k = 2 AND OTHER SOMETIMES HIDDEN ASSUMPTIONS IN CHEMICAL MEASUREMENT UNCERTAINTY INTERVAL

David L Duewer [1], Stephen LR Ellison [2], William F Guthrie [3], D Brynn Hibbert [4], Craig M Jackson [5], Anders Kallner [6], Stefan D Leigh [3], Reenie M Parris [1], Kenneth W Pratt [1], Michele M Schantz [1], Katherine E Sharpless [1]

[1] Analytical Chemistry Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8390 USA
[2] Analytical Technology, LGC Limited, Teddington, TW11 0LY, UK
[3] Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8980 USA
[4] School of Chemistry, University of New South Wales, Sydney, NSW 2052 Australia
[5] Hemosaga Diagnostics Corp, 5931 Seacrest View Rd, San Diego, CA 92121-4355, USA
[6] Department of Clinical Chemistry, Karolinska Hospital, Stockholm, SE-17176 Sweden
david.duewer@nist.gov

Disclaimer: All authors contributed to the calculations and/or the discussions upon which this work is based. However, the analysis and recommendations presented herein are those of David L. Duewer and may not reflect the views of the other authors or the institutions with which they are affiliated.

1. BACKGROUND

A recent interlaboratory study that required individual analysts to estimate uncertainty intervals for their results revealed that some experienced chemical analysts have difficulty with measurement uncertainty calculations. Several NIST analysts used somewhat similar internal standard chromatography measurement processes to quantify one or more measurands in the study material, a complex lipid-rich powder intended for use as a human nutritional supplement. These analysts successfully developed their own purpose-built "spreadsheet" tools to estimate measurand quantities from their experimental data. However, since these analysts routinely obtain expert statistical assistance when estimating measurement uncertainties, they did not themselves have the expertise for estimating uncertainty intervals for their quantity estimates.

The measurement equations underlying the spreadsheet calculations used for determination of the concentration of each measurand were identified. While differing in detail, all could be expressed as the product of a series of independent terms. A software tool appropriate to this form of measurement equation was developed to assist the analysts. To help validate the assumptions and calculations implemented in the measurement evaluation tool, the following model problem was constructed and posed to a number of chemical and statistical analysts with interest and expertise in the evaluation of measurement uncertainty.

The data that inform the problem are derived from one of the measurement systems used in the interlaboratory study.

2. MODEL PROBLEM

What is the 95% expanded uncertainty interval for the result of the measurement equation

$$\text{Result} = A \times B \times C$$

given the following:

Data for *A*: Six Independent Sets of Duplicate Measurements

Set	A_{i1}	A_{i2}
A1	2.476	2.540
A2	2.423	2.524
A3	2.334	2.422
A4	2.425	2.378
A5	2.444	2.498
A6	2.422	2.466

Data for *B*: Four Independent Measurements

Set	B_i
B1	0.968
B2	0.967
B3	0.973
B4	0.967

$C = 6.864$

For "Extra Credit": What is the 95% expanded uncertainty for the result if you regard the *A* values as 12 independent measurements rather than 6 sets of duplicates?

3. RESULTS

Several participants in the validation study provided more than one set of results, using either different mathematical approaches or different assumptions regarding the nature of the data. The results can be summarized (with caveats discussed below) as follows:

Summary of Results

		6 Duplicates			12 Independent		
Approach	*k*	#	*R*	*U*(*R*)	#	*R*	*U*(*R*)
PoU	2	3	16.26	0.27	5	16.26	0.23
PoU	t_s	4	16.26	0.35	5	16.26	0.25
MC	-	2	16.26	0.43	2	16.26	0.30

PoU	Propagation of uncertainty
MC	Monte Carlo
k	Coverage factor
t_s	Student's t
#	Number of results using same approach
R	Result
U(*R*)	Expanded uncertainty of the result

Caveats

It is somewhat embarrassing, but more than one of us either made minor data transcription or spreadsheet programming errors somewhere along the trail from initial to final result. The above Summary is of the results *after* these errors were corrected. Most of the errors can be attributed to each of us building our own single-use models, either spreadsheets or scripts, to analyze the model data (which itself argues for the development and use of appropriate, fit-for-purpose, and validated software tools for measurement uncertainty estimation).

In addition to evaluating the expanded uncertainty interval for the result of the measurement equation under the assumption that all of the *A* data come from a single measurement process, some of us also addressed: 1) what is the expanded uncertainty of the population of results that could be generated by this system? or 2) what is the expanded uncertainty assuming that each of the six sets of *A* duplicates came from different, independent measurement processes? While these are defensible interpretations of the problem as stated, these results are not included in the above Table.

4. PROPAGATION OF UNCERTAINTY

Most results were derived using the propagation of uncertainty (PoU) approach recommended in the Guide to the Expression of Uncertainty in Measurement (GUM) [1]. For the product of a series of quantities such as *Result* = *A* × *B* × *C*, the PoU model for the combined uncertainty is

$$u_c(\text{Result}) = \text{Result} \times \sqrt{\left(\frac{u(A)}{A}\right)^2 + \left(\frac{u(B)}{B}\right)^2 + \left(\frac{u(C)}{C}\right)^2}$$

where *u*(*A*), *u*(*B*), and *u*(*C*) are the standard uncertainties for the quantities *A*, *B,* and *C*.

Standard Uncertainties

For the model problem, all of us assumed that the standard uncertainty for the constant *C*, *u*(*C*), is zero and that the expected value for *B* is best estimated as the simple mean of the four values, $\overline{B}$, with the standard deviation of the mean, $u(\overline{B})$, as its standard uncertainty:

$$\overline{B} = \sum_{i=1}^{4} B_i \Big/ 4; \quad u(\overline{B}) = \sqrt{\sum_{i=1}^{4} \frac{(B_i - \overline{B})^2}{3}} \Big/ \sqrt{4}$$

The evaluation of the *A* data was less uniform. The main part of the model problem specifies that there are six independent sets of duplicate results but it does **not** specify the degree of dependence between the duplicates. Following a formal or informal analysis of variance, several of us concluded that these data are best considered to be

12 independent values. The expected value and its standard uncertainty for A are again the simple mean and the standard deviation of the mean, $\overline{A}$ and $u(\overline{A})$:

$$\overline{A} = \sum_{i=1}^{12} A_i \Big/ 12; \quad u(\overline{A}) = \sqrt{\sum_{i=1}^{12} \frac{(A_i - \overline{A})^2}{11}} \Big/ \sqrt{12}$$

Under this assumption, the main and "extra credit" parts of the model problem are identical.

The rest of us chose to estimate $\overline{A}$ and $u(\overline{A})$ using a within- and between-sample variance components model [2]. This model was evaluated using both purpose-built and commercial single-factor analysis of variance software; the two approaches yielded identical results.

Since the two replicate measurements are given for all six samples, the expected value is the same $\overline{A}$ as above. The within-sample standard deviation, s_{within}, for the data can be estimated by pooling the standard deviation estimates for the six sets of duplicates

$$s_{\text{within}} = \sqrt{\frac{\sum_{i=1}^{6}\left(\sum_{j=1}^{2}(A_{ij} - \overline{A}_i)^2\right)}{6}}; \quad \overline{A}_i = \frac{A_{i1} + A_{i2}}{2}$$

The among-sample standard deviation, s_{among}, can be estimated from the standard deviation of the sample averages corrected for the within-sample variance that "leaks through" the sample averages

$$s_{\text{among}} = \sqrt{\max\left(0, \frac{\sum_{i=1}^{6}(\overline{A}_i - \overline{A})^2}{5} - \frac{s_{\text{within}}^2}{2}\right)}$$

where "max(x,y)" is the function "take the maximum of the values x and y" and is used to ensure that s_{among} is non-negative. The standard uncertainty for the expected value, $u(\overline{A})$, is then calculated as the combination of the two variances divided by their respective degrees of freedom

$$u(\overline{A}) = \sqrt{\frac{s_{\text{among}}^2}{5} + \frac{s_{\text{within}}^2}{12-5}}$$

Combining the Standard Uncertainties

Most of the reported PoU combined uncertainties, u_c(Result), were evaluated from the standard uncertainties using the explicit GUM model given at the top of this section. However, two of us used Kragten-style numeric approximations [3],[4]. To two significant digits, the values from the two approaches are equivalent. Comparison of the results from the two approaches helps to validate both.

Expanded Uncertainty

The expanded uncertainty for the result, U(Result), is the product of the combined uncertainty and a coverage factor, k:

$$U(\text{Result}) = k \times u_c(\text{Result})$$

where k is chosen such that the interval Result ± U(Result) is expected to enclose a "true value" of the measurand with some specified level of confidence.

Roughly half of the reported PoU U(Result) values used the conventional k = 2 expansion factor with the remainder defining k from the two-tailed Student's t distribution, t_s(95% confidence, number of degrees of freedom). The number of effective degrees of freedom, ν_{eff}(Result), was calculated using the Welch-Satterthwaite approximation [5]

$$\nu_{\text{eff}}(\text{Result}) = \frac{(u_c(\text{Result})/\text{Result})^4}{\frac{(u_c(\overline{A})/\overline{A})^4}{\nu_{\text{eff}}(\overline{A})} + \frac{(u_c(\overline{B})/\overline{B})^4}{\nu_{\text{eff}}(\overline{B})}}$$

where $\nu_{\text{eff}}(\overline{B})$ is the number of effective degrees of freedom for $\overline{B}$ (and is the number of independent measurements minus the number of estimated parameters; here: 4 – 1 = 3) and $\nu_{\text{eff}}(\overline{A})$ is degrees of freedom for $\overline{A}$ (and is 12 – 1 = 11 when assuming that there are 12 independent

measurements. When assuming that there are six independent sets of duplicates, $\nu_{\text{eff}}(\overline{A})$ was similarly estimated from the s_{within} and s_{among} and their degrees of freedom (and rounds down to 5).

5. MONTE CARLO

While computationally intensive Monte Carlo (MC) methods are well established in other contexts [6], their use in the estimation of measurement uncertainties is a relatively new development [7]. The common thread among the diverse MC methods is that they directly estimate the distribution of results for a given set of assumptions and input values rather than the point-estimates provided by the PoU. Empirical Bayesian analysis is a MC-based technique where data are used to inform explicitly defined assumptions about the underlying sampling distribution(s) of the data and, after tossing about many random numbers, provide the expected posterior distribution as its result.

Two of us used the WinBUGS [8] freeware system to evaluate U(Result) for the Result = $A \times B \times C$ measurement equation, using the distributions $B_i \sim \text{Normal}(\beta, \sigma_B^2)$ where β is the true value of B and σ_B^2 is the true variance of the B_i measurements; $A_i \sim \text{Normal}(\alpha, \sigma_A^2)$ where α is the true value of A and σ_A^2 is the true variance of the A_i measurements when the A_i are considered as 12 independent values; and $A_i \sim \text{Normal}(\alpha, \sigma_{\text{among}}^2)$, $A_{ij} \sim \text{Normal}(A_i, \sigma_{\text{within}}^2)$ when they are considered as six independent sets of duplicates. The symbol "~" signifies "is distributed as".

When there is prior information about what the various distributional parameters should be, Bayesian evaluation can update the priors with new data to provide new and improved estimates for the parameters. For the model problem, only the general shape of the variances is known: σ_B^2, σ_A^2, σ_{within}^2, and σ_{among}^2 are all non-negative. Therefore, uninformative priors are used for all parameters, such as: very broad normal distributions centered on zero for the location parameters (β, α, and A_i) and uniform distributions from 0 to some value much larger than the largest plausible value for the variances.

Since the result of MC estimation is a distribution, the "expanded uncertainty" for any level of confidence is directly available from the distribution itself and does not require further computation.

6. ASSUMPTIONS THAT LEAD TO DIFFERENCE

While the 95% expanded uncertainty estimates for the data of the model problem are "roughly equivalent", they do range from 0.23 to 0.43. The smallest estimate comes from traditional PoU analysis assuming that there are 12 independent measurements of the A quantity and that the coverage factor $k = 2$ expands the combined uncertainty to an interval that covers the true value with an approximately 95% level of confidence. The largest estimate comes from the Bayesian MC analysis of the A data as six sets of duplicates, using very uninformative distributional priors. The three assumptions that lead to differences in the results for the model problem are thus: 1) the independence of replicated data, 2) the nature of the distributions from which the data were sampled, and 3) the coverage factor in PoU analysis. While explicit in the calculations for each result, these assumptions are not necessarily accessible to those who may wish to actually make use of the Result ±U(Result) values.

Given the (intentional) absence of information on the nature of the duplicate measurements, regarding all 12 of the A data as independent or as six independent sets of potentially dependent duplicates are both quite defensible assumptions. For the measurement system from which the model data were derived, the "12 independent value" model may be most appropriate since the measurements represent single (independent) injections of two independently prepared aliquots of six nominally identical units of the study material. However, the "6 sets of dependent duplicates" model would be appropriate if the values represented duplicate injections from six aliquots. In real life, the choice of models is informed by the experimental procedure and should be made by those familiar with the experiment. However, the nature of this choice is generally no more (if no less) relevant to the consumers of measurement results than are other experimental details.

Given that the use of MC techniques is relatively novel in chemical metrology and that their results can be strongly influenced by fairly subtle differences in the choice of the uninformative priors, measurement consumers may well need to be broadly aware that MC techniques were used.

The $k = t_s$ coverage factor reduces to the conventional $k = 2$ when the number of degrees of freedom is large. For the model problem, $t_s(95\%, v_{eff}=11)$ is about 2.20 and accounts for the 10% difference in PoU results when the 12 *A* values are regarded as fully independent. When the *A* values are considered as six independent sets of duplicates, $t_s(95\%, v_{eff}=5)$ is about 2.57 and accounts for the nearly 30% difference in the PoU results. Given that the MC 95% uncertainty intervals are 15 % to 20 % wider than the $k = t_s$ results, the conventional $k = 2$ does not appear to here provide a credible 95 % level of confidence coverage interval. Why then is the $k = 2$ factor used?

One obvious answer is that calculation of the effective number of degrees of freedom can be difficult for even simple measurement equations. There is also debate as to the appropriateness of the Welch-Satterthwaite estimation procedure [9],[10],[11]. However, the use of $k = 2$ has in some circles become institutionalized:

> "To be consistent with current international practice, the value of 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." [12]

7. MAKING THE "HIDDEN" ASSUMPTIONS LESS SO

The signatories to the International Committee for Weights and Measures (CIPM) Mutual Recognition Arrangement (MRA) bound themselves in 1999 to the transparent reporting of measurement uncertainty

> "Uncertainties are evaluated at a level of one standard uncertainty and information must be given on the number of effective degrees of freedom, required for a proper estimate of the level of confidence." [13]

and to the use of expanded uncertainty intervals having 95 % levels of confidence

> "The degree of equivalence of each national measurement standard is expressed quantitatively by two terms: its deviation from the key comparison reference value and the uncertainty of this deviation (at a 95 % level of confidence)." [14]

The use of the conventional $k = 2$ coverage factor will **not** give expanded uncertainty estimates providing coverage at the 95 % level of confidence for relatively small numbers of effective degrees of freedom. Without a separate and explicit statement of the effective number of degrees of freedom associated with the combined uncertainty, $k = 2$ does not enable estimation of the true level of confidence. Thus, for PoU methods, reporting an expanded uncertainty is at best redundant since knowledge of both u_c and v_{eff} are required for proper interpretation. However, MC methods directly estimate uncertainty intervals (for all desired levels of confidence) without explicitly estimating either u_c or v_{eff}.

The consumer of measurement results thus needs more information than is provided by "Result ± *U*(Result). For PoU methods, the "hidden assumptions" could made more visible by use of the alternate but GUM-defined symbol U_p:

> "Expanded uncertainty of output estimate *y* that defines an interval $Y = y \pm U_p$ having a high, specified level of confidence *p*, equal to coverage factor k_p times the combined standard uncertainty $u_c(y)$ of *y*: $U_p = k_p u_c(y)$" [1].

An uncertainty estimate providing approximately a 95 % level of confidence would thus be designated: U_{95}. This could be extended to encompass use of the conventional coverage factor: $U_{k=2}$. When accompanied by an explicit statement of the associated effective degrees of freedom, the approximate statistical level of confidence provide by the estimate can be evaluated and/or the estimate can be adjusted to provide a desired level of confidence. When there is no statement of the effective degrees of freedom, use of $U_{k=2}$ would serve to warn the user that the estimate may not provide the nominal 95 % level of confidence. When uncertainty is estimated using some MC method, the symbol could be further extended, perhaps to: U_{MCp}.

REFERENCES

[1] ISO. Guide to the expression of uncertainty in measurement. ISO, Geneva, Switzerland (1995)

[2] ISO. ISO 5725-2 Accuracy (trueness and precision) of measurement methods and results. Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method. ISO, Geneva, Switzerland (1994)

[3] Kragten J, Calculating standard deviations and confidence-intervals with a universally applicable spreadsheet technique. Analyst 1994:119(10):2161-2165.

[4] Kragten J, A standard scheme for calculating numerically standard deviations and confidence-intervals, Chemometrics And Intelligent Laboratory Systems 1995;28(1):89-97.

[5] Taylor BN, Kuyatt CE. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, NIST Technical Note 1297. NIST, Gaithersburg, MD 20899 USA (1994). http://physics.nist.gov/Pubs/guidelines/contents.html

[6] Diaconis P, Efron B. Computer-Intensive Methods In Statistics. Scientific American 1983;248(5):116+.

[7] Cox MG, Siebert BRL. The use of a Monte Carlo method for evaluating uncertainty and expanded uncertainty. Metrologia 2006;43:S178–S188.

[8] WinBUGS. Imperial College and the Medical Research Council, UK. http://www.mrc-bsu.cam.ac.uk/bugs/

[9] Ballico M. Limitations of the Welch-Satterthwaite approximation for measurement uncertainty calculations. Metrologia 2000;37(1):61-64.

[10] Hall BD, Willink R. Does "Welch-Satterthwaite" make a good uncertainty estimate? Metrologia 2001;38(1):9-15.

[11] Guthrie WF. Should (T1- T2) Have Larger Uncertainty Than T1?. Proceedings of the 8th International Conference on Temperature: Its Measurements and Control, Volume 2, pp. 887-892. http://www.itl.nist.gov/div898/pubs/author/guthrie/guthrie-2002-01.pdf

[12] NIST Quality Manual for Measurement Services QM-1, Appendix C, NIST, Gaithersburg, MD 20899 USA (2005). http://ts.nist.gov/qualitysystem/#qm-i

[13] Guidelines for CIPM key comparisons, Section 6. CIPM MRA, Paris, France (1999, Revised 2003). http://www.bipm.org/utils/en/pdf/guidelines.pdf

[14] CIPM revision of the technical supplement to the arrangement, Section T.2. CIPM MRA, Paris, France(2003). http://www.bipm.org/utils/en/pdf/mra_techsupp2003.pdf