Rapport BIPM-05/11

BUREAU INTERNATIONAL DES POIDS ET MESURES

Covariances for medians

Jörg W. Müller

November 2005

Pavillon de Breteuil, F-92312 SEVRES cedex

Rapport BIPM-2005/11

Covariances for medians

Jörg W. Müller

Abstract

A median is a simple measure of location, sometimes applied as an alternative or complement to the more common, but less robust, mean value. We propose to form a quantity which, in the context of medians, can be used in place of the traditional covariance. This will make it possible to adapt the general rules valid for the propagation of uncertainties to medians.

1. Introduction

Whenever we have to deal with quantities that are not independent of each other and for which we want to estimate the uncertainties, covariances come into play. This is already the case for a simple sum of the form (a, b = constants)

$$z = ax + by, \tag{1}$$

where the random variables x and y depend (or may be suspected to depend) on each other or could undergo a common influence (perhaps with opposite signs).

From repeated individual measurements of x and y it will be possible to obtain estimated values of their respective variance, say

$$\mathrm{Var}(x), \mathrm{Var}\,(y),$$

but the general expression for the variance of z is

$$\mathrm{Var}(z) = a^2\ \mathrm{Var}(x) + b^2\ \mathrm{Var}(y) + 2ab\ \mathrm{Cov}(x, y), \tag{2}$$

i.e. it also implies covariances.

It is therefore necessary to have available an estimate of the covariance. This is independent of what we choose for the "best" value of the sum z.

If we use mean values, then

$$\overline{z} = a\overline{x} + b\overline{y}, \tag{3a}$$

while for medians we would write

$$\tilde{z} = a\tilde{x} + b\tilde{y}. \tag{3b}$$

In both cases, the problem arises of knowing how the respective uncertainties (and covariances) for $\overline{z}$ or $\tilde{z}$ can be obtained. The answer will be of the form (2), but the estimates for the variances and covariances involved will be different.

As is well known [1], for mean values we have, for example, the relations (for a sample of size n)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4a}$$

and

$$\mathrm{Var}(\bar{x}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2. \tag{4b}$$

If one prefers to use the median, for instance because of its greater robustness, then one can write, as it has been shown previously [2],

$$\tilde{x} \equiv \mathrm{med}\{x_i\},$$

where $\tilde{x}$ is the solution of the condition

$$\sum_{i=1}^{n}\left|x_i - \tilde{x}\right| = \min, \tag{5}$$

and

$$\mathrm{Var}(\tilde{x}) = C^2\left\{\mathrm{MAD}(\tilde{x})\right\}^2,$$

with $$\mathrm{MAD}(\tilde{x}) \equiv \mathrm{med}\left\{\left|x_i - \tilde{x}\right|\right\} \quad \text{and} \quad C^2 \cong \frac{3.5}{n-1}. \tag{6}$$

2. Covariances

For mean values, a possibility for estimating the covariances is given by the relation

$$\mathrm{Cov}(\bar{x},\bar{y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right). \tag{7}$$

For medians, the expression for the covariance must be similar. If we require that

$$\mathrm{Cov}(a\tilde{x}, b\tilde{y}) = ab\ \mathrm{Cov}(\tilde{x},\tilde{y})$$

and

$$\mathrm{Cov}(\tilde{x},\tilde{y}) = \mathrm{Cov}(\tilde{y},\tilde{x}),$$

then essentially the only remaining possibility is (K = constant)

$$\mathrm{Cov}(\tilde{x},\tilde{y}) = K\ \mathrm{MAC}(\tilde{x},\tilde{y}),$$

with

$$\mathrm{MAC}(\tilde{x},\tilde{y}) \equiv \mathrm{med}\left\{(x_i - \tilde{x})(y_i - \tilde{y})\right\}. \tag{8}$$

If we put $K = C^2$ and since $\operatorname{med}\{x^2\} = [\operatorname{med}\{|x|\}]^2$, we also have, for $x = y$, the relation

$$\operatorname{Cov}(\tilde{x}, \tilde{x}) = \operatorname{Var}(\tilde{x}),$$

or

$$\operatorname{MAC}(\tilde{x}, \tilde{x}) = \operatorname{MAD}^2(\tilde{x}),$$

as is similarly the case for mean values.

Provided that reliable estimates of the statistical weights of the measured quantities are available, these can be readily incorporated [3]. In this case, (8) becomes

$$\operatorname{MAC}(\tilde{x}, \tilde{y}) = \operatorname{med}\{p_i q_i (x_i - \tilde{x})(y_i - \tilde{y})\}, \tag{9}$$

where p_i and q_i are the weights of x_i and y_i, respectively. It does not matter if they are normalized or not since the solution, MAC, of

$$\sum_i p_i q_i \left|(x_i - \tilde{x})(y_i - \tilde{y}) - \operatorname{MAC}\right| = \min \tag{10}$$

is the same*.

With these basic formulae at hand, it is now possible to deal readily with all practical cases for the propagation of uncertainties. The numerical example given in the Appendix illustrates the procedure.

3. Correlation coefficients

The correlation coefficient, usually defined by

$$\rho = \frac{\operatorname{Cov}(\bar{x}, \bar{y})}{\sqrt{\operatorname{Var}(\bar{x})\ \operatorname{Var}(\bar{y})}}, \tag{11}$$

is known to be limited by $\rho^2 \leq 1$. We note that the corresponding form with medians, i.e.

$$r = \frac{\operatorname{MAC}(\tilde{x}, \tilde{y})}{\operatorname{MAD}(\tilde{x})\ \operatorname{MAD}(\tilde{y})}, \tag{12}$$

is not necessarily within the range from –1 to + 1 since the Schwarz inequality

$$E^2(xy) \leq E(x^2)\, E(y^2), \tag{13}$$

on which the result for ρ relies, is no longer applicable.

4. Conclusions

A quantity, the median analogue to covariance (MAC), has been developed which, in the context of medians, can be used in place of the traditional covariance. This makes it possible to transform for medians, the general rules that are valid for the propagation of uncertainties. This enables medians to be used where traditionally means might have been used to analyze comparisons.

* MAC is formed in line with MAD, with C now referring to covariance.

Acknowledgements

Dr Maurice Cox (NPL) should be thanked for rightly insisting, in a recent discussion, that an expression for the covariance is needed before medians can be applied in a general way. The present report tries to fulfil this request.

Dr Guy Ratel (BIPM) has kindly read this text and proposed several improvements. Denise Müller is thanked for her usual editorial care.

References

[1] *Guide to the Expression of Uncertainty in Measurement* (ISO, Geneva, 1995).
[2] Müller J.W., Possible advantages of a robust evaluation of comparisons, 2000, *J.Res.Natl.Inst. Stand. Technol.*, **105**, 551.
[3] Müller J.W., Weighted medians, 2000, Rapport BIPM-2000/06.
[4] *Kleine Enzyklopädie Mathematik*, (Ed. Deutsch, Thun, 1977), p. 652.

1st version October 2001.
Revised version November 2005.

APPENDIX

A numerical application

The following example is taken from [4]. It concerns a sample of $n = 10$ pupils for whom height x_i cm and weight y_i kg have been measured.

The original data are

i	x_i	y_i
1	135	29.3
2	145	35.2
3	139	34.5
4	142	32.1
5	137	33.6

i	x_i	y_i
6	137	32.3
7	134	27.2
8	144	36.7
9	135	26.9
10	146	38.3

This readily yields

$$\tilde{x} = 138, \qquad \tilde{y} = 32.95 \quad \text{as well as}$$

$$\overline{x} = 139.4, \qquad \overline{y} = 32.61.$$

From the products

i	$(x_i - \tilde{x})(y_i - \tilde{y})$
1	10.95
2	15.75
3	1.55
4	-3.40
5	-0.65

i	$(x_i - \tilde{x})(y_i - \tilde{y})$
6	0.65
7	23.00
8	22.50
9	18.15
10	42.80

we see that $\mathrm{MAC}(\tilde{x}, \tilde{y}) = 13.35.$

Since

$$\mathrm{MAD}(\tilde{x}) = 3.5, \qquad \mathrm{MAD}(\tilde{y}) = 2.95 \quad \text{and} \quad C^2 \cong \frac{3.5}{9},$$

we find

$$\mathrm{Var}(\tilde{x}) = C^2 \mathrm{MAD}^2(\tilde{x}) \cong 4.76,$$

$$\mathrm{Var}(\tilde{y}) = C^2 \mathrm{MAD}^2(\tilde{y}) \cong 3.38$$

and

$$\mathrm{Cov}(\tilde{x}, \tilde{y}) = C^2 \mathrm{MAC}(\tilde{x}, \tilde{y}) \cong 5.19.$$

This leads to

$$r = \frac{\mathrm{Cov}(\tilde{x}, \tilde{y})}{\sqrt{\mathrm{Var}(\tilde{x})\ \mathrm{Var}(\tilde{y})}} \cong 1.29.$$

These values may be compared with the results based on mean values, i.e.

$$\mathrm{Var}(\bar{x}) \cong 2.03, \qquad \mathrm{Var}(\bar{y}) \cong 1.48 \quad \text{and} \quad \mathrm{Cov}(\bar{x}, \bar{y}) \cong 1.51,$$

thus

$$\rho \cong 0.87.$$

The agreement is not very good; the values based on the median are larger. However, this could be the price to pay for the improved stability.

Finally, another simple check may be appropriate. If we disregard the dimensions of x and y, they become two separate sequences of numbers on which mathematical operations can be applied. For the sums

$$z_i \equiv x_i + y_i$$

we readily obtain

$$\tilde{z} \cong 172 \qquad \text{and} \qquad \mathrm{MAD}(\tilde{z}) \cong 8,$$

thus

$$\mathrm{Var}(\tilde{z}) = C^2 \mathrm{MAD}^2(\tilde{z}) = 25.$$

This result would coincide with the expected value

$$\mathrm{Var}(\tilde{x}) + \mathrm{Var}(\tilde{y}) + 2\mathrm{Cov}(\tilde{x}, \tilde{y}) \quad \text{for} \quad \mathrm{MAC}(\tilde{x}, \tilde{y}) \cong 21.$$

However, since this is beyond our directly evaluated value (13.35), the latter can no longer be considered as too large.

It is likely, therefore, that the experimental values obtained are within their normal scatter.