LAL 00-20
August 2000

Combining Inconsistent Measurements

A. Höcker[a], S. Laplace[a] and F. Le Diberder[b]

[a] *Laboratoire de l'Accélérateur Linéaire,*
IN2P3-CNRS et Université de Paris-Sud,
BP 34, F-91898 Orsay Cedex, Campagne Française

[b] *Laboratoire de Physique Nucléaire et des Hautes Energies,*
IN2P3-CNRS et Université de Paris 6 & 7,
4 place Jussieu, Tour 33 r.d.c., F-75252 Paris Cedex 05, France

Abstract

A treatment of potentially inconsistent measurements is presented. The advocated method, termed the *Combiner* , is an extension of the classical weighted mean.

Contents

1 Introduction

When several measurements $x_{\rm exp}(i) \pm \sigma_{x_{\rm exp}}(i)$, with $i = 1,..,N$, of the same physical observable X are available, the question arises on how to combine them into a single measurement $\langle x_{\rm exp}\rangle$. Combining the measurements can serve two purposes: merely, it can be to provide a summary carrying the overall information within a conveniently easily-quoted global measurement, or, more ambitiously, it can be to provide a means to incorporate the set of measurements into a more involved analysis, where the physical observable X enters as one input among others.

The weighted mean (WM) method defined by:

$$\langle x_{\rm exp}\rangle = \sigma^2_{\langle x_{\rm exp}\rangle} \sum_{i=1}^{N} \sigma^{-2}_{x_{\rm exp}}(i)\, x_{\rm exp}(i) \;, \tag{1}$$

$$\sigma^{-2}_{\langle x_{\rm exp}\rangle} = \sum_{i=1}^{N} \sigma^{-2}_{x_{\rm exp}}(i) \;, \tag{2}$$

is the optimal method to merge the individual measurements, insofar all measurements are consistent the ones with the others. It leads to an easily-quoted global measurement: $X = \langle x_{\rm exp}\rangle \pm \sigma_{\langle x_{\rm exp}\rangle}$. Furthermore, because the underlying hypothesis is clear (the set of measurements is taken to be consistent) the WM method is statistically well-defined and its result is easy to use: the true value of the physical observable being assumed to be x, the probability for this value to yield for the χ^2:

$$\chi^2(\mathrm{x}) = \left(\frac{\langle \mathrm{x_{exp}}\rangle - \mathrm{x}}{\sigma_{\langle \mathrm{x_{exp}}\rangle}}\right)^2 \tag{3}$$

a value larger than the observed one, is given by:

$$\mathcal{P}(\mathrm{x}) = \mathrm{Prob}(\chi^2,\, 1) \quad , \tag{4}$$

where

$$\mathrm{Prob}(\chi^2, \mathrm{N_{dof}}) = \frac{1}{\sqrt{2^{\mathrm{N_{dof}}-1}}\Gamma((\mathrm{N_{dof}}-1)/2)} \int_{\chi^2}^{\infty} e^{-t/2} t^{(\mathrm{N_{dof}}-1)/2-1}\, \mathrm{d}t \;. \tag{5}$$

A measure of the consistency of the set of measurements is provided by the χ^2

$$\chi^2(\langle x_{\rm exp}\rangle) = \sum_{i=1}^{N} \left(\frac{x_{\rm exp}(i) - \langle x_{\rm exp}\rangle}{\sigma_{x_{\rm exp}}(i)}\right)^2 \;, \tag{6}$$

and, more conveniently, by its associated confidence level

$$\mathcal{P}(\langle x_{\rm exp}\rangle) = \mathrm{Prob}(\chi^2(\langle x_{\rm exp}\rangle),\, N-1) \tag{7}$$

If the value of $\chi^2(\langle x_{\rm exp}\rangle)$ is too large, one may suspect that some of the measurements in the set are flawed. If one insists on a democratic treatment of the $x_{\rm exp}(i)$, *i.e.*, if one refuses to remove the suspected ones, the PDG-recommended scheme[1], termed the rescaled weighted mean below (RWM), consists of rescaling the error $\sigma_{\langle x_{\rm exp}\rangle}$ of Eq. (2) by the scale factor

$$S = \sqrt{\chi^2(\langle x_{\rm exp}\rangle)/(N-1)} \ , \tag{8}$$

if the latter exceeds unity:

$$\Sigma_{\langle x_{\rm exp}\rangle} = \sigma_{\langle x_{\rm exp}\rangle} \ S \ . \tag{9}$$

The RWM method is simple and convenient, however, it suffers from two drawbacks:

1. Psychostatistics.

   On the average, the quoted error is necessarily enlarged with respect to the one of the un-rescaled weighted mean (WM), even for consistent data sets. Tampering with Eq (2) implies a departure from well-defined statistics to enter the realm of ill-defined (psycho)statistics, where working hypotheses are no longer fully explicited.

2. Schizostatistics.

   The average $\langle x_{\rm exp}\rangle$ may lie outside the range of values covered by the measurements. This is because the democratic treatment does not allow to detect measurements which obviously stick out from the set. Such a measurement, termed an outlier in the following, pulls the average toward its value, albeit it may remain inconsistent with the resulting weighted mean, eventhough the error of the latter is rescaled.

In this paper we advocate the use of a method, termed the *Combiner* , which is an extension of the weighted mean method. Although the *Combiner* does not provide for an escape from the first drawback, it is shown below to be more satisfactory with respect to the second drawback.

The discussion of the above drawbacks is further expanded below.

1.1 Psychostatistics

There is no way out of the first drawback: this is because accepting the *possibility* of having in the set of measurements some which are biased in an unspecified way implies a loss of information, which, furthermore, is ill-defined. When the standard deviation is rescaled following the RWM scheme, the (re)definition of Eq. (3) is to be taken, at best, as a test statistics. It is no longer a pure χ^2 term: that is to say that Eq. (4) does not hold. Moreover, the use of this test statistics is ill-defined. Its distribution cannot be determined, since the underlying hypothesis is now unspecified (the set of measurements is taken to be inconsistent, but this is not a precisely defined hypothesis).

However, the aim of the RWM method being to be conservative, the price to pay is to accept the use of ill-defined statistics and to deal with $\chi^2(\mathrm{x})$ as if it was a pure χ^2. Stated differently, applying Eq. (4) yields over-conservative confidence levels, which, after all, is precisely what one is looking for.

We use the term *psychostatistics* to draw the attention on the obvious fact that numerical values have henceforth to be taken with a grain of salt: this should not be perceived as an arrogant jest from our part (arrogant would use *Voodoostatistics*), but of realistic modesty.

1.2 Schizostatistics

The second drawback is worth being spelled out explicitly. If one is dealing with two measurements which are sufficiently apart for the rescaling of Eq.(9) to be enforced, namely, if

$$\Delta x^2_{\rm exp} \equiv (x_{\rm exp}(2) - x_{\rm exp}(1))^2 > {\sigma_{x_{\rm exp}}}^2(1) + {\sigma_{x_{\rm exp}}}^2(2)\,, \tag{10}$$

leading to a rescaled uncertainty (Eq. (9)):

$$\Sigma_{\langle x_{\rm exp}\rangle} = \frac{\sigma_{x_{\rm exp}}(1)\sigma_{x_{\rm exp}}(2) \mid \Delta x_{\rm exp} \mid}{{\sigma_{x_{\rm exp}}}^2(1) + {\sigma_{x_{\rm exp}}}^2(2)}\,, \tag{11}$$

then the RWM method is prone to contradict itself in the following way.

On the one hand, using $\Sigma_{\langle x_{\rm exp}\rangle}$ in place of $\sigma_{\langle x_{\rm exp}\rangle}$ in Eq.(3), yields:

$$\chi^2({\rm x} = {\rm x_{exp}}(1)) \;\; = \;\; \left(\frac{\sigma_{x_{\rm exp}}(1)}{\sigma_{x_{\rm exp}}(2)}\right)^2\,, \tag{12}$$

$$\chi^2({\rm x} = {\rm x_{exp}}(2)) \;\; = \;\; \left(\frac{\sigma_{x_{\rm exp}}(2)}{\sigma_{x_{\rm exp}}(1)}\right)^2\,, \tag{13}$$

independently on how far apart the two measurements are, provided Eq. (10) is fullfilled. Hence, if the two measurements have widely different uncertainties, the measurement with the largest uncertainty (e.g. the second one: $\sigma_{x_{\rm exp}}(2) \gg \sigma_{x_{\rm exp}}(1)$) is viewed as corresponding to a true value of the physical observable ${\rm x} = {\rm x_{exp}}(2)$ which is utterly ruled out by the data ($\chi^2({\rm x} = {\rm x_{exp}}(2)) \gg 1$), eventhough the weighted mean error is rescaled.

On the other hand, the weigthed mean is pulled away from both measurements in proportion of $\Delta x_{\rm exp}$. As a result, from the view point of both measurements $i = 1, 2$, the hypothesis that the true value of the physical observable is ${\rm x} = \langle {\rm x_{exp}}\rangle$ leads to

$$\chi^2({\rm x} = \langle {\rm x_{exp}}\rangle)({\rm i}) = \left(\frac{{\rm x_{exp}}({\rm i}) - \langle {\rm x_{exp}}\rangle}{\sigma_{\rm x_{exp}}({\rm i})}\right)^2 = \Delta {\rm x}^2_{\rm exp} \frac{{\sigma_{\rm x_{exp}}}^2({\rm i})}{({\sigma_{\rm x_{exp}}}^2(1) + {\sigma_{\rm x_{exp}}}^2(2))^2}\,, \tag{14}$$

and, thus, is liable to be ruled out as well, if $\Delta x_{\rm exp}$ is large enough. In particular, if $\sigma_{x_{\rm exp}}(2) \gg \sigma_{x_{\rm exp}}(1)$:

$$\chi^2({\rm x} \;\; = \;\; \langle x_{\rm exp}\rangle)(1) \simeq S^2 \left(\frac{\sigma_{x_{\rm exp}}(1)}{\sigma_{x_{\rm exp}}(2)}\right)^2 \ll 1\,, \tag{15}$$

$$\chi^2({\rm x} \;\; = \;\; \langle x_{\rm exp}\rangle)(2) \simeq S^2\,. \tag{16}$$

Therefore, if the second measurement has a much larger uncertainty than the first measurement, the conjunction of Eqs. (13-15) implies that when the rescaling is significant, the RWM result and the second measurement are mutually incompatible: this contradicts the use of the second measurement to define the weighted mean, especially considering its impact on $\Sigma_{\langle x_{\rm exp}\rangle}$ as displayed by Eq. (11).

For twin measurements with identical $\sigma_{x_{\rm exp}}$, the RWM method is not self-contradictory: whereas Eq. (14) indicates that from the view point of both measurements the RWM value may be unacceptable, Eqs. (12-13) guarantees that the rescaled uncertainty yield acceptable χ^2 for both.

If only two measurements enter into play, not much can be done to circumvent this second drawback, since there is no objective way to identify the flawed one. However, if more than two measurements are available, one may rely on the consistency of a subset of them to identify the possible outliers.

2 The *Combiner*

The *Combiner* method is explicitly build as an extension of the weighted mean: by construction, it tends to reproduce Eqs.(1-2) in the case of a consistent set of measurements. The (psycho)statistical point of view which is taken here is that some of the measurements might be uncorrect: if such measurements occur, they should be removed from the set. But other points of view can be chosen. For the sake of completeness, two alternative methods are discussed in section 4.

2.1 Principle

The removal of uncorrect measurements relies on the clustering of the others measurements around a common mean. Rather than removing abruptly a measurement if it meets some criteria, the *Combiner* does not use cut but it considers all possible hypotheses about the correctness of the measurements. A configuration being defined as a subset of measurements which are assumed to be consistent the ones with others, the *Combiner* weighs all possible configurations to build an overall likelihood. To reproduce the WM result, the *Combiner* favors the configurations involving the largest number of measurements, provided they have good probabilities.

2.2 Notations

We denote:

- c, an ordered list of N bits. It is referred to as a *configuration*, indicating which measurements are considered. For example, for $N = 3$, the configuration $c \equiv 101$ means that the two measurements $i = 1$ and $i = 3$ are to be merged, while disregarding the measurement $i = 2$. The void configuration being of no interest, the total number of configurations considered amounts to $2^N - 1$.

- n_c, the number of bits set to one. It is referred to as the *multiplicity* of the configuration.

- $c_{\rm all} = 11\ldots1$, the configuration where all measurements are considered ($n_{c_{\rm all}} = N$).

- χ^2_c, the χ^2 obtained from the weighted mean (*cf.*, Eqs. (1-6)) of the n_c measurements to be considered in the configuration c.

- $\mathcal{P}_c = \mathrm{Prob}(\chi^2_c,\, n_c - 1)$, the corresponding configuration *probability* (*cf.*, Eq. (6)).

- $\overline{\mathcal{P}}_c = 1 - \mathcal{P}_c$
- $\langle x_{\rm exp}\rangle(c)$ and $\sigma_{\langle x_{\rm exp}\rangle}(c)$, the results of the weighted mean for the configuration c.
- $G_c \equiv G_c(\mathrm{x})$, the Gaussian probability density function (p.d.f.) with mean value $\langle x_{\rm exp}\rangle(c)$ and standard deviation $\sigma_{\langle x_{\rm exp}\rangle}(c)$.
- w_c, a weight characterizing the configuration c. The sum of these weights over the $2^N - 1$ configurations is normalized to unity, *i.e.*, $\sum_c w_c = 1$.
- $c' > c$ denotes two configurations such that all the bits set at one in c' are also set at one in c, and there is at least one bit set at one in c' which is not set at one in c (*e.g.*, $1111 > 1101$, but $1101 \not> 1011$). The configuration c', embedding c, is said to be *larger* than c.

2.3 Definition

With these notations, the WM method consists of using only the configuration $c_{\rm all}$. In a likelihood analysis relying on those measurements, this treatment is equivalent to adding to the overall log-likelihood the term:

$$\chi^2(\mathrm{x}) = -2\ln \mathrm{G}(\mathrm{x}) \ , \tag{17}$$

where the global p.d.f. G is the fully combined Gaussian $G = G_{c_{\rm all}}$. The RWM method consists of using the same configuration, but, if $S > 1$, it modifies $G_{c_{\rm all}}$ by rescaling its standard deviation. As pointed-out in the Introduction (cf. section 1.1) eventhough in that case $-2\ln G_{c_{\rm all}}$ defines a pseudo-χ^2, it should be used as a pure χ^2.

Instead of using a single gaussian, the *Combiner* uses for the global p.d.f. a compound function:

$$G(\mathrm{x}) = \sum_{\mathrm{c}} \mathrm{w_c G_c(x)} \ , \tag{18}$$

which enters into the computation of the pseudo-χ^2 of Eq. (17) here also to-be-used as a pure χ^2. What remains to be done is to define appropriately the weights w_c. Obviously there is an infinite number of choices. Since the goal is to protect the analysis from biased measurements and over-optimistic $\sigma_{\langle x_{\rm exp}\rangle}$ values, which is admittedly a vague goal, one must rely on educated guesswork to pick-up a particular definition of the weights.

The weights advocated for the *Combiner* are:

$$w_c = \frac{\mathcal{P}_c}{\sum_{c'}^{n_c = n_{c'}} \mathcal{P}_{c'}} \left(1 - \prod_{c'}^{n_{c'}=n_c} \overline{\mathcal{P}}_{c'}\right) \prod_{c'}^{n_{c'}>n_c} \overline{\mathcal{P}}_{c'} \ . \tag{19}$$

The first term is a measure of the relative validity of the configuration c, with respect to configurations of the same multiplicity n_c. The second term weighs the validity of the configurations of multiplicity n_c, taken as a whole. The third term suppresses configurations of multiplicity n_c if any higher multiplicity configuration receives a large probability. Therefore, whether or not a configuration c' is larger than c, if $n_{c'} > n_c$, it is sufficient for c' to receive a large probability to suppress c. For example, if an outlier measurement is utterly incompatible with all the others, it is suppressed by the *Combiner* , if some of the others are mutually compatible.

For the configuration $c = c_{\rm all}$, the third term receives no contribution and is defined to be equal to one. The above expression ensures the normalization of the weights to unity.

3 Illustrations

3.1 Twin Measurements

To illustrate how the Combiner works, we first consider the evolution of the p.d.f. G as a function of the discrepancy between twin measurements $x_{\rm exp}(1)$ and $x_{\rm exp}(2)$, with $\sigma_{x_{\rm exp}}(1) = \sigma_{x_{\rm exp}}(2) = \sigma_0 = 1$. Figure 1 shows the combined p.d.f. for the *Combiner* and for the RWM method. The measurements, indicated by the vertical lines, develop a mutual disagreement which increases from the upper left figure to the lower right figure. One observes that the *Combiner* p.d.f.'s become broader with increasing discrepancy and eventually splits up into two Gaussian-like curves. In the limit of extreme incompatibility, the two curves become genuine Gaussians with central values $x_{\rm exp}(1)$ and $x_{\rm exp}(2)$, respectively, and width σ_0. On the contrary, the RWM gaussian stays on the central value while the error increases to always keep the two measurements within the (so-called) 68% probability limit. The unsatisfactory behavior of the RWM method becomes obvious at large inconsistencies. One observes on the four lower curves in Fig 1 that, whereas the *Combiner* treats the center value $x = 0$ as being unlikely — it is favored by the weighted mean, although there exist no supporting measurement, while the rescaled uncertainty gives rise to unduly broad tails which do not show up when using the *Combiner*.

This first exemple already exhibits advantages of using the *Combiner*. However, it does not allow to demonstrate fully the superiority of the method because it uses twin measurements. As discussed previously (cf. section 1.2), the behavior of the RWM method is even less satisfactory if the two measurements have very different $\sigma_{x_{\rm exp}}$. In addition, since only two measurements are available, the *Combiner* cannot use the clustering of correct measurements to suppress the flawed one(s). As shown in the next section, the behavior of the *Combiner* is markedly different for a set of more than two measurements among which a subset is consistent.

3.2 Information Loss

It was mentioned in the Introduction that enlarged errors (broader p.d.f's) are an unavoidable side-effect when taking into consideration the possibility of biased measurements. This entails a loss of information when the set is consistent. To quantify this loss of information, we perform a toy Monte Carlo simulation of a set of $N = 5$ measurements, each distributed following the same Gaussian, $\langle x_{\rm exp}\rangle(1-5) = \langle x_0\rangle = 5$ and $\sigma_{x_{\rm exp}}(1-5) = \sigma_0 = 1$. We use the RWM method and the *Combiner* to obtain the distributions of the mean value and the distributions of the root mean square (r.m.s) of the p.d.f. G. The results are shown on Fig. 2. The left plot shows the distribution of the mean value $\langle x[{\rm C}]\rangle$ as provided by the *Combiner* . The distribution of $\langle x[{\rm C}]\rangle$ is very close to be a Gaussian of width $\sigma = 0.47$. This is to be compared with the RWM gaussian distribution of width $\sigma_{WM:5} = 5^{-1/2} \simeq 0.45$. The figure on the right shows the distribution of the rms. σ[C] of the *Combiner* , to be compared with the hatched histogram which represents the distribution of the rms. σ[RWM] of the rescaled weighted mean. mean, the distribution of the maxima of the respective p.d.f.'s obeys a Gaussian shape with standard deviation $\sigma_0/\sqrt{N} \simeq 0.45$. The same holds for the rescaled weighted mean, as the center value of the p.d.f. is not affected by the enlargement of the width. However, this is not true for the Combiners. Although statistical outliers are suppressed, giving rise to a narrower effective width, the increase of the width due to the statistical occurrence of seeming inconsistencies superseeds the narrowing suppression effect. The left hand shows the resulting distributions of central values for the weighted mean (pure Gaussian) and the C and the T-Combiner. The increase of the width with respect to the weighted mean amounts to 10% for the C-Combiner and 9% for the T-Combiner. gives the corresponding distribution of the r.m.s. for the unscaled weighted mean (constant at threshold), the rescaled weighted mean as well as the C and the T-Combiner. As expected, the increase of the errors is strongest for the C-Combiner.

3.3 Inconsistent Set

The second example uses the following set of $N = 5$ measurements: $x_{\rm exp}(1-5) = 3.7,\ 4.2,\ 5.0,\ 5.5,\ 0.0$, all with identical errors $\sigma_{x_{\rm exp}}(1-5) = 1$. While the first four data points are mutually compatible with a Gaussian distribution ($\chi^2/{\rm N_{dof}} = 1.9/3$), the last measurement is an outlier leading to a large overall $\chi^2/{\rm N_{dof}} = 18.9/4$, translated into a scale factor of $S = 2.2$. The *Combiner* yields for the p.d.f. G (quoting only the leading terms):

$$G \simeq 0.580\ G_{1110} + 0.125\ G_{01110} + 0.118\ G_{10110} + 0.079\ G_{11010} + 0.077\ G_{11100}\,, \qquad (20)$$

The *Combiner* p.d.f. is shown in Fig. 3 together with the WM and the RWM functions. As pointed out in the Introduction, one observes that the WM and the RWM methods contradict themselves: the outlier pulls the mean value significantly but, even in the RWM method, the outlier and the mean value remain incompatible. Furthermore, in the exemple considered here, the measurement the farthest apart from the outlier, although correct, is also incompatible with the mean value.

4 Alternative Methods

4.1 The Cautious Combiner

For the Cautious Combiner, $G^{(\rm C)}$, the weights are defined as

$$w_c^{(\rm C)} = a\ \mathcal{P}_c \prod_{c'}^{c'>c} \overline{\mathcal{P}}_{c'}\ , \qquad (21)$$

where the constant a ensures the proper normalization of the weights. The first term is a measure of the validity of the combination, while the second term suppresses it, if any configuration larger than c receives a high probability. If a configuration c is such that all larger configurations are unlikely, it does not get suppressed by the above expression. For example, if an outlier measurement is utterly incompatible with all the others, it is kept by the C-Combiner with a weight equal to a. It is termed the *Cool-Combiner* because of that. However, if the number of consistent measurements grows, the outlier weight, a, decreases accordingly.

4.2 Bayesian Based Method

4.3 Comparisons

shown in Fig. 7. Whereas one may be satisfied by either the C-Combiner or the T-Combiner, the rescaled weighted mean is not satisfactory: the central value and the error are strongly affected by the outlier which, however, remains inconsistent with the weighted mean result. A sample of the most significant weights obtained for this example is given below:

Quoted are only the configurations with $w_c > 0.05$. One observes that configurations where the incompatible measurements are mixed, *i.e.*, bit five and another one are set to one, have negligible weights for both the C and the T-Combiner. The C-Combiner allows a sizable single weight for outlier (measurements five) which, while it is suppressed by the T-Combiner.

5 Applications

5.1 Combining Measurements of ϵ'/ϵ

As a practical application, we shall use the C and the T-Combiner to create a combined p.d.f. of the measurements of direct CP-violation in the K sector described by the quantity ϵ'/ϵ. At present, there exist four measurements which are assumed to be independent in the following:

$$\epsilon'/\epsilon \times 10^4 = \begin{cases} 23.0 & \pm & 6.5 & (\text{NA31 [3]}) \\ 7.4 & \pm & 5.9 & (\text{E731 [4]}) \\ 28.0 & \pm & 4.1 & (\text{KTeV [5]}) \\ 14.0 & \pm & 4.3 & (\text{NA48 [6]}) \end{cases} \tag{22}$$

for which we obtain a weighted mean of $\langle\epsilon'/\epsilon\rangle = (19.2 \pm 2.4) \times 10^{-4}$ and a scale factor $S = 1.9$, so that the rescaled weighted mean reads $\langle\epsilon'/\epsilon\rangle(\text{rescaled}) = (19.2 \pm 4.6) \times 10^{-4}$.

The resulting p.d.f.'s for the C and the T-Combiner together with the one standard deviation error bars for the simple and the rescaled weighted mean are plotted in Fig. 8. Also shown are the four measurements listed above. One observes the tendency of the T-Combiner to suppress the E731 measurement as an outlier, while it is retained by the C-Combiner. The weighted mean does not account for these details.

5.2 Combining Measurements of ΔM_d

6 Conclusions

We have developed a straightforward algorithm for combining potentially inconsistent measurements. The combiner behaves superior compared to the commonly used χ^2 rescaling of the error obtained from the standard weighted mean for the following reasons: (i), the high probability zones lie within the range of values covered by the measurements, and (ii), outliers are detected and can be suppressed, where the strength of the suppression is different for the C-Combiner and the T-Combiner. As an application, we show the probability density functions obtained for the present data set of ϵ'/ϵ measurements.

The FORTRAN program `combiner.f` together with a detailed description of its functionality can be copied from the Web page: http://www.slac.stanford.edu/~laplace/ckmfitter.html, *via* the link: *"What is CKM-Fitter"*.

Acknowledgements

We gratefully acknowledge stimulating discussions with Heiko Lacker.

References

[1] The Pudding Data Grinder,

[2] G. D'Agostini, *"Sceptical combination of experimental results: General considerations and application to ϵ'/ϵ"*, CERN-EP-99-139 (1999)

[3] NA31 Collaboration (G.D. Barr *et al.*), *Phys. Lett.* **B317** (1993) 233

[4] E731 Collaboration (L.K. Gibbons *et al.*), *Phys. Rev. Lett.* **70** (1993) 1203

[5] KTeV Collaboration (A. Alavi-Harati *et al.*), *Phys. Rev. Lett.* **83** (1999) 22

[6] G. Graziani for the NA48 Collaboration, *"A new result on direct CP violation from the NA48 experiment"*, Talk given at the XXXV Rencontres de Moriond - Electroweak Interactions and Unified Theories, Les Arcs, France, March 2000

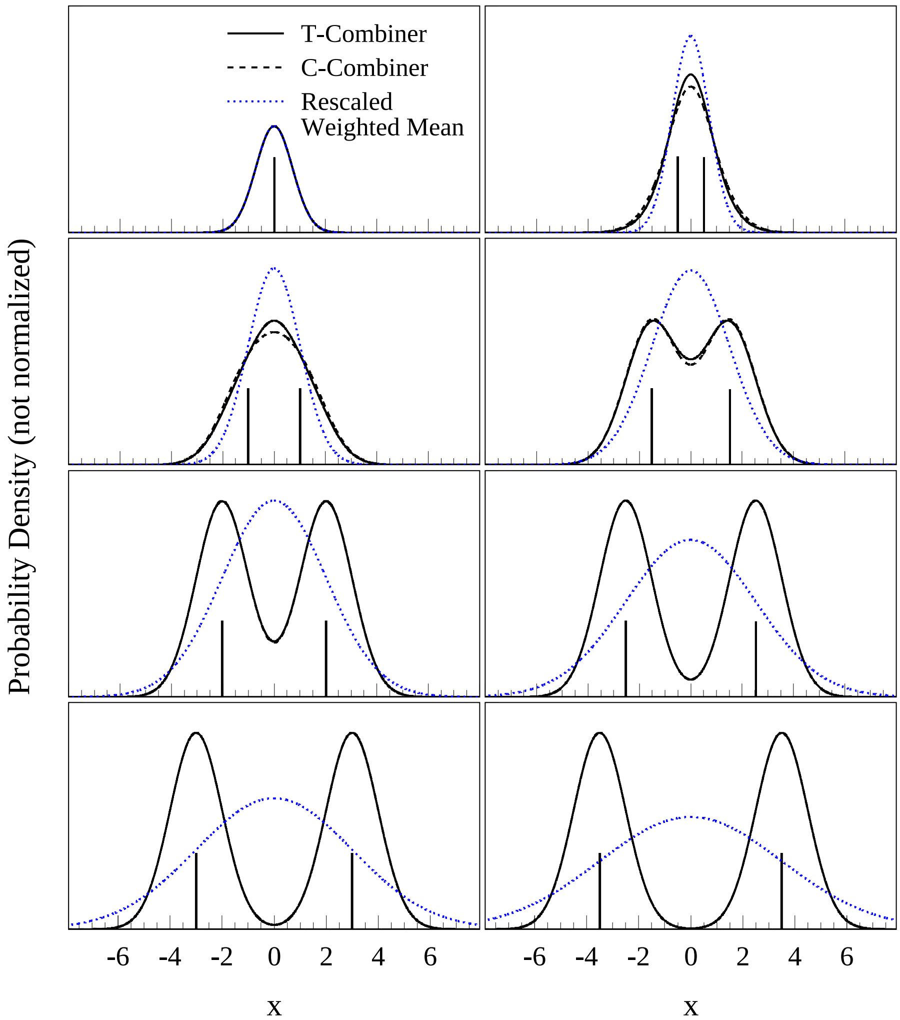

Figure 1: *Probability density functions G for the* Combiner *and for the RWM method. The vertical lines indicate the locations of the two individual measurements which develop a mutual disagreement rising from the upper left figure to the lower right figure. The dotted line corresponds to the p.d.f. of the Cautious*-Combiner *(cf. section 4.1).*

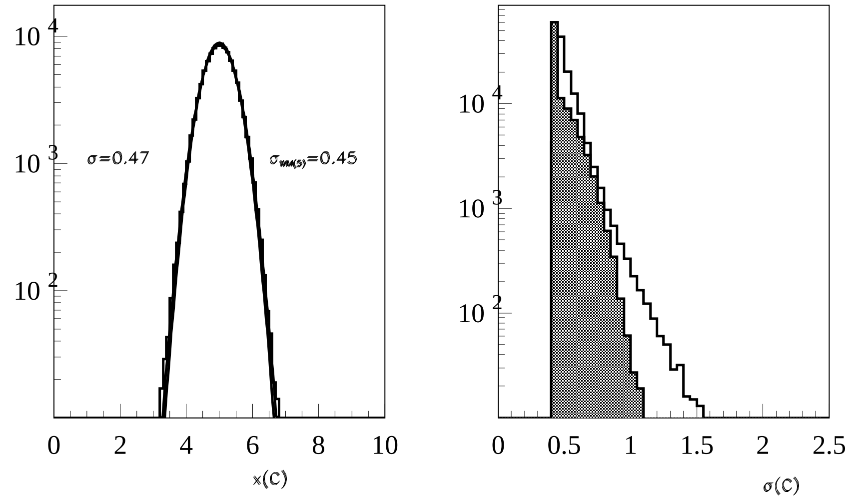

Figure 2: *The figure on the left shows the distribution of the mean value* $\langle x[\mathrm{C}]\rangle$ *as provided by the* Combiner *, for a set of five consistent measurements which are gaussian distributed around* $\langle x_{\mathrm{exp}}\rangle = 5$*, each with a standard deviation* $\sigma(i) = 1$*. Superimposed, but barely distinguishable, are a gaussian fit of the distribution which results in* $\sigma = 0.47$*, together with the weighted mean distribution characterized by* $\sigma_{WM:5} = 5^{-1/2} \simeq 0.45$*. The figure on the right shows the distribution of the rms.* $\sigma[\mathrm{C}]$ *of the* Combiner *, to be compared with the hatched histogram which represents the distribution of the rms.* $\sigma[\mathrm{RWM}]$ *of the rescaled weighted mean.*

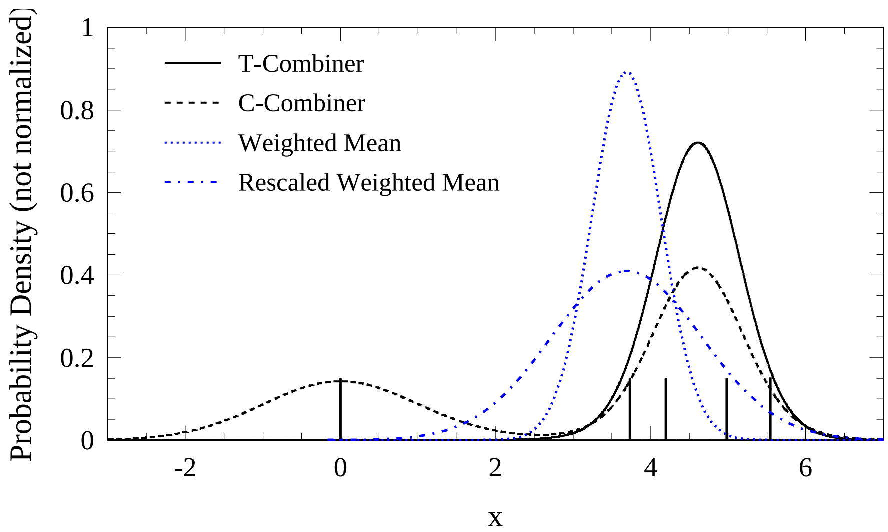

Figure 3: *Probability density functions G of the* Combiner *and of the WM and the RWM methods resulting from five individual measurements (indicated by the vertical lines).*

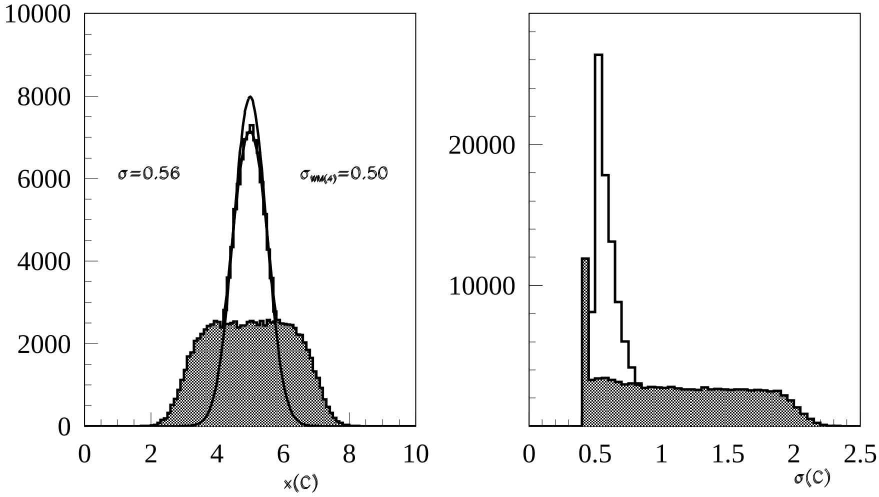

Figure 4: *The figure on the left shows the distribution of the mean value* $\langle x[\mathrm{C}]\rangle$ *as provided by the* Combiner *, for a set of inconsistent measurements, four of which are gaussian distributed around* $\langle x_{\rm exp}\rangle = 5$*, each with a standard deviation* $\sigma(i) = 1$*, while the fifth measurement is uniformly distributed between* $x_5 = -5$ *and* $x_5 = 15$*. Superimposed, are a gaussian fit of the distribution which results in* $\sigma = 0.47$ *(to be compared with the optimal WM result* $\sigma_{WM:4} = 0.5$*, obtained when the fifth measurement is removed from the set) together with the weighted mean distribution (the hatched histogram). The figure on the right shows the distribution of the rms.* $\sigma[\mathrm{C}]$ *of the* Combiner *, to be compared with the distribution of the rms. of the rescaled weighted mean (the hatched histogram).*

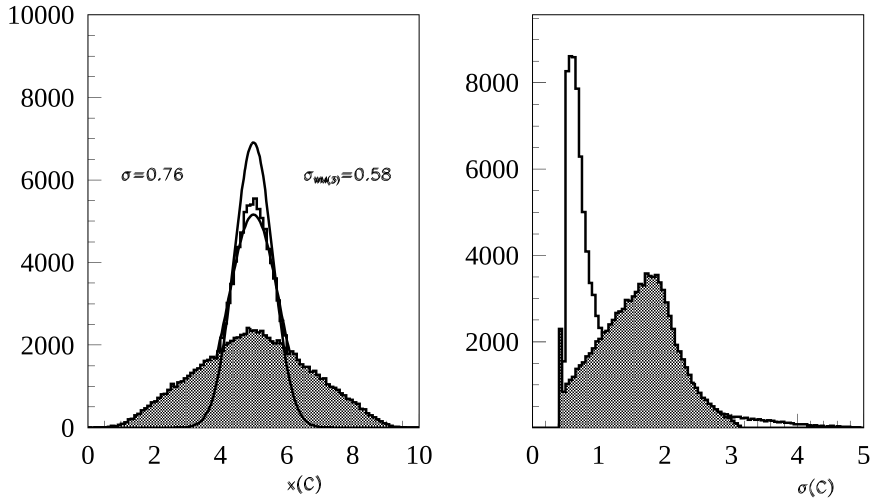

Figure 5: *The figure on the left shows the distribution of the mean value* $\langle x[\mathrm{C}]\rangle$ *as provided by the* Combiner *, for a set of inconsistent measurements, three of which are gaussian distributed around* $\langle x_{\mathrm{exp}}\rangle = 5$*, each with a standard deviation* $\sigma(i) = 1$*, while two measurements are uniformly distributed between* $x_5 = -5$ *and* $x_5 = 15$*. Superimposed, are a gaussian fit of the distribution which results in* $\sigma = 0.76$ *(to be compared with the optimal WM result* $\sigma_{WM:3} \simeq 0.58$*, obtained when the two latter measurements are removed from the set) together with the weighted mean distribution (the hatched histogram). The figure on the right shows the distribution of the rms.* $\sigma[\mathrm{C}]$ *of the* Combiner *, to be compared with the distribution of the rms. of the rescaled weighted mean (the hatched histogram).*

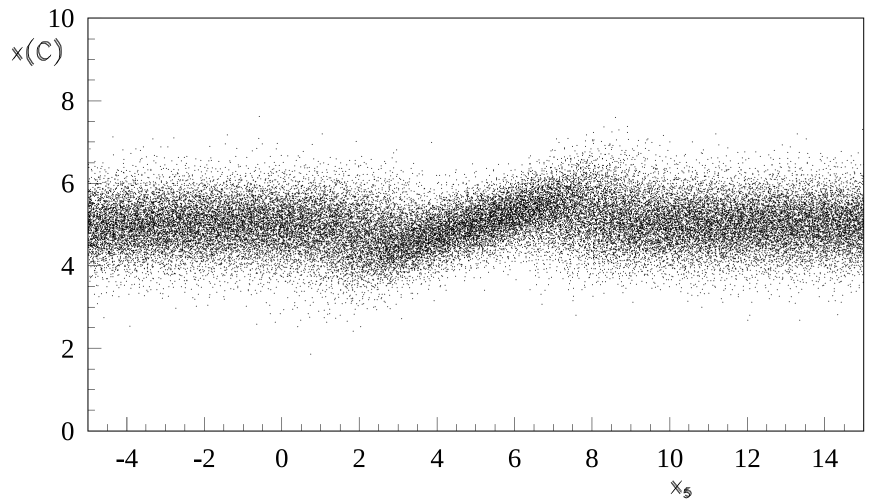

Figure 6: *Distribution of the mean value* $\langle x[\mathrm{C}]\rangle$ *as provided by the* Combiner *(on the vertical axis) for a set of inconsistent measurements, four of which are gaussian distributed around* $\langle x_{\mathrm{exp}}\rangle = 5$*, each with a standard deviation* $\sigma(i) = 1$*, while the fifth measurement is uniformly distributed between* $x_5 = -5$ *and* $x_5 = 15$ *(on the horizontal axis). The latter is effectively removed from the set when it departs from* $\langle x_{\mathrm{exp}}\rangle$ *by about* $1.5\sigma(i)$*.*

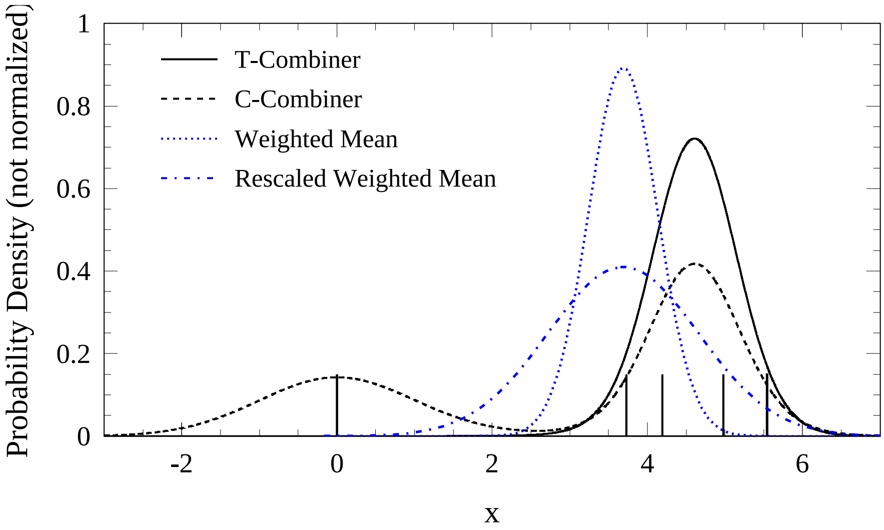

Figure 7: *Probability density functions G obtained by the three methods: T-Combiner, C-Combiner, weighted mean and the rescaled weighted mean according to Eq. (8). The vertical lines indicate the five individual measurements.*

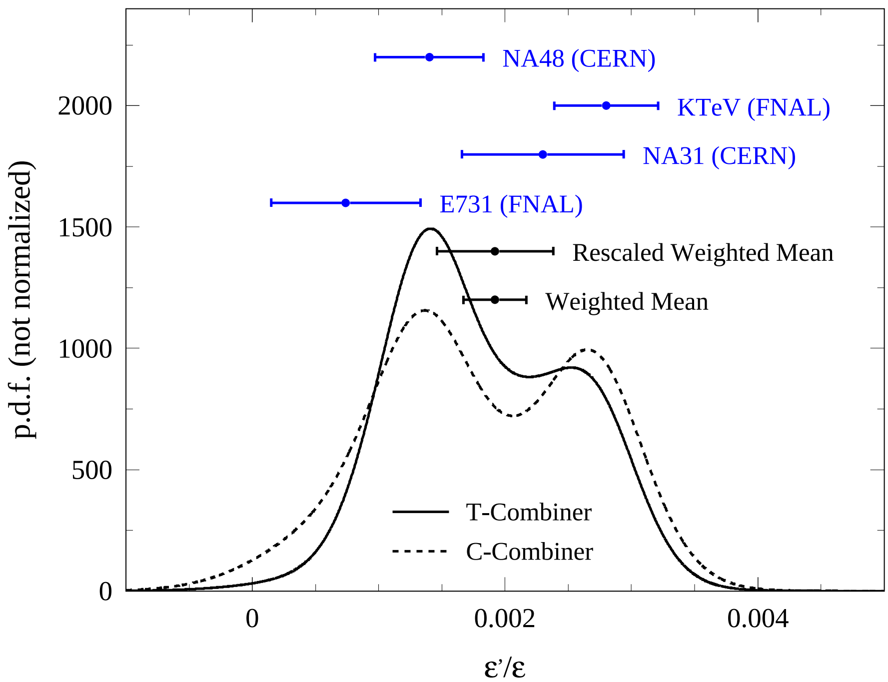

Figure 8: *Combined p.d.f. for ϵ'/ϵ using the C and the T-Combiner. Shown in addition are the one standard deviation error bars of the simple and the rescaled weighted mean as well as the four input measurements taken from Refs. [3, 4, 5, 6].*