degrees of freedom associated with $u_A(x_i)$ are $\nu_i = m_i - 1$.

A Type B evaluation of an input quantity X_j is commonly obtained by assigning a state-of-knowledge probability distribution for X_j. Then the estimate x_j is the expected value $E(X_j)$ and the standard uncertainty $u_B(x_j)$ is the standard deviation $S(X_j)$ of the assigned distribution [1, section 4.3]. The subscript B in $u_B(x_j)$ indicates that it is a Type B standard uncertainty. For example, if a rectangular distribution on the interval $(-a, a)$ is assigned to X_j, then $E(X_j) = x_j = 0$ and $S(X_j) = u_B(x_j) = a/\sqrt{3}$. The ISO-GUM [1, section G.4.2] suggests that subjective degrees of freedom ν_j may be assigned to a Type B uncertainty $u_B(x_j)$ to quantify uncertainty concerning the uncertainty $u_B(x_j)$ itself. When there is no uncertainty concerning the Type B uncertainty $u_B(x_j)$, it is assigned infinite degrees of freedom.

The estimate y is determined by substituting the estimates $x_1, \ldots, x_N$ for the input variables in the measurement equation $Y = f(X_1, \ldots, X_N)$. That is,

$$y = f(x_1, \ldots, x_N). \quad (2)$$

The standard uncertainties $u(x_1), \ldots, u(x_N)$ associated with the estimates $x_1, \ldots, x_N$ are components of uncertainty in determining the combined estimate y. The measurement equation (1) is approximated about y by a Taylor series as

$$Y \approx Y_{\text{linear}} = y + \sum_i c_i (X_i - x_i), \quad (3)$$

where $c_1, \ldots, c_N$ are partial derivatives of Y with respect to $X_1, \ldots, X_N$ evaluated at $x_1, \ldots, x_N$, respectively. The partial derivatives $c_1, \ldots, c_N$ are referred to as *sensitivity coefficients*. If we regard x_i and $u(x_i)$ as the expected value and standard deviation of a state-of-knowledge distribution for X_i, then the variance of Y_{linear} gives the following expression for propagating the uncertainties associated with the input values:

$$u^2(y) = \sum_i c_i^2 u^2(x_i) + 2 \sum_{(i<j)} c_i c_j u(x_i) u(x_j) r(x_i, x_j), \quad (4)$$

where $r(x_i, x_j)$ is the correlation coefficient between X_i and X_j for $i, j = 1, \ldots, N$ and $i \neq j$. The correlation coefficients are Type A or Type B depending on whether they are determined from statistical analyses or by other means. When x_i and $u(x_i)$ are the expected value and standard deviation of X_i, for $i = 1, \ldots, N$, the estimate y and the standard uncertainty $u(y)$ are the expected value and the standard deviation of Y_{linear}. The ISO-GUM regards y and $u(y)$ as approximate expected value and standard deviation of a state-of-knowledge probability distribution for Y. The estimate y and uncertainty $u(y)$ so determined differ from the unknown expected value $E(Y)$ and standard deviation $S(Y)$ to the extent that the distribution for Y differs from the distribution for Y_{linear}.

When it is necessary to express the uncertainty as an interval, multiply the standard uncertainty $u(y)$ by a *coverage factor* k to obtain the *expanded uncertainty* $U = ku(y)$ and the uncertainty interval $[y \pm U] = [y \pm ku(y)]$. The *coverage probability*[2] of an uncertainty interval $[y \pm ku(y)]$ is the fraction of a state-of-knowledge distribution for Y that is encompassed by this interval [1, sections 2.3 and 6.2.2]. The conventional value of k is two[3] [3]. To the extent that a state-of-knowledge probability distribution for Y represented by y and $u(y)$ is not determined, the coverage probability of $[y \pm ku(y)]$ cannot be stated.

The ISO-GUM [1, section 8, step 7] suggests selecting—when possible—the coverage factor k_p on the basis of the coverage probability p required of the interval $[y \pm k_p u(y)]$. In metrology, the required coverage probability p is generally set as 95%; coverage probabilities other than 95% are rarely used in metrology. The ISO-GUM [1, section 6.3.2] is very clear that the estimate y and standard uncertainty $u(y)$ are themselves insufficient for determining the coverage factor k_p for a required coverage probability p. An uncertainty interval $[y \pm k_p u(y)]$ with a stated (supposed) coverage probability p can be formed in very special conditions only.

The ISO-GUM's statement of the conditions in which k_p may be determined from a normal distribution. Suppose the following conditions are approximately met.

(i) The measurement equation is linear $Y = \sum_i c_i X_i$.
(ii) The state-of-knowledge distributions for $X_1, \ldots, X_N$ are mutually independent; therefore, $y = \sum_i c_i x_i$ and $u(y) = \sqrt{[\sum_i c_i^2 u^2(x_i)]}$, where $u^2(x_i)$ is regarded as the variance of X_i to a reasonable approximation.

[1] The ISO-GUM [1] uses the same symbols $X_1, \ldots, X_n$ for both the expected values of sampling distributions as well as the variables with state-of-knowledge distributions about the expected values. Likewise, the same symbol Y is used both for the value of the measurand as well as a variable with a state-of-knowledge distribution about the value of the measurand.

[2] The ISO-GUM uses the phrase 'level of confidence' as a synonym for coverage probability.

[3] The NIST policy on expression of uncertainty [3] states the following. 'To be consistent with current international practice, the value k to be used at NIST for calculating U is by convention $k = 2$. Values of k other than 2 are only to be used for specific applications dictated by established and documented requirements'.

(iii) The variance $u^2(y)$ is much larger than any single component $c_i^2u^2(x_i)$ from a non-normally distributed X_i.
(iv) The uncertainty $u(y)$ is not dominated by a standard uncertainty component obtained from a Type A evaluation based on just a few observations, or by a standard uncertainty component obtained from a Type B evaluation based on an assumed rectangular distribution.

Under conditions (i) through (iv), the ISO-GUM [1, section G.2.3] suggests that a reasonable first approximation to calculating an expanded uncertainty $U = k_pu(y)$ that provides an interval $[y \pm U] = [y \pm k_pu(y)]$ with coverage probability p is to use for k_p a value from the normal distribution. In support of this suggestion, the ISO-GUM refers to the central limit theorem and the calculations of tail probabilities for sums of various rectangular distributions and sums of rectangular and normal distributions as reported in [5, table 4.68]. For normal distribution, the value of k_p for the coverage probability p of 95% is 1.96, which is frequently approximated as two. Then the uncertainty interval $[y \pm k_pu(y)]$ with a stated (supposed) coverage probability p of 95% reduces to $[y \pm 2u(y)]$.

Comment 1: There is no single correct answer for the coverage probability associated with an uncertainty interval $[y \pm ku(y)]$, where y and $u(y)$ are determined according to the ISO-GUM. The coverage probability of $[y \pm ku(y)]$ is defined with respect to a specific probability distribution for $Y = f(X_1, \ldots, X_N)$ which in turn is a consequence of the joint probability distribution for $X_1, \ldots, X_N$. The estimate y and uncertainty $u(y)$ are identical for all those joint probability distributions for $X_1, \ldots, X_N$ that have the expected values $x_1, \ldots, x_N$, standard deviations $u(x_1), \ldots, u(x_N)$ and correlation coefficients $r(x_i, x_j)$, for $i, j = 1, \ldots, N$ and $i \neq j$. But the coverage probability of $[y \pm ku(y)]$ may be different for different such distributions for $X_1, \ldots, X_N$.

Comment 2: The central limit theorem requires that (i) $Y = \sum_i c_iX_i$, (ii) the probability distributions for $X_1, \ldots, X_N$ should be mutually independent, (iii) $c_1u(x_1), \ldots, c_Nu(x_N)$ should be the exact standard deviations of the probability distributions for $c_1X_1, \ldots, c_NX_N$ and (iv) the number N of summands in $Y = \sum_i c_iX_i$ should be sufficiently large for the given standard deviations $c_1u(x_1), \ldots, c_Nu(x_N)$ [4]. When $c_1u(x_1), \ldots, c_Nu(x_N)$ are unequal, which is often the case, the number N of summands required for an approximate normal distribution for $Y = \sum_i c_iX_i$ may be fairly large.

The central limit theorem does not apply when N is not sufficiently large for the given standard deviations $c_1u(x_1), \ldots, c_Nu(x_N)$ or when $c_1u(x_1), \ldots, c_Nu(x_N)$ are not the exact standard deviations of $c_1X_1, \ldots, c_NX_N$. A statement of coverage probability for the interval $[y \pm ku(y)]$ must be justified by the claimant.

Comment 3: (i) The degrees of freedom associated with a Type B evaluation are not comparable to the degrees of freedom associated with a Type A evaluation. The Type A degrees of freedom represent statistical-uncertainty in $u_A(x_i)$ arising from a limited number of independent measurements m_i available to evaluate $u_A(x_i)$. The degrees of freedom are m_i minus the number of statistical parameters estimated (restrictions) before evaluating $u_A(x_i)$. The Type B degrees of freedom associated with $u_B(x_j)$ represent subjective doubt about the parameters of a state-of-knowledge probability distribution assigned to the variable X_j.

(ii) It is superfluous in my opinion to quantify doubt about a state-of-knowledge probability distribution. For example in Bayesian statistics, one does not quantify doubt about a prior probability distribution which represents the state-of-knowledge about a statistical parameter before measurement. Hence the degrees of freedom associated with a Type B uncertainty are superfluous.

(iii) It is difficult to quantify doubt about a state-of-knowledge probability distribution.

(iv) I have never seen any metrologist assigning finite degrees of freedom to a Type B evaluation.

3. Approximate *t*-distribution based on the W–S formula

The ISO-GUM [1, sections G.3 and G.4] suggests that a better value of k_p than the one determined from normal distribution is provided by a t-distribution with effective degrees of freedom obtained from the W–S formula. Suppose Y is equal to $\sum_i c_iX_i$, for $i = 1, \ldots, N$, and $X_1, \ldots, X_N$ are mutually independent. Suppose the probability distributions for $X_1, \ldots, X_n$, for $n < N$, are Type A and the probability distributions for $X_{n+1}, \ldots, X_N$ are Type B. Suppose, for $i = 1, \ldots, n$, the expected value and standard deviation of a state-of-knowledge distribution for X_i are determined from m_i independent normally distributed measurements. Suppose $E(X_i) = x_i$ and $S(X_i) \approx u_A(x_i) = s(x_i)$ with degrees of freedom $\nu_i = m_i - 1$, respectively [1, sections 4.1.6 and 4.2]. Suppose $X_{n+1}, \ldots, X_N$ are evaluated from the assigned state-of-knowledge distributions with expected values $E(X_j) = x_j$ and standard deviations $S(X_j) = u_B(x_j)$ [1, section 4.3] with subjective degrees of freedom ν_j [1, section G.4.2].

Let us define

$$X_A = \sum_{i=1}^{n} c_iX_i \tag{5}$$

and

$$X_B = \sum_{i=n+1}^{N} c_iX_i. \tag{6}$$

Then $Y = \sum_i c_iX_i = X_A + X_B$, where X_A and X_B are the Type A and the Type B components of Y. Following the ISO-GUM [1, section 5.1] an estimate for X_A is

$$x_A = \sum_{i=1}^{n} c_ix_i \tag{7}$$

and its associated standard uncertainty is

$$u_A(x_A) = \sqrt{\sum_{i=1}^{n} c_i^2u_A^2(x_i)}. \tag{8}$$

The standard uncertainty $u_A(x_A)$ is an approximation for the standard deviation $S(X_A) = \sqrt{[\sum_i c_i^2S^2(X_i)]}$.

My understanding of the ISO-GUM [1, sections G.3 and G.4] is that it ascribes to the variable $(X_A - x_A)/u_A(x_A)$ an approximate state-of-knowledge t-distribution with effective degrees of freedom ν_{effA}, where

$$\nu_{\text{effA}} = \frac{[u_{\text{A}}(x_{\text{A}})]^4}{\sum_{i=1}^{n}([|c_i|u_{\text{A}}(x_i)]^4/\nu_i)}. \tag{9}$$

The expression (9) for ν_{effA} is referred to as the W–S formula. The effective degrees of freedom ν_{effA} quantify the statistical-uncertainty in $u_{\text{A}}(x_{\text{A}})$ arising from limited numbers $m_1, \ldots, m_n$ of measurements available for evaluating the parameters of the Type A variables $X_1, \ldots, X_n$, respectively.

The expected value $E(X_{\text{B}})$ and standard deviation $S(X_{\text{B}})$ are, respectively,

$$x_{\text{B}} = \sum_{j=n+1}^{N} c_j x_j \tag{10}$$

and

$$u_{\text{B}}(x_{\text{B}}) = \sqrt{\sum_{j=n+1}^{N} c_j^2 u_{\text{B}}^2(x_j)}. \tag{11}$$

The ISO-GUM suggests that the coverage factor k_p of an uncertainty interval $[y \pm k_p u(y)]$ for Y may be determined by ascribing to the variable $(Y - y)/u(y)$ an approximate t-distribution with effective degrees of freedom ν_{eff}, where

$$y = x_{\text{A}} + x_{\text{B}} = \sum_i c_i x_i, \tag{12}$$

$$u(y) = \sqrt{u_{\text{A}}^2(x_{\text{A}}) + u_{\text{B}}^2(x_{\text{B}})} = \sqrt{\sum_{i=1}^{n} c_i^2 u_{\text{A}}^2(x_i) + \sum_{j=n+1}^{N} c_j^2 u_{\text{B}}^2(x_j)} \tag{13}$$

and

$$\nu_{\text{eff}} = \frac{[u(y)]^4}{\sum_{i=1}^{n}([|c_i|u_{\text{A}}(x_i)]^4/\nu_i) + \sum_{j=n+1}^{N}([|c_j|u_{\text{B}}(x_j)]^4/\nu_j)}. \tag{14}$$

Thus the ISO-GUM suggests that k_p may be determined by ascribing to the variable Y an approximate state-of-knowledge scaled-and-shifted t-distribution with degrees of freedom ν_{eff} that have been scaled by $u(y)$ of equation (13) and shifted by y of equation (12). Suppose $t_p(\nu_{\text{eff}})$ is a percentile[4] of the *t-distribution* with degrees of freedom ν_{eff} such that the interval $[-t_p(\nu_{\text{eff}}), +t_p(\nu_{\text{eff}})]$ encompasses the fraction p of this distribution. Then according to the ISO-GUM, $[y \pm t_p(\nu_{\text{eff}})u(y)]$ is an uncertainty interval for Y with an approximate coverage probability p.

Comment 4: The W–S formula was developed for what are now called the Type A evaluations determined from sampling theory (frequentist statistics). Its origins are discussed in [6]. References [6–8] discuss the use of W–S formula from the viewpoint of sampling theory when $n = N = 2$.

Comment 5: The uncertainty interval $[y \pm t_p(\nu_{\text{eff}})u(y)]$ is wider than the interval $[y \pm 2u(y)]$ based on normal distribution for finite values of ν_{eff}. The coverage factor $t_p(\nu_{\text{eff}})$ and hence the width of the interval $[y \pm t_p(\nu_{\text{eff}})u(y)]$ increases as the effective degrees of freedom ν_{eff} decrease [1, table G.2]. Thus the lesser the degrees of freedom, the larger is the width of the interval $[y \pm t_p(\nu_{\text{eff}})u(y)]$. The larger width of the interval $[y \pm t_p(\nu_{\text{eff}})u(y)]$ reflects the uncertainty in $u(y)$ indicated by its effective degrees of freedom ν_{eff}.

Comment 6: The ISO-GUM does not explicitly state the conditions in which a user may determine the coverage factor k_p from an approximate t-distribution with effective degrees of freedom ν_{eff} determined from the W–S formula. Are these the same conditions as those stated in the ISO-GUM [1, section G.2] for determining k_p from normal distribution and restated here in section 2?

Comment 7: The ISO-GUM does not discuss the accuracy of the stated (supposed) coverage probability p of the interval $[y \pm t_p(\nu_{\text{eff}})u(y)]$ determined from an approximate t-distribution with effective degrees of freedom ν_{eff} determined from the W–S formula. The correct coverage probability would depend on the joint probability distribution for the input variables $X_1, \ldots, X_N$ and the corresponding probability distribution for $Y = \sum_i c_i X_i$ (see, comment 1).

Comment 8: It seems that the ISO-GUM's use of an approximate scaled-and-shifted t-distribution with effective degrees of freedom ν_{eff} for $Y = \sum_i c_i X_i$ requires two sets of assumptions.

(1) *Assumptions that underlie the W–S formula for the Type A variables* $X_1, \ldots, X_n$: Each input quantity X_i, for $i = 1, \ldots, n$, is (i) evaluated from a series of m_i mutually independent measurements having the same normal[5] sampling distribution with expected value X_i and variance σ_i^2 and (ii) the sets of $m_1, \ldots, m_n$ measurements for evaluating $X_1, \ldots, X_n$, respectively, are independent [10–12]. A consequence of the second part of this assumption is that the state-of-knowledge distributions for $X_1, \ldots, X_n$ are mutually independent. Here the number n of the Type A variables may be as little as two.

(2) *Assumptions similar to those that underlie the central limit theorem*: In particular, the number of summands in $Y = X_{\text{A}} + c_{n+1}X_{n+1} + \cdots + c_N X_N$ is sufficiently large for the central limit theorem to apply for the given standard deviations of the variables $X_{\text{A}}, c_{n+1}X_{n+1}, \ldots, c_N X_N$.

4. Approximate normal distribution based on a Bayesian uncertainty

Suppose the assumptions that underlie the W–S formula for the Type A variables $X_1, \ldots, X_n$, stated in comment 8, are reasonably satisfied. There are two statistical approaches to determine an estimate for an input quantity X_i: sampling theory (frequentist statistics) and Bayesian statistics [13]. Both approaches agree with the ISO-GUM's definition of a Type A evaluation for the input quantity X_i. An estimate for X_i

[4] The ISO-GUM's symbol $t_p(\nu_i)$ is $(1/2)(1 + p) \times 100$th percentile of the *t-distribution* with degrees of freedom ν_i.

[5] In sampling theory (frequentist statistics), an approximate t-distribution for $(x_{\text{A}} - X_{\text{A}})/u_{\text{A}}(x_{\text{A}})$ with effective degrees of freedom ν_{effA} obtained from the W–S formula requires that x_i be the arithmetic mean and s_i^2 be the sample variance of m_i independent normally distributed measurements with unknown variance σ_i^2 and that the sets of $m_1, \ldots, m_n$ measurements be independent. The ISO-GUM [1, section G.4.1] appears to imply that the normal distribution for the m_i independent measurements is not essential. This is incorrect. The W–S formula requires that x_i and s_i^2 have independent sampling distributions, for $i = 1, \ldots, n$. It turns out that normal distribution is the only probability distribution for which the sampling distributions of x_i and s_i^2 are mutually independent [9]. Therefore normal distribution is essential.

and its associated standard uncertainty determined through Bayesian statistics from only the m_i measurements without assuming any additional information are obtained by using non-informative *prior distributions* for X_i and σ_i^2. Non-informative prior distributions for X_i and σ_i^2 indicate that there is no prior knowledge about their values before (or in addition to) measurement[6]. A convenient pair of non-informative prior distributions[7] is as follows: the prior distribution for X_i is proportional to one and the prior distribution for σ_i^2 is proportional to $1/\sigma_i^2$ [13]. The m_i independent normal measurements provide a *likelihood function* for X_i and σ_i^2 given the measurements. It can then be shown using *the Bayes's theorem* that the Bayesian *posterior state-of-knowledge probability distribution* for $(X_i - x_i)/s(x_i)$ is the t-distribution with degrees of freedom $\nu_i = m_i - 1$ [13]. It follows that the Bayesian distribution for X_i based on only the m_i measurements is a *scaled-and-shifted t-distribution* with degrees of freedom ν_i that have been scaled by $u_A(x_i) = s(x_i)$ and shifted by x_i. The expected value and standard deviation of a *t-distribution* with degrees of freedom ν_i are, respectively, zero and $\sqrt{[\nu_i/(\nu_i - 2)]} = \sqrt{[(m_i - 1)/(m_i - 3)]}$ [14]. It follows that the expected value and standard deviation of the Bayesian posterior distribution for X_i are $E(X_i) = x_i$ and $S(X_i) = \sqrt{[(m_i - 1)/(m_i - 3)]} \times u_A(x_i)$, respectively. The standard deviation $S(X_i)$ is a Bayesian standard uncertainty, $u_{A,\mathrm{Bayes}}(x_i)$, associated with the estimate x_i [2][8]. That is,

$$u_{A,\mathrm{Bayes}}(x_i) = \sqrt{\frac{m_i - 1}{m_i - 3}} u_A(x_i). \qquad (15)$$

The ISO-GUM interprets the standard uncertainty $u_A(x_i) = s(x_i)$ determined from sampling theory (frequentist statistics) as an approximate standard deviation of a state-of-knowledge distribution for X_i with degrees of freedom $\nu_i = m_i - 1$ [1, section 4.1.6]. The standard uncertainty $u_{A,\mathrm{Bayes}}(x_i)$ is the standard deviation of a Bayesian posterior distribution for X_i, which is a state-of-knowledge distribution by definition. Thus $u_A(x_i) = s(x_i)$ may be regarded as an approximation for $u_{A,\mathrm{Bayes}}(x_i)$. The approximation is poor when the degrees of freedom ν_i are small but good when ν_i are large.

The Bayesian uncertainty $u_{A,\mathrm{Bayes}}(x_i)$ requires at least four independent normal measurements m_i. When $m_i = 2$ or $m_i = 3$, there are three options. First, investigate other reasonable prior distributions for X_i and σ_i^2 that might yield useful expressions for $u_{A,\mathrm{Bayes}}(x_i)$. Second, use a Type B probability distribution for X_i. Third, use ad hoc standard uncertainties. When $p = 95\%$, the ad hoc standard uncertainties for $m_i = 2$ and $m_i = 3$ are, respectively, $u_A^*(x_i) = 6.483 \times u_A(x_i)$ and $u_A^*(x_i) = 2.195 \times u_A(x_i)$ [2]. The ad hoc standard uncertainties were suggested in [2] by replacing the undefined factor $\sqrt{[(m_i - 1)/(m_i - 3)]}$, when m_i is less than four, with the ratio of the relevant percentiles of t-distribution and normal distribution for the chosen p of 95%.

We propose to approximate the Bayesian *scaled-and-shifted t-distribution* for X_i by a normal distribution, $N(x_i, u^2_{A,\mathrm{Bayes}}(x_i))$, with expected value x_i and variance $u^2_{A,\mathrm{Bayes}}(x_i)$. The two distributions agree up to the moments of order two [1, section C.2.13]; the moment of order three is zero because of symmetry. Then an approximate Bayesian posterior distribution for X_A is normal $N(x_A, u^2_{A,\mathrm{Bayes}}(x_A))$ where x_A is defined in equation (7) and

$$u_{A,\mathrm{Bayes}}(x_A) = \sqrt{\sum_{i=1}^{n} c_i^2 u^2_{A,\mathrm{Bayes}}(x_i)} = \sqrt{\sum_{i=1}^{n} c_i^2 \frac{m_i - 1}{m_i - 3} u_A^2(x_i)}. \qquad (16)$$

Now $Y = \sum_i c_i X_i = X_A + X_B$. Therefore, the expected value and standard deviation of Y are, respectively, $y = x_A + x_B = \sum_i c_i x_i$ as in equation (12), and

$$u_{\mathrm{Bayes}}(y) = \sqrt{\sum_{i=1}^{n} c_i^2 \frac{m_i - 1}{m_i - 3} u_A^2(x_i) + \sum_{j=n+1}^{N} c_j^2 u_B^2(x_j)}. \qquad (17)$$

Suppose the requirements that underlie the central limit theorem are reasonably satisfied for the sum $Y = X_A + X_B = X_A + c_{n+1}X_{n+1} + \cdots + c_N X_N$, where the distribution for X_A is approximately normal. Then the distribution for Y may be assumed to be approximately normal with expected value y and variance $u^2_{\mathrm{Bayes}}(y)$. Thus $[y \pm z_p u_{\mathrm{Bayes}}(y)]$ is an uncertainty interval for Y with approximate coverage probability p, where z_p is a percentile of the standard normal distribution such that the interval $[-z_p, +z_p]$ encompasses the fraction p of this distribution. For $p = 95\%$, $z_p = 1.96$ and the interval $[y \pm z_p u_{\mathrm{Bayes}}(y)]$ becomes $[y \pm 1.96 u_{\mathrm{Bayes}}(y)]$, which may sometimes be approximated as $[y \pm 2u_{\mathrm{Bayes}}(y)]$.

5. Benefits of a Bayesian uncertainty and the approximate normal distribution

The ISO-GUM is consistent because it interprets the Type A evaluations as parameters of state-of-knowledge distributions. Then Type A and Type B evaluations have a common probabilistic interpretation and they can be combined through a measurement equation. The ISO-GUM's interpretation is justified when either Bayesian statistics is used for the Type A evaluations or sampling theory estimates are regarded as approximations to Bayesian estimates. Therefore the ISO-GUM may be regarded as an extension of Bayesian statistics to incorporate non-statistical evaluations.

As indicated in the ISO-GUM [1, section E.4.3], the statistical-uncertainty in $u_A(x_i)$ arising from the limited number m_i of measurements may be large for practical values of m_i. Therefore, $u_A(x_i)$, defined in section 2, is an incomplete expression of the uncertainty associated with x_i without an accompanying statement of its degrees of freedom. Similarly, the statistical-uncertainty in $u(y)$, defined in section 3, may

[6] Prior distributions convey information in addition to the measurement data. In order to develop an approximate probability distribution for Y that is comparable to the ISO-GUM's t-distribution based on the W–S formula, we need non-informative prior distributions that do not carry additional information.

[7] These non-informative prior distributions are not proper probability distributions and hence they are termed improper prior distributions.

[8] A Bayesian standard uncertainty associated with x_i depends on the choice of prior distributions for X_i and σ_i^2. The posterior distribution for X_i, and hence a Bayesian uncertainty, is different for different prior distributions for the statistical parameters. The Bayesian uncertainty $u_{A,\mathrm{Bayes}}(x_i)$ of equation (15) is based on the particular improper non-informative prior distributions used for X_i and σ_i^2.

be large when its effective degrees of freedom ν_{eff} are small. Therefore, $u(y)$ is an incomplete expression of the uncertainty associated with y without an accompanying statement of its effective degrees of freedom. Unlike sampling theory, a Bayesian standard uncertainty has no statistical-uncertainty. Thus $u_{\text{A,Bayes}}(x_i)$ is a complete expression of the uncertainty associated with x_i, for $i = 1, \ldots, n$. Similarly, $u_{\text{Bayes}}(y)$ is a complete expression of the uncertainty associated with y.

The use of an approximate normal distribution for Y based on a Bayesian uncertainty $u_{\text{Bayes}}(y)$ greatly simplifies the expression of uncertainty associated with y by eliminating altogether the need for calculating effective degrees of freedom from the W–S formula.

The primary expression of uncertainty in the ISO-GUM is standard uncertainty. An uncertainty interval is a secondary expression of uncertainty obtained from standard uncertainty. The ISO-GUM accounts for the statistical-uncertainty in $u(y)$ arising from the limited numbers $m_1, \ldots, m_n$ of measurements when an uncertainty interval $[y \pm k_p u(y)]$ is computed by using a larger coverage factor $k_p = t_p(\nu_{\text{eff}})$ determined from a t-distribution rather than the coverage factor $k_p = z_p$ determined from the normal distribution. The factors $\sqrt{[(m_i - 1)/(m_i - 3)]}$, for $i = 1, 2, \ldots, n$, built in the Bayesian standard uncertainty $u_{\text{Bayes}}(y)$ enlarge it when one or more of the numbers $m_1, \ldots, m_n$ of measurements are small. Thus Bayesian statistics automatically accounts for the numbers of measurements in the standard uncertainty which is the primary expression of uncertainty.

6. Analytical comparison with respect to the Behrens–Fisher distribution for the difference between two means

Suppose the measurement equation for the value of the measurand is $Y = X_1 - X_2$, where X_1 and X_2 are independently distributed. Suppose the estimate and standard uncertainty for X_i are x_i and $u(x_i)$, respectively, for $i = 1$ and 2. Then $y = x_1 - x_2$ and $u(y) = \sqrt{[u^2(x_1) + u^2(x_2)]}$. This measurement equation applies when a common quantity is measured by two different instruments, methods or laboratories and the metrological interest lies in the difference between the two evaluations. In CIPM key comparisons [15] among national metrology institutes (NMIs), pair wise degrees of equivalence $d_{i,j}$ and their associated standard uncertainties $u(d_{i,j})$ are of the form y and $u(y)$, respectively.

We consider a special application of the measurement equation $Y = X_1 - X_2$, where X_i is evaluated from m_i independent and normally distributed measurements with unknown variance σ_i^2, for $i = 1$ and 2. Thus both X_1 and X_2 are Type A variables. Then, as discussed in section 2, result x_i is the arithmetic mean and uncertainty $u(x_i)$ is the standard deviation of the mean $s(x_i) = s_i/\sqrt{m_i}$. A Bayesian posterior distribution for $T_i = (X_i - x_i)/u(x_i)$, based on the non-informative prior distributions for X_i and σ_i^2 introduced in section 4, is the t-distribution with degrees of freedom $\nu_i = m_i - 1$. In this situation the probability distribution of $(Y - y)/u(y)$ is known to be the Behrens–Fisher distribution[9] with parameters $\nu_1 = m_1 - 1$, $\nu_2 = m_2 - 1$ and $\theta = \tan^{-1}[u(x_1)/u(x_2)]$ [18]. The percentiles of Behrens–Fisher distribution for the coverage probability p of 95% are tabulated in [19] for $\nu_1, \nu_2 = 1, 2, \ldots, 8, 10, 12, 24$ and ∞, and $\theta = 0^\circ$, 15°, 30°, 45°, 60°, 75° and 90°. These percentiles may be used to determine the correct coverage factor k_p of the uncertainty interval $[y \pm k_p u(y)]$ for Y for various values of ν_1, ν_2 and θ, and p of 95%.

This example is referred to in textbooks on statistics as 'the two-sample problem with unknown and possibly unequal variances'. For this case, reference [20, appendix B] discusses three other approximations proposed in [21] and [22] for the distribution of $(Y - y)/u(y)$. The approximations proposed in [21] and [22] are better than the approximate t-distribution based on the W–S formula. The approximation proposed in [22] applies to sums of many scaled t-distributions. In this section, we investigate the accuracies of the coverage factors determined from the approximate t-distribution based on the W–S formula and the approximate normal distribution based on a Bayesian uncertainty with respect to the correct coverage factor k_p based on the Behrens–Fisher distribution.

The correct coverage factor k_p of an uncertainty interval $[y \pm k_p u(y)]$ for Y with coverage probability p is $k_p = \text{BF}_p(\nu_1, \nu_2, \theta)$, where $\text{BF}_p(\nu_1, \nu_2, \theta)$ is a percentile of the Behrens–Fisher distribution such that the interval $[-\text{BF}_p(\nu_1, \nu_2, \theta), +\text{BF}_p(\nu_1, \nu_2, \theta)]$ encompasses the fraction p of this distribution. An approximate coverage factor determined from the t-distribution based on the W–S formula is $k_p(\text{W}) = t_p(\nu_{\text{eff}})$, where $\nu_{\text{eff}} = [u(y)]^4/[u^4(x_1)/\nu_1 + u^4(x_2)/\nu_2]$. An uncertainty interval for Y determined from the approximate normal distribution based on a Bayesian uncertainty is $[y \pm z_p u_{\text{Bayes}}(y)]$, where $u_{\text{Bayes}}(y) = \sqrt{[u^2_{\text{Bayes}}(x_1) + u^2_{\text{Bayes}}(x_2)]}$ and $u_{\text{Bayes}}(x_i) = \sqrt{[(m_i - 1)/(m_i - 3)]} \times u(x_i)$, provided m_i is greater than 3, for $i = 1$ and 2. To compare the interval based on Bayesian uncertainty with the interval $[y \pm k_p u(y)]$ determined from the Behrens–Fisher distribution, we express the former as $[y \pm k_p(\text{B}) u(y)]$, where $k_p(\text{B}) = z_p \times u_{\text{Bayes}}(y)/u(y)$. Then we compare the three coverage factors $k_p = \text{BF}_p(\nu_1, \nu_2, \theta)$, $k_p(\text{W}) = t_p(\nu_{\text{eff}})$ and $k_p(\text{B}) = z_p \times u_{\text{Bayes}}(y)/u(y)$ for various values of ν_1, ν_2 and θ, and $p = 95\%$. Here, k_p is the correct coverage factor and $k_p(\text{W})$ and $k_p(\text{B})$ are approximate coverage factors.

To determine the coverage factors k_p, $k_p(\text{W})$ and $k_p(\text{B})$, we require numerical values for the parameters $u(x_1) = s_1/\sqrt{m_1}$ and $u(x_2) = s_2/\sqrt{m_2}$. We arbitrarily set $s_1^2 = 100$ and determined the corresponding s_2^2 by solving the equation $\theta = \tan^{-1}[u(x_1)/u(x_2)]$ for various values of θ. The ratios $u(x_1)/u(x_2)$ corresponding to $\theta = 15^\circ$, 30°, 45°, 60° and 75° are, respectively, 27%, 58%, 100%, 173% and 373%. The value $\theta = 45^\circ$ represents the situation where the uncertainties $u(x_1)$ and $u(x_2)$ are equal. The values $\theta = 30^\circ$ and 60° represent the situation where the difference between $u(x_1)$ and $u(x_2)$ is small. The values $\theta = 15^\circ$ and 75° represent the situation where the difference between $u(x_1)$ and $u(x_2)$ is large.

The use of W–S formula is of greatest advantage when one or both the degrees of freedom ν_1 and ν_2 are small. Therefore, the smaller values of ν_1 and ν_2 are of greater interest.

[9] The distribution of $[T_1 \sin\theta - T_2 \cos\theta]$, where $0 \leqslant \theta \leqslant \pi/2$ radians (90°), is a Behrens–Fisher distribution [16, 17]. The distribution of $(Y - y)/u(y)$ has $\theta = \tan^{-1}[u(x_1)/u(x_2)]$. The Behrens–Fisher distribution was originally developed in what is called fiducial inference; however, it was subsequently identified as a Bayesian distribution [18].

Table 1. Relative errors Δk_p(W) and Δk_p(B) for (ν_1, ν_2, θ), where (ν_1, ν_2) = (1, 1), (2, 1) and (2, 2); and $\theta = 15°, 30°, 45°, 60°$ and 75°. Comparison for (1, 2, θ) is given by the comparison for (2, 1, 90° − θ).

ν_1	ν_2	θ	k_p(W)	k_p(B)	k_p	Δk_p(W)	Δk_p(B)
1	1	15°	12.71	12.71	15.56	−0.1834	−0.1834
2	1	15°	12.71	12.32	12.41	0.0239	−0.0070
2	2	15°	4.30	4.30	4.41	−0.0252	−0.0252
1	1	30°	12.71	12.71	17.36	−0.2681	−0.2681
2	1	30°	12.71	11.21	11.54	0.1011	−0.0284
2	2	30°	3.18	4.30	4.56	−0.3026	−0.0571
1	1	45°	4.30	12.71	17.97	−0.7606	−0.2929
2	1	45°	4.30	9.49	10.14	−0.5757	−0.0645
2	2	45°	2.78	4.30	4.62	−0.3996	−0.0695
1	1	60°	12.71	12.71	17.36	−0.2681	−0.2681
2	1	60°	4.30	7.37	8.34	−0.4843	−0.1173
2	2	60°	3.18	4.30	4.56	−0.3026	−0.0571
1	1	75°	12.71	12.71	15.56	−0.1834	−0.1834
2	1	75°	4.30	5.30	6.34	−0.3213	−0.1641
2	2	75°	4.30	4.30	4.41	−0.0252	−0.0252

Tables 1–6 display k_p(W), k_p(B), k_p, and the relative errors $\Delta k_p(\text{W}) = [k_p(\text{W}) - k_p]/k_p$ and $\Delta k_p(\text{B}) = [k_p(\text{B}) - k_p]/k_p$ for the triplets (ν_1, ν_2, θ), where ν_1, ν_2 = 1, 2, 3, 4, 5, 10 and 24; $\nu_1 \geqslant \nu_2$; and $\theta = 15°, 30°, 45°, 60°$ and 75°. Comparison for $\nu_1 \leqslant \nu_2$ is made by comparing the coverage factors for the triplets (ν_2, ν_1, 90° − θ) [18]. The relative errors Δk_p(W) and Δk_p(B) indicate the accuracies of k_p(W) and k_p(B) with respect to k_p, respectively. In tables 1–3, when ν_1, ν_2 = 1 or 2, we use for $u_{\text{Bayes}}(x_i)$ ad hoc standard uncertainties given in section 4.

Table 1 is for the case where both ν_1 and ν_2 are either 1 or 2. The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{W})|$ are 2.39% and 76.06%, respectively. The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{B})|$ are 0.70% and 29.29%, respectively. In the cases considered, $|\Delta k_p(\text{B})|$ is less than or equal to $|\Delta k_p(\text{W})|$.

Table 2 is for the case where either ν_1 or ν_2 is 1 and the other is 3, 4, 5, 10 or 24. The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{W})|$ are 3.39% and 66.06%, respectively. The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{B})|$ are 0.03% and 6.37%, respectively. In the cases considered, $|\Delta k_p(\text{B})|$ is substantially less than $|\Delta k_p(\text{W})|$.

Table 3 is for the case where either ν_1 or ν_2 is 2 and the other is 3, 4, 5, 10 or 24. The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{W})|$ are 1.48% and 29.63%, respectively. The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{B})|$ are 0.05% and 3.52%, respectively. In the cases considered, $|\Delta k_p(\text{B})|$ is substantially less than $|\Delta k_p(\text{W})|$.

Table 4 is for the case where ν_1, ν_2 = 3, 4, 5, 10 and 24 and the difference between $u(x_1)$ and $u(x_2)$ is large ($\theta = 15°$ or 75°). The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{W})|$ are 0.16% and 4.92%, respectively. The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{B})|$ are 0.01% and 6.77%, respectively. Thus the absolute relative error in both Δk_p(W) and Δk_p(B) is small when the difference between $u(x_1)$ and $u(x_2)$ is large ($\theta = 15°$ or 75°). Mostly, $|\Delta k_p(\text{B})|$ is less than $|\Delta k_p(\text{W})|$. In a few cases, $|\Delta k_p(\text{W})|$ is less than $|\Delta k_p(\text{B})|$.

Table 2. Relative errors Δk_p(W) and Δk_p(B) for (ν_1, ν_2, θ), where ν_1 = 3, 4, 5, 10 and 24; ν_2 = 1; and $\theta = 15°, 30°, 45°, 60°$ and 75°. Comparison for (ν_1, ν_2, θ), where $\nu_1 < \nu_2$ is given by the comparison for (ν_2, ν_1, 90° − θ).

ν_1	ν_2	θ	k_p(W)	k_p(B)	k_p	Δk_p(W)	Δk_p(B)
3	1	15°	12.71	12.30	12.29	0.0339	0.0012
4	1	15°	12.71	12.29	12.28	0.0347	0.0012
5	1	15°	12.71	12.29	12.28	0.0347	0.0009
10	1	15°	12.71	12.29	12.28	0.0347	0.0005
24	1	15°	12.71	12.28	12.28	0.0347	0.0004
3	1	30°	12.71	11.13	11.11	0.1437	0.0022
4	1	30°	12.71	11.09	11.06	0.1488	0.0028
5	1	30°	12.71	11.08	11.04	0.1509	0.0033
10	1	30°	12.71	11.06	11.03	0.1520	0.0026
24	1	30°	12.71	11.05	11.03	0.1520	0.0019
3	1	45°	3.18	9.30	9.30	−0.6579	−0.0003
4	1	45°	3.18	9.20	9.14	−0.6517	0.0066
5	1	45°	3.18	9.16	9.09	−0.6499	0.0078
10	1	45°	3.18	9.12	9.06	−0.6485	0.0069
24	1	45°	3.18	9.10	9.05	−0.6482	0.0060
3	1	60°	2.78	7.00	7.12	−0.6102	−0.0172
4	1	60°	2.78	6.79	6.77	−0.5899	0.0030
5	1	60°	2.57	6.72	6.64	−0.6126	0.0127
10	1	60°	2.31	6.63	6.51	−0.6458	0.0183
24	1	60°	2.20	6.60	6.49	−0.6606	0.0171
3	1	75°	3.18	4.64	4.96	−0.3584	−0.0637
4	1	75°	2.78	4.24	4.47	−0.3787	−0.0511
5	1	75°	2.57	4.10	4.22	−0.3906	−0.0286
10	1	75°	2.23	3.91	3.84	−0.4190	0.0198
24	1	75°	2.06	3.84	3.69	−0.4399	0.0413

Table 3. Relative errors Δk_p(W) and Δk_p(B) for (ν_1, ν_2, θ), where ν_1 = 3, 4, 5, 10 and 24, ν_2 = 2; and $\theta = 15°, 30°, 45°, 60°$ and 75°. Comparison for (ν_1, ν_2, θ), where $\nu_1 < \nu_2$ is given by the comparison for (ν_2, ν_1, 90° − θ).

ν_1	ν_2	θ	k_p(W)	k_p(B)	k_p	Δk_p(W)	Δk_p(B)
3	2	15°	4.30	4.25	4.24	0.0148	0.0019
4	2	15°	4.30	4.22	4.21	0.0232	0.0030
5	2	15°	4.30	4.21	4.19	0.0259	0.0032
10	2	15°	4.30	4.19	4.18	0.0284	0.0025
24	2	15°	4.30	4.19	4.18	0.0293	0.0023
3	2	30°	3.18	4.09	4.10	−0.2238	−0.0013
4	2	30°	3.18	3.98	3.96	−0.1972	0.0029
5	2	30°	3.18	3.94	3.91	−0.1859	0.0067
10	2	30°	3.18	3.88	3.85	−0.1725	0.0099
24	2	30°	3.18	3.86	3.83	−0.1686	0.0095
3	2	45°	2.78	3.88	3.90	−0.2886	−0.0071
4	2	45°	2.57	3.62	3.65	−0.2963	−0.0093
5	2	45°	2.57	3.53	3.54	−0.2728	−0.0015
10	2	45°	2.45	3.41	3.37	−0.2731	0.0143
24	2	45°	2.36	3.37	3.31	−0.2847	0.0191
3	2	60°	2.78	3.64	3.65	−0.2383	−0.0005
4	2	60°	2.57	3.22	3.31	−0.2239	−0.0267
5	2	60°	2.45	3.07	3.15	−0.2220	−0.0236
10	2	60°	2.20	2.87	2.87	−0.2339	−0.0015
24	2	60°	2.10	2.79	2.75	−0.2360	0.0137
3	2	75°	3.18	3.46	3.36	−0.0528	0.0307
4	2	75°	2.78	2.90	2.98	−0.0677	−0.0263
5	2	75°	2.57	2.69	2.78	−0.0767	−0.0352
10	2	75°	2.20	2.39	2.46	−0.1053	−0.0277
24	2	75°	2.06	2.27	2.31	−0.1065	−0.0154

Table 5 is for the case where ν_1, ν_2 = 3, 4, 5, 10 and 24 and the difference between $u(x_1)$ and $u(x_2)$ is small ($\theta = 30°$ or 60°). The minimum and maximum values of the absolute relative error $|\Delta k_p(\text{W})|$ are 1.63% and 16.00%, respectively.

Table 4. Relative errors Δk_p(W) and Δk_p(B) for (ν_1, ν_2, θ), where ν_1, $\nu_2 = 3, 4, 5, 10, 24$, $\nu_1 \geqslant \nu_2$; and $\theta = 15^\circ$ and 75°. Comparison for (ν_1, ν_2, θ), where $\nu_1 < \nu_2$ is given by the comparison for (ν_2, ν_1, $90^\circ - \theta$).

ν_1	ν_2	θ	k_p(W)	k_p(B)	k_p	Δk_p(W)	Δk_p(B)
3	3	15°	3.18	3.39	3.19	−0.0027	0.0639
4	3	15°	3.18	3.36	3.15	0.0106	0.0660
5	3	15°	3.18	3.34	3.13	0.0155	0.0670
10	3	15°	3.18	3.33	3.12	0.0210	0.0676
24	3	15°	3.18	3.32	3.11	0.0230	0.0677
4	4	15°	2.78	2.77	2.77	0.0016	−0.0001
5	4	15°	2.78	2.76	2.75	0.0082	0.0009
10	4	15°	2.78	2.74	2.73	0.0155	0.0010
24	4	15°	2.78	2.73	2.73	0.0181	0.0009
5	5	15°	2.57	2.53	2.56	0.0026	−0.0131
10	5	15°	2.57	2.51	2.54	0.0116	−0.0126
24	5	15°	2.57	2.50	2.53	0.0148	−0.0127
10	10	15°	2.20	2.19	2.22	−0.0099	−0.0142
24	10	15°	2.20	2.18	2.21	−0.0059	−0.0145
24	24	15°	2.05	2.05	2.06	−0.0049	−0.0072
3	3	75°	3.18	3.39	3.19	−0.0027	0.0639
4	3	75°	2.78	2.82	2.82	−0.0140	0.0007
5	3	75°	2.57	2.60	2.63	−0.0211	−0.0109
10	3	75°	2.20	2.29	2.31	−0.0480	−0.0087
24	3	75°	2.06	2.16	2.16	−0.0492	0.0008
4	4	75°	2.78	2.77	2.77	0.0016	−0.0001
5	4	75°	2.57	2.55	2.58	−0.0044	−0.0135
10	4	75°	2.20	2.23	2.27	−0.0295	−0.0146
24	4	75°	2.06	2.10	2.12	−0.0295	−0.0068
5	5	75°	2.57	2.53	2.56	0.0026	−0.0131
10	5	75°	2.20	2.22	2.25	−0.0209	−0.0144
24	5	75°	2.06	2.08	2.10	−0.0202	−0.0071
10	10	75°	2.20	2.19	2.22	−0.0099	−0.0142
24	10	75°	2.05	2.06	2.07	−0.0097	−0.0072
24	24	75°	2.05	2.05	2.06	−0.0049	−0.0072

Table 5. Relative errors Δk_p(W) and Δk_p(B) for (ν_1, ν_2, θ), where ν_1, $\nu_2 = 3, 4, 5, 10, 24$, $\nu_1 \geqslant \nu_2$; and $\theta = 30^\circ$ and 60°. Comparison for (ν_1, ν_2, θ), where $\nu_1 < \nu_2$ is given by the comparison for (ν_2, ν_1, $90^\circ - \theta$).

ν_1	ν_2	θ	k_p(W)	k_p(B)	k_p	Δk_p(W)	Δk_p(B)
3	3	30°	2.78	3.39	3.23	−0.1391	0.0527
4	3	30°	2.78	3.25	3.09	−0.1009	0.0526
5	3	30°	2.57	3.20	3.03	−0.1505	0.0577
10	3	30°	2.57	3.14	2.94	−0.1262	0.0665
24	3	30°	2.57	3.11	2.91	−0.1175	0.0687
4	4	30°	2.45	2.77	2.78	−0.1195	−0.0026
5	4	30°	2.45	2.71	2.72	−0.0994	−0.0013
10	4	30°	2.45	2.64	2.63	−0.0689	0.0041
24	4	30°	2.45	2.61	2.59	−0.0567	0.0060
5	5	30°	2.31	2.53	2.56	−0.0999	−0.0124
10	5	30°	2.31	2.45	2.47	−0.0668	−0.0085
24	5	30°	2.31	2.42	2.44	−0.0534	−0.0071
10	10	30°	2.12	2.19	2.22	−0.0438	−0.0116
24	10	30°	2.12	2.16	2.18	−0.0267	−0.0100
24	24	30°	2.02	2.05	2.06	−0.0163	−0.0053
3	3	60°	2.78	3.39	3.23	−0.1391	0.0527
4	3	60°	2.45	2.94	2.91	−0.1600	0.0093
5	3	60°	2.36	2.77	2.76	−0.1420	0.0058
10	3	60°	2.18	2.55	2.50	−0.1278	0.0193
24	3	60°	2.07	2.45	2.38	−0.1279	0.0321
4	4	60°	2.45	2.77	2.78	−0.1195	−0.0026
5	4	60°	2.36	2.59	2.63	−0.0992	−0.0123
10	4	60°	2.16	2.35	2.37	−0.0888	−0.0089
24	4	60°	2.06	2.25	2.25	−0.0855	−0.0007
5	5	60°	2.31	2.53	2.56	−0.0999	−0.0124
10	5	60°	2.14	2.28	2.31	−0.0715	−0.0126
24	5	60°	2.05	2.18	2.19	−0.0635	−0.0059
10	10	60°	2.12	2.19	2.22	−0.0438	−0.0116
24	10	60°	2.03	2.08	2.10	−0.0303	−0.0066
24	24	60°	2.02	2.05	2.06	−0.0163	−0.0053

The minimum and maximum values of the absolute relative error $|\Delta k_p(\mathrm{B})|$ are 0.07% and 6.87%, respectively. In the cases considered, $|\Delta k_p(\mathrm{B})|$ is substantially less than $|\Delta k_p(\mathrm{W})|$.

Table 6 is for the case where ν_1, $\nu_2 = 3$, 4, 5, 10 and 24 and $u(x_1)$ and $u(x_2)$ are equal ($\theta = 45^\circ$). The minimum and maximum values of the absolute relative error $|\Delta k_p(\mathrm{W})|$ are 2.21% and 24.57%, respectively. The minimum and maximum values of the absolute relative error $|\Delta k_p(\mathrm{B})|$ are 0.02% and 6.02%, respectively. In the cases considered, $|\Delta k_p(\mathrm{B})|$ is substantially less than $|\Delta k_p(\mathrm{W})|$.

The relative errors displayed in tables 1–6 for $p = 95\%$ show that generally the correction factors k_p(B) determined from the approximate normal distribution based on a Bayesian uncertainty are more accurate than the correction factors k_p(W) determined from the approximate t-distribution based on the W–S formula. We conclude that the approximate normal distribution based on a Bayesian uncertainty is both simpler and better in this non-trivial metrology application.

Table 6. Relative errors Δk_p(W) and Δk_p(B) for (ν_1, ν_2, θ), where ν_1, $\nu_2 = 3, 4, 5, 10, 24$, $\nu_1 \geqslant \nu_2$; and $\theta = 45^\circ$. Comparison for (ν_1, ν_2, θ), where $\nu_1 < \nu_2$ is given by the comparison for (ν_2, ν_1, $90^\circ - \theta$).

ν_1	ν_2	θ	k_p(W)	k_p(B)	k_p	Δk_p(W)	Δk_p(B)
3	3	45°	2.45	3.39	3.24	−0.2457	0.0465
4	3	45°	2.45	3.10	3.01	−0.1876	0.0289
5	3	45°	2.36	2.99	2.90	−0.1838	0.0335
10	3	45°	2.26	2.86	2.72	−0.1680	0.0508
24	3	45°	2.23	2.80	2.64	−0.1573	0.0602
4	4	45°	2.31	2.77	2.79	−0.1726	−0.0054
5	4	45°	2.31	2.65	2.68	−0.1379	−0.0079
10	4	45°	2.20	2.50	2.50	−0.1189	0.0002
24	4	45°	2.16	2.44	2.42	−0.1077	0.0064
5	5	45°	2.23	2.53	2.57	−0.1313	−0.0135
10	5	45°	2.16	2.37	2.39	−0.0961	−0.0097
24	5	45°	2.12	2.30	2.31	−0.0831	−0.0046
10	10	45°	2.09	2.19	2.22	−0.0583	−0.0107
24	10	45°	2.05	2.12	2.14	−0.0410	−0.0073
24	24	45°	2.01	2.05	2.06	−0.0221	−0.0043

7. Parameters for comparing the two approximate distributions by numerical methods

A thorough investigation of the relative performance of the approximate t-distribution based on the W–S formula as suggested in the ISO-GUM and the simpler approximate normal distribution based on a Bayesian uncertainty as proposed in this paper would require comparison for a range of measurement equations by numerical methods. Suppose Y is equal to $\sum_i c_i X_i$, for $1, \ldots, N$, where $X_1, \ldots, X_n$ have the Type A distributions, and $X_{n+1}, \ldots, X_N$ have the Type B distributions. A number of values for n and a number of values of N would have to be considered for $n = 1, \ldots, N$ and $N = 2, 3, \ldots$. One would have to consider various negative and positive sensitivity coefficients $c_1, \ldots, c_N$ associated with the variables $X_1, \ldots, X_n$ and $X_{n+1}, \ldots, X_N$. A useful comparison may assume that the expected value and standard deviation of X_i, for $i = 1, \ldots, n$, are determined from m_i independent normally distributed measurement and $X_{n+1}, \ldots, X_N$ have rectangular distributions. For each Type A variable X_i one would consider various values for the number m_i of independent measurements with normal sampling distributions. For each X_i one would consider various values of the standard deviations σ_i, for $i = 1, \ldots, n$. For each X_j one would consider a number of rectangular distributions of various widths for $j = n + 1, \ldots, N$. The required coverage probability p may be set as 95% or several different values of p may be considered. Numerical simulation would be required for each combination of the aforementioned parameters. The number of simulations for each combination of parameters should be sufficiently large for a meaningful difference between the approximate coverage factors k_p determined from the two approximate distributions for Y. Numerical simulation would indicate the accuracies of the coverage factors determined from each of the two approximate distributions.

8. Conclusion

The ISO-GUM interprets the Type A evaluations (commonly determined from sampling theory) as parameters of state-of-knowledge probability distributions. It has previously been shown that the ISO-GUM's interpretation is justified when the Type A evaluations are either determined from Bayesian statistics or are regarded as approximations to Bayesian estimates determined from sampling theory [2]. In certain disciplines, uncertainty is traditionally expressed as an interval about an estimate for the value of the measurand. Expression of uncertainty as an interval with a stated (supposed) coverage probability requires a description of the probability distribution for the value of the measurand. The ISO-GUM suggests that under certain conditions, motivated by the central limit theorem, the coverage factor of an uncertainty interval that yields the required coverage probability may be determined from an approximate t-distribution with effective degrees of freedom obtained from the W–S formula. As a sequel to [2], this paper proposes an approximate normal distribution based on a Bayesian uncertainty as an alternative to the approximate t-distribution based on the W–S formula. The use of an approximate normal distribution based on a Bayesian uncertainty greatly simplifies the expression of uncertainty by eliminating altogether the need for calculating effective degrees of freedom from the W–S formula.

When the measurand is the difference between two means, each evaluated from independent normally distributed measurements, the probability distribution for the value of the measurand is known to be a Behrens–Fisher distribution. We compared the accuracy of correction factors determined from the two approximate distributions with respect to the correct coverage factor from the Behrens–Fisher distribution for the coverage probability of 95%. Our investigation shows that in this special case the coverage factor determined from the approximate normal distribution is generally more accurate. This example suggests that perhaps the simpler approximate normal distribution may be preferable to the approximate t-distribution for other measurement equations as well. A thorough investigation of the relative performance of the approximate t-distribution based on the W–S formula as suggested in the ISO-GUM and the simpler approximate normal distribution based on a Bayesian uncertainty as proposed in this paper would require comparison for a range of measurement equations by numerical methods.

The assumptions that underlie the approximate t-distribution based on the W–S formula and the approximate normal distribution based on a Bayesian uncertainty are essentially identical. Unless these assumptions are validated, one is not fully justified in using the approximate distributions. Some tend to use the approximate t-distribution based on the W–S formula as a general rule [23]. To the extent that one is willing to use the approximate t-distribution without validating the underlying assumptions, one should be willing to use the simpler approximate normal distribution.

Acknowledgments

Suggestions by the referees have greatly improved the paper. The following provided useful comments on several earlier drafts of this paper: Bob Frenkel, Ron Boisvert, Bonita Saunders and Al Jones.

References

[1] International Organization for Standardization 1995 *Guide to the Expression of Uncertainty in Measurement* 2nd edn, ISBN 92-67-10188-9

[2] Kacker R N and Jones A T 2003 On use of Bayesian statistics to make the *Guide to the Expression of Uncertainty in Measurement* consistent *Metrologia* **40** 235–48

[3] Taylor B N and Kuyatt C E 1994 Guidelines for evaluating and expressing the uncertainty of NIST measurement results *NIST Technical Note* 1297, US Department of Commerce, National Institute of Standards and Technology http://physics.nist.gov/Document/tn1297.pdf

[4] Stuart A and Ord J K 1987 *Kendall's Advanced Theory of Statistics, Distribution Theory* vol 1 (Oxford: Oxford University Press)

[5] Dietrich C F 1991 *Uncertainty, Calibration, and Probability* 2nd edn (New York: Adam Hilger)

[6] Ballico M 2000 Limitations of the Welch–Satterthwaite approximation for measurement uncertainty calculations *Metrologia* **37** 61–4

[7] Hall B D and Willink R 2001 Does 'Welch–Satterthwaite' make a good uncertainty estimate? *Metrologia* **38** 9–15

[8] Guthrie W G 2003 Should ($T_1 - T_2$) have larger uncertainty than T_1? *TEMPMEKO: Proc. 8th Int. Symp. on Temperature and Thermal Measurements in Industry and Science (Berlin, Germany, 19–21 June 2001)* ISBN 3-8007-2676-9

[9] Mathai A M and Pederzoli G 1977 *Characterizations of the Normal Probability Law* (New York: Wiley)

[10] Welch B L 1947 The generalization of 'Student's' problem when several different population variances are involved *Biometrika* **34** 28–35
Welch B L 1937 The significance of the difference between two means when the population variances are unequal *Biometrika* **29** 350–62

[11] Satterthwaite F E 1946 An approximate distribution of estimates of variance components *Biometrics Bull.* **2** 110–14

[12] Brownlee K A 1960 *Statistical Theory and Methodology in Science and Engineering* (New York: Wiley)

[13] Lee PM 1997 *Bayesian Statistics* 2nd edn (Oxford: Oxford University Press)

[14] Evans M, Hastings N and Peacock B 2000 *Statistical Distributions* 3rd edn (New York: Wiley)

[15] CIPM 1999 Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, http://www1.bipm.org/utils/en/pdf/mra_2003.pdf

[16] Behrens W V 1929 Ein Betrag zur Fehlenberechnung bei wenigen Beobachtungen *Landwirtschaftliche Jahrücher* **68** 807–37

[17] Fisher R A 1935 The fiducial argument in statistical inference *Ann. Eugenics* **6** 391–8

[18] Box G E P and Tiao G C 1973 *Bayesian Inference in Statistical Analysis* (Reading, MA: Addison-Wesley)

[19] Lindley D V and Scott W F 1984 *New Cambridge Elementary Statistical Tables* (Cambridge: Cambridge University Press)

[20] Willink R 2003 On the interpretation and analysis of a degree-of-equivalence *Metrologia* **40** 9–17

[21] Patil V H 1965 Approximation to the Behrens–Fisher distribution *Biometrika* **52** 267–71

[22] Davis A W and Scott A J 1973 An approximation to the k-sample Behrens–Fisher distribution *Sankhyā* B **35** 45–50

[23] Singapore Institute of Standards and Industrial Research 1995 *Guidelines on the Evaluation and Expression of the Measurement Uncertainty* SINGLAS Technical Guide 1