remplace doc préliminaire sans N°

BIPM
F-92310 SEVRES

Rapport BIPM-78/4

Some second thoughts on error statements*

Jörg W. Müller

1. Introduction

For a metrologist, the numerical result of a measurement has hardly any value if it is not followed by a statement indicating the precision (or accuracy) it is supposed to have. Since this has been realized for a long time, most experimentalists have taken the habit of supplementing their data with such an estimate. However, the lack of uniformity in these statements is a constant source of trouble for all those who try to use or compile data.

The absence of a standard "format" may be somewhat reminiscent of the similarly unsatisfactory situation which existed for the mean values before the introduction of a generally accepted system of units. Therefore, this could again be a good subject for some international agreement, although we realize that the problem is quite different and more complex. As a number of standardizing bodies have already advanced conflicting recommendations (or are going to do so), clarifying this situation has become an urgent task.

The basic idea lying behind the present approach to these problems is an attempt to relate the prescriptions for error statements to some elementary physical or statistical facts and to use then the restrictions they impose on the possible structure of acceptable general rules. On the other hand, we deliberately avoid the usual discussion of such items as nomenclature, historical development and philosophical background as they have little or no effect on practical applications.

It is not the objective of the following remarks, which indicate the author's personal opinion (and should therefore in no way be taken as a BIPM view), to impose on the reader some possibly new guidelines for error statements. It is rather to convince him that an open-minded look at some basic problems will be worth-while prior to proposing for general use rules which may, at best, have proved useful in a very limited area. It is hoped,

* The basic ideas developed in this report have been briefly presented on September 20, 1976, at an informal discussion on "Probleme bei der Behandlung systematischer Messunsicherheiten" organized by S. Wagner at the PTB in Braunschweig (Germany). I am deeply indebted to Prof. Wagner for his kind invitation to take part in this meeting.

therefore, that some doubts can be cast into the mind of one or the other, even if this concerns concepts and habits hitherto taken as safe and well established. Where at first sight the choice between various proposals may seem to be a matter of personal preference, some simple arguments of common sense can sometimes facilitate a decision.

Apart from the well-known and strong influence of habit, limited background and taste (which are usually poor guides), the opinion necessarily depends also on the use a person intends to make of an error statement. Here two extreme attitudes may be characterized (although in a necessarily oversimplified way) by the different positions we would probably take when put in the situation of a buyer or a seller of a product*. Our own position is intermediate inasmuch as it just looks for an objective characterization of uncertainty which should be both simple and general.
A typical metrological application we have in mind is, for example, the continuous search for improved fundamental physical constants. Here - and in many other cases - progress is equivalent to detecting previously unknown ("systematic") errors. For a general background, see the relevant discussions at the Gaithersburg Conference [1] .

We shall confine ourselves in what follows to a few basic questions which regularly turn up in this context, as for instance:

- Which quantity should be used to characterize the precision (or accuracy) of a measurement?
- Is there a basic distinction between "random" and "systematic" errors?
- How should they be combined, if at all?

After this lengthy introduction let us now be more specific.

2. Standard deviation versus confidence interval

There exists an impressive body of literature to give us advice on how to determine the uncertainty of a physical quantity from a set of experimental data. The exact way this can be achieved depends on the form in which information concerning the quantity looked for is available. The appropriate techniques are well understood and we shall not be concerned with them here. The final result then usually appears either in the form of an estimated standard deviation (or a multiple of it) or as a confidence interval which is supposed to cover the "true value" with a given probability (e.g. 95%). In both cases such a limit is clearly a useful piece of information on the uncertainty and it seems difficult to advance objective criteria for a preference. A closer look at the way these estimates have been obtained reveals, however, that the basic quantity resulting from the calculation is nearly always the variance

* In a loose sense, this applies also to any kind of calibration service.

(or second central moment). Transforming this value into a confidence interval not only requires some arbitrary choice of the corresponding confidence level, but - and this is often more debatable - relies on some explicit analytic form the uncertainties (or errors) are supposed to follow (usually a Gaussian). In the majority of practical cases, the important question whether this assumption is justified or not cannot be adequately tested for lack of data and the validity of the hypothesis becomes a matter of belief. However, the need for making any such assumption may be doubted, and it is indeed unnecessary for evaluating the variance. We may also recall that, for instance, any application of a t-factor deprives the resulting uncertainty of the desirable property of being an unbiased estimate of the standard deviation (as actually delivered by most formulae). Whereas this may be a minor drawback or none at all for a final result, it is often a major inconvenience for intermediate results which need further processing, as the way to handle them correctly is then much more complicated or even unknown.

In order to further clarify this point, let us go back for a moment to the basic and generally accepted law of error propagation. Let y be a known function F of the n variables x_i. The classical statement then is that small random displacements Δx_i of the variables result for y in a corresponding uncertainty Δy which is given approximately by

$$(\Delta y)^2 \cong \sum_{i=1}^{n} (\frac{\partial F}{\partial x_i} \cdot \Delta x_i)^2 + \sum_{i \neq k} (\frac{\partial F}{\partial x_i} \cdot \Delta x_i) (\frac{\partial F}{\partial x_k} \cdot \Delta x_k) \; .$$

By applying the usual abbreviations for the expectation values, i.e.

$$E \{(\Delta y)^2\} = \sigma_y^2 \, , \qquad E \{(\Delta x_i)^2\} = \sigma_i^2$$

and $E \{\Delta x_i \cdot \Delta x_k\} = \mathrm{Cov}\,(x_i\,,\,x_k) = \sigma_{ik}$,

we arrive at the well-known expression ($F_i \equiv \frac{\partial F}{\partial x_i}$)

$$\sigma_y^2 \cong \sum_{i=1}^{n} (F_i \cdot \sigma_i)^2 + 2 \sum_{i<k} F_i \cdot F_k \cdot \sigma_{ik} \, , \qquad (1)$$

where the quantities σ_i^2 and σ_{ik} now stand for the variances and covariances of the random variables x_i. In practice, the corresponding empirical sample variates (denoted by s) are normally used instead.
A more compact notation would be possible by applying matrix notation, but this will not be needed in what follows. In real situations, the covariances are often not very well known; however, they can usually be

determined experimentally, at least in principle. They vanish if the quantities x_i are mutually independent.

For the sake of simplicity, let us now restrict ourselves to the very special case where $y = x_1 + x_2$, with x_1 and x_2 supposed to be independent of each other. The law of error propagation then simply leads to

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 . \tag{2}$$

This relation now holds rigorously since $y = F(x_1 , x_2)$ is a linear function. It is familiar to everybody as the rule of summing errors quadratically.

A slightly different point of view may be of interest here. Let us recall that the addition of variances is quite a general property of independent random quantities which are added (or subtracted). Let again be

$$y = x_1 + x_2 .$$

If x_1 has the density $f_1(x_1)$, and likewise for x_2, then the density for the sum is known to be given by the convolution

$$f(y) = \int_{-\infty}^{\infty} f_1(y - \alpha) \cdot f_2(\alpha) \, d\alpha , \tag{3}$$

which is usually written in a symbolic shorthand notation as $f = f_1 * f_2$. A similar form exists for the difference $x_1 - x_2$. It is easy to show that (in both cases) the corresponding variances σ^2 are then linked by (2). A relation of this type is obviously still valid if some multiple or fraction of the standard deviation is preferred as this just amounts to replacing σ by $\sigma' = \gamma \sigma$, provided that the same value of γ is taken for both x_1 and x_2.

However, if instead of σ a specific confidence interval δ is chosen which refers to a given probability p, no formula equivalent to (2) exists*; hence in general

$$\delta_y^2 \neq \delta_1^2 + \delta_2^2 . \tag{4}$$

This basic fact seems to be often ignored by physicists. Since we are looking for simple and general relations, this should be a strong argument for using standard deviations rather than confidence intervals as a measure of uncertainty. A simple numerical example illustrating this situation is given in Appendix A.

* The only noteworthy exception is the normal distribution which is self-reproducing and where a given ratio δ / σ implies a certain confidence level.

3. On the concepts of "random" and "systematic" errors

One of the few generally accepted facts concerning the statement of an uncertainty seems to be the need to distinguish clearly between so-called "random" and "systematic" errors. Most students in physics in their first contact with practical laboratory work are advised that these have to be evaluated separately and must not be mixed up. The seemingly good foundation of this prescription is substantiated by some carefully chosen examples rendering it fully convincing. Thus, for instance, measuring a distance with a given (and correctly subdivided) meter bar gives a series of results x_i. Their mean value $\bar{x}$ can be improved by augmenting the number n of measurements since the corresponding standard deviation $s_{\bar{x}}$, which is clearly a random quantity, is proportional to $1/\sqrt{n}$. If it happens, however, that the standard used for the comparison is erroneous for some reason, this implies a certain "systematic" error which cannot be reduced by taking more measurements. A similar situation holds when we use a balance for weighing a piece of metal: any calibration error of the balance (i.e. of the masses used) results in a systematic deviation of all readings.

However, even for such simple situations a somewhat different view is possible, although this might at first sight seem less "natural", namely that calibration errors correspond to a certain transformation of the measurements, or more simply to the use of other units. To illustrate this general idea, let us assume that all the n measurements x_i one has made are possibly shifted and taken in a unit which may be incorrect. Although this is clearly not the most general situation one can imagine, it may be sufficient to explain the principle. If we denote the corresponding "ideal" results (i.e. those one might have got with instruments which are free of "systematic" errors) by y_i, they are related to the ones actually obtained by

$$y_i = \alpha + \beta \cdot x_i \, . \tag{5}$$

Here α is a (common) shift and β adjusts the scale factor. If we now wish to evaluate some function of the "improved" observations y_i, all we have to do is to apply the law of error propagation (1) in order to obtain the corresponding uncertainty. If this function is, for example, just the arithmetic mean, the rearrangements needed for this case can be found in Appendix B.

For the still simpler case where allowance is only made for a possible shift α, hence assuming

$$y_i = \alpha + x_i \, , \tag{5'}$$

the result obtained for the variance of $\bar{y} = \sum_i y_i/n$ is (compare eq. B6)

$$s_{\bar{y}}^2 = s_{\alpha}^2 + \frac{1}{n} \cdot s_x^2 \, , \tag{6}$$

where the variance s_x^2 for a single measurement x_i has been estimated in the usual way by forming $s_x^2 = \sum (x_i - \bar{x})^2/(n-1)$, with $\bar{x} = \sum x_i/n$.

Obviously equation (6) is in no way a surprise. As a matter of fact, it states exactly what we all would have certainly expected in this trivial example, namely that the "total" variance $s_{\bar{y}}^2$ is composed of a "random" contribution s_x^2/n and a "systematic" contribution s_α^2. Nevertheless, the result (6) is of interest as it shows that the various contributions are automatically and correctly (note the factor 1/n for the random part) taken into account by simple application of the rules of error propagation, provided that the initial relations (in our case eq. 5) are conveniently formulated. It is left to the reader to convince himself that this approach also works for more complicated situations. In addition, the example indicates that it is practical to express the "systematic" uncertainty s_α in terms of a quantity which is an estimate of the corresponding (usually unknown) standard deviation since this is what we need in (1); we shall come back to this point later. At the same time it answers the question of how to combine the various contributions in a natural way.

As a matter of fact, the traditional approach would lead to the same contributions as given in (6), but they would have been quoted separately. It has been recommended (for instance in [2]) that "random" and "systematic" errors should always be kept apart as they are not of the same nature. In practice, this rule is rarely followed, especially in more complicated situations. This, it would seem, is not only due to the practical need for an "overall" uncertainty, but might also have deeper reasons. In particular, it presupposes that in any given situation such a classification be unambiguous, a view which is difficult to support. In fact, it is easy to find examples where an error which originally should clearly be considered as "random", becomes "systematic" when the result is applied in another context, and vice versa. It therefore seems that Vigoureux's remark that "one has to remember that some errors are random for one person and systematic for another" ([1], p. 524) must be close to the truth.

To avoid this ambiguity, other people have recently suggested classifying errors on the basis of the method used for their estimation. In this case, one would, for example, talk of uncertainties derived from (repeated) measurements on the one hand and of uncertainties determined by estimation (the latter being known as "guesstimates"). Such "measurable" and "estimated" errors - while permitting, for the simpler practical cases, a definite attribution - by their very definition waive the claim of establishing a qualitative difference. In fact, the generally supposed large difference in the precision of the two types of estimates is rarely justified. This can be seen by an evaluation of the statistical uncertainty which has to be attributed to the numerical value of a standard deviation determined on the basis of a finite number n of measurements. Since a sketch of the relevant calculation is given in Appendix C, we restrict ourselves here to the result valid for a sample taken from a normal population, where the relative random uncertainty of the calculated standard deviation (i.e. the "error of the error") is given by the simple expression $1/\sqrt{2\,(n-1)}$. As can be seen from Table 1, this uncertainty is far from being negligible for realistic values of n.

n	$\frac{1}{\sqrt{2\,(n-1)}}$	n	$\frac{1}{\sqrt{2\,(n-1)}}$
2	71 %	10	24 %
3	50	20	16
4	41	30	13
5	35	50	10

Table 1 - Some numerical values indicating the relative uncertainty of a standard deviation determined from a sample of n measurements (taken from a normal population)

The Table not only shows that indicating an error with say three significant figures is practically never justified (in most cases one would be enough), but it also illustrates - and this is particularly important for the present discussion - that a clear-cut distinction between "measurable" and "estimated" errors based on their different degree of reliability is impossible because such a limit does not exist.

It therefore seems that the traditional view of the completely different nature of "random" and "systematic" errors is at least difficult to support in a quantitative way: each time we try to substantiate an apparently decisive difference, it eludes our grip and becomes shadowy.

Let us add that in most real-life situations (which are usually much more involved than the artificial ones mentioned before) it is practically impossible to subdivide an overall error (or its various contributions, if available), in an unambiguous way into "random" and "systematic" parts. Those who doubt this statement may, for example, try to perform such a separation for the uncertainties attributed to the fundamental physical constants as they result from a least-squares adjustment.

To be clear, we are not suggesting the complete abandonment of the use of the adjectives "random" and "systematic" in connection with uncertainties: they are too deeply rooted in our habits and their use can occasionally be quite practical. However, one should begin to doubt whether this distinction is of a fundamental nature.

4. On the measure for indicating "systematic" errors

For those who accept the view that any subdivision of errors into different classes is, in general, an artificial and unnecessary complication, the question of how to express a "systematic" uncertainty is redundant, for in this case the arguments given above for the "random" parts would obviously be valid here too, favouring thus the general use of a quantity which is an estimate of the corresponding standard deviation.

However, experience shows that some users might still hesitate to adopt this position and for them the question seems to be an important one, as can be judged by the length of discussion devoted to this problem. Essentially, there are two main suggestions. One of them is to use "maximum (possible) limits" which should practically never be exceeded, as implied by the word maximum; the other asks the experimenter to make an attempt at estimating an uncertainty which would correspond as closely as possible to something like a "standard deviation". Since both proposals have some merits and drawbacks, the choice is not quite obvious and calls for some second thoughts.

As for the "maximum limit", the main advantage is its simplicity; in case of doubt it can always be enlarged to become "safer". This also reveals its weakest point, namely the fact that for physical situations it is an ill-defined concept and therefore nearly void of useful information. In some way, characterizing a random quantity by its extreme value is about the poorest possible choice. In this case we are in the region where the distribution depends entirely on the exact shape of its "tail" which is normally not very well known. Obviously, the theory of extremes is a very valid branch of mathematical statistics which has found important applications (see e.g. [3] or [4] for more details), but this fact does not imply that maximum or minimum values (an example of which is the "range") are useful estimators for characterizing a random quantity.

At some stage in the interpretation of the experimental data there normally arises the question of what the given error limits actually can tell us, e.g. in terms of probability. Again one is in the unpleasant situation to confess that one simply cannot tell. The situation becomes still worse when the problem of error propagation is raised. Since the maximum is more a mathematical than a statistical concept, it should be logical - at least for a simple situation as the one discussed in section 2 - to use linear addition. This (and only this) would guarantee that the characteristic of the maximum be maintained. This rule has indeed been applied for quite some time (and occasionally still is), but nowadays most users prefer to replace it by the "geometric" or quadratic addition, not only because this is more in line with the treatment of "random" errors, but primarily because the resulting intervals were found to be unrealistically large (- and who could sell such a product!). In addition, such a practice would run a great risk of hiding the presence of unknown errors. Considering all these unpleasant features of a "maximum error", it seems advisable to choose a more suitable measure, and here the standard deviation offers an obvious alternative.

However, this proposal also raises some problems. Firstly, there may be an objection of a somewhat philosophical nature, namely the question of whether a "systematic" error can be said to have a second moment, as this implies (at least implicitly) that there exists a corresponding probability distribution. At first sight this may look like a rather serious obstacle since no repeated (independent) measurements can be taken for verifying this. But here one should remember that the possibility of performing experimental checks is not a condition for the existence of a probability density. After all, "subjective" and "objective" probabilities [5] are governed in all their essential points by the same basic rules. The fact that we often cannot avoid some degree of arbitrary or subjective judgment is no valid argument for abandoning statistical reasoning, but rather a challenge to incorporate them in the best possible way.

An intermediate way out of the problem has been suggested repeatedly (see for instance [6]). While accepting the notion of a random distribution for "systematic" errors, it proposes for their probability density a definite rectangular form where the limits are identified with the maximum error bounds. This choice looks like an artificial device suggested in desperation; it seems to be taken more seriously by some recent adherents than intended by the inventors. As is well known since the lengthy historical discussions following Bayes' original proposal (reproduced in [7]; compare also [8]), identifying complete ignorance with the hypothesis of a constant probability density (within finite limits, for normalization) leads inevitably to logical contradictions; for a clear and detailed discussion compare [9]. This can be readily seen if we remember that ignorance of x implies also ignorance of any function $f(x)$, but the corresponding densities cannot both be represented by rectangulars. For an explicit example, see e.g. [10]. In addition, the usual claim that such a rectangular density would always be a "pessimistic" (and therefore "safe") hypothesis is doubtful. For this purpose, let us consider for example a quantity (temperature T, say) which oscillates periodically in time t between the limits $\pm s$, thus (compare Fig. 1a)

$$T(t) = s \cdot \cos \omega t , \tag{7}$$

where the mean value has been taken as zero for the sake of simplicity. As the density of T is proportional to the time spent in a given region, or inversely proportional to the speed of the temperature change, we have

$$f(t) = c \cdot \left| \frac{\partial T}{\partial t} \right|^{-1} = c \cdot \left| s\omega \cdot \sin \omega t \right|^{-1} ,$$

where c is a normalizing constant. Since $f(T)$ is expected to be symmetrical, it will be sufficient to consider the range $0 \lesssim t \lesssim \pi/2\,\omega$, where

$$f(T) = \frac{c}{\omega s} \cdot \frac{1}{\sin \omega t} = \frac{c}{\omega s}\left[1 - (\frac{T}{s})^2\right]^{-1/2} .$$

Normalization demands that (putting $T/s = x$)

$$1 = \int_{-s}^{s} f(T)\ dT = \frac{2c}{\omega s}\int_0^1 \frac{s \cdot dx}{\sqrt{1 - x^2}} = \frac{c\pi}{\omega} ,$$

hence $c = \omega/\pi$. The required probability density for the temperature T is therefore

$$f(T) = \frac{1}{s\pi}\left[1 - (\frac{T}{s})^2\right]^{-1/2} , \qquad \text{for } |T| \leqslant s , \tag{8}$$

the behaviour of which is sketched in Fig. 1b. We note in particular that f is infinite for T = $\underline{+}$ s. In this case the supposed rectangular (with limits at $\underline{+}$ s) would be neither a good nor a "safe" substitute.

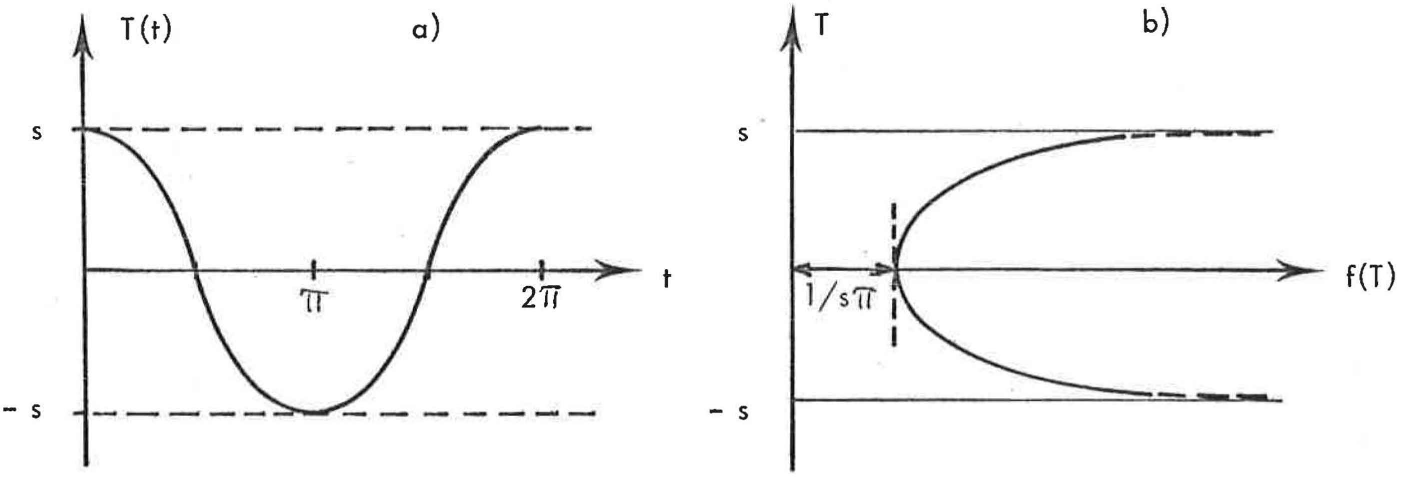

Figure 1 - Schematic behaviour of the density for a temperature T which oscillates periodically (see text).

Another unwanted feature of the rectangular density stems from the fact that multiples of the standard deviation s which are beyond $\sqrt{3}$ are meaningless (as they exceed the limits); one would have to refrain, therefore, from the familiar habit of taking e.g. 2 or 3 times s.

It therefore seems that such a rectangular density for describing "systematic" errors has more pitfalls than merits to offer and is probably best avoided.

But then the question arises by which other density it should be replaced. Our answer would be "by none", and this for the simple reason that such a specific hypothesis is not really needed. All that is in fact required for applying the general error propagation formula is an estimate of s, the corresponding standard deviation. As is well known, the mere demand for the existence of the moments (here of order two) does not specify the corresponding density.

In our opinion it should be the duty of the experimenter to estimate, to the best of his knowledge, such a quantity. As this will by necessity be a rather crude estimate, we should not be too hesitant in slightly modifying the description of the quantity looked for if this can be of any help. Thus, one could for instance demand that the range covered by the interval ± s should contain the "true" value in about 70% of the cases, i.e. with a probability approaching 2/3.

If the estimation of a 50% confidence level is considered to be more practical, such a "probable error" would have to be multiplied by a factor 3/2 to obtain an approximate value of the corresponding standard deviation.

Apparently, some people are not very happy with such a suggestion, and a few even claim that an experimenter is unable to indicate such a quantity, stating that a "maximum limit" (as discussed above) is all he can actually produce. To our mind such a claim should not be taken too seriously. In any case, a look at the literature shows that experienced people in fact can produce such estimates (see, for example, a recent paper from the NPL [11]). We do not say that this is always an easy task, on the contrary, but one should encourage people to make a serious effort to do so. In any case, this seems to be the only way to obtain a quantity which is of real use for further processing.

It is not the place here to discuss methods for evaluating "systematic" errors which are not based on pure guessing; the interested reader will find many of them e.g. in the publications of Youden (a good collection of which can be found in [12] or [13]). In general, estimates based on such indirect methods give also values which correspond to a standard deviation rather than to a "maximum value".

5. On the combination of "random" and "systematic" errors

Here again, there are essentially two schools: those who refuse to combine errors of different "types" at all and the others who do not see a problem at all in doing so. In keeping the errors apart, one immediately comes across the problems mentioned earlier, for example, the fact that in practice a clear-cut separation is often impossible.

In general, however, most people accept the idea of combining them somehow and such a procedure is also often strongly imposed by the practical need for a simple characterization of the total uncertainty. The main discussion therefore is about the exact way this should best be done. Various proposals have been made for this purpose, but most of them look so artificial or are in contradiction with elementary rules that they can be safely discarded. As an example let us have a look at a prescription (actually used in certificates) of the form

$$S_{tot} = \sqrt{\sum (t \cdot S_{rand})^2} + \sum S_{syst} ,$$

where t is a Student-type factor corresponding to a confidence level of 99.7%, say, and S_{rand} a standard deviation, whereas S_{syst} gives the "maximum limit" for a "systematic" error component. Here a number of steps are combined, all of which have been recognized as doubtful in previous sections. Thus not only are confidence levels added quadratically (for the random contributions), in contradiction with (4), but also the total "systematic" error is obtained by adding linearly the various contributions (and likewise in forming the total error). Such a value S_{tot} has no clear statistical (or other) meaning. It is therefore of little practical use and should be avoided. A number of other suggestions can be found which are slight variants of the one given above; the conclusion would be essentially the same.

The only solution which offers itself in a natural way is the one based on the law of error propagation. Depending on the functional dependence, this may or may not correspond to a simple explicit form. However, if we consider again the case of an addition of variables, this will result in a quadratic sum of the contributions of the variance.

Such a "total", "effective" or "final" standard deviation is often considered by experimentalists to be a measure of the error which is not "safe", especially in connection with certificates. However, anybody should feel free, of course, to enlarge such a value, e.g. by multiplying it by a factor of 2 or 3. The important point here is that such an arbitrary augmentation of the region of uncertainty should only be made at the very end and not for intermediate results, since otherwise no clear use can be made of the data if further processing is needed. It will be obvious that application of such a factor (and its numerical value) should always be clearly stated.

The frequent case (e.g. in compilations of data) where uncertainties which have a common origin play an important role can only be correctly dealt with if a detailed list of the numerical parameters and their uncertainties is given. When better values become available, the necessary adjustments are readily made and the incidence of common uncertainties can be taken into account (for instance for assigning statistical weights). Such a listing is an absolute necessity for any serious measurement; the mere distinction between "random" and "systematic" errors is clearly insufficient.

6. Concluding remarks

It seems that most of the problems usually raised in connection with error statements are created in a somewhat artificial way. On the one hand, complications are introduced by the apparent need to distinguish between errors which are of a "random" or a "systematic" nature, and on the other hand by a frequent mixture of concepts which relate either to the point estimation of a parameter (like variance) or to the evaluation of confidence intervals.

By an exclusive and systematic application of the measured standard deviations or quantities which are believed to best approximate them, most of the problems raised vanish. A careful application of the general propagation law of errors then leads to a natural and unambiguous evaluation of the overall uncertainty to be associated with an experimentally determined quantity.

This very personal review of some problems has been provoked by a number of recent discussions on the assignment of uncertainties. It is intended to cast doubt on some existing practices and to provoke discussion. The author would be happy to receive comments and criticisms and he is prepared to adjust his present opinion in the light of a better knowledge.

It is a great pleasure to acknowledge the kind interest shown by several members of the BIPM staff in the questions treated in this report, in particular Drs. P. Giacomo, T.J. Quinn, P. Carré and A. Rytz. The pertinent critical remarks of Miss M.-T. Niatel on a draft version have led to the elaboration of Appendix B which, I hope, is free of the previous shortcomings. Her judicious remarks deserve my best thanks.

Finally, I am deeply indebted to Prof. A. Allisy with whom I have had the privilege of discussing matters related to this report many times over the past few years. His often quite different practical approach to many problems has been a continuous challenge. His constant interest in these questions and the frequent exchange of information with him have been vital to the present short review.

APPENDICES

A. An explicit example for illustrating eq. (4)

This first appendix, which describes some simple numerical properties of rectangular or related densities, is clearly of little general interest and in particular it may be skipped by all those readers who do not need more details to be convinced that summing the squares of confidence intervals is in general an illegitimate or useless operation.

For the sake of simplicity, let us choose for both random variables x_1 and x_2 the common rectangular density function centered at the origin (Fig. A1a)

$$f(x) = \begin{cases} \frac{1}{2a} & \text{for} \quad -a \leqslant x \leqslant a \\ 0 & \text{outside.} \end{cases} \tag{A1}$$

The probability p corresponding to a given confidence interval δ (i.e. ranging from $-\delta$ to $+\delta$) is then obviously

$$p_x(\delta) = \int_{-\delta}^{\delta} f(x)\,dx = \frac{\delta}{a}, \qquad \text{for} \quad 0 \leqslant \delta \leqslant a . \tag{A2}$$

The variance is readily derived as

$$\sigma_x^2 = \int_{-a}^{a} x^2 \cdot f(x)\,dx = a^2/3 . \tag{A3}$$

The sum $y = x_1 + x_2$ is then known to have a triangular density (Fig. A1b) described by

$$f(y) = \begin{cases} \frac{1}{b^2}(b - |y|) & \text{for} \quad |y| \leqslant b \\ 0 & \text{otherwise ,} \end{cases} \tag{A4}$$

with $b = 2a$.

A simple calculation gives here for the probability corresponding to a confidence interval δ' the value

$$p_y(\delta') = \frac{\delta'}{b^2}(2b - \delta') , \qquad \text{for} \quad 0 \leqslant \delta' \leqslant b . \tag{A5}$$

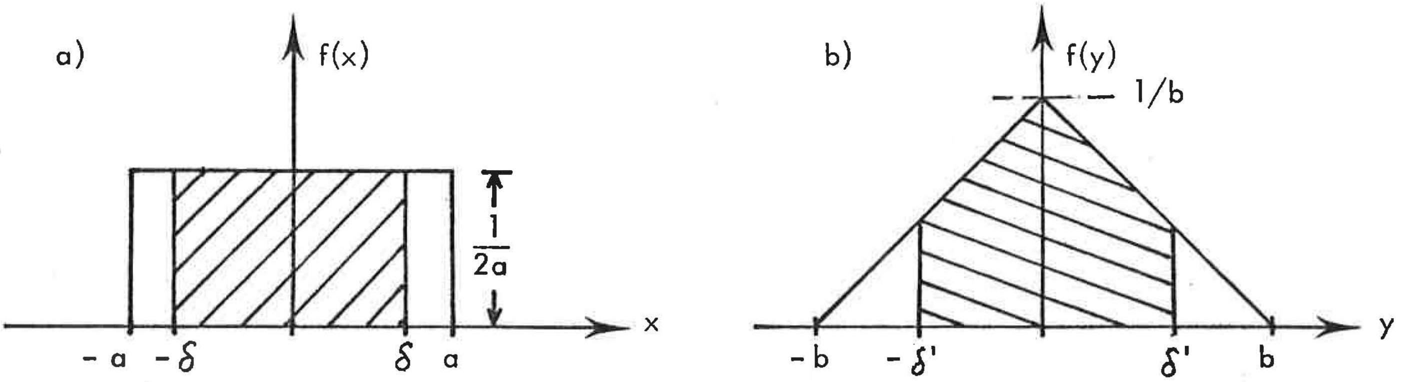

Figure A1 - The centered rectangular (a) and triangular (b) probability densities

In this case, the variance is easily shown to be

$$\sigma_y^2 = b^2/6 . \tag{A6}$$

As for the variances, it is clear that they are additive as expected, i.e.

$$\sigma_1^2 + \sigma_2^2 = 2 \cdot \frac{a^2}{3} = \frac{b^2}{6} = \sigma_y^2 , \tag{A7}$$

since $b = 2a$.

Let us now look at the confidence intervals, for which we choose quite arbitrarily the two probabilities p = 50% and 95%.

1. The case $p = 0.5$

For the rectangular, we get here clearly $\delta = a/2$; the corresponding value for the triangular is derived, according to (A5), from the condition

$$p_y(\delta') = \frac{\delta'}{b^2} (2b - \delta') = 0.5 , \tag{A8}$$

which yields (remembering that $\delta' \leqslant b$) the result

$$\delta' = b\,(1 - \frac{1}{\sqrt{2}}) = a\,(2 - \sqrt{2}) \cong 0.586\,a . \tag{A9}$$

On the other hand, adding in quadrature the intervals for a rectangular density gives

$$\delta'' = \sqrt{\delta^2 + \delta^2} = \delta\sqrt{2} = \frac{a}{\sqrt{2}} \cong 0.707\, a\ , \tag{A10}$$

a value which is some 21% higher than δ' .

This shows that the supposed equation (4) is not very well followed.

2. The case p = 0.95

It may be of some interest to see whether the situation improves at a more popular probability level, as e.g. $p = 0.95$. Since the calculations are very much the same, we confine ourselves to giving the results which are

- for the rectangular : $\delta = 0.95\, a$,
- for the triangular : $\delta' = 2a\,(1 - \sqrt{0.05}) \cong 1.553\, a$. (A11)

However, application of (4) would give

$$\delta'' = \delta\sqrt{2} = 0.95\sqrt{2}\, a \cong 1.344\, a\ ; \tag{A12}$$

hence a value which this time is about 13% too low.

(For $p = 0.99$, the corresponding value would even be too low by 22%).

In fact, the simple quadratic addition of confidence intervals is permitted here only for a specific value of the probability. Indeed if the condition $\delta' = \delta\sqrt{2}$ is used in equating p_x with p_y, we obtain from (A2) and (A5), since $b = 2a$,

$$\frac{\delta}{a} = \frac{\delta\sqrt{2}}{4a^2}\,(4a - \delta\sqrt{2})\ , \tag{A13}$$

from which results

$$\delta = 2a\,(\sqrt{2} - 1) \cong 0.828\, a\ . \tag{A14}$$

Thus, formula (4) is only applicable to rectangular densities if one chooses $p \cong 0.83$.

Similar restrictions would clearly have to be respected for any other combination of densities, which shows that a relation of the type (4) is not useful for practical purposes.

B. An explicit example illustrating the way to handle "systematic" errors

Let, as in (5), the actual measurements x_i and the corresponding unbiased values y_i be related by

$$y_i = \alpha + \beta \cdot x_i \ , \qquad i = 1, 2, \ldots, n \ . \qquad (B1)$$

If all known corrections have already been applied to x_i, we can put

$$\alpha = 0 \pm s_\alpha \qquad \text{and} \qquad \beta = 1 \pm s_\beta \ , \qquad (B2)$$

where s_α and s_β indicate the known or guessed uncertainties of the parameters α and β.

For the sake of simplicity, let us suppose that the function f we want to evaluate (and for which the "error" should be determined) is simply the mean value, i.e. we have

$$f(x_i ; \alpha, \beta) \equiv \bar{y} = \frac{1}{n} \sum_i y_i = \alpha + \frac{\beta}{n} \sum_{i=1}^{n} x_i \ . \qquad (B3)$$

Assuming that the measurements x_i are uncorrelated, we have $\sigma_{ik} = 0$ in (1). Evaluation of the partial derivatives leads readily to

$$\begin{aligned} \frac{\partial f}{\partial \alpha} &= 1, \\ \frac{\partial f}{\partial \beta} &= \frac{1}{n} \sum x_i = \bar{x} \quad \text{and} \\ \frac{\partial f}{\partial x_i} &= \frac{\beta}{n} = \frac{1}{n} \ , \end{aligned} \qquad (B4)$$

Insertion into the error-propagation formula (1) then gives directly the expression looked for

$$\begin{aligned} s_{\bar{y}}^2 &\cong (1 \cdot s_\alpha)^2 + (\bar{x} \cdot s_\beta)^2 + \sum_i \left(\frac{1}{n} \cdot s_x\right)^2 \\ &= s_\alpha^2 + (\bar{x} \cdot s_\beta)^2 + \frac{1}{n} \cdot s_x^2 \ . \end{aligned} \qquad (B5)$$

In the traditional terminology, the first two terms would be called the "systematic", and the third the "random" contribution to the "total" variance. Thereby, it has been tacitly assumed that s_α and s_β are estimates of the corresponding (unknown) standard deviations. From (B5) we can readily obtain the following two special cases

$$\begin{aligned} &- \text{for } s_\alpha = 0, \quad \text{i.e. } y_i = \beta \cdot x_i : \quad s_{\bar{y}}^2 \cong (\bar{x} \cdot s_\beta)^2 + \frac{1}{n} \cdot s_x^2 \ , \\ &- \ \text{"} \ s_\beta = 0, \quad \text{i.e. } y_i = \alpha + x_i : \quad s_{\bar{y}}^2 \cong s_\alpha^2 + \frac{1}{n} \cdot s_x^2 \ . \end{aligned} \qquad (B6)$$

C. On the uncertainty of a standard deviation

In this digression (which can be omitted at first reading) I would like to remind the reader first of some general relations describing the statistical properties of an estimated standard deviation and then apply them to the case of a normal population. These remarks are based on Cramér's well-known approach (chapter 27 of [4]) and for the ease of comparison his notation will be adopted in what follows.

If we denote by

$$m_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{(C1)}$$

the second central sample moment of n measurements x_i, where $\bar{x} = (1/n) \sum_i x_i$, and write

$$\mu_\nu = E \left\{ \left[x - E(x) \right]^\nu \right\} \qquad \text{(C2)}$$

for the central moment of order ν of the population, then the variance of m_2 is known (see [4], p. 348) to be given exactly by (D^2 stands for variance)

$$D^2(m_2) = \frac{1}{n} (\mu_4 - \mu_2^2) - \frac{2}{n^2} (\mu_4 - 2 \mu_2^2) + \frac{1}{n^3} (\mu_4 - 3 \mu_2^2) . \qquad \text{(C3)}$$

Hence, for a sample taken from a normal population we have

$$E(m_2) = \frac{n-1}{n} \cdot \mu_2 \quad \text{and}$$
$$D^2(m_2) = \frac{2(n-1)}{n^2} \cdot \mu_2^2 , \qquad \text{(C4)}$$

since then $\mu_4 = 3 \mu_2^2$.

Applying the rules of error propagation we find

$$D^2(\sqrt{m_2}) = \frac{1}{4 m_2} \cdot D^2(m_2) = \frac{\mu_2}{2n} . \qquad \text{(C5)}$$

If the unbiased estimate of the experimental standard deviations is denoted by

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}} \, , \qquad (C6)$$

with $E(s_x^2) = \mu_2$,

we find for its variance

$$D^2(s_x) = \frac{n}{n-1} \cdot D^2(\sqrt{m_2}\,) = \frac{\mu_2}{2\,(n-1)} \; . \qquad (C7)$$

Hence, it may also be expressed in the form

$$\begin{aligned} s_x &= E(s_x) \pm D(s_x) \\ &= \sqrt{\mu_2} \left[1 \pm \frac{1}{\sqrt{2\,(n-1)}} \right] . \end{aligned} \qquad (C8)$$

The term $[2\,(n - 1)]^{-1/2}$ indicates the relative uncertainty of the experimental standard deviation ("error of the error") which is due to the purely statistical reason of limited sampling. It can therefore be considered as an unavoidable minimum contribution to the error of the real standard deviation.

Similar expressions may be obtained for other error laws [14] .

References

[1] "Precision Measurement and Fundamental Constants", Proceedings of the International Conference held at the National Bureau of Standards, Gaithersburg, D.N. Langenberg and B.N. Taylor, Eds. (NBS Special Publication 343, 1971), 493-525

[2] P.J. Campion, J.E. Burns, A. Williams: "A Code of Practice for the Detailed Statement of Accuracy" (National Physical Laboratory, H.M. Stationery Office, London, 1973)

[3] H. Cramér: "Mathematical Methods of Statistics" (Princeton University Press, Princeton, 1946)

[4] E.J. Gumbel: "Statistics of Extremes" (Columbia University Press, New York, 1958)

[5] L.J. Savage (and other contributors): "The Foundations of Statistical Inference" (Methuen, London, 1962)

[6] S. Wagner: "Zur Behandlung systematischer Fehler bei der Angabe von Messunsicherheiten", PTB Mitteilungen 79, 343-347 (1969). An English version is available as PTB-Bericht FMRB 31/69.

[7] "Thomas Bayes's essay towards solving a problem in the doctrine of chances" (with a biographical note by G.A. Barnard), Biometrika 45, 293-315 (1958); original publication in Phil. Trans. Roy. Soc. 53, 370-418 (1763)

[8] W. Feller: "An Introduction to Probability Theory and its Applications, Vol. I" (Wiley, New York, 1968^3)

[9] I. Hacking: "Logic of Statistical Inference" (Cambridge University Press, Cambridge, 1965), chapter 12

[10] D.J. Hudson: "Statistics Lectures II: Maximum Likelihood for Least Squares Theory", CERN 64-18 (Geneva 1964), p. 167 ff.

[11] T.G. Blaney et al.: "Measurement of the speed of light, part II. Wavelength measurements and conclusions", Proc. Roy. Soc. London A355, 89-114 (1977)

[12] "Precision Measurement and Calibration", Selected NBS Papers on Statistical Concepts and Procedures, H.H. Ku, Ed. (NBS Special Publication 300, Vol. I, 1969); contains 14 papers by W.J. Youden

[13] Journal of Quality Technology 4, 1-67 (1972). The entire issue is devoted to the memory of W.J. Youden and contains 10 of his papers.

[14] J.W. Müller: "Incertitude d'un écart-type", Rapport BIPM-69/10 (1969)

(June 1978*)

* In view of a possible unification of error statements, BIPM has recently organized among the national laboratories an enquiry concerning their opinion on this matter. The answers to a questionnaire which was distributed had to be sent in by May 15, 1978. In order to avoid any influence on these replies, the present report, although written in August/September 1977, has not been distributed before.

This report is dedicated to the memory of my dear mother

Alice Müller-Schmid (1902-1978)

who, however, might have found it of little use as she always tried to adhere to the principle of avoiding errors rather than estimating them.