How to combine independent data sets for the same quantity

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 $P_1(x), P_2(x), P_3(x), \dots, P_n(x)$ where the subscripts denote the sources of the data sets. Below, we define the *conflation* of these distributions, denoted by $\&(P_1(x), P_2(x), \dots, P_n(x))$, and give some of the main results for the special but rather general case of normal distributions.

The idea of *conflation* was proposed by my colleague 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.

This method for combining independent data sets is especially pertinent today as progress is made in creating highly precise measurement standards and reference values for basic physical quantites. Ted and I were originally concerned with the value of Avogadro's number [American Scientist] and later with a re-definition of the kilogram [Archives]. This endeavor brought us into contact with the researchers at the National Institute of Standards and Technology, NIST. Through them and their foreign counterparts, it became apparent that an objective method for combining data sets measured in different laboratories was a pressing need. *Conflation* is the result produced by an objective analysis of this question.

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

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

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

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

where

$$\frac{1}{\Sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \tag{3}$$

and

$$M = \Sigma^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) \tag{4}$$

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

$$exp\left[-\frac{(x-m_1)^2}{2\sigma_1^2} - \frac{(x-m_2)^2}{2\sigma_1^2}\right] = exp\left[-\left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right)x^2 + \left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right)x - \frac{m_1^2}{2\sigma_1^2} - \frac{m_2^2}{2\sigma_2^2}\right] \tag{5}$$

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

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

$$N(x, M_n, \Sigma_n^2) \tag{6}$$

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).

M_n is the same as the "weighted least squares" mean with weights σ_i^{-2} for m_i, $i = 1, 2, \ldots, n$.

Proof: Minimize $E(S)$, the expectation of S defined by

$$S = \sum_{i=1}^{n} \frac{(M - x_i)^2}{\sigma_i^2} \tag{9}$$

with respect to M. This readily yields the same formula asin Eq.(8). QED

A. Aitken showed that this is also BLUE, the **b**est **l**inear **u**nbiased **e**stimator. Consider two normally distributed densities with means m_1 and m_2 and variances σ_1^2 and σ_2^2 respectively. Let X denote a linear combination of the random variables with these densities.

$$X = px_1 + (1 - p)x_2 \tag{10}$$

And let the mean of X, M, be given by

$$M = pm_1 + (1 - p)m_2 \tag{11}$$

The value of p that minimizes the variance of X, Σ^2, is given by the minimum of

$$\Sigma^2 = E((M - X)^2) = p^2\sigma_1^2 + (1 - p)^2\sigma_2^2 \tag{12}$$

Therefore p is given by

$$p\sigma_1^2 - (1-p)\sigma_2^2 = 0 \tag{13}$$

The results

$$p = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{14}$$

$$(1-p) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \tag{15}$$

$$\Sigma^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{16}$$

Note that eqs.(16) and (3) agree. Moreover, extention to three or more densities is straight-forward based on the provable associativity property of *conflation*. To do three, do any two first and then conflate the result with the remaining density.

Σ_n is the same as the "weighted least squares" variance.

Proof: Define S_V by

$$S_V = \sum_{i=1}^{n} \frac{\left(\Sigma^2 - \frac{1}{n}\Delta x_i^2\right)^2}{\sigma_i^2} \tag{17}$$

where $\Delta x_i = x_i - m_i$. Minimize $E(S_V)$ with respect to Σ . This implies

$$2 \times\ 2 \times\ \Sigma \times \sum_{i=1}^{n} \left(\frac{\Sigma^2}{\sigma_i^2} - \frac{1}{n}\frac{E(\Delta x_i^2)}{\sigma_i^2}\right) = 0 \tag{18}$$

Since $E(\Delta x_i^2) = \sigma_i^2$, the result is

$$\Sigma^2 \times \sum_{i=1}^{n} \frac{1}{\sigma_i^2} = 1 \tag{19}$$

This equivalent to the formula, Eq.(7), for Σ_n^2 , i.e.

$$\Sigma^{-2} = \sum_{i=1}^{n} \sigma_i^{-2} \tag{20}$$

QED

While we believe the kilogram should be defined in terms of a predetermined theoretical value for Avogadro's number [Archives], the NIST approach is based instead on a more precise value for Planck's constant determined in the laboratory using a Watt balance. In fact this approach may result in a defined exact value for Planck's constant in parallel with the speed of light and the second (these two determine the meter exactly as well). As of a year or so ago, the different laboratories making measurements with Watt balances did not know (personal communication) how their data was combined with the NIST values to determine the published values [CODATA]. We strongly urge the use of *conflation*.

The basic idea behind conflation was published earlier and independently in:
"Hilbert Space of Probability Density Functions Based on Aitchison Geometry"
by J. J. Egozcue, J. L. Diaz-Barrero and V. Pawlowsky-Glahn
Acta Mathematica Sinica, English Series Jul., 2006, Vol. 22, No. 4, pp. 1175–1182
Published online: Jan. 19, 2006 DOI: 10.1007/s10114-005-0678-2
Http://www.ActaMath.com