APPLICATION OF THE *GUIDE TO THE EXPRESSION OF UNCERTAINTY IN MEASUREMENT* AT THE INDUSTRIAL LEVEL

W. Bich
CNR - IMGC, Istituto di Metrologia «G. Colonnetti»,
Str. delle cacce, 73, 10135 Torino, Italia
Tel: + 39 011 3977 445 – Fax: + 39 011 3977 437 – email: w.bich@imgc.to.cnr.it

Abstract: The *Guide to the Expression of Uncertainty in Measurement* is reviewed and its main features are discussed. The content of innovation of the Guide is emphasised, and the multivariate case is addressed. Application to low-end measurements is sketched. A final part is devoted to future developments in the field of uncertainty.

INTRODUCTION

The rapidly increasing diffusion of quality systems in the industrial world has raised the need of evaluating measurement uncertainty. This need was already well recognised at the scientific level, although until 1993 its fulfilment was by no means generally accepted.

The reference document in this field is, since 1993, the *Guide to the Expression of Uncertainty in Measurement* [1], also known as the *GUM*, published by ISO in the name of an impressive list of international organisations, namely: the International Electrotechnical Commission (IEC), the International Federation of Clinical Chemistry (IFCC), the International Organisation for Standardisation (ISO) itself, the International Union of Pure and Applied Chemistry (IUPAC), the International Union of Pure and Applied Physics (IUPAP) and the International Organisation of Legal Metrology (OIML). The *GUM* states precise rules for assigning the uncertainty in a wide range of experimental situations, and therefore also in the industrial context.

However, its application in many cases may result difficult, and special guidance supplementary documents may be necessary. As a matter of fact, a number of such documents has been and is being developed by various institutions.

THE MEASUREMENT MODEL

The measurement model we adopt is such that the quantity *Y* to be measured, *the measurand*, is supposed constant during measurement. This condition is by no means mandatory, but it simplifies analysis.

Indeed, in many cases measurands are subject to time evolution, which can be both random (noise) and/or deterministic (drift). In such cases, uncertainty treatment should be appropriately extended in order to take these effects into account.

We assume that the measurand itself is not directly observed, but rather measured by observing *n* other quantities X_i, on which the measurand depends according to a function

$$Y = f(X_1, X_2, \ldots X_n) \ . \qquad (1)$$

This scheme is sometimes called the *indirect measurement* and encompasses most of the experimental situations. For example, even the simple case of direct reading of an instrument is described in terms of Eq. 1. The output reading *X* depends on the value of the measurand *Y* at the input of the instrument and on the transfer characteristic of the instrument according to a functional relationship $X = g(Y)$, so that, inverting it, one may write

$$Y = f(X,S) = \frac{X}{S} \ , \qquad (2)$$

in which *S*, the *sensitivity*, is defined as the reciprocal of the slope of $X = g(Y)$.

The observed quantities X_i, i.e., the independent variables in Eq. 1, are the *input quantities*, whereas the measurand *Y* is often called the *output quantity*.

RANDOMNESS

In eq. (1), everything is deterministic. Equation (1) describes an exact physical law relating the measurand to a number of other physical input quantities. In the real world, however, we do not know exactly the values of the quantities, only approximate *estimates* being available.

For example, let us consider the well-known geometric relation $C = 2pR$, relating the length of a circle *C* to its diameter *2R*. This is an exact, theoretical relation. Let us suppose now that we wont to use this relation to determine the circle length by means of a measurement, or a number of repeated measurements, of the diameter.

A first difficulty is that no artefact will be a perfect circle, so that deviations from circularity should be taken into account in a physical model. The second is that diameter position(s) are impossible to determine exactly, so that also this approximation should be taken into account. Incidentally, this situation occurs in primary metrology: a possible new definition of the kilogram, based on the Avogadro constant and the molar mass of silicon, implies the knowledge of the volume of a silicon sphere. This is determined by geometric measurements of a number of diameters of the sphere, so that the same difficulties occur in three dimensions.

But let us assume here for the sake of simplicity that both the circle and our ability to determine diameter positions are perfect. We are still faced with the incomplete knowledge of the diameter, due to experimental difficulties. In addition, also our knowledge of the numerical value of π is incomplete. This geometric constant has an infinite number of digits, and calculations so far carried out have determined some 10^8 to 10^9 digits for it [2]. Therefore, beyond that limit a vagueness remains. In fact, the next digit cannot be predicted with certainty. In the *GUM* approach, the estimate available for π is considered a *random* quantity, and as such is affected by an *uncertainty*.

A consequence of this approach, which treats quantities as random, is that statistical techniques can be applied in their entirety, with no need for developing further specific mathematical tools for treating measurement uncertainty.

It is to be remarked that, in a more traditional statistical framework, by no means would π be considered random, but rather an estimate affected by a *systematic*, unknown error, as those coming from a bias in an instrument, or from an imperfect correction, or from an imperfect model. The motivation for this approach is that all these errors, by their very nature, cannot have zero mean. The concept itself of mean can be meaningless for some of these errors.

However, this different approach, widely adopted before the *GUM*, gave rise to a serious difficulty, since systematic errors, being non-statistical quantities, needed different tools. The problem is that the proposed tools were not unique and often ambiguous, thus generating different and generally questionable solutions.

In the *GUM* approach, the solution adopted lies in a broader interpretation of randomness: random is whatever cannot exactly be known, or predicted, and the probability of an event quantifies the degree of belief in its occurrence.

Therefore, developing the above expressed ideas, It is assumed that each X_i is estimated by a corresponding *input variable* x_i, which is considered a random quantity, related to X_i by

$$x_i = X_i + e_i \quad . \tag{3}$$

In eq. (3), the term $å_i$ represents an additive, random error subject to the condition of zero mean. This implies that the constant-valued physical quantity X_i and the random input variable x_i differ by a random amount, which should be, in average, zero. To put it in formulas,

$$\overline{x}_i = X_i \tag{4}$$

where the notation $\overline{x}$ indicates the mean of x. Another notation for the mean, or *expectation*, is $E(x)$.

Therefore, the random outcome of an observation is considered as the sum of the constant value of the quantity and an unknown random error with zero mean (or, equivalently, zero expectation). A concise way to express this feature is that input estimates have to be *unbiased*.

THE MEASURAND ESTIMATE

The y value resulting from a set of input estimates x_i,

$$y = f(x_1, x_2, ...x_n) \ , \tag{5}$$

is the *output estimate*. It is obviously a random quantity, and it can be shown that, if input quantities x_i are unbiased estimates of the corresponding quantities X_i, that is, if condition (4) holds, the output quantity y is au unbiased estimate of the measurand Y.

This is an important result, since it ensures that if our observations are "good", the resulting estimate of the measurand will accordingly be "good".

UNBIASEDNESS

The unbiasedness condition has consequences on the measurand estimate itself. In fact, the requirement of zero-mean errors has to be re-interpreted in subjective terms. The estimate is the “best” estimate not in absolute terms, but simply within the knowledge of the experimenter. Accordingly, since the knowledge of the experimenter is usually based not only on “objective” data, coming from the observation of the physical world, but also

on “subjective” judgement, such as that sometimes adopted for evaluating a systematic effect, these two components should both contribute to the estimate. A good example is the so-called cosine error, that is, the error arising, for example, from misalignment of a reference length-measuring device with respect to the length to be measured. This results in an overestimation, whose correction is frequently based on subjective evaluation. Another example is a reference gas in a bottle, accompanied by a composition certificate stating that the molar fraction ($1-\chi_{ref}$) of contaminants is lower than a given value. Also in this case, a correction is needed to the nominal (1) or effective (χ_{ref}) molar fraction, in order to avoid overestimation or underestimation, respectively. In the *GUM* approach, these correction are necessary.

The need for such procedures is far from being fully realised in the metrological community. Most of us are reluctant to correct an estimate for an effect which is judged to influence it, but about which objective data are not available. It is often preferred to include the subjective evaluation as an uncertainty component. This practice should be avoided.

STANDARD UNCERTAINTY

As we said, the advantage of the *GUM* approach is that well-established statistical tools can be adopted. The basic idea of the *GUM* is to identify the uncertainty of measurement of a quantity x, $u(x)$ with the standard deviation $\sigma(x)$ of the estimate of that quantity.

$$u(x) \equiv s(x). \tag{6}$$

Measurement uncertainty defined in such a way is called by analogy *standard uncertainty*. Standard deviation is a well-known measure of the dispersion of a random variable. It is defined as the positive square root of the variance $V(x)$:

$$s(x) = \sqrt{V(x)}, \tag{7}$$

where

$$V(x) = E\left[\left(x - \bar{x}\right)^2\right]. \tag{8}$$

This is by no means a new idea, as the old error theory, dating back to Gauss, is based on the very same idea.

Therefore, to each of the x_i input estimates, a corresponding standard uncertainty $u(x_i)$ has to be associated. The *GUM* contains a practical guidance to evaluation of these input uncertainties. In a nutshell, two different ways can be followed. If a sample is available, for example a number of repeated, independent observations, the appropriate statistical tools can be adopted, typically (but not exclusively) the *experimental standard deviation*. If on the contrary no sample is available, such as in the case of a single observation, or a value taken from a textbook, the corresponding uncertainty cannot be evaluated from sample statistics, and different methods must be used. The former way is known as Type A evaluation, the latter as Type B.

To briefly comment on the two different input uncertainty evaluations, it is worth remarking that this distinction has no special importance. It has to be intended just as a useful classification of the different possible evaluation methods. In particular, it would be a mistake to associate type A evaluation with random errors, and type B with systematic. It may happen that uncertainty contribution from a random effect is evaluated by a type B evaluation, and *viceversa*.

As concerns type A evaluation, it is understood that the experimental standard deviation is not the only allowed statistic for estimating input uncertainty. Any other recognised statistic can be used, according to the specific application. For example, in the presence of *outliers*, robust estimators can be used instead of the experimental standard deviation. What matters is that the estimated standard uncertainty has the size of a standard deviation. For example, in the analysis of the results of laboratory intercomparisons, in which outliers are likely to exist, the *Median Absolute Deviation* from the median (MAD) is sometimes used, because of its better resistance to outliers. However, since it is a different dispersion measure, it needs to be suitably normalised in order to be comparable with standard deviation [3].

A further comment concerns the fact that type A evaluations are not necessarily better than type B evaluations. Especially when the sample is small, a type B evaluation based on previous data can be more reliable than a type A evaluation, provided that the experimental process is under statistical control.

OUTPUT UNCERTAINTY, ONE MEASURAND

Once the input uncertainties have been estimated, they are combined in the appropriate way to form the *combined standard uncertainty* $u(y)$ of the measurand estimate, namely:

$$u(y) = \sqrt{\sum_{i=1}^{n} c_i^2 u^2(x_i)} \tag{9}$$

in which each *sensitivity coefficient* c_i is the weight of the i-th input estimate on the output, and is given by

$$c_i = \frac{\partial y}{\partial x_i} \ . \qquad (10)$$

There may be some difficulty in the analytical calculation of the sensitivity coefficients. In some cases, especially when the model is complicated or can only be expressed implicitly, it may be convenient to evaluate them by means of numerical techniques, such as Monte Carlo simulation.

THE MULTIVARIATE CASE

The *GUM* does not treat explicitly the multivariate case, that is, the case in which more that one measurand are determined simultaneously by means of a common set of input quantities. The multivariate case occurs when *m* measurands are obtained from a common set of *n* input variables

$$\begin{aligned} y_1 &= f_1(x_1, x_2, ...x_n) \\ y_2 &= f_2(x_1, x_2, ...x_n) \\ &..... \\ y_m &= f_m(x_1, x_2, ...x_n) \end{aligned} \ . \qquad (11)$$

In the multivariate case one cannot ignore the common statistical properties of random variables, that is, their degree of mutual dependence. A good measure for this degree of statistical interdependence between two random variables x_i, x_j is the *covariance* $\mathrm{cov}(x_i, x_j)$, which is defined as

$$\mathrm{cov}(x_i, x_j) = E\left[\left(x_i - \overline{x_i}\right)\left(x_j - \overline{x_j}\right)\right] \ . \qquad (12)$$

Covariance is a generalisation of variance, in the sense that variance can be interpreted as the covariance of a quantity with itself. Covariance is invariant by change of variables, that is,

$$\mathrm{cov}(x_i, x_j) = \mathrm{cov}(x_j, x_i) \ . \qquad (13)$$

As an alternative, the dimensionless *correlation coefficient* r_u may be defined, which is related to covariance by

$$r_{i,j} = \frac{\mathrm{cov}(x_i, x_j)}{u(x_i)u(x_j)} \ . \qquad (14)$$

It can easily be demonstrated that the correlation coefficient can take values between –1 and +1. A correlation coefficient equal to zero means that the two random variables are uncorrelated. If it is equal to +1 or -1, the two variables are linearly dependent.

Covariances may exist between input as well as between output quantities. However, input covariances can be often avoided by an appropriate change in the model. Output covariances, on the contrary, can hardly be avoided, and can exist even when input quantities are uncorrelated.

The general expression giving the covariance $\mathrm{cov}(x_h, x_k)$ between the *h*-th and the *k*-th output estimate has the form

$$\mathrm{cov}(y_k, y_h) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial y_k}{\partial x_i}\frac{\partial y_h}{\partial x_j}\mathrm{cov}(x_i, x_j) \ . \qquad (15)$$

With $h = k$, eq. 15 gives the variance of the relevant output estimate, from which

$$u(y_k) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial y_k}{\partial x_i}\frac{\partial y_k}{\partial x_j}\mathrm{cov}(x_i, x_j)} \ . \qquad (16)$$

Obviously, eq. (16) is also the appropriate equation in the case of one measurand when some of the input quantities are correlated.

More interesting is considering the multivariate case with uncorrelated input quantities. In this case, in the right-hand-term of eq. (15) all covariances vanish, and only variances remain. It can be easily checked that, in this case, eq. (15) simplifies to

$$\mathrm{cov}(y_k, y_h) = \sum_{i=1}^{n} \frac{\partial y_k}{\partial x_i}\frac{\partial y_h}{\partial x_i}u^2(x_i), \qquad (17)$$

which shows that, even with uncorrelated input estimates, correlations may arise between output estimates. This fact is well known, but the metrologist is usually reluctant to believe that *his* particular field needs multivariate treatment. I will sketch here a simple example, showing the influence of the reference standard [4, 5], in the attempt to shake our consolidated habits.

THE CORRELATION DUE TO TRACEABILITY

Let for simplicity assume that two standards of unknown values y_i are calibrated by comparison to one reference standard of value *R* and standard uncertainty u_R. The model (11) has here the simple form

$$\begin{aligned} y_1 &= R - x_1 \\ y_2 &= R - x_2 \end{aligned} \ , \qquad (18)$$

in which x_i are the differences observed between the reference standard and each unknown. Let also assume that the observations are uncorrelated and have the same standard uncertainty u_C. The

appropriate equation for this case is eq. (17), which yields

$$\begin{aligned} u(y_1) &= u(y_2) = \sqrt{u_C^2 + u_R^2} \\ \mathrm{cov}(y_1, y_2) &= u_R^2 \end{aligned} \quad . \qquad (19)$$

Therefore, the two output estimates are correlated even if the input quantities are independent.

In case of subsequent use of the sum of two standards, the appropriate uncertainty of the sum, according to eq. (16), is

$$u(y_1 + y_2) = \sqrt{u^2(y_1) + u^2(y_2) + 2\mathrm{cov}(y_1, y_2)} \qquad (20)$$

which yields

$$u(y_1 + y_2) = \sqrt{2u_C^2 + 4u_R^2} \, . \qquad (21)$$

Neglecting covariances would yield the underestimate

$$u(y_1 + y_2) = \sqrt{2\left(u_C^2 + u_R^2\right)} \, . \qquad (22)$$

From eq. (19), by applying eq. (14), the correlation coefficient between the two output estimates can be calculated:

$$r_{1,2} = 1 - u_C^2 / \left(u_C^2 + u_R^2\right). \qquad (23)$$

This simple example demonstrates the quite obvious fact that the traceability uncertainty is common to both estimates, that is to say, in statistical terms, is their covariance. This basic result holds also for comparison schemes more complicated than that described by eq. (18), such as those involving least-squares adjustments. Actually, since any calibration of a set of artefact standards, such as weights, resistors, capacitors, gauge blocks may be represented by a model such as that in Eq. (18), the presence of covariances between output estimates due to traceability is much more pervasive than it is usually believed.

Often, the traceability contribution to uncertainty u_R is dominant with respect to comparison repeatability u_C, so that $r_{y_{1,2}} \approx 1$. This is, for example, typical for mass standards. In this case, ignoring covariance leads to a dramatic underestimation of the combined uncertainty of the sum.

THE CONSTRUCTION OF A CONFIDENCE REGION

In many applications, specially at the industrial level, a confidence interval for the estimate is required. Also in primary metrology, the Mutual Recognition Arrangement makes explicit reference to a confidence interval having a 95% coverage probability.

The *GUM* is primarily concerned with the estimation of the standard uncertainty, and the problem of constructing a confidence interval having a specified probability level is only addressed in a specific Annex.

A nice feature of standard uncertainty is that, for its evaluation, no assumptions are needed about the statistical distribution of the involved quantities. This is no longer true when a confidence interval is required. Indeed, a confidence interval at a specified probability level implies full knowledge of the probability distribution. In practice, an approximate knowledge of the probability distribution makes it possible to construct an approximate confidence interval, which is sufficient in most application.

In most cases, some knowledge can be claimed on the output distribution. The central limit theorem ensures that, if there is a sufficient number of input components, if these are uncorrelated and if there is no dominant non-gaussian distribution, then the output distribution approaches to a good degree a gaussian, or normal distribution, for which a confidence interval can easily be constructed. For example, a confidence interval having a 95% coverage probability can be obtained by multiplication of the output uncertainty $u(y)$ by a coverage factor $k = 2$ (actually, this is a rounded value generally adopted instead of 1,96).

To be more accurate, it must be noted that the standard uncertainty of the output estimate is only an approximation of the true standard deviation of the corresponding output distribution. This approximation has a reliability, which results from the reliability of each individual input variance. A well-established statistical measure of the reliability of the experimental standard deviation is the *degrees of freedom* ν, defined as $\nu = n - m$, where n is the number of observations in the sample and m is the number of parameters (therefore, for one measurand $m = 1$). This measure can be extended to standard uncertainties arising from type B evaluations, in such a way that every input uncertainty has an associated number of degrees of freedom.

The reliability of the output uncertainty is obtained according to the classical Welch-Satterthwaite formula [6, 7]

$$n_{\text{eff}} = u^4(y) \Big/ \sum_{i=1}^{n} \frac{\left(\frac{\partial y}{\partial x_i} u(x_i)\right)^4}{n_i} \,, \qquad (24)$$

in which n_i is the degrees of freedom of the *i*-th input uncertainty and n_{eff} is the *effective degrees of freedom* of the output uncertainty. The appropriate coverage factor *k* can be obtained as a function of n_{eff} from the table of Student's *t*-distribution.

As a consequence, the practice of simply multiplying the measurand standard uncertainty by a coverage factor $k = 2$ should be discouraged. In fact, This value could yield an underestimated confidence interval. The correct practice involves the assignment of a probability distribution to each of the input quantities and, if appropriate, the evaluation of the effective degrees of freedom of the output uncertainty.

There is a concern about the degrees of freedom of a subjectively estimated input uncertainty. This is a somewhat new concept, whose (subjective) evaluation yields invariably very high values which in some cases tend to dominate in Eq. (24), thus making the effective degrees of freedom of the output largely insensitive to the actual number of observations.

A coverage factor $k = 2$ could also yield an overestimated confidence interval, if for example there is a dominant input uncertainty having uniform (rectangular) distribution. In this case, the output distribution will tend to be trapezoidal, and a coverage factor $k = 2$ could encompass a fraction of the real axis larger than the distribution bounds.

In the general multivariate case, the analytical difficulties inherent in the construction of a confidence region in the space of the output, could be overcome by using computer-intensive methods such as Monte Carlo simulation or bootstrap resampling [8, 9, 10].

COST OPTIMISATION

Cost is an important parameter, especially at the industrial level. On one side, uncertainty evaluation may be perceived as a costly extra operation, for which time should be minimised. On the other side, an incorrect uncertainty evaluation may be even more costly. An approach aiming at a rough, "upper limit" evaluation, just to be on the safe side, might be tempting. The *GUM* deprecates this still widespread practice of “safe” uncertainty evaluation and strongly recommends *realistic* evaluation. This viewpoint should not be misunderstood. Safe confidence intervals are vital in industry, but should be obtained by multiplication of realistic standard uncertainties by “safe” coverage factors *k*, rather than by “safe” standard uncertainty evaluations.

It is worth noting that uncertainty needs not be more accurate than few percent, yet this task may be difficult. The main obstacle is probably the complete modelling of the specific measurement. The experimenter should take every possible care in order to identify and take into account every possible systematic effect. These effects should be enclosed in the measurement model in the form of appropriate corrections. As it has already mentioned, unbiasedness is a severe condition, that cannot be violated, so that unbiased estimates should be adopted for these corrections. Within these limits, some degrees of freedom remain in order to optimise the balance between cost and risk. Highly-demanding applications will require an accurate experimental evaluation of a given correction, with the benefit of an accordingly low uncertainty. For low-end applications a low-cost, approximated correction (perhaps even a correction equal to zero) will be sufficient, at the price of a higher uncertainty.

As an example, let us consider the case of mass comparisons. The most important correction in this case concerns the buoyancy effect, that is, the effect due to aerostatic force. This, for the difference of two standards having volumes V_1 and V_2 respectively, is equal to $r(V_1 - V_2)$. One can choose between measuring the actual volumes by hydrostatic weighing and the actual air density by means of an expensive experimental set-up on one hand, or setting the volume difference equal to zero on the other hand. For two 1 kg stainless-steel standards, the typical standard uncertainty of the correction will be few micrograms in the first option and as much as some milligrams in the second.

The example is useful in that it clearly shows that uncertainty cannot be separated by the estimate. One can choose a rough estimate with an accordingly high uncertainty or a sophisticated estimate with a corresponding small uncertainty. What is not possible, is to have an estimate and to decide *a posteriori* the level of the corresponding uncertainty.

THE FUTURE

The *GUM* is a living document, and as such needs care and maintenance. To these purposes a *Joint Committee for Guides in Metrology* (JCGM) has been created by the BIPM and the organisations supporting the *GUM*, already mentioned in this paper. A further one joined the Committee, that is, the

International Laboratory Accreditation Cooperation (ILAC). Within JCGM two Working Groups have been established. Working Group 1, “Expression of Uncertainty in Measurement”, has the task to promote the use of the *GUM* and to prepare supplemental guides for its broad application. The WG has collected a bibliography on the topic of uncertainty [11]. At the moment, two documents are under preparation. One will deal with the multivariate case. The second will concern the use of numerical techniques, such as Monte Carlo, in the evaluation of uncertainty in those cases in which an analytical treatment such as that presented in the *GUM* would be impossible or too complicated, or would give unacceptably approximated results.

References

[1] ISO, *Guide to the Expression of Uncertainty in Measurement*, Switzerland, First ed. (1993).
[2] S. James Press, *Bayesian Statistics: Principles, Models, and Applications* (Wiley, New York, 1989).
[3] W. Bich, *Estimation and uncertainty in metrology*, to be printed in: *Recent Advances in Metrology and Fundamental Constants, Proceedings of the International school of Physics “Enrico Fermi”, Course CXLVI*, edited by T. J. Quinn, S. Leschiutta and P. Tavella (IOS Press, Amsterdam) 2001
[4] W. Bich, *Metrologia* **27**, 111 (1990).
[5] W. Bich, *Metrologia* **33**, 181 (1996).
[6] B. L. Welch, *Biometrika* **34**, 28 (1947).
[7] F. E. Satterthwaite, *Biometrics Bull*. **2**(6), 110 (1946).
[8] B. F. J. Manly: *Randomization and Montecarlo Methods in Biology* (Chapman and Hall, London, 1991).
[9] E. W. Noreen: *Computer Intensive Methods for Testing Hypotheses* (Wiley, New York, 1989).
[10] M. G. Cox et al., *Use of Monte Carlo Simulation for Uncertainty Evaluation in Metrology*, in *Advanced Mathematical and Computational Tools in Metrology*, Edited by P. Ciarlini, M. G. Cox, E. Filipe, F. Pavese, & D. Richter (World Scientific, Singapore, 2001), pp.93 – 105.
[11] JCGM-WG1: http://www.bipm.org/CC/documents/JCGM/bibliography_on_uncertainty.html