Measurement 46 (2013) 3725–3729

Contents lists available at SciVerse ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Rounding and notation, namely when using stipulations in the definition of measurement units

F. Pavese *

Istituto Nazionale di Ricerca Metrologica, Strada delle Cacce 73, 10139 Torino, Italy

ARTICLE INFO

Article history:
Available online 24 April 2013

Keywords:
Rounding
Stipulation
Notation
SI units

ABSTRACT

This paper intends to tackle, in the context of measurement and the definition of measurement units, a problem well known in computing science, the inherent propagation and accumulation of rounding errors throughout the intermediate steps of numerical calculation, and some issues in notation, namely of integer numbers.

1. Introduction

This paper intends to tackle, in the context of measurement, a problem well known in computing science [1], the inherent propagation and accumulation of rounding errors throughout the intermediate steps of numerical calculation, and some issues in notation, namely of integer numbers.

In this context, the use of the so-called 'stipulated values', or 'defined values', is intrinsic in definitions, namely those aiming to establish regulatory conditions of all kinds. Contrary to 'consensus values', which are measured values with an associated uncertainty, stipulated values are rounded numbers—either real or integer—deemed exact by definition and have zero uncertainty. The propagation effect of rounding or truncation will occur when more than one stipulated value is combined in an algebraic expression. This may happen in measurement, e.g., when computing the values of multidimensional quantities and having to use more than one unit containing in its definition a stipulated value.

The issue deserves general attention of the experimentalist and of the metrologist, and, in particular, it places intriguing questions concerning the current debate on a more extensive use of stipulated values of "fundamental constants" in the definition of measurement units of the International System of Units (SI) [2–4], a field where missing a single digit of defined values can make the difference in the accuracy between using them and making them useless. The origins of the exact stipulated values are the measurements of those constants at their best accuracy *at the moment of stipulation*. There will clearly be some degree of rounding error involved in such a procedure involving what are essentially truncated values, and some subsequent propagation problem.

2. Rounding and truncating

Let us start form the simplest example. Assume to have two rational numbers: $A = 5.6$ and $B = 4.6$. If rounded to integer numbers, they become $A_r = 6$ and $B_r = 5$, if truncated $A_t = 5$ and $B_t = 4$. The result of their sum is $R_S = A + B = 10.2$ exactly, $R_{Sr} = A_r + B_r = 11$, $R_{St} = A_t + B_t = 9$. The result of their difference is $R_D = A - B = 2.0$ exactly, $R_{Dr} = A_r - B_r = 1$, $R_{Dt} = A_t - B_t = 1$. The result of their product is $R_P = A \cdot B = 25.76$ exactly, $R_{Pr} = A_r \cdot B_r = 30$, $R_{Pt} = A_t \cdot B_t = 20$. The result of their ratio is $R_R = A/B = 1.2173\ldots$ (rational or real number), $R_{Rr} = A_r/B_r = 1.2$, and $R_{Rt} = A_t/B_t = 1.25$.

Large errors may obviously occur and be propagated and expanded in the communication of results in rounded and truncated forms. If a long calculation can safely be

* Tel.: +39 3488130101.
E-mail address: frpavese@gmail.com

0263-2241/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.measurement.2013.04.010

rounded off to N decimals, it is not valid to round off intermediate steps to the same number of digits because round-off errors may accumulate. A larger number of digits (say M) is required at intermediate steps and the difference $M - N$ are called the "guard digits".

In measurement, a first additional problem arises from the fact that an *algebraic combination of stipulated values* is often said to be a stipulated value, so requiring to also be exact by definition.

However, after stipulation, one might no longer take into account the fact that these numbers were *originally* in actuality estimates of real numbers, and affected by an experimental uncertainty. Therefore, one might not compute them from the *originally* imprecise numbers, and afterwards stipulate their values, either as R_r or R_t, in order to compensate for the rounding error. Nor could one take into account anymore the effects of the original uncertainty. It has been abolished by definition, so that, in general, "guard digits" are not admitted in stipulation.

As an example, this is the case of the molar gas constant $R = k_B \cdot N_A$, where k_B is the Boltzmann constant, $k_B = 1.380\,6488(13) \times 10^{-23}$ J K^{-1} (CODATA 2010 [6]),[1] and N_A the Avogadro number, $N_A = 6.022\ 141\ 29(27) \times 10^{23}$ mol^{-1} (CODATA 2010 [6], see later Section 4 for a distinct problem for N_A). Should they become stipulated (exact) numbers, for the definition of the kelvin and mole unit respectively, but R not be stipulated, the result of the product of the two stipulated numbers would be a rational number with a larger number of decimal digits (or even a real number in other circumstances).

R has also been measured directly: its CODATA 2010 value is 8.314 4621(75) J mol^{-1} K^{-1}, to be compared with the result of the above product: 8.314 462 145 468 95 J mol^{-1} K^{-1}.

However, to which digit should be truncated the latter, certainly having more digits than the significant ones? To the corresponding CODATA digit for the uncertain R? It does not seem correct.

The above latter value of R is obviously consistent with the former, because all digits reported for $k_B \cdot N_A$ have been used. However, being the uncertainties of k_B and N_A reported with two digits, the second one is obviously a "guard digit" that should not be used in stipulation. See Section 3.1 for a consequence of this fact.

Two more problems arise in measurement. First, let us modify the initial example by adding a digit to the rational numbers: $A = 5.66$ and $B = 4.66$. If rounded in the usual way one obtains $A_r = 5.7$ and $B_r = 4.7$, now also rational numbers; if truncated, they become $A_t = 5.6$ and $B_t = 4.6$. The result of their ratio is now $R_R = A/B = 1.21459\ldots$, $R_{Rr} = A_r/B_r = 1.21276\ldots$, and $R_{Rt} = A_t/B_t = 1.21739\ldots$: in general, they all are *real* numbers now. Thus, one might *not* expect that the result of a ratio operation is still a rounded number with a manageable number of digits, but this is in fact not generally true. A common case is when $R = 1/A$.

Secondly, one is not always dealing originally with real numbers. In the case of an integer number (typically, the result of a counting), is rounding (stipulation) admitted, being rounding a concept usually related to real numbers? A corollary of this problem is: which is the correct notation for an integer value of a discrete quantity of which not all digits (either some of the most significant or some of the least significant) are known? See Sections 3.2 and 4: this problem among others was initially discussed in [4].

3. An application to measurement: stipulation in measurement units

The consequences of the previous considerations can be applied to the case of an extensive use of stipulated values in the definition of SI units, as is currently being proposed. They are significant also in the context of the documented conflict between the SI and the requirements of many data systems and informatics particularly evident in sensor and instrumentation technologies [5].

3.1. More than one value stipulated

If the value of more than one "fundamental constant" is stipulated, should the values of other constants that are algebraic expressions of them be computed as a combination of the stipulated values, or of the original values?

For example, the Stefan–Boltzmann constant $\sigma = 2\pi^5 k_B{}^4/15h^3c_0{}^2$ is given the value $5.670\ 373(21) \times 10^{-8}$ W m^{-2} K^{-4} [6]. This value is computed from the values [6] of the three constants $k_B = 1.380\ 6488 \times 10^{-23}$ J K^{-1}, $h = 6.626\ 069\ 57(29) \times 10^{-34}$ J s and $c_0 = 299\ 792\ 458$ m s^{-1} (*the latter already a stipulated value*), using all the reported digits, including the uncertain ones—$\sigma = 5.670\ 372\ 623 \cdots 10^{-8}$ W m^{-2} K^{-4} before rounding. Using instead the stipulated values for all three constants, rounded by excluding both the uncertain digits ($k_B = 1.380\ 65 \times 10^{-23}$ J K^{-1}, $h = 6.626\ 069 \times 10^{-34}$ J s), one obtains $\sigma = 5.670\ 393\ 80 \cdots 10^{-8}$ W m^{-2} K^{-4}. An identical result is obtained in this example by rounding to the first uncertain digit ($k_B = 1.380\ 649 \times 10^{-23}$ J K^{-1}, $h = 6.626\ 0696 \times 10^{-34}$ J s). An obvious rounding effect occurs.

Similarly, in the case of R in Section 2 the following stipulated (rounded) values should be used when limited to the first uncertain digit: $k_B = 1.380\ 649 \times 10^{-23}$ J K^{-1} and $N_A = 6.022\ 1413 \cdot 10^{23}$ mol^{-1}. Consequently, $R = 8.314\ 463\ 363\ 703\ 70$ J mol^{-1} K^{-1}, *not compatible* with the CODATA value for R.

When using values already having been stipulated, no uncertainty can be associated to the value of σ, a real number, nor to R, a rational number: in fact, in [2] the fundamental constants obtained from algebraic operations using stipulated constants are said to have zero associated uncertainty. However, the questions already placed in Section 2, still arise. In addition, the use for the stipulation of all uncertain digits, typically two, looks inconsistent with the very concept of stipulation: the less significant digit is generally allowed in the notation of uncertainty only

[1] The CODATA values are used here. However, note that the CODATA values have been elaborated using a "Least Squares Adjustment" procedure that *alters* the values of the constants, in the meantime that obtains the best *consistency* and lower uncertainties of those values for all constants considered. They are *not* the simple mean (or weighted mean) of the *measured* values, and the obtained uncertainty is in general *better* than can be obtained experimentally, and should not be confused with the latter.