USE OF THE MEDIAN IN THE ANALYSIS OF EXPERIMENTAL DATA

ENRIC LEON GRIGORESCU, CAMELIA AVADANEI

"Horia Hulubei" National Institute of Physics and Nuclear Engineering, 077125, Magurele, Romania

Abstract. *The concept of median is defined. The advantage of the median in special cases and the comparison with the classical arithmetical mean and weighted mean are presented. The case of discrepant results is analyzed. The calculation of the median uncertainty, including a type B component, is discussed. The calculation of the supplementary uncertainty s_B is illustrated for two applications concerning nuclear data. A suggestion for environment is indicated.*

Keywords: *median, uncertainty, discrepant data.*

1. INTRODUCTION

Each set of experimental result must be characterized by a sort of mean which must represent, as well as possible, the real value of the measured quantity.

The arithmetical mean is defined as:

$$\bar{x} = \frac{\sum x_i}{n} \tag{1}$$

The $\bar{x}$ ensures the minimum for the expression

$$\sum (x_i - \bar{x})^2 \tag{2}$$

Uncertainty of the mean is calculated with

$$s(\bar{x}) = t\sqrt{\frac{\sum (x_i - \bar{x})^2}{n(n-1)}} \tag{3}$$

where t is the Student Factor.

If results are of heterogeneous origins, the weighted mean $\bar{x}_W$ is used.

$$\bar{x}_W = \sum \text{pixi} \tag{4}$$

where p_i is the relative weight of the x_i result, calculated with the formula

$$p_i = \frac{w_i}{\sum w_i} \tag{5}$$

and $w_i = (1/s_i^2)$ is the absolute weight.

The symbol s_i is the uncertainty for x_i.

The $\bar{x}_W$ ensures the minimum for the expression:

$$\sum p_i (x_i - \bar{x})^2 \tag{6}$$

There are two types of uncertainties for $\bar{x}_W$:

- internal uncertainty

$$\mathbf{s}(\bar{x}_W)_{\text{int}} = \sqrt{\frac{1}{\sum w_i}} \tag{7}$$

- external uncertainty

$$\mathbf{s}(\bar{x}_W)_{ext} = \sqrt{\sum p_i (x_i - \bar{x}_W)^2} \tag{8}$$

The weighted mean is used, for instance, for the evaluation of Nuclear Data (half time particle energies branching ratios etc.) as well as for International Comparisons of Radioactive Standard Solutions. “Horia Hulubei” National Institute of Physics and Nuclear Engineering (IFIN-HH) participated with success in the comparisons since 1960.

In such a comparison, samples from solutions with unknown concentrations (Bq/g) are distributed to 20-30 involved laboratories. The results of the laboratories provide an arithmetical mean and deviations from it are considered; they ought to correspond to the declared uncertainties.

2. CONCEPT OF MEDIAN

Some times, one or two of the results deviate much from the corresponding mean. They are called discrepant or outliers results. The general theory of measurement considers that, usually, the results are distributed around the mean as described by the normal distribution (Gauss). A result, far from the mean, has a low probability of appearance. One may suspect that a discrepant result represents a wrong measurement, an error (the operator or the equipment).

Even if the χ^2 test confirms the Normal Distribution, one tries to avoid the influence on the mean of the “apparent” discrepant result.

There are two ways of doing this:

- the elimination of the results, using the Chauvenet criterion;
- to enlarge the uncertainty of the result which gives a low p_i

The second action is also applied in the case of suspected low uncertainties.

In this case the experts in nuclear data use different procedures [1]:

- the limitation of p_i (LSRW), recommended by the International Atomic Energy Agency;
- normalization of residuals;
- Rajeval procedure;
- use of median.

If the uncertainty of $\bar{x}_W$ is great compared with that of $\bar{x}$, some evaluators preferred $\bar{x}$.

Sometimes there are a few authors providing false uncertainties; too low for prestige, too high for prudence. Elimination of a result, as well as enlarging an uncertainty is, of course, an offense for the author.

The concept of "median", which appeared long ago, was seldom used; the median is insensitive to discrepant results but neglects the p_i values [2].

If the results are in an ascending order, the median is defined as it follows:

$$\hat{x} = x_{k+1} \qquad \text{for n=2k+1} \tag{9}$$

$$\hat{x} = \frac{1}{2}(x_k + x_{k+1}) \qquad \text{for n = 2k} \tag{10}$$

So the median has an equal number of results higher or lower than it.

The median ensures the minimum for the expression

$$\sum |x_i - \bar{x}| \tag{11}$$

It is supposed that the results are symmetrically distributed around the real value, which is valid only with the probability calculated with the Binomial Distribution (BD).

3. MEDIAN UNCERTAINTY

For the median uncertainty one may use the Median of the Absolute Deviation (MAD).

$$\text{MAD} = \text{med}\ \left\{|x_i - \hat{x}|\right\} \text{ for i = 1 ... n} \tag{12}$$

The following formula can be demonstrated:

$$\text{MAD} = 0.674\left(\sqrt{\frac{n-1}{n}}\right)s(x) \tag{13}$$

where s(x) is the uncertainty of the individual result x_i.

One may find in literature a complicated calculus which, for $n \to \infty$, provides the uncertainty of $\hat{x}$ [3]

$$s(\hat{x}) = \sqrt{\frac{\pi}{2n}}s(x) = \frac{1.858}{\sqrt{n-1}}MAD \tag{14}$$

where (13) formula was used. It can be seen that $s(\hat{x})$ is a little higher than $s(\bar{x})$.

In the real cases *n* has small values 6, …, 10, …, 20 and the distribution of results around the real value is given by the BD probability P (n, k):

$$P(6, 3) = 0.31 \tag{15}$$

$$P(10, 5) = 0.25 \quad (16)$$

$$P(20,10) = 0.176 \quad (17)$$

P (n, n/2) is lower and lower with an increasing n. The result is only apparently wrong because the relative asymmetry, defined as (high-low /n) tends to zero.

For $n = 6$, the other probabilities have the values

$$P(6, 4) = P(6, 2) = 0.235 \quad (18)$$

$$P(6, 5) = P(6, 1) = 0.09 \quad (1)$$

$$P(6, 6) = P(6, 0) = 0.015 \quad (2)$$

for $n = 10$ values are:

$$P(10, 6) = 0.21 \quad (3)$$

$$P(10, 7) = 0.12 \quad (4)$$

$$P(10, 8) = 0.044 \quad (5)$$

$$P(10, 9) = 0.01 \quad (6)$$

$$P(10, 10) = 0.001 \quad (7)$$

The above values were calculated with the „success" probability of 1/2.

In literature can be found the concept of weighted median $\hat{x}_W$. [4] It ensures the minimum value for the expression

$$\sum p_i |x_i - \hat{x}_W| \quad (8)$$

Its value is obtained by trials using all the x_i values; MAD is also obtained by trials. The position of $\hat{x}_W$ in the row of values is no more central and may be far from it. The weighted median uses all the available information.

4. UNCERTAINTY ASSOCIATED TO ASSIMMETRY

In formula (14) $s(\hat{x})$ represents the fluctuation of $\hat{x}$, when the experiment is repeated, around the $\hat{x}_0$ value obtained with $n \to \infty$.

But in every experiment, the probability for the real value to be situated outside the central interval is constant and equal with

$$1\text{-}P(n, n/2) \quad (9)$$

To take into account the possible deviation of the median from the real value, for experiments hard to repeat, one may try the estimation of a supplementary uncertainty s_B.

For each interval where the real value may be situated, associated with a BN probability of some value, the difference between the centre of the interval and the median is calculated (in absolute value). The deviation is multiplied by the associated BD probability.

Summing the contributions of all intervals, a mean deviation is obtained, which represents s_B.

The s_B uncertainty characteristic is a mixed one; statistical (the difference) and systematical (the constant BD probability).

5. EXAMPLES

Six results from the period 1962 – 1985 for the branching ratio of the 6891 KeV Alpha Ray of Po-211 are presented in 10^{-4} units

51.0 ± 5 52.4 ± 0.9 54.6 ± 1.9 57.0 ± 3 60.0 ± 1 70.0 ± 14

The discrepant result is 70.0 ± 14. The following values are obtained:

$\bar{x} = 57.5$ $s(x) = 7$ $s(\bar{x}) = 2.86$
$\bar{x}_W = 57.4$ $s(\bar{x}_W)_{\text{int}} = 0.6?$ $s(\bar{x}_W)_{ext} = 3.8$
$\hat{x} = 55.8$ $s(\hat{x}) = 3.9$ [with formula (14) using s(x)]

It is noticed that formula (13) for MAD is not verified. It results

$$\frac{s(\hat{x})}{s(\bar{x})} = \frac{3.9}{2.86} = 1.36 \tag{10}$$

The big value for s(x) is due to the discrepant result.

To calculate s_B the following scheme is considered. It contains the results and the centers of the intervals (Fig. 1).

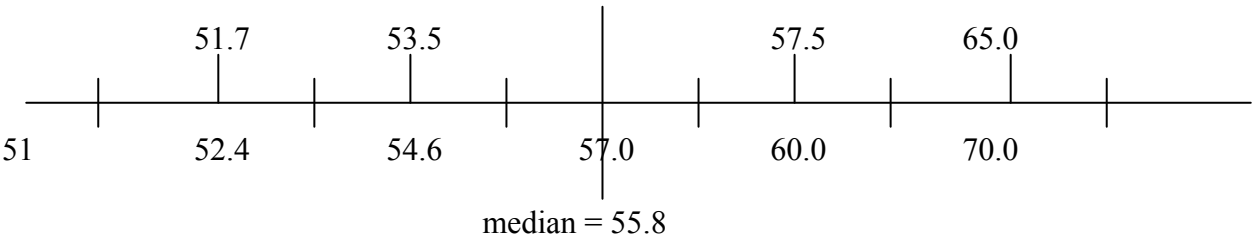

Fig. 1. Calculus of s_B.

For instance, for 57.0 – 60.0 interval, the diference is 1.7. The value is multiplied with the binomial probability of 0.23. The final value is 0.39. If the calculus is repeated for the other intervals, one obtains by summing a final value of 2.12.

The compound uncertainty $s_c(\hat{x})$ is calculated with the usual formula:

$$s_c(\hat{x}) = \sqrt{3.9^2 + 2.12^2} = \sqrt{19.7} = 4.514 \tag{11}$$

The s_B contribution is small because s(x) is big always when a very discrepant result is present.

The same procedure, applied to six results in the paper [2], produces a $s_B = 0.048$, value comparable with the $s(\hat{x}) = 0.05$.

It must be mentioned that s_B has a sense for experiments hard to repeat as international comparisons and nuclear data.

If *n* is great enough, a group of narrow intervals accumulate around the median, and the probabilities associated with distant intervals are negligible. So, s_B is negligible too.

6. CONCLUSIONS

It is not easy to establish the situations which involve the median. The median is usually applied in international comparison where the elimination of the result is forbidden. The Nuclear Data analysts calculates $\bar{x}, \bar{x}_W$ and $\hat{x}$.

If $s(\bar{x}_W)$ is big, the other two means are considered, or a mean of them. It is found in literature the definition of weighted median.

For small values of *n* and special experiments an additional uncertainty s_B is calculated. Anyway, the use of median is the choice of specialists, at least as a comparative value.

For the environment control the measured parameters presents great variations. One has to control large surfaces or a great number of samples of different origins. The median is useful as a comparative value.

REFERENCES

[1] MacMahon, D., Pearce, A., Harris, P., *Applied Radiation and Isotopes*, **60**, 275-281, 2004.
[2] Müler, J. W., *Journal of Research of the National Institute of Standards and Technology*, **105**, 551-555, 2000.
[3] Wilks, S. S., *Mathematical statistics*, Wiley, New York, 1962.
[4] Müler, J. W., *Weighted medians*, Rapport BIPM-2000/6.

Manuscript received: 28.04.2010
Accepted paper: 30.08.2010
Published online: 04.10.2010