On estimating systematic errors from repeated measurements

Louis Lyons

Nuclear Physics Laboratory, University of Oxford, Keble Road, Oxford OX1 3RH, UK

Abstract. We discuss various methods for estimating the bias from a repeated set of measurements that exhibit a spread larger than expected on the basis of their statistical errors. Monte Carlo simulations are used to test how well the methods perform.

1. Introduction

Experimental measurements are usually subject to two types of errors, statistical and systematic. The former arise from fluctuations in finite data samples and from limited resolution of measurement devices. They are such that the average of a large number of repeated measurements should tend to the correct value. On the other hand systematic errors are potential sources of bias, arising, for example, from incorrectly calibrated instruments, samples contaminated by background, external effects not taken into account, inadequate theories, etc. It is a general feature of error estimation that the contribution from statistical effects is relatively easy to obtain, while systematic errors tend to be assessed in a much more subjective manner.

One approach to investigating systematic effects is to calculate the quantity of interest x in several different ways, and to see how much scatter there is among the various answers x_i. Thus, for example, different selection criteria or cuts could be used to obtain samples from which the x_i are determined. Alternatively the identical data could be used, but analysed in several different methods. Yet another possibility is that there could be several variants of the theory with which the data are compared in order to extract x.

In this paper, we discuss the problem of deriving a numerical estimate of the systematic error from several estimates of the same quantity. Monte Carlo tests on the reliability of the suggested procedures are presented.

The methods we describe could also be applied to the problem of extracting the best value and its error from several different experimental measurements of the same quantity, in the situation where the scatter of the measurements is larger than expected from the quoted errors (and hence some source of bias in the quoted results may be suspected).

The methods discussed here are applicable only to those forms of bias which are distributed with zero mean, and which vary from measurement to measurement. Any bias which is common to the separate results would escape detection by repeated measurement. However, the methods we describe can readily be extended to the situation of correlated statistical errors, such as arise from overlapping data samples, or from different analysis techniques applied to the same data (see [1]).

2. Techniques for estimating bias

2.1. Two measurements

We assume that we have two measurements $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$ with uncorrelated errors. In the absence of systematic effects, the expected (RMS) difference between the measurements is $\sqrt{\sigma_1^2 + \sigma_2^2}$.

The effect of systematics is to shift the measurements by some unknown amount. We assume that the shift in each result can be regarded as being derived from a distribution of mean zero and RMS width s. We furthermore assume that the possible biases on the two measurements are uncorrelated. This may well not be true. However if they were completely correlated, the two measurements would be shifted in the same sense, and this approach of looking for differences in the x_i would break down.

Since the systematic and the statistical errors on a given measurement are certainly uncorrelated, the total error on each x_i is $\sqrt{\sigma_i^2 + s^2}$, and so, in the presence of statistical and systematic errors of the type discussed in the previous paragraph, the expected difference in measurements becomes $\sqrt{(\sigma_1^2 + s^2) + (\sigma_2^2 + s^2)}$. If we set this equal to the observed difference, we obtain our estimate

$$s^2 = \tfrac{1}{2}\left[(x_1 - x_2)^2 - (\sigma_1^2 + \sigma_2^2)\right]. \qquad (1)$$

Because of statistical fluctuations, $|x_1 - x_2|$ can be smaller than $\sqrt{\sigma_1^2 + \sigma_2^2}$. In that case, one would conclude that there is no evidence for any systematic effect. This in fact will result in a small bias when the true value of s is small; if our estimates of s^2 are larger than the true value we accept them, while for smaller values we are likely to set our estimate to zero.

An alternative is to assume that the effect of systematic errors is to increase the statistical error by a factor $\kappa \geqslant 1$. Then the expected difference becomes $\kappa\sqrt{\sigma_1^2 + \sigma_2^2}$, and so our estimate of κ is

$$\kappa = \frac{|x_1 - x_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \qquad (2)$$

Then $\kappa > 1$ is a measure of any systematic effects. In analogy with the previous example, if our estimate $\kappa < 1$, we would assume there were no systematic effects.

2.2. Several measurements

We now extend the discussion of section 2.1 to the case of several measurements $x_i \pm \sigma_i$ with uncorrelated errors. Again the bias is assumed to be distributed with mean zero and RMS s, and to be uncorrelated between measurements. We discuss four different approaches for estimating s.

2.2.1. The χ^2 method. If we assume that the distribution of bias values is not too different from normal, we can define a χ^2 for the measurements to be consistent with a single value $\hat{x}$:

$$\chi^2 = \sum \frac{(x_i - \hat{x})^2}{\sigma_i^2 + s^2} \qquad (3)$$

where $\hat{x}$ is chosen to minimize the χ^2 for a given s i.e.

$$\hat{x} = \left\{ \sum x_i / (\sigma_i^2 + s^2) \right\} \bigg/ \left\{ \sum 1 / (\sigma_i^2 + s^2) \right\}. \tag{4}$$

For N measurements, the expected value of χ^2 is $N-1$, and so we determine s as that value which results in

$$\sum \frac{(x_i - \hat{x})^2}{\sigma_i^2 + s^2} = N - 1 \tag{5}$$

where $\hat{x}$ is a function of s (see (4)). The value of s that satisfies (5) we obtain numerically. This is our estimate of the systematic error. The statistical error we obtain in the usual way as $\{ \sum 1/\sigma_i^2 \}^{-1/2}$.

For the case where s is large compared with all the σ_i, equation (5) gives us our estimate of s^2 as $\sum (x_i - \hat{x})^2 / (N-1)$, with $\hat{x}$ being the ordinary unweighted average of the x_i. Thus when systematic errors dominate over statistical ones, s is estimated simply as the RMS of the observed x_i values.

In analogy with section 2.1, here we can also adopt the alternative of considering that the effect of systematics is to multiply all statistical errors by a factor κ. In this case,

$$\hat{x} = \left\{ \sum x_i / \kappa^2 \sigma_i^2 \right\} \bigg/ \left\{ \sum 1 / \kappa^2 \sigma_i^2 \right\} \tag{6}$$

and hence is independent of κ. The value of κ is determined analytically from

$$\chi^2 = \sum \frac{(x_i - \hat{x})^2}{\kappa^2 \sigma_i^2} = N - 1. \tag{7}$$

As discussed in section 2.1, we would probably not allow $s^2 < 0$ or $\kappa < 1$.

It is easy to check that, for the case where $N = 2$, equations (5) or (7) give the same value for s or κ as obtained from (1) or (2).

We can obtain error estimates for s or κ as follows. Even when theory and experiment are in agreement, χ^2 need not be exactly $N-1$. Indeed the χ^2 distribution has RMS width of $\sqrt{2(N-1)}$. Thus for N not too small, we can obtain upper and lower error estimates for s from

$$\sum \frac{(x_i - \hat{x})^2}{(\sigma_i^2 + s^2)} = (N-1) \pm \sqrt{2(N-1)} \tag{5'}$$

(and similarly for κ). For small values of N, it is better to choose values of χ^2 which correspond to appropriate values (16 and 84%) of the integrated area under the χ^2 distribution. For small N, the magnitudes of the errors on s or κ are large.

2.2.2. The probability approach. Assuming again that the biases come from a normal distribution, the probability of observing the given set of x_i values is

$$\prod \frac{1}{\sqrt{2\pi(\sigma_i^2 + s^2)}} \exp \left\{ -\frac{(x_i - \hat{x})^2}{2(\sigma_i^2 + s^2)} \right\}. \tag{8}$$

We then maximize the logarithm of this probability in order to obtain the best estimates of $\hat{x}$ and s. The former is given as before by (4), and the best value of s is determined numerically. The error on s can be determined by finding the values required to reduce the logarithm of the probability by $\frac{1}{2}$ (as compared with its maximum value).

When $s \gg \sigma_i$, this procedure results in

$$s^2 = \sum (x_i - \hat{x})^2 / N \tag{9}$$

with $\hat{x}$ being the unweighted average. Thus in this limit s^2 agrees with the value of section 2.2.1, except for the denominator being N rather than $N - 1$. This is the well-known effect that likelihood methods give a biased estimate of the variance of a normal distribution.

For the situation where all the errors σ_i are equal, the χ^2 method and the probability one also agree (except for the difference between N and $N - 1$ in the denominator).

Finally for the alternative where the systematic effects are assumed to result in a scaling of the statistical errors by a factor κ, this approach gives

$$\kappa^2 = \frac{1}{N} \sum \frac{(x_i - \hat{x})^2}{\sigma_i^2} \tag{10}$$

which again agrees (up to a factor of $(N - 1)/N$) with the corresponding approach in section 2.2.1.

2.2.3. Individual bias method.

This method, which does not require any explicit assumption about the shape of the distribution of biases, defines the χ^2 for the agreement of observed values x_i as

$$\chi^2 = \sum \frac{(x_i - \hat{x} - b_i)^2}{\sigma_i^2} \tag{11}$$

where b_i is the estimate of the bias for the ith measurement, and the weighted mean is

$$\hat{x} = \frac{\sum (x_i - b_i)/\sigma_i^2}{\sum 1/\sigma_i^2}. \tag{12}$$

The b_i are constrained by

$$\sum b_i^2 = (N - 1)s^2 \tag{13}$$

i.e. s is the RMS of the b_i. Then the χ^2 of (11) is minimized with respect to the b_i, subject to the constraint (13) for a given value of s. Our estimate of the overall bias is that value of s for which the minimum value of χ^2 is $N - 1$ (unless χ^2 is already smaller than $N - 1$ when $s = 0$, in which case we set the estimate of the bias to zero).

This procedure can be illustrated graphically for the case of two measurements. In figure 1, the star represents the true values of x_1 and x_2, while the measured
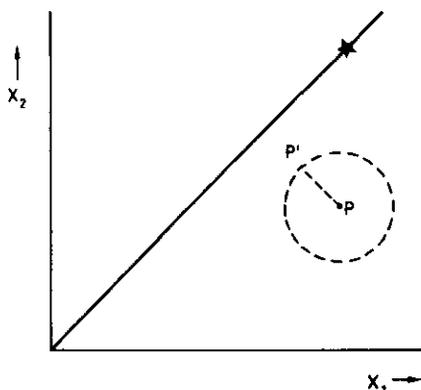
Figure 1. Diagram illustrating the 'individual bias' technique for the case of two measurements x_1 and x_2, denoted by the point P. The true (unknown) value is represented by the star, but any value on the line $x_1 = x_2$ would constitute a satisfactory solution $\hat{x}$. The constraint $b_1^2 + b_2^2 = s^2$ requires $(x_1 - b_1, x_2 - b_2)$ to lie on the broken circle. Then we calculate the minimum χ^2 for agreement between any point on the circle and any point on the line; this will select the point P', such that the gradient of P'P is -1. Finally the radius s of the broken circle is chosen so that the minimum χ^2 is 1.

values are shown by the point P. Since the position of the star is unknown, we will accept as our solution for the best value $\hat{x}$ any point on the line $x_1 = x_2$. If we correct for the biases b_1 and b_2, the measurements would be moved to $(x_1 - b_1, x_2 - b_2)$. Given the constraint (13), this is required to lie on the broken circle. The point on the circle which gives the smallest value of χ^2 for consistency with any point on the 45° line is given by $b_1 + b_2 = 0$ i.e. the closest point to the line. Finally the radius s of the circle is chosen such that the minimum χ^2 is $N - 1 = 1$.

For the case where the two measurements are $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$, the estimated bias is

$$s = \frac{1}{\sqrt{2}} \left[|x_1 - x_2| - \sqrt{\sigma_1^2 + \sigma_2^2} \right] \tag{14}$$

provided that this gives $s > 0$.

The errors on s are obtained by letting the minimum χ^2 be the appropriate values, rather than $N - 1$ (compare section 2.2.1).

If we believe that the biases on some measurements are inherently likely to be larger than those of others, we can incorporate this by modifying (13) appropriately. Thus if the biases are assumed proportional to the statistical errors, we could write

$$\sum \left(\frac{b_i}{\sigma_i} \right)^2 = (N - 1)\kappa^2 \tag{15}$$

where now κ is our unknown factor to be determined, as in sections 2.2.1 and 2.2.2.

For the general case of N measurements, minimizing the χ^2 subject to the constraint (13) results in the following linear equations for the b_i:

$$R b_i (1 + \lambda \sigma_i^2) - \sum (b_j / \sigma_j^2) = R x_i - \sum (x_j / \sigma_j^2) \tag{16}$$

where

$$R = \sum 1 / \sigma_j^2 \tag{17}$$

and λ is the unknown Lagrangian multiplier. The procedure for finding s is thus to choose a value of λ, and to solve equations (16) for the b_i. This enables us to calculate $\hat{x}$ from (12) and then χ^2 from (11), as well as s from (13). Then we have to iterate to find the value of λ which makes the minimized χ^2 equal to $N - 1$; the corresponding value of s is our estimated overall bias.

2.2.4. Frohner's method. Frohner [2] has described an elegant Bayesian approach to this problem. He writes the joint probability for the answer $\hat{x}$ and the bias b_i as

$$p(\hat{x}, b|x, \sigma, \tau) \sim \int_0^\infty p(c) \prod_i \sqrt{c} \exp \left\{ -\frac{(x_i - \hat{x} - b_i)^2}{2\sigma_i^2} - \frac{cb_i^2}{2\tau_i^2} \right\} dc \qquad (18)$$

where as usual the possible biases in each experiment are assumed to be Gaussian distributed about zero with unknown variances. The extra assumption here is that these variances are taken as τ_i^2/c, where τ_i is our best guess of the width of the distribution of bias for the ith experiment, and c is a common scale factor. The prior for this scale factor is $p(c)dc$, for which Frohner suggests the form $\exp(-c)dc$.

Then integration over the b_i gives

$$p(\hat{x}|x, \sigma, \tau) \sim \int_0^\infty e^{-c} \prod_i \left[\exp \left\{ -\frac{1}{2} \frac{(x_i - \hat{x})^2}{\sigma_i^2 + \tau_i^2/c} \right\} \bigg/ (\sigma_i^2 + \tau_i^2/c)^{1/2} \right] dc. \qquad (19)$$

Similarly integration over $\hat{x}$ and also over c results in

$$p(b|x, \sigma, \tau) \sim \exp \left\{ -\tfrac{1}{2}(\tilde{x} - \tilde{b}) \mathbf{A}^{-1}(x - b) \right\} \left[1 + \tfrac{1}{2} \tilde{b} \mathbf{B}^{-1} b \right]^{-n/2-1} \qquad (20)$$

where we have written x and b for the vectors of the x_i and b_i respectively, and the matrices $\mathbf{A}$ and $\mathbf{B}$ are defined by

$$(\mathbf{A}^{-1})_{ij} = \delta_{ij}/\sigma_i^2 - 1/\{\sigma_i^2 \sigma_j^2 (\Sigma 1/\sigma_k^2)\} \qquad (21)$$

and

$$(\mathbf{B}^{-1})_{ij} = \delta_{ij}/\tau_i^2. \qquad (22)$$

The maxima of equations (19) and (20) give the estimates $\hat{x}$ and b_i respectively. For the latter we can differentiate (20) to obtain the b_i as the solutions of

$$\mathbf{A}^{-1}(b - x) + \frac{n+2}{2} \frac{\mathbf{B}^{-1}b}{1 + \tilde{b}\mathbf{B}^{-1}b/2} = 0. \qquad (23)$$

This reduces to

$$b_i = x_i - \bar{x} - \bar{b} - \frac{n+2}{2} \frac{b_i \sigma_i^2/\tau_i^2}{1 + \Sigma(b_j^2/2\tau_j^2)} \qquad (24)$$

where $\bar{x}$ and $\bar{b}$ are the $1/\sigma_i^2$ weighted averages. According to Frohner, (24) can be solved iteratively for the b_i, starting with $b_i = x_i - \bar{x}$.

We finally estimate s^2 as $\Sigma b_i^2/(N - 1)$.

The performance of these four methods is described in the next two sections.

3. Two uncorrelated measurements

We first discuss the bias estimates provided by these methods for the very simple situation of two measurements with equal statistical errors. We assume that these measured values were 10 ± 1 and 14 ± 1.

3.1. The χ^2 method

Since the two σ_i are the same, our best estimate $\hat{x}$ is 12, independent of s. Then from (5)

$$\chi^2 = \frac{2^2}{1 + s^2} + \frac{2^2}{1 + s^2} = 1$$

which gives $s^2 = 7$.

   As previously mentioned in section 2.2.1, this gives the identical answer to that obtained from a simple approach: the estimate of the standard deviation of the two measurements is $\sqrt{8}$, of which 1 is contributed by the statistical error, leaving a systematic error of $\sqrt{7}$.

   Now we estimate how accurately we determine s. We first require the values of χ^2 for which the fractional area in the remaining tail is 16 or 84%, corresponding to those beyond the 1 standard deviation points of a Gaussian. For one degree of freedom, the relevant values of χ^2 are 1.98 and 0.041. Substituting these instead of the $N - 1 = 1$ on the right-hand side of (5), we determine the possible range of s as 1.7 to 13.9, compared with the central value of $\sqrt{7} = 2.6$. As we have already commented, estimating a spread from two values is not very precise.

3.2. The probability approach

For this particular example with equal statistical errors, the 'probability' approach gives the same bias estimates as the χ^2 method, except that they are scaled down by $\sqrt{(N-1)/N} = 1/\sqrt{2}$. Thus the bias s is given as 1.8, and the range of acceptable values is 1.2 to 9.8.

3.3. Individual bias method

Equation (14) determines the separate biases b_i as ∓ 1.3. For the $N = 2$ case, the estimate of the bias $s = \sqrt{2}|b_1|$, and the 68% probability range is 1.4 to 2.6. (The agreement of the bias estimate with that from the probability approach is coincidental.) Compared with the previous two techniques, this provides a significantly lower estimate of the upper error limit This is due to the fact that, in order to obtain the lowest acceptable value of χ^2 (0.041) in (11), the b_i in the numerator are limited to be below half the difference in the observed x_1 and x_2, and hence so is $s/\sqrt{2}$. In contrast for (3) s is in the denominator and has to become large in order to make χ^2 suitably small. An equivalent remark applies also to (8) of the 'probability' method.

3.4. Frohner's method

In this case we investigate the more general situation of two measurements 10 ± 1 and $x_2 \pm 1$, where x_2 is in the range 11 to 20. As in the 'individual bias' method, $b_1 = -b_2$ and $\hat{x} = (x_2 + 10)/2$.

Several problems become apparent:

(i) Iteration with (24) gives problems concerning convergence. This is acute when $x_2 - 10$ is small, and for small τ (see table 1). It is not significantly alleviated by choosing a better starting value of b.

(ii) For this simple problem, it is not necessary to use (24), but we can evaluate p in (20) as a function of b_2 in order to find the maximum. Then we can plot b_2 as a function of τ. This is shown for $x_2 = 12, 14$ or 15 in figure 2(a). We see that our estimate of the bias varies with τ; at low τ, $b_2 \sim 0$, while for large values it tends to $(x_2 - 10)/2$.

(iii) At larger values of x_2 and at small τ, there are three solutions of (24). Of these two correspond to maxima of p, and one is a minimum (see figure 3). We select as the bias the one which gives the larger maximum. The iterative solution of (24) in this region often converges on the wrong maximum. Furthermore at some value of τ, the solution for b_2 jumps from the lower branch to the upper one (see figure 2(b)).
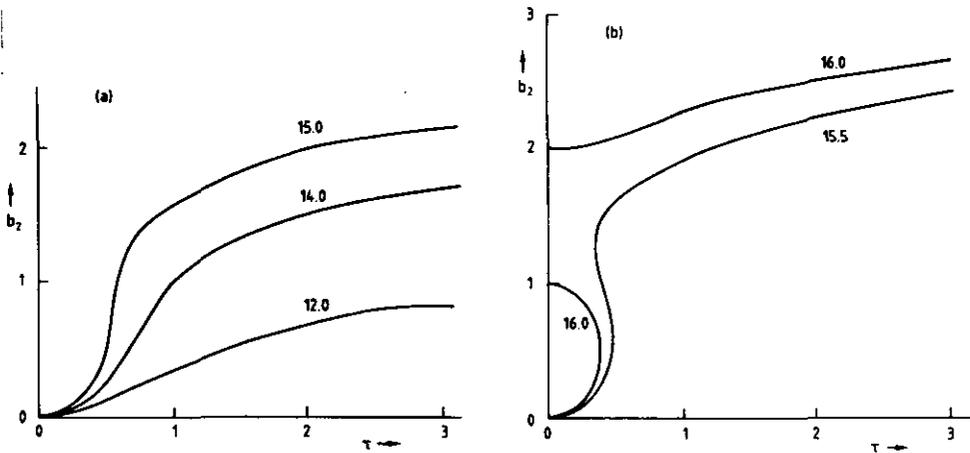

Figure 2. The bias b_2 as a function of τ, for measurements 10 ± 1 and $x \pm 1$, where x is as specified on the curves. The curves correspond to the positions of $\partial p / \partial b = 0$, with p as given in (20). At large τ, b_2 always tends to $(x_2 - x_1)/2$. For the values of x_2 as shown in (a), b_2 grows monotonically with τ. For $x_2 = 15.5$, the function p displays two maxima for $\tau \sim 0.4$, and for $x_2 > 10 + 4\sqrt{2}$, this region extends down to $\tau = 0$, with the curve of b_2 as a function of τ splitting into two. The larger maximum starts on the lower curve and then jumps to the upper one for larger τ; this occurs at $\tau \sim 0.42$ for $x_2 = 15.5$, and at $\tau \sim 0.31$ for $x_2 = 16$. For x_2 larger than 16, the area of the lower portion of the curve shrinks, and the range of b_2 values covered by the upper part of the curve also decreases.

In fact the situation is not quite as bad as at first appears, provided $|x_2 - x_1|$ and τ are not too small. Thus for $x_2 = 16$ and $\tau > 0.5$, the range of b_2 is from 2.05 to 3 (for $\tau = \infty$). For larger x_2, the range of b_2 is even smaller.

The results of these methods are summarized in table 2.

Table 1. Convergence of the Frohner method. The Frohner method, applied to two measurements 10 ± 1 and $x_2 \pm 1$. Equation (24) is solved iteratively for the bias, with values of τ/D as shown, where $D = x_2 - 10$. The region shown by the Ns and below is where iteration fails to converge, while the Ys and above correspond to satisfactory convergence. The Ws indicate combinations of x_2 and τ where iteration converges to the wrong maximum.

x_2	11	12	13	14	15	16	17	18	19
τ/D									
2	Y								
1	N	Y							
0.5		N	Y						
0.2			N	Y					
0.1				N	Y				
0.05						N	Y	Y	
0.02						N	W	W	Y
0.01						N	W	W	Y
0.005					N	W	W	W	Y
0.001					N	W	W	W	W

Table 2. Estimated biases for the different techniques, for measurements 10 ± 1 and 14 ± 1, with uncorrelated statistical errors.

Method	Best value of s	Lower limit	Upper limit
χ^2	2.6	1.7	13.9
Probability	1.8	1.2	9.8
Individual bias	1.8	1.4	2.6
Frohner	See figure 2 ($s = \sqrt{2}b_2$)		

4. Monte Carlo tests

In order to investigate how these procedures behave in more complicated cases, we have performed Monte Carlo simulations of N independent measurements of the same quantity, each subject of random and systematic errors. The true value of the quantity is assumed to be unity, and the separate measurements are smeared by Gaussian distributions of widths specified by their individual random errors σ_i. Finally each of these results is moved by its bias, which we have taken from another Gaussian distribution of width B_{in}. Then we have estimated the bias by using the four methods described above. For each method, we also calculated the best estimate $\hat{x}$.

The generation of the set of N independent results was then repeated a further 49 times, in order to find the average value of the square of the calculated bias and its error for each technique. We have also obtained for each method the average value of the weighted average, and the width of its distribution, in order to see which technique provides the best estimate of the quantity we are trying to measure. In these 50 repetitions, the numerical values of each of the N assumed biases could either be kept at the values they had for the first set; or else they could be recalculated from new random numbers for each of the 50 sets. We have adopted the second alternative.

Finally the whole procedure was performed for a variety of B_{in} values, in order
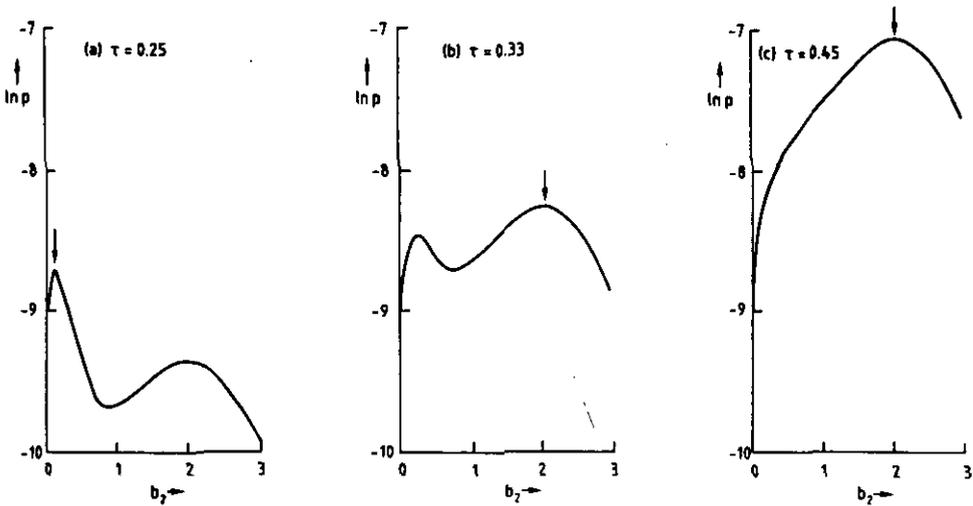
Figure 3. The logarithm of the probability p as a function of the bias b_2, for different values of τ, and for two measurements $x_1 = 10 \pm 1$ and $x_2 = 16 \pm 1$. (The absolute scale of $\ln p$ is arbitrary.) At low τ there are three stationary values. In (*a*) the largest value is the maximum at lower b_2, while in (*b*) it has moved to the higher b_2. At even larger τ, there is a single maximum, which moves towards $b_2 = 3$ as τ increases.

Table 3. Statistical errors used in Monte Carlo calculation.

Measurement no	Number of events	Statistical error
1	1000	0.032
2	2000	0.022
3	300	0.058
4	500	0.045
5	100	0.100
6	1000	0.032
7	20	0.224
8	700	0.038
9	400	0.050
10	50	0.141

to see how the performance of each method depends on the magnitude of the bias.

We now present the results for $N = 10$, and with the magnitudes of the statistical errors as shown in table 3. For the chosen values, the statistical error on the weighted mean is 0.013. In figure 4, the estimates of the output bias for the tested methods are shown. Several features are apparent:

(i) For very large B_{in}, all methods converge to approximately the correct value of the bias (to within a factor of $\sqrt{(N-1)/N}$ for the probability method). The 'individual bias' method has the slowest approach to the correct value.

(ii) For very small B_{in}, all methods as expected overestimate the bias. This is simply because we are unwilling to accept estimates of the bias that are negative.

(iii) For intermediate B_{in}, the 'individual bias' method produces estimates of the bias that are significantly below B_{in}. This seems to be related to the fact that for two measurements, the bias is given by equation (14), rather than the expected form of (1).
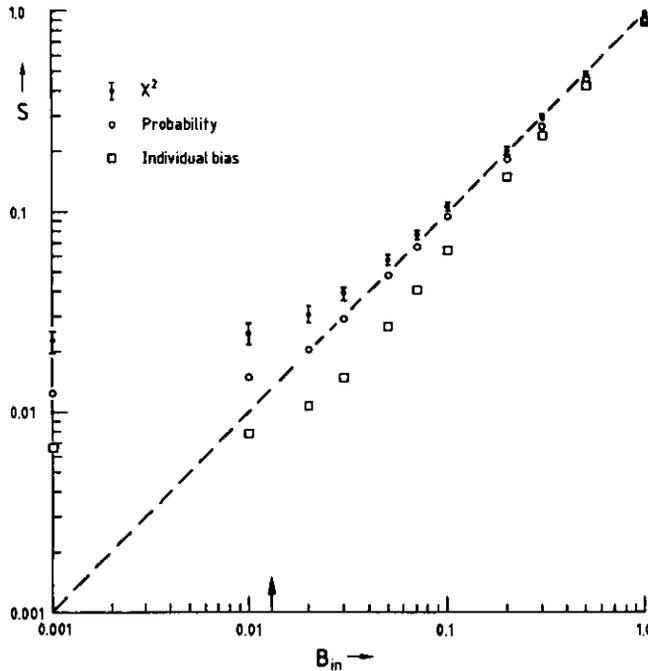
Figure 4. The calculated bias s as functions of the input bias B_{in} for the different techniques. The broken line is the ideal situation $s = B_{in}$. The magnitude of the statistical error on the weighted average is denoted by the arrow. The error bars on s for the χ^2 method correspond to those on the square root of the average s^2; errors on individual determinations of s are some seven times larger. The errors on the other methods are similar. The values of s for the Frohner method are not shown because of the problems of choosing a suitable value of τ. However, with τ fixed at 0.4, the agreement between s and B_{in} is good for large B_{in} ($s = 0.110 \pm .003$ for $B_{in} = 0.1$), but not surprisingly $s > B_{in}$ for small values (e.g. $s = 0.057 \pm .002$ for $B_{in} = 0.01$).

(iv) For intermediate B_{in}, the 'probability' method produces better estimates of the bias than does the χ^2 method. This is in part a reflection of the $\sqrt{(N-1)/N}$ bias of the 'probability' method, which becomes increasingly important at small B_{in}, and which compensates partially for the positive bias discussed in (ii).

(v) With τ fixed at 0.4, the Frohner method behaves well for $B_{in} > 0.1$. Below that, $\tau = 0.4$ makes the bias estimate significantly larger than B_{in}, while smaller values of τ give problems with the convergence of the iterative solution.

At any given value of B_{in}, the widths of the distributions of the best estimates $\hat{x}$ of each of the methods are barely distinguishable, provided that in the Frohner method τ is reduced to be comparable to B_{in}. (The problem of convergence applies only to the bias estimates, and not to that of $\hat{x}$.) In figure 5, they are compared with those for the unweighted mean, and the usual weighted mean (i.e. (4), with $s = 0$). It is seen that, as expected, the best estimates' widths lie either on or below those for the unweighted and the usual weighted mean. They tend to the unweighted mean at large B_{in} (when the individual statistical errors are irrelevant) and to the weighted mean at small B_{in}.

We show in figure 6 how our estimates of the individual biases b_i for the Frohner method compare with the input values, in a couple of typical cases, for large B_{in}. The
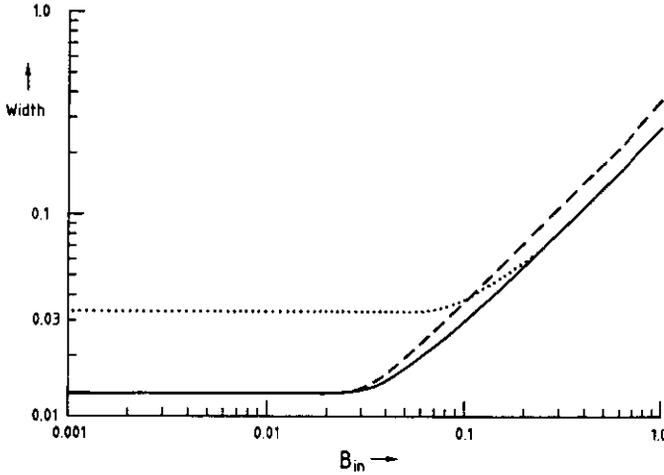
Figure 5. The widths of the distributions of best values $\hat{x}$, as functions of the input bias B_{in}, for the unweighted average (dotted curve), weighted average (broken curve) and the methods of section 2.2. The latter methods give indistinguishable widths, and are shown as the full curve.

fact that the points tend to lie along a line at $\sim 45°$ is as expected. For smaller values of the bias B_{in}, the correlation between the individual output and input biases is not so pronounced as for larger values of B_{in}, because of the relatively larger importance of the statistical errors.
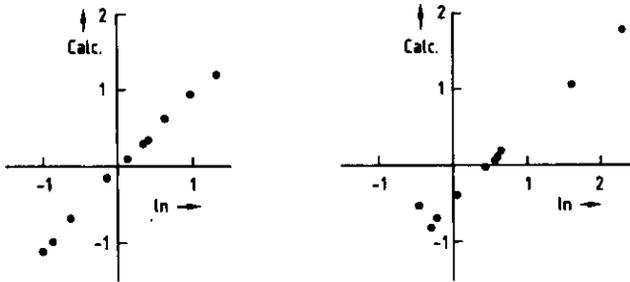

Figure 6. Plots of the ten individual output biases ('Calc') against the actual values of the corresponding individual biases ('In'), for two separate sets of ten measurements. The 'In' values are from a random Gaussian distribution with $B_{in} = 1$; the 'Calc' values are as obtained for the Frohner method with $\tau = 0.4$. The strong correlation arises because the biases are much larger than the statistical errors. Because a common bias cannot be detected by any of our methods, the line along which the points tend to lie will not pass through the origin if the average of the 'In' values happens to be significantly non-zero (see right-hand diagram).

We have repeated the Monte Carlo tests of the four methods for the situation where all ten measurements of a single experiment have equal statistical errors of ± 0.032. The major difference as compared with the case of unequal errors discussed above is that s provides a better estimate of B_{in} down to smaller values. Thus for $B_{in} = 0.01$, which is equal to the statistical error on the mean, $s = 0.016 \pm .002$ for the χ^2 method, and $0.033 \pm .001$ for the Frohner technique with $\tau = 0.4$.

5. Conclusions

We have discussed in some detail the problem of extracting estimates of the systematic error from a repeated set of measurements of the same quantity, where the statistical errors are specified.

Four different methods for extracting the bias are described (the χ^2 method, the probability approach, the individual bias technique and the Frohner method). Their analytic properties in simple situations are derived, and their performance for a more general situation is investigated in detail by a Monte Carlo method. It is found that the 'probability' method described in section 2.2.2 provides a reasonable estimate of the bias for the case of ten independent measurements, provided the bias is larger than the statistical error. For lower values of the bias, the estimates for all methods are larger than the input values of the bias; this arises simply because the estimated biases are forced to be positive or zero.

For smaller numbers of measurements, the bias of the 'probability' method is underestimated by a factor of $\sim \sqrt{(N-1)/N}$, and then the 'χ^2 method' is preferable.

The Frohner method, which is very appealing because of its more rigorous approach, requires a value of τ for each bias. With τ small, there are problems of convergence and of multiple solutions; large τ produces estimates of s which are higher than the input values B_{in}, unless B_{in} is large compared with the statistical errors.

All methods give very similar accuracies for their estimates of the best value of the parameter of interest.

It is hoped that anyone attempting to estimate biases from the spread of a repeated set of measurements will be encouraged to test the reliability of their method by using the Monte Carlo methods described here, but with the parameters adapted to their particular results.

References

[1] Lyons L *Preprint* Oxford OUNP–90–20
[2] Frohner F H 1989 *Maximum Entropy and Bayesian Methods* ed J Skilling (Deventer: Kluwer) p 467