Rapport BIPM-91/7

Harry, Rolf and Richard - or Is there anything new in uncertainties ?

by Jörg W. Müller

Bureau International des Poids et Mesures, F-92312 Sèvres Cedex

Abstract

An attempt is made to show that a recent proposal concerning the statement of uncertainties, suggested by H.H. Ku in a publication of the NIST, is based on disputable assumptions and leads to unrealistic conclusions.

Harry H. Ku, the well-known statistician (now retired from the National Institute of Standards and Technology), published recently as *NIST Special Publication 805* (December 1990) a report entitled "Uncertainty and accuracy in physical measurements". It is no doubt because of the subject treated, the influential position held by the author and the surprising conclusion to which the document leads that I have been asked by a number of colleagues, including some from the United States, to comment on the paper. This puts me in a difficult position indeed and there has been a great temptation to pass over to someone else the unpleasant task of answering the queries raised. As no such replacement could be found and since pressure continued, I finally accepted, although with much reluctance.

There are several reasons for this. On the one hand, I have known Harry Ku for more than twenty years and I much appreciate several of his statistical papers. (No doubt I am aware of only a small fraction of his production). On the other hand, I have also been familiar with the ongoing debate on measurement uncertainties from its very beginning. As a result of this, I cannot consider the paper in question simply as a contribution to statistics. Rather, I see it as a revealing and therefore interesting document in applied psychology and I think it is only through such an approach that one may hope to understand it. The author ends with a proposal which is not only startling but - at least for a metrologist - also deeply unsatisfactory. How could such a thing happen?

Harry Ku and many (perhaps the majority) of his fellow-statisticians live in a world of random events. They realize, of course, that there are many things which happen outside these limits, but try to ignore them as much as possible. Thus, for example, the concept of a "systematic error" has always caused them headaches, for this is no longer part of statistics - at least in the traditional sense. A cursory look at the relevant NBS papers on this subject (among them excellent ones by Youden and Eisenhart) will easily confirm this.

However, the steady pressure brought to bear from the surrounding world of practical measurements has made it increasingly difficult in the past decade for a statistician to maintain a rigid psychological position of rejection. Somehow the "systematic" influences have to be accepted as real, and since they are often intimately intertwined with random effects, even a clear separation may be problematic. The alternative to incorporate them on an equal basis in the universe of statisticians would be an unacceptable proposal. This has led them to the seemingly clever Salomonic judgment that they should be treated "separately". In reality this is not a useful solution but just a provisional escape which arranges nothing. It may leave the traditional attitude relatively untouched, but it has the serious drawback of being at variance with the real world since all practical applications require a single quantity which can be used as a statement of uncertainty. The need for a unification of what has been kept apart becomes inevitable and clearly shows that there is a real problem to be faced. Apparently a new generation with looser links to tradition was required to recognize that all the possible influences affecting the result of a measurement should be seen in a more comprehensive way and that a formal description corresponding to this view had to be elaborated. This explains why the decisive steps did not come from statisticians, but from experimentalists who were confronted daily with the problems and the unsatisfactory solutions available to them.

Dr. Ku's paper may be seen as a desperate attempt to defend the orthodox method against the increasing number of barbarian incursions into reserved territory. It is a rear-guard action and the author probably feels it. One can expect that the adopted strategy of splitting the practical problems into two categories will soon have to be abandoned as unrealistic. Would it therefore not be more appropriate and useful to consider the mutual penetration of fields as a lucky event which, by abandoning artificial borderlines, allows one to see a subject in a wider and more general context ? Can anyone seriously hope to solve a problem by merely suggesting two terms for one, as is done in Ku's paper with "uncertainty" and "accuracy" ? Such a proposal ignores the fact that the debate is about concepts not words.

The basic step one has to make is conceptual. It is difficult for most people since it asks them to jump over their own shadow: they have to realize that the traditional distinction they make between a random uncertainty and a systematic one (or "error", as adherents to an unfortunate vocabulary still call it) has no objective basis. An uncertainty, once numerically estimated, has no special character in itself. Whether it will be diminished or not by making more measurements, i.e. whether it "behaves" as a random or as a systematic uncertainty component, depends only on the context in which it is used. In well-defined circumstances, the context is given by a mathematical model and the relationship between the quantities is described by mathematical functions. Once this is established, all the rest, and in particular the way in which the various uncertainty contributions entering the problem affect the output quantity in which we are finally interested, is a simple algebraic matter of uncertainty propagation, where both variances and covariances have to be considered. If this is not the case, then the measurement situation has not been correctly modelled. The traditional distinction between the "error types" is actually a false problem. What we have just sketched is the obvious way out of a situation which, when seen in a wrong perspective, brings in endless complications that have no real solution. The requirement to jump over one's own shadow - trivial from the mathematical point of view - is thus recognized as a psychological hurdle. If we are not willing to make the effort, the problem of assessing experimental uncertainties cannot find a satisfactory solution.

Once the obstacle is identified clearly all is simple. In particular, one realizes that the pseudo-philosophical distinction of two "schools of thought" is without foundation. Indeed, the realistic estimation of uncertainties is largely a matter of clear and critical insight in the experimental methods used (including their possible pitfalls) rather than an exercise on principles; for purists the playground is therefore ill chosen. One of the indispensable prerequisites for arriving at such a clarification is obviously the distinction between "error" and "uncertainty": all errors, as far as they are known, have disappeared by applying appropriate corrections; what remains - including "unknown errors" - are just uncertainties and they can be treated as such.

The subdivision of uncertainties into "type A" and "type B", as suggested in the BIPM Recommendation, has often been misunderstood and taken as new labels for "random" and "systematic". There is no basis in the recommendation for such an interpretation. The distinction, proposed for educational purposes, should remind the experimentalist of the very different situations in which he can find himself when asked to "estimate" a standard deviation. They may range from a simple routine operation in the case of repeated measurements (type A) to the delicate task of making such an assessment, for lack of information, on a subjective basis. A deep insight and long experience may help him to arrive even in such a difficult situation at a realistic estimate (type B). Clearly, the distinction between the types is a gradual one and may occasionally be disputed, but this is of no consequence since all contributions will subsequently be used "in the same way". Obviously one still has to take full account of their relation to the other input quantities, as described in a quantitative way by the respective covariances. Therefore, the usual reproach (repeated by Dr. Ku) that the BIPM approach, by systematically applying a quadratic addition, gives low values, is based on poor comprehension and is without foundation.

Once these things are clarified - and recognized as a matter of great simplicity -, our main task is accomplished. It is therefore hardly worth going in much detail through Dr. Ku's paper which, as we shall see, takes a bad turn right from the start. Whereas for those who have grasped the basic idea a critical look is no longer necessary, some remarks may be helpful for readers still hesitating to accept the evidence.

The troubles begin immediately with the discussion of the "historical background". After sketching briefly the "traditional method" as practiced by the "orthodox school" (the expressions are his), Dr. Ku arrives at the conclusion that "the ideas underlying the traditional method are sound". This is a surprising statement for those who know that this includes such disparate proposals as the ones made by Eisenhart and by Campion *et al.* It becomes even more puzzling if one learns on what this opinion is based. What does the author really mean by "traditional method" ? After mentioning that "random components" of "experimental errors" are expressed as variances whereas "systematic components" are characterized by "credible (maximum) bounds", it is said that these two components are "propagated separately through the necessary steps, and combined at the end by linear addition if needed". Such a hotchpotch procedure is neither "sound" nor reasonable. It has no scientific basis, is arbitrary and leads to a result (or even to two) which has no clear meaning and therefore cannot be used in a subsequent step. In such a context it is difficult to follow an author who tells us that the traditional method requires "a thorough understanding of some of the basic principles, e.g.: what is a realistic 'repetition' of a measurement?" In our opinion the real problems are elsewhere. This is why the BIPM, from the late seventies onwards, found it necessary to become involved in this matter and to pave the way for a solution.

It is obvious that numerical estimates of uncertainties can be used for quite different purposes. This shows that we are in the presence of a very important quantity. From this to conclude, however, that different types or concepts of uncertainty have to be used in various applications is a most unfortunate extrapolation. We are not too far from the well-known story where the employee, called by his master to bring him a meter stick, first asks if he wants the one by which they buy or the one by which they sell.

I have always thought that it was one of the basic and undisputed requirements of a quantity used in metrology to be neutral, i.e. not to give any prejudice to the purpose for which it will be applied. Now I read that "any unique way of expressing uncertainty for one group of users cannot be satisfactory to the other group, and any compromise would fail both groups". The statement is followed by the stories in which "Rolf" and "Richard" are starring. This belongs to a literary genre on which I do not feel competent to comment. As for the assumed statistical background, it is largely misleading. Once more it is repeated that "systematic errors", since they remain constant, "do not behave in the same manner as the random part" and therefore "should be kept separate". As discussed above, this is a popular but unfounded statement. Even if for repeatable measurements the mean value, as a function of the number of data taken, may change position with respect to the (unknown) "true" value, any final value is either "high" or "low", but we do not know which possibility holds. The argument that for "systematic errors", even after correction, the value adopted is always systematically "high" or "low" is therefore no argument for a separate treatment. Some of the other claims are equally disturbing, e.g. when it is said that "physicists are interested in differences only" or that BIPM has a "preference for a 'small' uncertainty". Similarly unconvincing is the attempt to characterize the "two main groups of users" by claiming that one looks for the "sameness" of their results while the other looks for "differences".

One might also mention in passing that it is somewhat puzzling when the proposal to offer to anybody the yardstick he is calling for comes from a national laboratory, for it is normally assumed to be among their main duties to ensure uniformity in a given field of measurement.

I am still fully convinced that a general and uniform way of expressing an experimental uncertainty which is attached to the result of a measurement is possible and necessary. It should be clear that such a quantity can in no way be taken as a prejudice for the application a user wants to make of it. Obviously, the requirements and the methods used by those who want to protect themselves against "outliers" (with respect to some guaranteed limits) or by others who try to decide on the presence or absence of a suspected small effect will not be the same. However, this does not point to any contradiction or conflict; the aims are just different. Whatever the user may wish to do with his data, he must have a reliable and objective basis to start from, and I can see no reason why this cannot be the same for everybody. This is what the uncertainty as sketched in the CIPM Recommendation INC-1 attempts to provide. The possibility of choosing the factor k to accommodate to the required purpose of the measurement brings in the flexibility needed in practice.

To summarize our position, we cannot help thinking that Harry Ku's proposal, unfortunately, is not a very useful contribution to the ongoing discussion. The aim must be to arrive at a clear, general and useful concept. For achieving these requirements, it must be uniformly applicable and contain a single numerical value. If any of these conditions is abandoned, the goal sought cannot be achieved. I fully agree with the opinion of Mosteller and Tukey, quoted by Dr. Ku at the very end of his paper, namely that "obtaining a valid measure of uncertainty is not just a matter of looking up a formula" and I feel tempted to add "but also of calling in question cherished personal habits".

(December 1991)