Rapport BIPM-2000/6

Weighted medians

by Jörg W. Müller

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

Abstract

The evaluation of a median is extended to the case where (reliable) statistical weights are available. An estimate of the corresponding uncertainty is also given and the approach is illustrated by a numerical example. A special study explains why a weighted median always coincides with one of the measurements and why a similar rule holds for MAD, a measure of its uncertainty.

1. Introduction

Some time ago we proposed [1] to replace the mean value, which is traditionally used as location parameter, by the median as it has improved statistical "robustness". While suggesting this we had in mind the analysis of international comparisons, where the data come from participants who may have used different measuring techniques. The results also include estimations of the uncertainties. Although the principles used for their evaluation are now accepted on a worldwide basis [2], it would be naive to believe that all problems have therefore disappeared. Since the statistical weights of measurement results are in most cases derived directly from the uncertainties, any problem that may exist with uncertainties will necessarily also occur in the context of weights, normally even in a more pronounced way.

Experience shows that laboratories may have particular "habits". While some have a tendency to enlarge uncertainties, probably as a measure of "protection", others take the opposite position. If high precision is linked to prestige, there is a danger that stated values are sometimes unrealistically small. The analyzer of the submitted data then is in a most uncomfortable position. He notes the great spread in the uncertainties and suspects that some are too large, others too small, but obviously he cannot change them. On the other hand, if he takes them as stated to infer statistical weights, he well realizes that in this way he would favour the rash and punish the cautious experimenter. This looks like a situation without any satisfactory issue. All he can do (without endangering his relations with colleagues) is to ignore the stated uncertainties. Then all submitted values receive the same weight. This is not an uncommon situation for intercomparisons. Considering this, the fact that weights had been disregarded in [1] is perhaps not so great a loss.

Nevertheless, there clearly exist situations where the relative uncertainties, and therefore also the statistical weights of the input values, deserve confidence. In these cases it would be desirable to have available an algorithm which allows us to take them into account.

It is well known that the use of weights does not automatically improve the results. Wrong weights can even severely distort the data, and this can sometimes be seen afterwards. Thus, if the uncertainty of a weighted result is clearly larger than for unweighted data, we were probably misled in accepting the weights. In this case it may be wise to go back to the unweighted case. Obviously, this also applies to mean values.

2. The formal introduction of weights

Before we try to generalize medians, it may be useful to recall briefly what we know already [1].

Let us consider a set of n measurements. In the general case they will be noted z_i . This may correspond to the order in which they were obtained. However, for many situations it is preferable to assume that they are ordered. In this case, we shall denote them by x_i ,

$$\text{with} \quad x_i < x_{i+1} , \quad \text{for} \quad 1 \leq i \leq n-1 . \tag{1}$$

If two input values z coincide numerically, they may be combined into a single ordered value x, with summed weight. For simplicity, all values will be assumed positive.

For a series of n measurements z_i , the median, written as $\tilde{m} = \text{med}\{z_i\}$, can be considered as the solution of the condition

$$\sum_{i=1}^{n} |z_i - \tilde{m}| = \min . \tag{2}$$

The median $\tilde{m}$, therefore, plays a similar role as the better-known mean value $\bar{m}$ which is known to solve the analogous least-squares condition

$$\sum_i (z_i - \bar{m})^2 = \min . \tag{3}$$

As a measure of the precision of a median, it is practical to evaluate first a quantity called MAD (for "median of the absolute deviations"), defined by

$$\text{MAD} = \text{med}\{|z_i - \tilde{m}|\} , \tag{4}$$

which is the solution of

$$\sum_i ||z_i - \tilde{m}| - \text{MAD}| = \min. \tag{5}$$

We now want to apply statistical weights. It is not always quite clear on what they are based; an innocent view is to consider them just as experimentally "given". A better way would be to link them to the number of (equivalent) measurements, if these are known. More frequent will be the case where they are based on (supposedly reliable) uncertainties s_i, which themselves usually come from an evaluated variance or an associated quantity; $s^2(\tilde{m})$ will only rarely be available [1]. Whatever the exact origin of s_i, the weight $w_i > 0$ of a result z_i will be taken as

$$w_i = 1 / s_i^2 , \tag{6}$$

or also in its normalized form

$$p_i = \frac{w_i}{\sum_i w_i} , \tag{7}$$

with $\sum_i p_i = 1$.

It does not matter for the evaluation of the median if we use w_i or p_i since the minimum occurs for the same argument.

Although we assume here for convenience that the weights are positive, this is not a condition. As is well known [3], weights may become negative for (strongly) correlated data, but their formal treatment does not change. If we can suppose the results to be independent of each other, this complication does not occur.

The introduction of statistical weights is now straightforward. When we go back to the basic approach (2), it is readily seen that this can be generalized to

$$\sum_i p_i |z_i - \tilde{m}| = \min . \tag{8}$$

Likewise, relation (5), when including weights p_i, now becomes

$$\sum_i p_i ||z_i - \tilde{m}| - \mathrm{MAD}| = \min . \tag{9}$$

Equations (8) and (9) are the required generalizations for weighted medians.

The practical evaluation of $\tilde{m}$ and MAD is somewhat more cumbersome. Whereas previously the median could be easily found by ranking the available data (cf. eq. 2 in [1]), this is no longer possible with weights. The numerical solutions $\tilde{m}$ for (8) and MAD for (9) now have to be determined empirically, and a simple program (even on a pocket computer) will soon prove useful.
Once MAD is known numerically, the wanted uncertainty $s(\tilde{m})$ of the weighted sample median $\tilde{m}$ is still given by (cf. eq. 9 in [1])

$$s(\tilde{m}) \cong \frac{1.9}{\sqrt{n-1}} \mathrm{MAD} , \tag{10}$$

where n is the number of data z_i and with MAD now obtained from (9).

The practical application of the recipe sketched above to a particular case may serve as an illustration of the procedure. This is done below.

3. A numerical example

To illustrate how the determination of a median and its uncertainty can be obtained in the case of statistical weights, we use data which have been provided by the IAEA. They are in the form $z_i \pm s_i$, from which the weights were deduced.

i	z_i	s_i	$w_i = s_i^{-2}$	p_i
1	35.03	0.21	22.68	0.350
2	34.15	0.4	6.25	0.096
3	34.15	0.4	6.25	0.096
4	35.44	0.61	2.69	0.042
5	35.14	0.7	2.04	0.032
6	34.03	0.4	6.25	0.096
7	34.23	0.4	6.25	0.096
8	34.13	0.4	6.25	0.096
9	34.20	0.4	6.25	0.096
		sum :	64.91	1.000

a) Mean value

If we use the trial value $\tilde{m} = 34.2$, the 9 contributions to

$$Q = \sum_i p_i |z_i - \tilde{m}|$$

are, for the above data,

i	$p_i \|z_i - 34.2\|$
1	0.290 50
2	0.004 80
3	0.004 80
4	0.052 08
5	0.030 08
6	0.016 32
7	0.002 88
8	0.006 72
9	0.000 00
	$Q = 0.408\ 18$

For some other trial values chosen for $\tilde{m}$, the numerical results obtained for Q are assembled below.

$\tilde{m}$	Q	$\tilde{m}$	Q
34.0	0.542 90	34.18	0.412 82
34.1	0.456 34	34.22	0.407 38
34.2	0.408 18	34.23	0.406 98
34.3	0.417 62	34.24	0.408 50
34.4	0.432 82	34.25	0.410 02

We see that the minimum looked for occurs at

$$\tilde{m} = 34.23 .$$

b) Uncertainty

We form the quantity

$$Q' = \sum_i p_i \, ||z_i - \tilde{m}| - MAD| ,$$

with $\tilde{m} = 34.23$ and various trial values for MAD. The numerical results are

MAD	Q'	MAD	Q'
0.15	0.345 30	0.201	0.343 45
0.18	0.344 10	0.21	0.344 82
0.19	0.343 70	0.22	0.346 34
0.199	0.343 34	0.25	0.350 90
0.200	0.343 30	0.30	0.358 50

The minimum looked for occurs at MAD = 0.20 .

It follows from (10) that

$$s(\tilde{m}) \cong \frac{1.9}{\sqrt{8}} \, 0.20 \cong 0.13 .$$

Hence, we can write for the required result

$$\tilde{m} = 34.23 \pm 0.13 .$$

This may be compared with the unweighted case, for which we find

$$\tilde{m} = 34.20 \pm 0.05 .$$

Obviously, the fact that the weighted median has a lower precision than the unweighted one is not a very encouraging result. It may indicate that the weights used, which are of heterogeneous origin, are unreliable and should not be trusted. On the other hand, since the uncertainty can only assume a limited number of values (see below), caution is required in the interpretation. Alternatively, a comparison with the corresponding mean values may be of interest. They are

$$\begin{aligned} \bar{m} &= 34.50 \pm 0.18 , \quad \text{without weights,} \\ &= 34.54 \pm 0.17 , \quad \text{with weights.} \end{aligned}$$

These results seem to show that the uncertainty of the unweighted median is abnormally small, but the statistical basis is clearly too narrow for drawing a valuable conclusion. Dr. Ratel has studied several larger samples, especially international comparisons. It follows from them that weighting normally has a very limited influence on the median and its uncertainty. The situation is therefore similar to the one which applies to mean values, with little or no improvement by using weights given by the participants.

4. Possible values for median and MAD

A critical look at the numerical values obtained for the median in our example (as well as in others) reveals that they apparently always coincide with measured values z_i . While this is a well-established feature [1] for unweighted data, it is somewhat surprising to find that this peculiarity should still hold when arbitrary weights are applied. A keen observer may even find evidence that a similar principle seems to apply to the numerical values of MAD which always appear to be of the form $|\tilde{m} - z_i|$, i.e. they correspond to the difference between two measured values - a fact which apparently has not been noted before.

The question is whether the above observations are correct and, if so, whether this can be understood. As the treatment of this problem would sidetrack the present discussion, we shall discuss it later. It is explicitly solved in Appendix A for the median, and in Appendix B for MAD.

The result is not a mere curiosity, but is of real interest to the calculator. Without knowledge of this simplification, the numerical search for the parameters $\tilde{m}$ and MAD would have to be made for a considerable range of trial values. The following

two simple rules are obtained:

a) $\tilde{m}$ must be equal to a specific value z_i ,occasionally to two (or their mean);

b) MAD must be equal to $|\tilde{m} - z_i|$, i.e. to the difference between two input values, one of them being the median (occasionally to the mean of two differences).

Note that the complication with two possible solutions can only occur in the unweighted case, as shown in Appendix A.

Application of these rules greatly shortens the search. Another consequence is that there is no need to use trial values which have more decimals than the input values. On the other hand, the use of weights may result in a clear enlargement of the range of values.

Whereas, for the case of no weights, the median was always close to the central value, this is not necessarily so when weights are used. This fact is illustrated in Appendix C by an instructive simple case. It follows that, in principle, any input value, even an extreme one, can become a median; the decision depends entirely on the chosen statistical weights.

5. Final remarks

My interest in the evaluation of a weighted median was initiated by a request from Dr. P. Andreo, of the IAEA in Vienna. In December 1998 he sent me numerical data for a number of stopping-power ratios for protons, usually denoted by W and measured in eV. He had determined their median, but wondered if there was a possibility of doing this by using statistical weights, since their empirical uncertainties differed greatly. In January 1999 I sent him, as an answer, an outline of the suggested general approach, together with an application to his data. This document corresponds to a large extent to the present report. Apart from some additional text, the only new developments are those presented here as Appendices.

As in the meantime I have been approached by several other colleagues who asked me for a comprehensive text explaining how to use weights for a median, I decided to assemble the available information in a document that can be distributed to those interested. So much for the origin of this report, which may also explain some of its shortcomings.

I am grateful to Pedro Andreo (IAEA) for his stimulus, Guy Ratel (BIPM) for help with numerical calculations and Barry Taylor (NIST) for his interest. As is the case for all previous reports, this one could not have been issued either in its present form without the constant assistance of my wife Denise, who should be thanked for her patience.

APPENDICES

A. Position of the weighted median

One of the basic questions is why not only the ordinary median, but also its weighted form always seem to coincide with one of the original measurement values (the case without weights was treated in [1]).

With weights p_i and ordered values x_i, the problem is to find for which value $t = \tilde{m}$ the quantity

$$Q(t) = \sum_{i=1}^{n} p_i |t - x_i| \tag{A1}$$

assumes its minimum.

Let us consider for t the range $x_k \le t \le x_{k+1}$. We then have

$$\begin{aligned} Q(t) &= p_1(t-x_1) + \ldots + p_k(t-x_k) + p_{k+1}(x_{k+1}-t) + \ldots + p_n(x_n-t) \\ &= -p_1x_1 - \ldots - p_kx_k + p_{k+1}x_{k+1} + \ldots + p_nx_n + t(p_1 + \ldots + p_k - p_{k+1} - \ldots - p_n) . \end{aligned} \tag{A2}$$

Let us look at the values for the lower and upper limits of t.

$$\text{- for } t = x_k: \qquad Q_- = -\sum_{i=1}^{k} p_ix_i + \sum_{i=k+1}^{n} p_ix_i + x_k \left(\sum_{i=1}^{k} p_i - \sum_{i=k+1}^{n} p_i \right), \tag{A3}$$

$$\text{- for } t = x_{k+1}: \qquad Q_+ = -\sum_{1}^{k} p_ix_i + \sum_{k+1}^{n} p_ix_i + x_{k+1} \left(\sum_{1}^{k} p_i - \sum_{k+1}^{n} p_i \right).$$

We now form the difference

$$Q_+ - Q_- = (x_{k+1} - x_k) \left[\sum_{1}^{k} p_i - \sum_{k+1}^{n} p_i \right]. \tag{A4}$$

Since $x_{k+1} > x_k$, the sign of $Q_+ - Q_-$ is determined by the sign of

$$\Delta P \equiv \sum_{1}^{k} p_i - \sum_{k+1}^{n} p_i . \tag{A5}$$

Therefore

$$\begin{aligned} &\text{- if } \Delta P > 0: \quad Q_+ - Q_- > 0 \quad \text{and} \quad Q \text{ is minimal for} \quad \tilde{m} = x_k \,, \\ &\text{- if } \Delta P < 0: \quad Q_+ - Q_- < 0 \quad \text{and} \quad Q \text{ is minimal for} \quad \tilde{m} = x_{k+1} \,. \end{aligned} \tag{A6}$$

The only ambiguity for $\tilde{m}$ occurs if $\Delta P = 0$.

However, for $n = 2k$ and equal weights p for all values x_i, we always have $\Delta P = kp - (2k - k)\, p = 0$. This explains why, for an even number of data and no weights, the median can be $\tilde{m} = x_k$, x_{k+1} or any intermediate value.

With weights, the case $\Delta P = 0$, although possible, is unlikely to occur. For this reason, $\tilde{m}$ then practically always agrees with a (single) measured value.

B. In search of the structure of MAD

We have seen in (9) that MAD, the basic quantity used for the evaluation of uncertainties, is defined as the solution of

$$\tilde{Q} = \sum_i p_i \, ||x_i - \tilde{m}| - MAD| = \min \,. \tag{B1}$$

It is practical to use instead of MAD another variable. Without loosing anything in generality, we can, for example, put

$$MAD = |t - \tilde{m}| \,, \tag{B2}$$

introducing thereby t as the new variable. This should make it easier to see if MAD is indeed of the form suggested in section 4. The reasoning to be presented is similar to that outlined in Appendix A, although somewhat more involved.

For the transformation of absolute values into positive or negative ones, as required for determining a minimum, it is necessary to know the relative order of the three quantities x_i, $\tilde{m}$ and t, for which there are $3! = 6$ possibilities.

We shall see that it is sufficient to distinguish between two cases, namely

- α if both x_i and t are on the same side of $\tilde{m}$, i.e. $x_i, t < \tilde{m}$, or $x_i, t > \tilde{m}$;
- β if x_i and t are on different sides of $\tilde{m}$, i.e. $x_i < \tilde{m} < t$, or $t < \tilde{m} < x_i$.

Let us briefly discuss all possibilities.

Case α

Here we have

$$|x_i - \tilde{m}| - |t - \tilde{m}| = \tilde{m} - x_i - (\tilde{m} - t) = t - x_i , \qquad \text{for } x_i , t < \tilde{m} ,$$

$$\text{or} \qquad = x_i - \tilde{m} - (t - \tilde{m}) = x_i - t , \qquad \text{for } x_i , \ t > \tilde{m} .$$

Hence, for case α we always have

$$||x_i - \tilde{m}| - MAD| = |t - x_i| , \tag{B3}$$

and therefore also

$$\tilde{Q}(t) = \sum_i p_i |t - x_i| . \tag{B4}$$

Since (B3) is identical with (A1), it also assumes its minimum value at $t = x_k$ or x_{k+1}, as shown in (A6).

Case β

Here we have

$$|x_i - \tilde{m}| - |t - \tilde{m}| = x_i - \tilde{m} - (\tilde{m} - t) = x_i + t - 2\tilde{m} , \qquad \text{for } t < \tilde{m} < x_i$$

$$\text{or} \qquad = \tilde{m} - x_i - (t - \tilde{m}) = 2\tilde{m} - (x_i + t) , \quad \text{for } x_i < \tilde{m} < t .$$

Therefore, we always have for case β

$$||x_i - \tilde{m}| - MAD| = |x_i + t - 2\tilde{m}| , \tag{B5}$$

and thus also

$$\tilde{Q}(t) = \sum_i p_i |x_i + t - 2\tilde{m}| . \tag{B6}$$

We now look for the value of t which makes $\tilde{Q}$ minimal. There are again two possibilities.

Let us suppose that $x_i + t < 2\tilde{m} , \quad \text{for } t < \tilde{t} ,$

$$\text{but} \qquad x_i + t > 2\tilde{m} , \quad \text{for} \quad t > \tilde{t} .$$

Hence

$$|x_i + t - 2\tilde{m}| = 2\tilde{m} - x_i - t, \qquad \text{for } t < \tilde{t},$$

$$\text{but} \qquad = x_i + t - 2\tilde{m}, \qquad \text{for } t > \tilde{t}.$$

Since the critical value $\tilde{t}$ must necessarily fall in an interval bounded by measured values, say between x_k and x_{k+1}, we can also write for (B6)

$$\tilde{Q}(t) = \sum_{i=1}^{k} p_i (2\tilde{m} - x_i - t) + \sum_{i=k+1}^{n} p_i (x_i + t - 2\tilde{m})$$

$$= (2\tilde{m} - t) \left[\sum_{1}^{k} p_i - \sum_{k+1}^{n} p_i \right] - \sum_{i}^{k} p_i x_i + \sum_{k+1}^{n} p_i x_i . \qquad \text{(B7)}$$

With the abbreviation ΔP, introduced in (A5), we now obtain for the two limiting values of t

- for $t = x_k$: $$\tilde{Q}_- = (2\tilde{m} - x_k)\,\Delta P - \sum_{1}^{k} p_i x_i + \sum_{k+1}^{n} p_i x_i , \qquad \text{(B8)}$$

- for $t = x_{k+1}$: $$\tilde{Q}_+ = (2\tilde{m} - x_{k+1})\,\Delta P - \sum_{1}^{k} p_i x_i + \sum_{k+1}^{n} p_i x_i . \qquad \text{(B9)}$$

In order to find the minimum, we form the difference

$$\tilde{Q}_- - \tilde{Q}_+ = (x_{k+1} - x_k)\,\Delta P .$$

From this we conclude, since $x_{k+1} > x_k$, that

- for $\Delta P > 0$: $\tilde{Q}_- < \tilde{Q}_+$.

  Hence the minimum of $\tilde{Q}$ is reached at $t = x_k$, and therefore

$$\text{MAD} = |x_k - \tilde{m}| ; \qquad \text{(B10)}$$

- for $\Delta P < 0$: $\tilde{Q}_+ < \tilde{Q}_-$.

  Hence the minimum of $\tilde{Q}$ is reached at $t = x_{k+1}$, and therefore

It follows from all this that, for both cases α and β, MAD is indeed of the form $|\tilde{m} - z_i|$, as claimed in section 4.

Again, there can only be an ambiguity in the solution t for $\Delta P = 0$, as it had already occurred for the median $\tilde{m}$. For a discussion we refer to the end of Appendix A.

This result for the form of MAD is no doubt new and rather unexpected.

C. Extreme medians

Medians, according to their definition, are central values. While this is clearly true in the unweighted case, the question arises if it is still so when weights are used. After all, the purpose of weights is to give more credibility to some measurements than to others. If an extreme value has a high weight, can it become a median?

In order to decide this question, we consider a special case of n (ordered) results x_i . It is practical to use non-normalized weights, and we choose in particular

$$w_1 = M; \quad w_i = 1, \quad \text{for} \quad 2 \le i \le n, \tag{C1}$$

Since our interest is focused on the minimal value x_1 , we can write, for the range $x_1 \le t \le x_2$,

$$\begin{aligned} Q(t) &= M(t - x_1) + (x_2 - t) + \ldots + (x_n - t) \\ &= -Mx_1 + x_2 + \ldots + x_n + t[M - (n-1)] . \end{aligned} \tag{C2}$$

Hence, we have

$$\begin{aligned} &\text{- for } t = x_1 : \quad Q_- = -Mx_1 + (x_2 + \ldots + x_n) + x_1(M + 1 - n) \\ &\qquad = x_1(1 - n) + (x_2 + \ldots + x_n) , \\ &\text{- for } t = x_2 : \quad Q_+ = -Mx_1 + (x_2 + \ldots + x_n) + x_2(M - n + 1) \\ &\qquad = -Mx_1 + x_2(M - n + 2) + (x_3 + \ldots + x_n) . \end{aligned} \tag{C3}$$

The difference is

$$\begin{aligned} Q_+ - Q_- &= x_1(-M + n - 1) + x_2(M - n + 1) \\ &= (x_2 - x_1)(M + 1 - n) . \end{aligned} \tag{C4}$$

The minimum of Q is reached at x_1 if the difference is positive. For the median, this means that

$$\tilde{m} = x_1 , \quad \text{if } M > n-1 . \tag{C5}$$

This example therefore shows that also an extreme value (here the minimum x_1) may become a median, provided that its weight is sufficiently high.

In a similar way it is possible to show for these data that

$$\begin{aligned} \tilde{m} &= x_2 , \quad \text{if} \quad n-3 < M < n-1 , \\ &= x_3 , \quad \text{if} \quad n-5 < M < n-3 , \quad \text{etc.} \end{aligned} \tag{C6}$$

References

[1] J.W. Müller: "Possible advantages of a robust evaluation of comparisons", Rapport BIPM-95/2 (1995). This article has just been reprinted in J. Res. Natl. Inst. Stand. Technol. **105**, 551 (2000).

[2] "Guide to the Expression of Uncertainty in Measurement" (ISO, Geneva, 1995)

[3] J.W. Müller: "Some mathematical problems in counting statistics", in "Advanced Mathematical Tools in Metrology" (ed. by p. Ciarlini, M.G. Cox, R. Monaco and F. Pavese), World Scientific, Singapore, 1994, 197-203.

(October 2000)