Accred Qual Assur (2010) 15:601–602
DOI 10.1007/s00769-010-0715-4

EDITORIAL

Measurement results should not be tortured until they confess

Paul De Bièvre

Published online: 20 October 2010

During a recent holiday in the mountainous center of Kriti (in Axos for the interested), a sentence struck me in an interesting book [1]:

"Admittedly, one reason why statistical arguments sometimes fail to persuade is that different statistical methods may produce varying results and the investigators are suspected of choosing the method most favorable to their arguments. The range of statistical techniques available to the econometrician is so wide that the zealous advocate can often "torture the data[1] until they confess."

These sentences are worth being reflected upon and, maybe, they should be propagated and applied a bit more in other contexts than the purely economical, because they could be of general applicability.

A case in point may be the treatment of 'measurement results' (see entry 2.9 in [2]) of 'interlaboratory comparisons' (ILCs) in chemistry where such results are often "assumed to be normal" but frequently are not [3]. Sometimes (or often?) they are trimmed, combed, submitted to selective "cutting" procedures, or other kinds of treatment, until they fit the a priori model of "data distribution", mostly "assumed to be normal". Any value not conforming to that distribution -or disturbing it somehow- is looked at with suspicion. Apparently, we prefer to see what they should look like according to an a priori conceived picture rather than see them as they are.

One can make the following observations.

Mostly people are interested in the value "closest to the truth", or at least "close to the truth" and they assume that such a value is located at -or lying close to- the central location of the distribution of the results. Hence, they look for the average of the 'measured quantity values' (see entry 2.10 in [2]). When the own measurement result also finds itself near that location, that generates the comfortable feeling of "being where most of the values are found" and certainly the conclusion that "the best value cannot be far away from that", and also "that could not be wrong, could it?"

Hence, 'measurement results' close to the center of a distribution are important and deserve full attention, not the (very) low or (very) high values. Remarkably enough, that does not happen. Rather much of the attention is concentrated on the low and high values, and much effort is spent in finding reasons to eliminate them. Why? Clearly, because that makes the calculated standard deviation of the average smaller. However, eliminating extreme values does not change very much the location of the average. Rather the spread around the "comfortable" average value is reduced, which gives the average still more authority. This approach is all the more interesting if the distribution of the values becomes more "normal".

Are we then looking at a self-fulfilling prophecy"? Or, better, at a "self-fulfilling reasoning" tweaked to confirm an assumption (of normal distribution) already made on beforehand?

Is it logical to proceed along a reasoning which reduces the standard deviation per se in order to increase the "trust" in the average? Since the most centrally located 'quantity value' is determined by the most centrally located measured values and not by a few extraneous ones, the suspicion arises that we eliminate the extreme values in

P. De Bièvre (✉)
Kasterlee, Belgium
e-mail: paul.de.bievre@skynet.be

[1] The original text reads "it" which, of course, is wrong since "data" is the plural of the Latin "datum"; this error is very often made in English texts. This point has been definitively made in the new 2008 VIM.

order for the others to strengthen the more centrally located average of a pre-conceived (normal) distribution.

There is another remarkable observation to be made: it is a well-known boundary condition in statistics that a minimum of 40 or more individual points of a homogeneous sample of data (in our case 'measured quantity values' in a 'measurement result') are necessary to draw statistical conclusions, e.g. about the type of distribution. In cases where only a small number of 'measured quantity values' are available in a 'measurement result', no such conclusions can be drawn. That applies even more when 'measured quantity values' are obtained by different 'measurement procedures' (see entry 2.6 in [2]).

We look at that a little closer in the case of 'measurement results' of chemical measurements in an ILC, obtained for the same 'measurand' in the same material:

- first, we rarely if ever have to do with big homogeneous populations obtained under 'repeatability conditions of measurement' (see 2.20 in [2]): conditions require the same 'measurement procedures' for all measurements; hence, the boundary condition is not fulfilled, and
- second, we have rarely, if ever, to do with 40 or more different 'measurement procedures' measuring the same 'measurand' in order to be able to talk about a sufficiently large enough number of "independent" results.

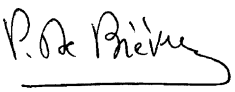
These seem to be two solid reasons for not applying statistical techniques to arrive at whatever conclusion about the distribution of the results in ILCs. Yet, we see so many times that results are "tortured" by means of statistical techniques "until they confess" what is wanted: a scientific ground to treat them as "normally distributed". Is that wise? Is that consistent? Is that allowed altogether?

Shouldn't we better realize that measurement results in general, including in ILCs, also give us the means to evaluate the "quality" of our thinking: we want to learn from an ILC (an a posteriori event i.e. occurring after the ILC), rather than wanting to confirm an assumption already made before that ILC (an a priori event i.e. occurring before the ILC). Thus, we can derive conclusions from the 'measurement results' as they are, without … torturing them.

Normal distribution of ILC results can only be "assumed" if such results indeed show such a distribution, e.g., on a simple graphical presentation. If such a proof is not given, or cannot be given, no conclusion can be drawn which is based on a "normal distribution".

Maybe, instead of torturing our 'measured quantity values', we should torture our mind by trying to interpret the 'measured quantity values' we really see.

Final comment: "torturing" the measurement results yields an essentially *variable* "average" in any case. Just taking the average of a presumed normal distribution does not give us this other important element needed in an ILC: a reliable, independent, and *stable* reference value. To evaluate the results of an ILC, why should we not take just one metrologically traceable reference value, chosen because of its clearly established metrological traceability chain? VIM offers us three possible "references" for such a chain (see entry 2.41, Note 1, in [2]). But that is another story.



Paul De Bièvre
Editor-in-Chief

By the way, looking for justification of Metrology in Chemistry?

Drug screening in urine is performed in over 400 million samples per year, at a cost of over $20 billion per year and this is only one type of analysis out of many.

Source: I Kuselman at the CITAC Workshop in Ness Ziona (Rehovoth, Israel) on 2010-01-21.

Assuming a measurement uncertainty of an incredibly optimistic 1%, that reflects a cost of unreliable or useless measurements of between an estimated 25–40% of the cases, equivalent to a loss of at least USD $5 \cdot 10^9$ per annum.

References

1. Olson M (1982) The rise and decline of nations, p 96, lines 5–10, Yale University Press, New Haven
2. BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML (2008) The international vocabulary of metrology—basic and general concepts and associated terms (VIM), 3rd edn. JCGM 200:2008. http://www.bipm.org/vim
3. De Bièvre P (2009) "Comparing Results of Chemical Measurements: time to raise some basic questions from practice", Chap 8. In: Pavese F, Forbes AB (eds) "data modelling for metrology and testing in measurement science". Birkhauser, Boston, pp 255–273. ISBN 978-0-8176-4592-2, e-ISBN 978-0-8176-4804-6. doi: 10.1007/978-0-8176-4804-6