ISSN 1757- 5958

amc technical briefs

Editor: Michael Thompson Analytical Methods Committee AMCTB No 30 April 2008

The standard deviation of the sum of several variables

Usually we consider the variance of the sum of several variables to be the sum of the variances of the individual variables, but that is true only if the variables are not correlated. If they are correlated we should take account of the covariance, which is related to the correlation coefficients as shown below.

When the variables are results for concentrations of several analytes measured in a single analytical operation (for example in a single injection into an HPLC column) the results may be correlated because some of the uncontrolled variations, such as different recovery in each run, will affect all analytes to roughly the same extent.

Example

In the analysis of aflatoxins in foodstuffs, it is normal to report separately the results for the four analytes, aflatoxins B1, B2, G1, and G2, but the total aflatoxin content is also important, as there are regulatory limits for total aflatoxin. Table 1 shows an example, in which several laboratories have reported separate results for the four aflatoxins in a particular material. Their variances and standard deviations are also shown. Can we estimate the standard deviation for total aflatoxin directly from these four standard deviations?

Suppose that we assumed that the results for the four aflatoxins are uncorrelated. We would take the standard deviation of the sum of the four results to be the square root of the sum of the individual variances, namely

$$\sqrt{1.881+0.438+0.259+0.129}=1.65.$$

But if we calculate the individual total aflatoxin results for the laboratories, we get values shown in the last columns of Table 1. These have a standard deviation of 2.53, considerably larger than the calculation above. What has gone wrong?

Covariance

The answer is that the observations are not independent—they show appreciable correlation. This can be seen by calculating their covariances cov(x,y) (Table 2) and the related correlation coefficients $r(x,y)$ (Table 3), using the formulae

$$\operatorname{cov}(x,y)=\sum_i\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\Big/(n-1)$$
$$\equiv r(x,y)s_x s_y,$$

where x_i, y_i are the i-th pair of the variables x, y, and s_x, s_y are individual standard deviations. Several of the correlation coefficients are well over 0.5, indicating that they will substantively affect the combined standard deviation. (This should be taken as a rule-of-thumb—0.5 is not a critical level!)

***Warning**: The covariance calculation in at least one popular spreadsheet uses n instead of n-1 as the denominator in the calculation. Take care to adjust the result accordingly.*

Table 1. Replicated results from the determination of aflatoxins (ppb mass fraction)

Laboratory	afb1	afb2	afg1	afg2	Total
1	8.5	4.3	3.5	1.6	17.9
2	4	2.5	1.7	2.1	10.3
3	6.6	3.6	2.1	2	14.3
4	5.9	3.4	2.3	2.2	13.8
5	4.2	2.2	1.8	1.6	9.8
6	6.2	3.5	2.6	2.7	15.0
7	7.1	3.8	2.6	2.5	16.0
8	5.2	3.4	2.1	2.2	12.9
9	4.9	2.45	2.15	1.8	11.3
10	6.3	3.3	2.3	1.9	13.8
Variance	1.881	0.438	0.259	0.129	6.40
Standard deviation	1.371	0.662	0.509	0.359	2.53

Table 2. Covariance matrix

	afb1	afb2	afg1	afg2
afb1	**1.881**	0.848	0.635	0.032
afb2	0.848	**0.438**	0.277	0.066
afg1	0.635	0.277	**0.259**	0.000
afg2	0.032	0.066	0.000	**0.129**

Table 3. Correlation coefficients

	afb1	afb2	afg1	afg2
afb1	1	0.934	0.909	0.064
afb2	0.934	1	0.823	0.276
afg1	0.909	0.823	1	0.000
afg2	0.064	0.276	0.000	1

Taking proper account of covariance

With some correlation between the variables, as we have in our example data, the correct standard deviation of the sum is the square root of the *sum of the variances and the covariances*, that is:

$$\sqrt{\begin{pmatrix}1.881+0.438+0.259+0.129\\+2(0.848+0.635+0.032\\+0.277+0.066+0.000)\end{pmatrix}}=2.53.$$

As we saw above, this result can be obtained directly as the simple standard deviation of the calculated total aflatoxin contents above. (Notice that both $\mathrm{cov}(x,y)$ and $\mathrm{cov}(y,x)$ have to be included in the sum, hence the factor of 2 for the covariance terms).

Uncertainty and correlation

Exactly the same principles as above apply when estimating the standard uncertainty for sums of variables. The standard deviation above provides the standard uncertainty associated with random effects for the total aflatoxin content reported by a single laboratory. For an average of *n* results (for example, from a series of observations within a single laboratory run), the calculated standard deviation should be divided by $\sqrt{n}$ in the usual way.

Conclusion

If there is any correlation between variables which are to be added together (or, indeed, combined in any way), it is important to take proper account of that correlation in estimating the standard deviation of the result. If the raw data are available, this can either be done by calculating the individual results and taking their standard deviation or by calculating the necessary covariances and summing those. If only the standard deviations and correlation coefficients (or covariances) are available, it is necessary to calculate the combined standard deviation from the covariances.

Further reading

Shall we consider covariances?
W Bremser, W Haesselbarth
Accred Qual Assur, 1998, **3,** 106-110.

Including correlation effects in an improved spreadsheet calculation of combined standard uncertainties
S L. R. Ellison
Accred Qual Assur, 2005, **10** 338-343.

This Technical Brief was prepared for the Analytical Methods Committee by the Statistical Subcommittee (Chair M Thompson).

The Analytical Methods Committee (AMC) is the Committee of the Analytical Division Council that handles matters of technical importance to the members of the RSC Analytical Division and the analytical science community in general. The broad aim of the AMC is to participate in national and international efforts to establish a comprehensive framework for appropriate quality in chemical measurement, and to keep the analytical science community informed of developments. It achieves this aim through the activities of its expert sub-committees.

The aim of the Statistical Methods Subcommittee is to optimise the usage of statistical methods within the analytical community. That aim is addressed by:

- providing information about good basic statistical practice
- investigating the benefits and limitations of both traditional and more modern statistical methods in analytical science
- facilitating the use of newer statistical techniques by providing the necessary information in a readily usable form
- applying statistical principles to undecided problems relating to quality in analytical data.

Comments and queries about
AMC Technical Briefs
can be posted on the
AMC Web-board.

Find out more
about the AMC on:
www.rsc.org/amc

CPD Certification I certify that I have studied this document as a contribution to Continuing Professional Development.

Name……………………...
Signature…………………………….…….Date………

Name of supervisor………………………………..
Signature…………………….……………Date…………