How to combine independent data sets for the same quantity

Theodore P. Hill, and Jack Miller

View Online   **Export Citation**

ARTICLES YOU MAY BE INTERESTED IN

Scilight
Summaries of the latest breakthroughs
in the **physical sciences**

AIP Publishing

How to combine independent data sets for the same quantity

Theodore P. Hill[1] and Jack Miller[2]
[1]*School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*
[2]*Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

This paper describes a new mathematical method called conflation for consolidating data from independent experiments that measure the same physical quantity. Conflation is easy to calculate and visualize and minimizes the maximum loss in Shannon information in consolidating several independent distributions into a single distribution. A formal mathematical treatment of conflation has recently been published. For the benefit of experimenters wishing to use this technique, in this paper we derive the principal basic properties of conflation in the special case of normally distributed (Gaussian) data. Examples of applications to measurements of the fundamental physical constants and in high energy physics are presented, and the conflation operation is generalized to weighted conflation for cases in which the underlying experiments are not uniformly reliable. © *2011 American Institute of Physics*. [doi:10.1063/1.3593373]

When different experiments are designed to measure the same unknown quantity, how can their results be consolidated in an unbiased and optimal way? Given data from experiments made at different times, in different locations, with different methodologies, and perhaps differing even in underlying theory, is there a straightforward, easily applied method for combining the results from all of the experiments into a single distribution? This paper describes a new mathematical method called conflation for consolidating data from independent experiments that measure the same physical quantity.

I. INTRODUCTION

The consolidation of data from different sources can be particularly vexing in the determination of the values of the fundamental physical constants. For example, the U.S. National Institute of Standards and Technology (NIST) recently reported "two major inconsistencies" in some measured values of the molar volume of silicon $V_m(\text{Si})$ and the silicon lattice spacing d_{220}, leading to an *ad hoc* factor of 1.5 increase in the uncertainty in the value of Planck's constant h (Refs. [1], p. 54, and [2]). (One of those inconsistencies has since been resolved[3]).

Input data distributions that happen to have different means and standard deviations are not necessarily "inconsistent" or "incoherent" (Ref. [4], p. 2249). If the various input data are all normally (Gaussian) or exponentially distributed, for example, then every interval centered at the unknown positive true value has a positive probability of occurring in every independent measurement. Ideally, of course, *all* experimental data, past and present, should be incorporated into the scientific record. But in the case of the fundamental physical constants, this could entail listing scores of past and present experimental datasets, each of which includes results from hundreds of experiments with thousands of data points, for *each one of the fundamental constants*. Most experimentalists and theoreticians who use Planck's constant, however, need only a concise summary of its current value rather than the complete record. Having the mean and estimated standard deviation (e.g., via weighted least squares) does give some information, but without any knowledge of the distribution, knowing the mean within two standard deviations is only valid at the 75% level of significance, and knowing the mean within four standard deviations is not even significant at the standard 95% confidence level. Is there an objective, natural and optimal method for consolidating several input-data distributions into a single posterior distribution *P*? In this paper, we describe a new such method called *conflation*.

Note that this is not the standard statistical problem of producing point estimates and confidence intervals, but rather a method for simply summarizing all of the experimental data with a single distribution.

First, it is useful to review some of the shortcomings of standard methods for consolidating data from several different input distributions. For simplicity, consider the case of only two different experiments in which independent laboratories Lab I and Lab II measure the value of the same quantity. Lab I reports its results as a probability distribution P_1 (e.g. via an empirical histogram or probability density function) and Lab II reports its findings as P_2.

A. Averaging the probabilities

One common method of consolidating two probability distributions is to simply average them—for every set of values A, set $P(A) = \dfrac{P_1(A) + P_2(A)}{2}$. If the distributions both have densities, for example, averaging the probabilities results in a probability distribution with density, the average of the two input densities (Figure 1). This method has several significant disadvantages. First, the mean of the resulting distribution P is always exactly the average of the means of P_1 and P_2, independent of the relative accuracies or variances
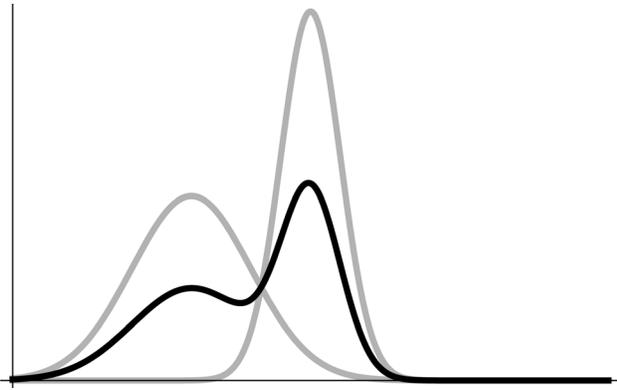
FIG. 1. Averaging the probabilities. (The black curve is the average of the gray (input) curves. Note that the variance of the average is larger than the variance of either input.)

of each. (Recall that the variance is the square of the standard deviation.) But if Lab I performed twice as many of the same type of trials as Lab II, the variance of P_1 would be half that of P_2, and it would be unreasonable to weight the two respective empirical means equally.

A second disadvantage of the method of averaging probabilities is that the variance of P is always *at least as large* as the minimum of the variances of P_1 and P_2 (see Figure 1), since

$$V(P) = \frac{V(P_1) + V(P_2)}{2} + \frac{[\text{mean}(P_1) - \text{mean}(P_2)]^2}{4}.$$

If P_1 and P_2 are nearly identical, however, then their average is nearly identical to both inputs, whereas the standard deviation of a reasonable consolidation P should probably be strictly less than that of both P_1 and P_2. The method of averaging probabilities completely ignores the fact that two laboratories independently found nearly the same results. Figure 1 also shows another shortcoming of this method—with normally-distributed input data, it generally produces a multimodal distribution, whereas one might desire the consolidated output distribution to be of the same general form as that of the input data—Gaussian, or at least unimodal.

B. Averaging the data

Another common method of consolidating data—one that does preserve normality—is to average the underlying input data itself. That is, if the result of the experiment from Lab I is a random variable X_1 (i.e. has distribution P_1) and the result of Lab II is X_2 (independent of X_1, with distribution P_2), take P to be the distribution of $\frac{X_1 + X_2}{2}$. As with averaging the distributions, averaging the data also results in a distribution that always has exactly the average of the means of the two input distributions, regardless of the relative accuracies of the two input data-set distributions (see Figure 2). With this method, on the other hand, the variance of P is *never larger* than the maximum variance of P_1 and P_2 since $V(P) = \frac{V(P_1) + V(P_2)}{4}$, whereas some input data distributions that differ significantly should sometimes
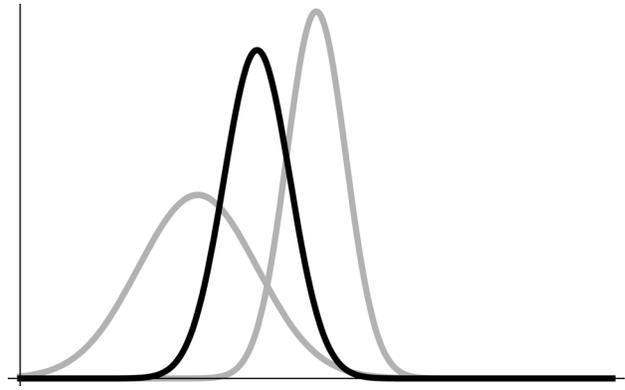

FIG. 2. Averaging the data. (The black curve is the average of the gray data curves. Note that the mean of the averaged data is exactly the average of the means of the two input distributions, even though they have different variances.)

reflect a higher uncertainty. A more fundamental problem with this method is that in general it requires averaging data that were obtained using very different and even indirect methods, for example, as with the watt balance and x-ray/optical interferometer measurements used in part to obtain the 2006 CODATA recommended value for Planck's constant.[2]

The three main goals of this paper are: to describe conflation and derive important basic properties of conflation in the special case of normally-distributed data (perhaps the most common class of experimental data); to provide concrete examples of conflation using real experimental data; and to introduce a new method for consolidating data when the underlying data sets are not uniformly weighted.

II. CONFLATION OF DATA SETS

For consolidating data from different independent sources,[5] introduced a mathematical method called *conflation* as an alternative to averaging the probabilities or averaging the data. In practice, it is impossible to ensure that there are absolutely no unidentified correlations between two measurements, no matter how different in methodology they may be. However, whether data can be treated as truly independent is a common problem and is a hypothesis for the experimenters to decide. This paper addresses how to combine the data, once that determination has been made.

The hypothesis of independence of the underlying experiments or data sets in conflation is exactly analogous to the hypothesis of independence in the usual statistical analysis of data. For example, when the sample average of repeated measurements of a quantity is used to estimate the unknown true value of the quantity, independence of the underlying repetitions of the experiment is assumed in applying the strong law of large numbers or the central limit theorem. In practice, formal mathematical independence of those experiments is usually impossible to ascertain. But if independence of the repetitions is accepted as a reasonable assumption, the sample average will be a good estimate of the unknown value. Similarly, if independence of the underlying experiments of different types seems like a reasonable assumption, then conflation will provide a good consolidation of the data from those experiments.

Conflation (designated with the symbol "&" to suggest consolidation of P_1 *and* P_2) has none of the disadvantages of the two averaging methods described above and has many advantages that will be described below.

In the important special case that the input distributions $P_1, P_2, \ldots, P_n$ all have densities (e.g. Gaussian or exponential distributions), then the conflation $\&(P_1, P_2, \ldots, P_n)$ of $P_1, P_2, \ldots, P_n$ is simply the probability distribution with density, the normalized product of the input densities. That is,

( )   If $P_1 \ldots, P_n$ have densities $f_1, \ldots, f_n$, respectively, and the denominator is not 0 or ∞, then $\&(P_1, P_2, \ldots, P_n)$ is continuous with density

$$f(x) = \frac{f_1(x)f_2(x)\cdots f_n(x)}{\int_{-\infty}^{\infty} f_1(y)f_2(y)\cdots f_n(y)dy}.$$

(Especially note that the product in ( ) is taken for the densities evaluated at the same point, x. Note also that conflation is easy to calculate and to visualize; see Figure 3.)

Remark: This normalized product of density functions also arises in other stochastic contexts, such as in log-opinion polls for combining subjective expert opinions,[6] in the conditional distribution of independent random variables given they are equal,[5] and in statistical inference calculating the posterior distribution based on the prior distribution and the likelihood function. Thus, some of the properties in Sec. III below, such as the fact that conflation of normal distributions is normal, may also be derived in those settings. In contrast to those frameworks, however, the notion of conflation does not require external random variables or underlying parametric statistical models. For discrete input distributions, the analogous definition of conflation is the normalized product of the probability mass functions, and for more general situations the definition is more technical.[5] For the purposes of this paper, it will be assumed that the input distributions are continuous, and that the integral of their product is not 0 or ∞. This is always the case, for example, when the input distributions are all Gaussian.

As can easily be seen from (∗) and elementary conditional probability, the conflation of distributions has a natural heuristic and practical interpretation—gather data from the
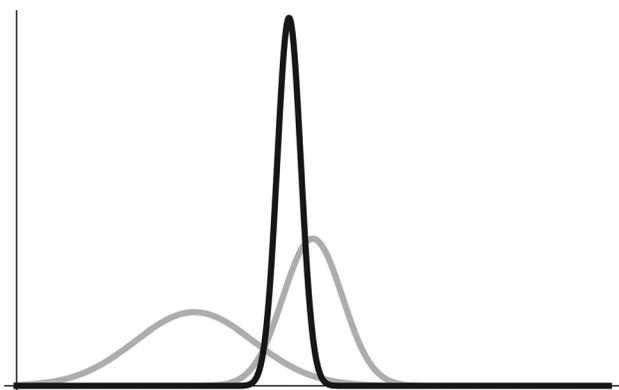

FIG. 3. Conflating distributions. (The black curve is the conflation of the gray curves. Note that the mean of the conflation is closer to the mean of the input distribution with smaller variance, i.e. with greater accuracy.)

independent laboratories sequentially and simultaneously and record the values only at those times when the laboratories (nearly) agree. This observation is readily apparent in the discrete case—if two independent integer-valued random variables X_1 and X_2 (e.g., binomial or Poisson random variables) have probability mass functions $f_1(k) = \Pr(X_1 = k)$ and $f_2(k) = \Pr(X_2 = k)$, then the probability that $X_1 = j$ given that $X_1 = X_2$, is simply

$$\frac{\Pr(X_1 = X_2 = j)}{\Pr(X_1 = X_2)} = \frac{f_1(j)f_2(j)}{\sum_k f_1(k)f_2(k)}.$$

The argument in the continuous case follows similarly.

At first glance, it may seem counterintuitive that the conflation of two relatively broad distributions can be a much narrower one (Figure 3). However, if both measurements are assumed equally valid, then with relatively high probability, the true value should lie in the overlap region between the two distributions. Looking at it statistically, if one lab makes 50 measurements and another lab makes 100, then the standard deviations of their resulting distributions will usually be different. If the labs' methods are also different, with different systematic errors, or their methods rely on different fundamental constants with different uncertainties, then the means will likely be different too. But the bottom line is that the total of 150 valid measurements is substantially greater than either lab's data set, so the standard deviation should indeed be smaller.

III. PROPERTIES OF CONFLATION

Conflation has several basic mathematical properties with significant practical advantages, and to describe these properties succinctly, it will be assumed throughout this section that X_1 and X_2 are independent normal random variables with means m_1, m_2 and standard deviations σ_1, σ_2, respectively. That is, for $i = 1, 2$,

( ) $X_i \quad N(m, \sigma_i^2)$ has *density function*
$$f_i(x) = \frac{1}{\sigma_i\sqrt{2\pi}}\exp\left[-\frac{(x-m_i)^2}{2\sigma_i^2}\right] \text{ for all } -\infty < x < \infty$$
and *distribution* P_i given by $P_i(A) = \int_A f_i(x)dx$.

Remark: The generalization of the properties of conflation described below to more than two distributions is routine; the generalization to non-normal distributions can be found in Ref. 5.

Some of the basic properties of conflation are as follows:

(1) Conflation is commutative and associative:

$$\&(P_1, P_2) = \&(P_2, P_1) \text{ and } \&(\&(P_1, P_2), P_3)$$
$$= \&(P_1, \&(P_2, P_3)).$$

Proof: Immediate from ( ) and the commutativity and associativity of real numbers, which implies that $f_1(x)f_2(x) = f_2(x)f_1(x)$ and $(f_1(x)f_2(x))f_3(x) = f_1(x)(f_2(x)f_3(x))$.

(2) Conflation is iterative:

$$\&(P_1, P_2, P_3) = \&(\&(P_1, P_2), P_3).$$

Proof: Immediate from ( ).

Thus from property (2), to include a new data set in the consolidation, simply conflate it with the overall conflation of the previous data sets.

(3) Conflations of normal distributions are normal:

If P_1 and P_2 satisfy ( ), then

$$\&(P_1, P_2) \text{ is normal with } m = \frac{\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\sigma_2^2 m_1 + \sigma_1^2 m_2}{\sigma_1^2 + \sigma_2^2}$$

$$\text{and } \sigma^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Proof: By $(*)$ and $(**)$, $\&(P_1, P_2)$ is continuous with density proportional to

$$f_1(x) f_2(x) = \frac{1}{\sigma_1 \sigma_2 2\pi} \exp\left(\left[\frac{(x - m_1)^2}{2\sigma_1^2}\right]\left[\frac{(x - m_2)^2}{2\sigma_2^2}\right]\right).$$

Completing the square of the exponent gives

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

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

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

$$+ \left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right)^2 / \left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right) \left(\frac{m_1^2}{2\sigma_1^2} + \frac{m_2^2}{2\sigma_2^2}\right)\right],$$

which is easily seen to be the exponent of the density of a normal distribution with the mean and variance in property (3).

By properties (2) and (3), conflations of any finite number of normal distributions are always normal (see Figure 3, and the dashed curve in Figure 4(b)). Similarly, many of the other important classical families of distributions, including gamma, beta, uniform, exponential, Pareto, Laplace, Bernoulli, zeta, and geometric families, are also preserved under conflation (Ref. 5, Theorem 7.1).

(4) Means and variances of conflations of normal distributions coincide with those of the weighted-least-squares method.

Sketch of proof: Given two independent distributions with means m_1, m_2 and standard deviations σ_1, σ_2, respectively, the weighted-lease-squares mean m is obtained by minimizing the function

$$f(m) = \frac{(m - m_1)^2}{\sigma_1^2} + \frac{(m - m_2)^2}{\sigma_2^2},$$

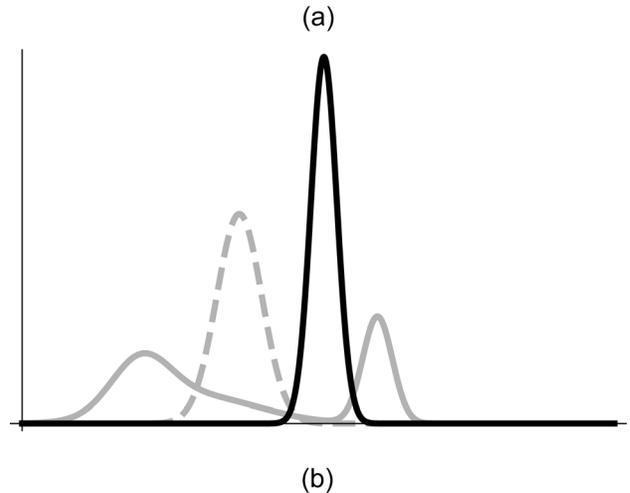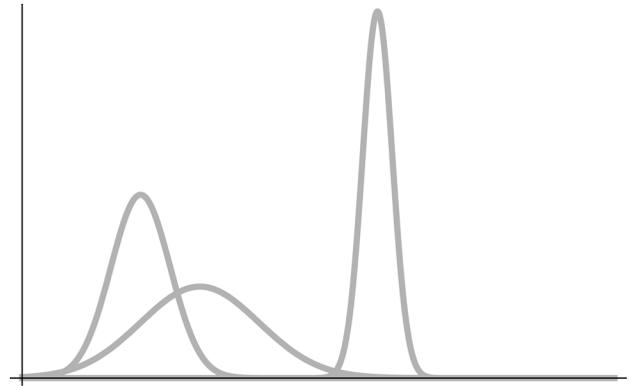with respect to m. Setting

(a)

(b)

FIG. 4. Comparison of averaging probabilities, averaging data, and conflating (The gray curve in Figure 4(b) is the average of the three input distributions in Figure 4(a), the dashed curve is the average of the three input datasets, and the black curve is the conflation.)

$$f'(m) = \frac{2(m - m_1)}{\sigma_1^2} + \frac{2(m - m_2)}{\sigma_2^2} = 0$$

and solving for m yields $m = \frac{\sigma_2^2 m_1 + \sigma_1^2 m_2}{\sigma_1^2 + \sigma_2^2}$, which, by property (3), is the mean of the conflation of two normal distributions with means m_1, m_2 and standard deviations σ_1, σ_2. The conclusion for the weighted-least-squares variance follows similarly.

Whenever data from several (input) distributions are consolidated into a single (output) distribution, this will typically result in some loss of information, however that is defined. A classic measure of information is the Shannon information. Recall that the Shannon information obtained from observing that a random variable X is in a certain set A is $\log_2$ of the probability that X is in A. That is, the Shannon information is the number of binary bits of information obtained by observing that X is in A. For example, if X is a random variable uniformly distributed on the unit interval $[0, 1]$, then observing that X is greater than $1/2$ has Shannon information exactly $\log_2 \Pr X > \frac{1}{2} = -\log_2 \frac{1}{2} = 1$, so one unit (binary bit) of Shannon information has been obtained, namely, that the first binary digit in the expansion of X is 1.

The Shannon information is also called the *surprisal* or *self-information*—the smaller the value of $\Pr(X \in A)$, the greater the information or surprise—and the (combined) Shannon information obtained by observing that independent random variables X_1 and X_2 are *both* in A is simply the sum of the information obtained from each of the datasets X_1 and X_2, that is,

$$S_{P_1,P_2}(A) = S_{P_1}(A) + S_{P_2}(A) = -\log_2 P_1(A)P_2(A).$$

Thus, the *loss in Shannon information* incurred in replacing the pair of distributions P_1, P_2 by a single probability distribution Q is $S_{P_1,P_2}(A) - S_Q(A)$ for the event A.

(5) Conflation minimizes the loss of Shannon information:

If P_1 and P_2 are independent probability distributions, then the conflation $\&(P_1, P_2)$ of P_1 and P_2 is the unique probability distribution that minimizes, over all events A, the maximum loss of Shannon information in replacing the pair P_1, P_2 by a single distribution Q.

Sketch of proof: First, observe that for an event A, the difference between the combined Shannon information obtained from P_1 and P_2 and the Shannon information obtained from a single probability Q is $S_{P_1,P_2}(A) - S_Q(A) = -\log_2 \frac{Q(A)}{P_1(A)P_2(A)}$. Since $\log_2(x)$ is strictly increasing, the maximum (loss) thus occurs for an event A where $\frac{Q(A)}{P_1(A)P_2(A)}$ is maximized.

Next, note that the largest loss of Shannon information occurs for small sets A, since for disjoint sets A and B,

$$\frac{Q(A \cup B)}{P_1(A \cup B)P_2(A \cup B)} = \frac{Q(A) + Q(B)}{P_1(A)P_2(A) + P_1(B)P_2(B)}$$
$$\leq \max\left\{\frac{Q(A)}{P_1(A)P_2(A)}, \frac{Q(B)}{P_1(B)P_2(B)}\right\},$$

where the inequalities follow from the inequalities $(a+b)(c+d) \geq ac + bd$ and $\frac{a+b}{c+d} \leq \max\left\{\frac{a}{c}, \frac{b}{d}\right\}$ for positive numbers a, b, c, and d. Since P_1 and P_2 are normal, their densities $f_1(x)$ and $f_2(x)$ are continuous everywhere, so the small set A may in fact be replaced by an arbitrarily small interval, and the problem reduces to finding the probability density function f that makes the maximum, over all real values x, of the ratio $\frac{f(x)}{f_1(x)f_2(x)}$ as small as possible. But, as is seen in the discrete framework, the minimum over all non-negative $p_1, \ldots, p_n$ with $p_1 + \cdots + p_n = 1$ of the maximum of $\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}$ occurs when $\frac{p_1}{q_1} = \cdots = \frac{p_n}{q_n}$ (if they are not equal, reducing the numerator of the largest ratio, and increasing that of the smallest, will make the maximum smaller). Thus, the f that makes the maximum of $\frac{f(x)}{f_1(x)f_2(x)}$ as small as possible is when $f(x) = cf_1(x)f_2(x)$, where c is chosen to make f a density function, i.e., to make f integrate to 1. But this is exactly the definition of the conflation $\&(P_1, P_2)$ in ( ).

Remark: The proof only uses the facts that normal distributions have densities that are continuous and positive everywhere, and that the integral of the product of every two normal densities is finite and positive.

(6) Conflation is a best linear unbiased estimate (BLUE):

If X_1 and X_2 are independent unbiased estimates of θ with finite standard deviations σ_1, σ_2, respectively, then $\Theta = \text{mean}[\&(N_1, N_2)]$ is a best linear unbiased estimate for θ, where N_1 and N_2 are independent normal probability distributions with (random) means X_1 and X_2 and standard deviations σ_1 and σ_2, respectively.

Sketch of proof: Let $X = pX_1 + (1-p)X_2$ be the linear estimator of θ based on X_1 *and* X_2 and weight $0 \leq p \leq 1$. Then the expected value $E(X)$ of X is $E(X) = pm_1 + (1-p)m_2$, and since X_1 and X_2 are independent, the variance $V(X)$ of X is $V(X) = p^2\sigma_1^2 + (1-p)^2\sigma_2^2$. To find the p that minimizes $V(X)$, setting $\frac{dV}{dp} = 2p\sigma_1^2 - 2(1-p)\sigma_2^2 = 0$ yields $p = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$, so $X = \frac{\sigma_2^2 X_1}{\sigma_1^2 + \sigma_2^2} + \frac{\sigma_1^2 X_2}{\sigma_1^2 + \sigma_2^2}$ is BLUE for θ. But by property (3), X is the mean of $\&(N_1, N_2)$.

(7) Conflation yields a maximum likelihood estimator (MLE):

If X_1 and X_2 are independent normal unbiased estimates of θ with finite standard deviations σ_1, σ_2, respectively, then $\Theta = \text{mean}[\&(N_1, N_2)]$ is a MLE for θ, where N_1 and N_2 are independent normal probability distributions with (random) means X_1 and X_2 and standard deviations σ_1 and σ_2, respectively.

Sketch of proof: The classical likelihood function in this case is

$$L = f(X_1; \theta)f(X_2; \theta)$$
$$= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[\frac{-(X_1 - \theta)^2}{2\sigma_1^2}\right] \left(\frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[\frac{-(X_2 - \theta)^2}{2\sigma_2^2}\right]\right)$$

so to find the θ that maximizes L, take the partial derivative of $\log L$ with respect to θ and set it equal to zero

$$\frac{\partial \log L}{\partial \theta} = \frac{X_1 - \theta}{\sigma_1^2} + \frac{X_2 - \theta}{\sigma_2^2} = 0.$$

This implies that the critical point (and maximum likelihood) occurs when

$$\theta\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) = \frac{X_1}{\sigma_1^2} + \frac{X_2}{\sigma_2^2}.$$

Thus

$$\theta = \left(\frac{X_1}{\sigma_1^2} + \frac{X_2}{\sigma_2^2}\right) \bigg/ \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right).$$

By property (3), this implies that the MLE θ is the mean of $\&(N_1, N_2)$.

Remark: Note that the normality of the underlying distributions is used in property (7), but it is not required for properties (5) or (6). Properties (4), (6), and (7) in the general cases use Aiken's generalization of the Gauss-Markov theorem and related results, see, e.g., Refs. 7 and 8.

In addition to properties (6) and (7), conflation is also optimal with respect to several other statistical properties. In classical hypotheses testing, for example, a standard technique to decide from which of n known distributions given data actually came is to maximize the likelihood ratios, that is, the ratios of the probability density or probability mass functions. Analogously, when the objective is how best to consolidate data from those input distributions into a single (output) distribution P, one natural criterion is to choose P so as to make the ratios of the likelihood of observing x under P to the likelihood of observing x under *all* of the (independent) distributions $\{P_i\}$ as close as possible. The conflation of the distributions is the unique probability distribution that makes the variation of these likelihood ratios as small as possible (Ref. 5, Theorem 5.2).

The conflation of the distributions is also the unique probability distribution that preserves the proportionality of likelihoods (Ref. 5, Theorem 5.5). A criterion similar to likelihood ratios is to require that the output distribution P reflect the relative likelihoods of identical individual outcomes under the $\{P_i\}$. For example, if the likelihood of all the experiments $\{P_i\}$ observing the identical outcome x is twice that of the likelihood of all the experiments $\{P_i\}$ observing y, then $P(x)$ should also be twice as large as $P(y)$.

Conflation has one more advantage over the methods of averaging probabilities or data. In practice, assumptions are often made about the form of the input distributions, such as an assumption that underlying data are normally distributed.[1] But the true and estimated values for Planck's constant are clearly never negative, so the underlying distribution is certainly not truly normally distributed—more likely, it is truncated normal. Using conflation, the problem of truncation essentially disappears—it is automatically taken into account. If one of the input distributions is summarized as a true normal distribution and the other excludes negative values, for example, then the conflation will exclude negative values, as is seen in Figure 5.

IV. EXAMPLES IN MEASUREMENTS OF PHYSICAL CONSTANTS AND HIGH-ENERGY PHYSICS

As described in the Introduction, methods for combining independent data sets are especially pertinent today as pro-
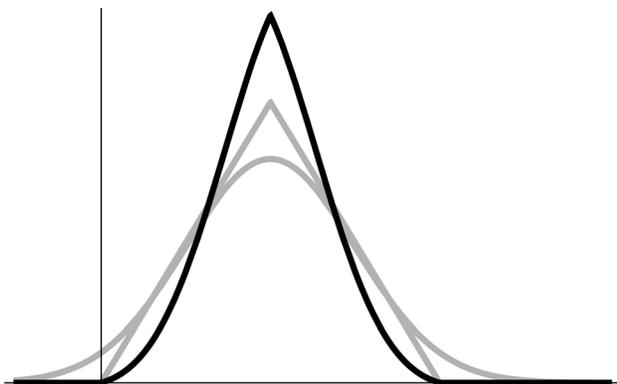
gress is made in creating highly precise measurement standards and reference values for basic physical quantities. A suggestion by the authors for a re-definition of the kilogram[9] brought them into contact with the researchers at NIST and their counterparts outside the U.S. and, as suggested in the Introduction, it became apparent that there is a pressing need for an objective method for combining data sets measured in different laboratories.

In Ref. 9, it is proposed that the kilogram be defined in terms of a predetermined theoretical value for Avogadro's number. In contrast, the NIST approach is based 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). Since *conflation* is the result produced by an objective analysis of exactly this question—how to consolidate data from independent experiments—perhaps, conflation can be employed to obtain better consolidations of experimental data for the fundamental physical constants. In this section, we illustrate, using experimental data, how conflation may be used in this way.

Example 1: ({220} Lattice spacing measurements) The input data used to obtain the CODATA 2006 recommended values and uncertainties of the fundamental physical constants includes the measurements and inferred values of the absolute {220} lattice spacing of various silicon crystals used in the determination of Planck's constant and the Avogadro constant. The four measurements came from three different laboratories and had values 192,015.565(13), 192,015.5973(84), 192,015.5732(53), and 192,015.5685(67), respectively, Ref. 2, Table XXIV, where the parenthetical entry is the uncertainty. The CODATA task force viewed the second value as "inconsistent" with the other three (see gray curves in Figure 6) and made a consensus adjustment of the uncertainties. Since those values "are the means of tens of individual values, with each value being the average of about ten data points,"[2] the central limit theorem suggests that the underlying datasets are approximately normally distributed as is shown in Figure 6 (gray curves). The conflation of those four input distributions, however, requires no consensus
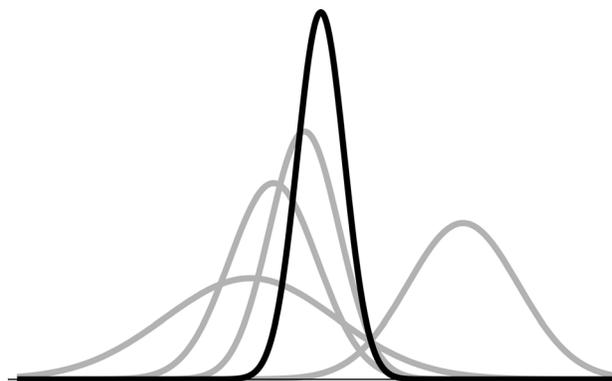

FIG. 5. The black curve is the conflation of the gray curves. Note that the conflation has no negative values, since the triangular input had none.

FIG. 6. The four gray curves are the distributions of the four measurements of the {220} lattice spacing underlying the CODATA 2006 values; the black curve is the conflation of those four distributions and requires no *ad hoc* adjustment.

adjustment and yields a value essentially the same as the final CODATA value, namely, 192,015.5762 (Ref. 2, Table LIII), but with a much smaller uncertainty. Since uncertainties play an important role in determining *the values of the related constants* via weighted least squares, this smaller and theoretically justifiable uncertainty is a potential improvement to the current accepted values.

Example 2: (top quark mass measurements) The top quark is a spin-1/2 fermion with charge two thirds that of the proton, and its mass is a fundamental parameter of the standard model of particle physics. Measurements of the mass of the top quark have been made by two different detector groups at the Fermi National Accelerator Laboratory (FNAL) Tevatron: the Collider Detector at Fermilab (CDF) collaboration using a multivariate-template method, a *b*-jet decay-length likelihood method, and a dynamic-likelihood method; and the D0 collaboration using a matrix-element-weighting method and a neutrino-weighting method. The mass of the top quark was then "calculated from 11 independent measurements made by the CDF and D0 collaborations" yielding the 11 measurements: 167.4(11.4), 168.4(12.8), 164.5(5.5), 178.1(8.3), 176.1(7.3), 180.1(5.3), 170.9(2.5), 170.3(4.4), 186.0(11.5), 174.0(5.2), and 183.9(15.8) GeV (Figure 4 in Ref. 10). Again assuming that each of these measurements is approximately normally distributed, the conflation of these 11 independent input distributions is normal with mean and uncertainty (standard deviation) 172.63(1.6), which has a slightly higher mean and a lower uncertainty than the average mass of 171.4(2.1) reported in Ref. 10. (Top quark measurements are being updated regularly, and the reader interested in the latest values should check the most recent FNAL publications; these particular (*c.* 2006) values were used simply for illustrative purposes.)

V. WEIGHTED CONFLATION

The conflation $\&(P_1, \ldots, P_n)$ of n probability distributions of independent random variables (experimental datasets) $P_1, \ldots, P_n$ described above and in Ref. 5 treated all the underlying distributions equally, with no differentiation between relative perceived validities of the experiments. A related statistical concept is that of a *uniform prior*, that is, a prior assumption that all the experiments are equally likely to be valid.

If, on the other hand, additional assumptions are made about the reliability or validity of the various experiments—for instance, that one experiment was supervised by a more experienced researcher, or employed a methodology thought to be better than another—then a consolidation of the data from the independent experiments should probably be adjusted to account for this perceived non-uniformity.

More concretely, suppose that in addition to the independent experimental distributions $P_1, \ldots, P_n$, non-negative weights $w_1, \ldots, w_n$ are assigned to each of the distributions to reflect their perceived relative validity. For example, if P_1 is considered twice as reliable as P_1, then $w_1 = 2w_2$. Without loss of generality, the weights $w_1, \ldots, w_n$ are nonnegative, and at least one is positive. How should this additional infor-

mation $w_1, \ldots, w_n$ be incorporated into the consolidation of the input data? That is, what probability distribution $Q = \&((P_1, w_1), \ldots, (P_n, w_n))$ should replace the uniform-weight conflation $\&(P_1, \ldots, P_n)$

For the case where all the underlying datasets are assumed equally valid, it was seen that the conflation $\&(P_1, \ldots, P_n)$ is the unique single probability distribution Q that minimizes the loss of Shannon information between Q and the original distributions $P_1, \ldots, P_n$. Similarly, for weighted distributions $(P_1, w_1), \ldots, (P_n, w_n)$, identifying the probability distribution Q that minimizes the loss of Shannon information between Q and the weighted data distributions leads to a unique distribution $\&((P_1, w_1), \ldots, (P_n, w_n))$ called the *weighted conflation*.

Given n weighted (independent) distributions $(P_1, w_1), \ldots, (P_n, w_n)$, the *weighted Shannon Information* of the event A, $S_{((P_1, w_1), \ldots, (P_n, w_n))}(A)$, is

$$S_{((P_1, w_1), \ldots, (P_n, w_n))}(A) = \sum_{j=1}^{n} \frac{w_j}{w_{\max}} S_{P_j}(A)$$

$$= \sum_{j=1}^{n} \frac{w_j}{w_{\max}} \log_2 P_j(A),$$

where, here and throughout, $w_{\max} = \max\{w_1, \ldots, w_n\}$.

Note that $S_{((P_1, w_1), \ldots, (P_n, w_n))}$ is continuous and symmetric in both $P_1, \ldots, P_n$ and $w_1, \ldots, w_n$, and that $S_{((P_1, w_1), \ldots, (P_n, w_n))}(A) = 0$ if all the probabilities of A are 1, for all $P_1, \ldots, P_n$ and $w_1, \ldots, w_n$. That is, no matter what the distributions and weights, no information is attained by observing any event that is certain to occur.

Remarks:

(i)   Dividing by $w_{\max}$ reflects the assumption that only the *relative weights* are important, so for instance if one experiment is considered twice as likely to be valid as another, then the information obtained from that experiment should be exactly twice as much as the information from the other, regardless of the absolute magnitudes of the weights. Thus in this latter case, for example,

$$S_{((P_1, 2), (P_2, 1))}(A) = S_{((P_1, 4), (P_2, 2))}(A)$$
$$= S_{P_1}(A) + \frac{1}{2} S_{P_2}(A).$$

In general, this means simply that for all $P_1, \ldots, P_n$ and $w_1, \ldots, w_n$,

$$S_{((P_1, w_1), \ldots, (P_n, w_n))}(A) = S_{((P_1, w_1/w_{\max}), \ldots, (P_n, w_n/w_{\max}))}(A).$$

(ii)   If all the weights are equal, the weighted Shannon information coincides with the classical combined Shannon information, i.e.,

$$S_{((P_1, w_1), \ldots, (P_n, w_n))}(A) = \sum_{j=1}^{n} S_{P_j}(A)$$
$$\text{if } w_1 = \quad = w_n > 0.$$

(iii)   The weighted Shannon information is at least the Shannon information of the single input distribution

with the largest weight and no more than the classical combined Shannon information of $P_1, \ldots, P_n$, that is,

$$S_{P_1}(A) \leq S_{((P_1,w_1),\ldots,(P_n,w_n))}(A) \leq S_{P_1,\ldots,P_n}(A),$$

with equality if $w_1 > w_2 = \quad = w_n = 0$ or $w_1 = \quad = w_n > 0$, respectively.

Next, the basic definition of conflation ( ) is generalized to the definition of weighted conflation, where $\&((P_1,w_1),\ldots,(P_n,w_n))$ designates the *weighted conflation of $P_1, \ldots, P_n$ with respect to the weights $w_1, \ldots, w_n$.*

( )   If $P_1, \ldots, P_n$ have densities $f_1, \ldots, f_n$, respectively, and the denominator is not 0 or ∞, then $\&((P_1,w_1),\ldots,(P_n,w_n))$ is continuous with density

$$f(x) = \frac{f_1^{\frac{w_1}{w_{\max}}}(x) f_2^{\frac{w_2}{w_{\max}}}(x) \cdots f_n^{\frac{w_n}{w_{\max}}}(x)}{\int_{-\infty}^{\infty} f_1^{\frac{w_1}{w_{\max}}}(y) f_2^{\frac{w_2}{w_{\max}}}(y) \cdots f_n^{\frac{w_n}{w_{\max}}}(y) dy}.$$

Remarks:

(i)   The definition of weighted conflation for discrete distributions is analogous, with the probability density functions and integration replaced by probability mass functions and summation.

(ii)   If $P_1, \ldots, P_n$ are all normal distributions, then $\&((P_1,w_1),\ldots,(P_n,w_n))$ is normally distributed (calculation analogous to property (4)).

(iii)   The weighted conflation depends only on the relative, not the absolute, values of the weights, that is,

$$\&((P_1,w_1),\ldots,(P_n,w_n))$$
$$= \& \left(P_1, \frac{w_1}{w_{\max}} , \ldots, P_n, \frac{w_n}{w_{\max}} \right).$$

(iv)   If all the weights are equal, the weighted conflation coincides with the standard conflation, that is,

$$\&((P_1,w_1),\ldots,(P_n,w_n)) = \&(P_1,\ldots,P_n)$$
$$\text{if } w_1 = \cdot \quad = w_n > 0.$$

(v)   Updating a weighted distribution with an additional distribution and weight is straightforward: compute the weighted conflation of the pre-existing weighted conflation distribution and the new distribution, using weights $w_{\max} = \max\{w_1, \ldots, w_n\}$ and w_{n+1}, respectively. That is, the analog of property (2) for weighted conflation is

$$\&((P_1,w_1),\ldots,(P_n,w_n),(P_{n+1},w_{n+1}))$$
$$= \&((P_1,w_1),\ldots,(P_n,w_n),w_{\max}),(P_{n+1},w_{n+1})).$$

(vi)   Normalized products of density functions of the forms ( ) and (   ) have been studied in the context of "log opinion polls" and, more recently, in the setting of Hilbert spaces—see Refs. 6 and 5 and the references therein.

(8) Weighted conflation minimizes the loss of weighted Shannon information:

If $(P_1,w_1),\ldots,(P_n,w_n)$ are weighted independent distributions, then the weighted conflation $\&((P_1,w_1),\ldots,(P_n,w_n))$ is the unique probability distribution that minimizes, over all events A, the maximum loss of weighted Shannon information in replacing $(P_1,w_1),\ldots,(P_n,w_n)$ by a single distribution Q.

The proofs of the above conclusions for weighted conflation follow almost exactly from those for uniform conflation; the details are left for the interested reader.

VI. CONCLUSION

The conflation of independent input-data distributions is a probability distribution that summarizes the data in an optimal and unbiased way. The input data may already be summarized, perhaps as normal distributions with given means and variances, or may be the raw data themselves in the form of empirical histograms or densities. The conflation of these input distributions is easy to calculate and visualize and affords easy computation of sharp confidence intervals. Conflation is also easy to update, is the unique minimizer of loss of Shannon information, is the unique minimal likelihood ratio consolidation, and is the unique proportional consolidation of the input distributions. Conflation of normal distributions is always normal, and conflation preserves truncation of data. Perhaps, the method of *conflating* input data will provide a practical and simple, yet optimal and rigorous method to address the basic problem of consolidation of data.

[1]P. Mohr, B. Taylor, and D. Newell, Phys.Today **60**(7), 52 (2007).
[2]P. Mohr, B. Taylor, and D. Newell, Rev. Mod. Phys. **80**, 633 (2008).
[3]P. Mohr, private communication (2008).
[4]R. Davis, Philos. Trans. R. Soc.London, Ser.A **363**, 2249 (2005)
[5]T. Hill, Trans. Am. Math. Soc. **363**, 3351 (2011).
[6]C. Genest and J. Zidek, Stat. Sci. **1**, 114 (1986).
[7]J. Aitchison, J. R. Stat. Soc. Ser. B (Stat. Methodol.) **44**, 139 (1982).
[8]A. Rencher and G. Schaalje, *Linear Models in Statistics* (Wiley-Interscience, Hoboken, NJ, 2008).
[9]T. Hill, J. Miller, and A. Censullo, Metrologia **48**, 83 (2011).
[10]A. Heinson, Top quark mass measurements, DØ Note 5226, Fermilab-Conf-06/287-E (2006), http://arxiv.org/ftp/hep-ex/papers/0609/0609028.pdf.