Pooled variances

by Jörg W. Müller

Bureau International des Poids et Mesures, F-92310 Sèvres

Experimenters in general are well advised to subdivide lengthy measurements into a number of shorter ones, whenever feasible. This gives them the possibility - should anything have gone wrong - at least to be aware of the problem and to know roughly when it occurred. If the results do not indicate an anomaly, the partial results must be combined in such a way that the outcome characterizes the entire measuring period.

Problems with recent measurements have led us to re-consider the question of how experimental values for the variance, obtained in sequences of runs, should be used to arrive at a "best" overall estimate.

This problem occurs rather frequently and, no doubt, has been treated many times. A few years ago we looked at the special case of two samples [1]. In what follows these earlier findings are generalized to include an arbitrary number m of samples. It is shown that the contributions arising from the differences between the individual mean values can be incorporated in those terms that involve only the measured variances.

Suppose that we have to deal with a situation which is "under statistical control". What we actually require is that the first two moments of the number x of events, counted in a given time interval, show no significant trend with time.

Now consider measurements that have been performed on m samples, of size n_j (j = 1, 2, ... , m), taken randomly from some stable population. Let the results be available in the form of mean values $\bar{x}_j$ and variances s^2_j (for a single measurement), which have been obtained by forming from the measurements x_{ij} the quantities

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \qquad (1)$$

and

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \; . \qquad (2)$$

This allows us to form the weighted mean value

$$\bar{\bar{x}} = \frac{1}{N} \sum_{j=1}^{m} n_j \, \bar{x}_j \, ,$$

where

$$N = \sum_{j=1}^{m} n_j$$

is the total number of measurements performed. Note that all the variances (2) refer to a single measurement x_{ij}, not to a mean value $\bar{x}_j$.

Our problem is to evaluate, on the basis of the data given in (1) and (2), a value for the variance of a single measurement, denoted by $s^2(x)$. According to the definition of a variance, we have, considering all the N measurements performed, the estimation

$$s^2(x) = \frac{1}{N - 1} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{x}})^2 \, . \qquad (3)$$

This expression will now be rearranged. In a first step we easily find

$$(N\text{-}1) \cdot s^2(x) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left[(x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{\bar{x}}) \right]^2$$

$$= \sum_j \sum_i \left[(x_{ij} - \bar{x}_j)^2 + (\bar{x}_j - \bar{\bar{x}})^2 + 2 \, (x_{ij} - \bar{x}_j) \, (\bar{x}_j - \bar{\bar{x}}) \right] . \qquad (4)$$

Since according to (1)

$$\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j) = 0 \quad ,$$

we arrive, by using (2), at the form

$$s^2(x) = \frac{1}{N-1} \sum_{j=1}^{m} \left[(n_j-1)\, s_j^2 + n_j\, (\bar{x}_j - \bar{\bar{x}})^2 \right]^2 . \tag{5}$$

This is a useful identity. For the special case of m = 2 samples, it follows from (1) that

$$\bar{x}_1 - \bar{\bar{x}} = \bar{x}_1 - \frac{1}{N} (n_1\, \bar{x}_1 + n_2\, \bar{x}_2) = \frac{n_2}{N} (\bar{x}_1 - \bar{x}_2),$$

and similarly

$$\bar{x}_2 - \bar{\bar{x}} = \frac{n_1}{N} (\bar{x}_2 - \bar{x}_1) \; .$$

Hence, the relation (5) takes the form

$$s^2(x) = \frac{1}{N-1} \left[(n_1-1)s_1^2 + (n_2-1)s_2^2 + \frac{n_1 n_2}{N} (\bar{x}_1 - x_2)^2 \right] , \tag{6}$$

in agreement with a result given in [1].

This procedure can be taken a step further. To see this, let us recall the general formula for the variance $s^2(\bar{y})$ of a weighted mean value $\bar{y}$, based on m measured results y_k, with statistical weights g_k. This expression is known to be given by

$$s^2(\bar{y}) = \frac{\sum_{k=1}^{m} g_k \; (y_k - \bar{y})^2}{(m-1) \sum_{k=1}^{m} g_k} . \tag{7}$$

Since in our case, i.e. for the measurement $\bar{x}_j$, the sample sizes n_k correspond to the weights g_k, we have the relation

$$\begin{aligned} \sum_{j=1}^{m} n_j \; (\bar{x}_j - \bar{\bar{x}})^2 &= (m-1) \sum_j n_j \cdot s^2 (\bar{\bar{x}}) \\ &= (m-1) \cdot s^2(x) , \end{aligned} \tag{8}$$

as $\sum_j n_j = N$ and $s^2(\bar{\bar{x}}) = s^2(x)/N$.

Note, however, that the last equation is not an identity: it is a statistical relation. Thus, it does not always hold numerically. It is true on the average and requires that experimental conditions do not change.

If we take advantage of (8), (5) can be brought into the form

$$(N-1) \cdot s^2(x) = \sum_j (n_j - 1) \cdot s_j^2 + (m-1) \cdot s^2(x) ,$$

or

$$s^2(x)[N-1 - (m-1)] = \sum_j (n_j - 1) \cdot s_j^2 .$$

Hence, we arrive at the general formula

$$s^2(x) = \frac{\sum_{j=1}^{m} (n_j - 1) \cdot s_j^2}{N - m} , \tag{9}$$

which allows us to obtain the required variance from those measured in the m samples. This form is quite remarkable in that the experimental mean values $\bar{x}_j$, present in (5), have disappeared.

For samples of equal size ($n_j = n$) we are readily led to the expression

$$s^2(x) = \frac{1}{m} \sum_{j=1}^{m} s_j^2 \, , \tag{10}$$

for which even a knowledge of the sample size is no longer required.

A comparison of (10) with (3) shows that, at least for equal groups of measurements obtained in stable conditions, pooled estimates of both the expectation value and the variance are obtained by simply forming arithmetical means. The reader, in hindsight, should feel free to find this result either trivial or quite remarkable. It is likely, although not proven, that analogous relations hold for the central moments of higher order.

References

[1] J.W. Müller: "Moments d'échantillons superposés", BIPM WPN-215 (1980).

(April 1993)