

On the Propagation of Statistical Errors

Lara De Nardo

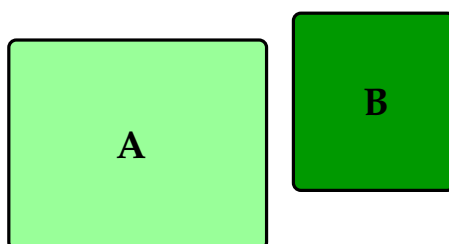
March 28, 2002

1 Statistical errors

1.1 Introduction.

The problem of calculating the error of an expression containing quantities that are correlated, i.e. measured on a common data sample, is common in data analysis. The way to handle the propagation of errors is not always clear, though. The case when one wants to see whether the measured asymmetry in one particular time period is compatible with the asymmetry calculated over the whole data set (which includes all time periods) is an example in which one needs to propagate the error in a function of two quantities (the asymmetries) that are totally correlated. If one wants to check that the asymmetry calculated in a certain z_{vertex} bin agrees with the asymmetry in a different z_{vertex} bin then the error for independent variables should be used. In the cross-check between two different analysis the problem becomes very complicated, as one needs to consider that there are events that are not common to the two data sets, so the propagation of errors in partially correlated variables should be used. In this brief report I will review the error calculation for the three cases of independent, partially correlated and totally correlated variables. I will always assume that the quantities under consideration, that I will call E_A and E_B , represent *the same physical entity*, as for example an asymmetry.

1.2 Independent variables.



Given M independent measurements E_i of the same quantity E with different standard deviations σ_i , the best estimate for E is given by the weighted mean:

$$E = \frac{\sum_{i=1}^M \frac{E_i}{\sigma_i^2}}{\sum_{i=1}^M \frac{1}{\sigma_i^2}}, \quad \frac{1}{\sigma^2} = \sum_{i=1}^M \frac{1}{\sigma_i^2} \quad (1.1)$$

where σ is the standard deviation in the weighted mean.

For example, if two quantities $E_A \pm \sigma_A$ and $E_B \pm \sigma_B$ are calculated from different data sets, namely A and B , with $A \cap B = \emptyset$, then they are independent, and the best estimate of E , over the whole data set $A+B$ is given by the weighted mean.

The error of any function $f(E_A, E_B)$ has the form:

$$\sigma_f = \sqrt{\left(\frac{\partial f}{\partial E_A} \sigma_A\right)^2 + \left(\frac{\partial f}{\partial E_B} \sigma_B\right)^2} \quad (1.2)$$

This is, for example, the case when one wants to compare the asymmetries in two different bins of z_{vertex} , were the function f may be the difference $E_A - E_B$ or the ratio E_A/E_B .

1.3 Correlated variables.

When the statistics involved in calculating E_A and E_B are not independent, the error for a function $f(E_A, E_B)$ has the expression:

$$\sigma_f = \sqrt{\left(\frac{\partial f}{\partial E_A} \sigma_A\right)^2 + \left(\frac{\partial f}{\partial E_B} \sigma_B\right)^2 + 2 \frac{\partial f}{\partial E_A} \frac{\partial f}{\partial E_B} \text{cov}(E_A, E_B)}, \quad (1.3)$$

where the last term takes care of the correlations between E_A and E_B .

Given a large number N of measurements E_{A_i} , the standard deviation σ_A is empirically defined as:

$$\sigma_A^2 = \frac{1}{N-1} \sum_{i=1}^N (E_{A_i} - E_A)^2, \quad (1.4)$$

while the covariance between E_A and E_B is given by:

$$\text{cov}(E_A, E_B) = \frac{1}{N-1} \sum_{i=1}^N (E_{A_i} - E_A)(E_{B_i} - E_B) \quad (1.5)$$

where E_A and E_B are the averages of E_{A_i} and E_{B_i} ¹. When E_A and E_B are independent, over a large number N of measurements they will fluctuate around their average in an uncorrelated way, so that the covariance is zero and one recovers the usual formula for the propagation of errors in a function of independent variables. From eq.(1.4) it follows that

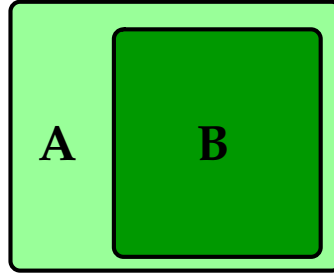
$$\text{cov}(E_A, E_A) = \sigma_A^2, \quad (1.6)$$

while the linearity properties of the covariance follow from eq.(1.5):

$$\text{cov}(aE_A + bE_B, E_c) = a \text{cov}(E_A, E_c) + b \text{cov}(E_B, E_c) \quad (1.7)$$

that will prove to be useful later (here a and b are constants).

It is worth noting that the covariance is a property only of E_a and E_B , and not of the specific form of the function f .



1.3.1 Totally correlated variables.

Let us suppose that our data (the set A) is divided into M disjunct samples (as bins in time, or in z_{vertex} , or θ_y).

In this case the relation

$$\frac{E_A}{\sigma_A^2} = \sum_{i=1}^M \frac{E_i}{\sigma_i^2} \quad (1.8)$$

holds (see eq.(1.1)), where the indices i indicate the independent subsets of A .

We need to check whether E_B agrees with E_A , where A is the total set that includes all M bins, and B is a subset of A .

Then for a particular bin corresponding to set B , the relation between E_A and E_B is given by the weighted mean:

$$E_A = \frac{\sigma_A^2}{\sigma_B^2} E_B + \frac{\sigma_A^2}{\sigma_{A-B}^2} E_{A-B} \quad (1.9)$$

where remaining terms containing the values of E_i in the other bins i have gone into $A - B$. The relation between E_A and E_B is then linear and one can apply eqs.(1.6) and (1.7) to get the covariance $\text{cov}(E_A, E_B)$:

$$\text{cov}(E_A, E_B) = \frac{\sigma_A^2}{\sigma_B^2} \text{cov}(E_B, E_B) + \frac{\sigma_A^2}{\sigma_{A-B}^2} \text{cov}(E_{A-B}, E_B) = \frac{\sigma_A^2}{\sigma_B^2} \sigma_B^2 = \sigma_A^2, \quad (1.10)$$

where we used the independence of E_{A-B} and E_B , which gives $\text{cov}(E_{A-B}, E_B) = 0$.

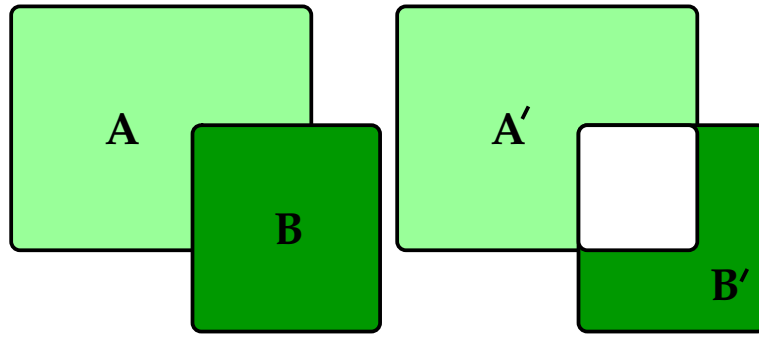
The standard deviation in f will be:

$$\sigma_f = \sqrt{\left(\frac{\partial f}{\partial E_A}\right)^2 \sigma_A^2 + \left(\frac{\partial f}{\partial E_B}\right)^2 \sigma_B^2 + 2 \frac{\partial f}{\partial E_A} \frac{\partial f}{\partial E_B} \sigma_A^2} \quad B \subseteq A. \quad (1.11)$$

1.3.2 Partially correlated variables.

A more difficult case is when the two quantities E_A and E_B under consideration are calculated using two data sets that have a non zero intersection, as it could happen for example if one wants to see the agreement between the asymmetry calculated in $12 \text{ cm} < z_{vertex} < 18 \text{ cm}$ and one calculated for $10 \text{ cm} < z_{vertex} < 15 \text{ cm}$. In this case the two quantities are only partially correlated.

¹The averaged value of E is supposed to be a good approximation of the *true* value, so they are assumed to be equal, and no distinction is going to be made between the two.



To calculate the covariance let us introduce two sets A' and B' such that $A = A' + A \cap B$ and $B = B' + A \cap B$. It must be:

$$\begin{aligned} \frac{E_A}{\sigma_A^2} &= \frac{E_{A'}}{\sigma_{A'}^2} + \frac{E_{A \cap B}}{\sigma_{A \cap B}^2} \\ \frac{E_B}{\sigma_B^2} &= \frac{E_{B'}}{\sigma_{B'}^2} + \frac{E_{A \cap B}}{\sigma_{A \cap B}^2} \end{aligned} \tag{1.12}$$

so that

$$E_A = \frac{\sigma_A^2}{\sigma_{A'}^2} E_{A'} + \frac{\sigma_A^2}{\sigma_B^2} E_B - \frac{\sigma_A^2}{\sigma_{B'}^2} E_{B'} \tag{1.13}$$

The covariance $\text{cov}(E_A, E_B)$ is:

$$\begin{aligned} \text{cov}(E_A, E_B) &= \frac{\sigma_A^2}{\sigma_{A'}^2} \text{cov}(E_{A'}, E_B) + \frac{\sigma_A^2}{\sigma_B^2} \text{cov}(E_B, E_B) - \frac{\sigma_A^2}{\sigma_{B'}^2} \text{cov}(E_{B'}, E_B) \\ &= \sigma_A^2 - \frac{\sigma_A^2}{\sigma_{B'}^2} \text{cov}(E_{B'}, E_B) \\ &= \sigma_A^2 - \frac{\sigma_A^2}{\sigma_{B'}^2} \sigma_B^2 \end{aligned} \tag{1.14}$$

where we used the fact that $\text{cov}(E_{A'}, E_B) = 0$ (A' and B are independent) and $\text{cov}(E_B, E_{B'}) = \sigma_B^2$ as follows from eq.(1.10), since $B' \subseteq B$. Using the relation $\frac{1}{\sigma_B^2} = \frac{1}{\sigma_{B'}^2} + \frac{1}{\sigma_{A \cap B}^2}$, we get the covariance in the case of partially correlated variables:

$$\text{cov}(E_A, E_B) = \frac{\sigma_A^2 \sigma_B^2}{\sigma_{A \cap B}^2} \tag{1.15}$$

This expression recovers both the errors for the case of independent variables than the one for totally correlated, in the two limits of $\sigma_{A \cap B} = \infty$ and $\sigma_B = \sigma_{A \cap B}$. In this case however the knowledge of $E_A \pm \sigma_A$ and $E_B \pm \sigma_B$ alone is not enough to calculate the error in any expression including A and B , since one also needs the error in the intersection $A \cap B$.

1.3.3 A useful table.

Table 1.1 contains a compilation of errors for some functions, for the three cases of independent, completely and partially correlated quantities.


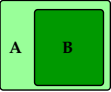
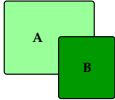
 Independent		 Completely Correlated		 Partially Correlated	
f	σ_f	f	σ_f	f	σ_f
$E_A - E_B$	$\sqrt{\sigma_A^2 + \sigma_B^2}$	$E_A - E_B$	$\sqrt{ \sigma_A^2 - \sigma_B^2 }$	$E_A - E_B$	$\sqrt{\sigma_A^2 + \sigma_B^2 - 2\frac{\sigma_A^2\sigma_B^2}{\sigma_{A\cap B}^2}}$
$\frac{E_A - E_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}$	1	$\frac{E_A - E_B}{\sqrt{ \sigma_A^2 - \sigma_B^2 }}$	1	$\frac{E_A - E_B}{\sqrt{\sigma_A^2 + \sigma_B^2 - 2\frac{\sigma_A^2\sigma_B^2}{\sigma_{A\cap B}^2}}}$	1
$\frac{E_A}{E_B}$	$\frac{E_A}{E_B} \sqrt{\frac{\sigma_A^2}{E_A^2} + \frac{\sigma_B^2}{E_B^2}}$	$\frac{E_A}{E_B}$	$\frac{E_A}{E_B} \sqrt{\frac{\sigma_A^2}{E_A^2} + \frac{\sigma_B^2}{E_B^2} - \frac{2}{E_A E_B} \sigma_A^2}$	$\frac{E_A}{E_B}$	$\frac{E_A}{E_B} \sqrt{\frac{\sigma_A^2}{E_A^2} + \frac{\sigma_B^2}{E_B^2} - \frac{2}{E_A E_B} \frac{\sigma_A^2\sigma_B^2}{\sigma_{A\cap B}^2}}$

Table 1.1: This table shows the errors for some simple functions, useful to check the agreement between two quantities E_A and E_B . The cases of complete independence, complete and partial correlation (that is one of the two, either E_A or E_B , is calculated over a data set that is included in the other) are considered.