How to combine data sets from different laboratories

Sometimes data sets used to determine the values of physical quantities are gathered from experiments or measurements performed at geographically different laboratories. This circumstance presents the problem of how to combine these data sets in the fairest manner for all concerned. Assuming that the different groups providing the data are essentially equally well trained and, therefore, essentially equally competent to do the measurements, the question arises as to whether or not an objective method for combining their data sets exists. In this paper a proposal that describes and justifies such a method is presented. We assume that the data sets can be presented as probability distributions $P_1(x), P_2(x), P_3(x), \ldots, P_n(x)$ where the subscripts denote the sources of the data sets. Below, we define the *conflation* of these distributions, denoted by $\&(P_1(x), P_2(x), \ldots, P_n(x))$, and give some of the main results for the special but rather general case of normal distributions. If the data sets are given unequal weights, say as adjudged by one of the laboratories, then a formula with weights exists and is

$$\& \left(P_1^{\frac{w_1}{w_{max}}}(x), P_2^{\frac{w_2}{w_{max}}}(x), \ldots, P_n^{\frac{w_n}{w_{max}}}(x) \right)$$

in which w_{max} is the maximum weight in the set $(w_1, w_2, \ldots, w_n)$. This definition reduces to the ordinary conflation, denoted above, when the weights are equal.

The idea of *conflation* was proposed by Ted Hill, I provided the name, as well as several explicit results for normal distributions, the distributions of most interest to experimental physicists. Ted proposed using the ampersand, $\&$, to denote conflation. Ted explored a very general setting for this subject using modern probability theory and established several theorems [T. P. Hill]. Because the experimentalist would benefit from a much more transparant presentation, I am writing this document.

Normal probability densities.

Frequently, measured data is reported in the form of probability density that approximates a normal density, $N(x, m, \sigma^2)$, defined by

(1)

$$N(x, m, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$$

in which x is the independent variable and runs over the entire real line (these constraints are relaxed in [Hill]). The mean value of x is given by m and the variance, σ^2, is the square of the standard deviation, σ.

The conflation of two normal densities is defined by the left equality below while the right equality will be proved

(2)

$$\&(N(x,m_1,\sigma_1^2)N(x,m_2,\sigma_2^2)) = \frac{N(x,m_1,\sigma_1^2)N(x,m_2,\sigma_2^2)}{\int dy\ N(y,m_1,\sigma_1^2)N(y,m_2,\sigma_2^2)} = \frac{1}{\Sigma\sqrt{2\pi}}exp\left[-\frac{(x-M)^2}{2\Sigma^2}\right]$$

$$= N(x,M,\Sigma^2)$$

where

(3)

$$\frac{1}{\Sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

and

(4)

$$M = \Sigma^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right)$$

Proof: Ignore the normalization factors and complete the square in the numerator of the conflation in eq.(2):

(5)

$$exp\left[-\frac{(x-m_1)^2}{2\sigma_1^2} - \frac{(x-m_2)^2}{2\sigma_1^2}\right] = exp\left[-\left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right)x^2 + \left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right)x - \frac{m_1^2}{2\sigma_1^2} - \frac{m_2^2}{2\sigma_2^2}\right]$$

$$= exp\left[-\frac{1}{2\Sigma^2}\left(x - \Sigma^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right)\right)^2 + \frac{\Sigma^2}{2}\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right)^2 - \left(\frac{m_1^2}{2\sigma_1^2} + \frac{m_2^2}{2\sigma_1^2}\right)\right]$$

where Σ^2 is given above in eq.(3). The x-independent terms can be ignored while determining the normalization, that for the remaining gaussian form is given by $\Sigma\sqrt{2\pi}$. QED Especially note that the product in Eq.(2) is taken for both distributions evaluated at the same point, x.

By similar arguments the conflation of n normal densities is associative, straight-forward and results in the normal density

(6)

$$N(x,M_n,\Sigma_n^2)$$

in which

$$\Sigma_n^{-2} = \sigma_1^{-2} + \sigma_2^{-2} + \cdots + \sigma_n^{-2} \tag{7}$$

and

$$M_n = \Sigma_n^2(m_1\sigma_1^{-2} + m_2\sigma_2^{-2} + \cdots + m_n\sigma_n^{-2}) \tag{8}$$

The standard deviation, σ, determines how sharply peaked the normal distribution is. The smaller σ is the sharper the peak [Hill and Miller]. All else being equal one feels justified in giving higher weight to means determined by sharper distributions. This is exactly what eq.(4) implies. More weight is given to the mean with the smaller standard deviation. This is generalized to n distributions in Eq.(8).

For the weighted case, the corresponding formulae are

$$\Sigma_n^{-2} = \frac{w_1}{w_{max}}\sigma_1^{-2} + \frac{w_2}{w_{max}}\sigma_2^{-2} + \cdots + \frac{w_n}{w_{max}}\sigma_n^{-2} \tag{9}$$

$$M_n = \Sigma_n^2\left(\frac{w_1}{w_{max}}m_1\sigma_1^{-2} + \frac{w_2}{w_{max}}m_2\sigma_2^{-2} + \cdots + \frac{w_n}{w_{max}}m_n\sigma_n^{-2}\right) \tag{10}$$

Perhaps of greater interest is the case of n laboratories and N variables that may have correlations. In the Gaussian case all statistics are determined by the mean values and the correlation matrix [Fox, 1978 Section I.1]. Let the N variables be denoted by $x_1, x_2, \ldots, x_N$. The normal distribution for the case in which the mean value of x_i is m_i and in which the correlation matrix is C_{ij} is given by

$$N(x_1, x_2, \ldots, x_N, m_1, m_2, \ldots, m_N, \boldsymbol{C}) = \left(\frac{det(\boldsymbol{C}^{-1})}{(2\pi)^N}\right)^{1/2} exp\left(-\frac{1}{2}(x_i - m_i)C_{ij}^{-1}(x_j - m_j)\right) \tag{11}$$

in which the repeated indices in the exponential are summed. The mean values satisfy

$$m_i = \int dx_1 dx_2 \ldots dx_N \ x_i \ N(x_1, x_2, \ldots, x_N, m_1, m_2, \ldots, m_N, \boldsymbol{C}) \tag{12}$$

for $i = 1,2, \ldots, N$ and the correlations satisfy

$$\tag{13}$$

$$C_{ij} = \int dx_1 dx_2 \dots dx_N \, (x_i - m_i)(x_j - m_j) \, N(x_1, x_2, \dots, x_N, m_1, m_2, \dots, m_N, \mathbf{C})$$

for $i, j = 1, 2, \dots, N$. Clearly, $\mathbf{C}$ is a real symmetric matrix. Consequently so is $\mathbf{C}^{-1}$. The verification of the preceeding results follows from diagonalization of the correlation matrix that results in N independent Gaussian variables.

In the one variable case presented at the beginning of this essay, the subscrips referred to the laboratory providing the data. In the multi-variable case just discussed the subscripts refer to the N variables measured in each laboratory. We need to find a new location for the laboratory label in the conflation formulae for N variables. We will use superscripts in parentheses for this purpose. Therefore the conflation of two data distributions for N variables is defined by

$$(14)$$

$$\&\left(N\left(x_1, x_2, \dots, x_N, m_1^{(1)}, m_2^{(1)}, \dots, m_N^{(1)}, \mathbf{C}^{(1)}\right) N\left(x_1, x_2, \dots, x_N, m_1^{(2)}, m_2^{(2)}, \dots, m_N^{(2)}, \mathbf{C}^{(2)}\right)\right)$$

$$= \frac{N\left(x_1, x_2, \dots, x_N, m_1^{(1)}, m_2^{(1)}, \dots, m_N^{(1)}, \mathbf{C}^{(1)}\right) N\left(x_1, x_2, \dots, x_N, m_1^{(2)}, m_2^{(2)}, \dots, m_N^{(2)}, \mathbf{C}^{(2)}\right)}{\int dy_1 dy_2 \dots dy_N \, N\left(y_1, y_2, \dots, y_N, m_1^{(1)}, m_2^{(1)}, \dots, m_N^{(1)}, \mathbf{C}^{(1)}\right) N\left(y_1, y_2, \dots, y_N, m_1^{(2)}, m_2^{(2)}, \dots, m_N^{(2)}, \mathbf{C}^{(2)}\right)}$$

$$= \left(\frac{\det(\boldsymbol{\chi}^{-1})}{(2\pi)^N}\right)^{1/2} \exp\left(-\frac{1}{2}(x_i - M_i)(\boldsymbol{\chi}^{-1})_{ij}(x_j - M_j)\right)$$

$$= N(x_1, x_2, \dots, x_N, M_1, M_2, \dots, M_N, \boldsymbol{\chi})$$

in which

$$(15)$$

$$M_j = (\boldsymbol{\chi})_{jk}\left(\left(\mathbf{C}^{(1)^{-1}}\right)_{kl} m_l^{(1)} + \left(\mathbf{C}^{(2)^{-1}}\right)_{kl} m_l^{(2)}\right)$$

and

$$(16)$$

$$\boldsymbol{\chi}^{-1} = \mathbf{C}^{(1)^{-1}} + \mathbf{C}^{(2)^{-1}}$$

It is obvious how to generalize this result to n laboratories. The results are (sum over repeated indices)

$$(17)$$

$$M_j = (\boldsymbol{\chi})_{jk}\left(\left(\mathbf{C}^{(1)^{-1}}\right)_{kl} m_l^{(1)} + \left(\mathbf{C}^{(2)^{-1}}\right)_{kl} m_l^{(2)} + \dots + \left(\mathbf{C}^{(n)^{-1}}\right)_{kl} m_l^{(n)}\right)$$

and

$$(18)$$

$$\chi^{-1} = C^{(1)^{-1}} + C^{(2)^{-1}} + \cdots + C^{(n)^{-1}}$$

The correspondence between Eqs.(7,8) and Eqs.(17,18) is manifest. Finally, it is possible to incorporate weights in a fashion paralleling Eqs.(9,10). The weights will be denoted by $w^{(i)}$ in order to maintain the new use of superscripts to denote laboratories. Thus, we obtain (sum over repeated indices)

$$M_j = (\chi)_{jk} \left(\frac{w^{(1)}}{w_{max}} \left(C^{(1)^{-1}} \right)_{kl} m_l^{(1)} + \frac{w^{(2)}}{w_{max}} \left(C^{(2)^{-1}} \right)_{kl} m_l^{(2)} + \cdots + \frac{w^{(n)}}{w_{max}} \left(C^{(n)^{-1}} \right)_{kl} m_l^{(n)} \right)$$

and

$$\chi^{-1} = \frac{w^{(1)}}{w_{max}} C^{(1)^{-1}} + \frac{w^{(2)}}{w_{max}} C^{(2)^{-1}} + \cdots + \frac{w^{(n)}}{w_{max}} C^{(n)^{-1}}$$

Note carefully the use of bold face to indicate matrices without their indices.

Ronald F. Fox
Smyrna, Georgia
August 25, 2010