CDF/ANAL/PUBLIC/5776
Version 2.10
August 20, 2002

Everything you always wanted to know about pulls

Luc Demortier[1] and Louis Lyons[2]
[1] *The Rockefeller University,* [2] *University of Oxford*

Abstract

This note explains various ways to define a "pull" or "stretch". It discusses applications of this concept in problems of parameter estimation (constrained and unconstrained fits) and hypothesis testing. Monte Carlo methods are described to characterize pull distributions in situations involving small samples.

1 Introduction

If a random variable x is generated repeatedly with a Gaussian distribution of mean μ and width σ, then it is almost a tautology that the pull

$$g \;=\; \frac{x-\mu}{\sigma} \tag{1}$$

will be distributed as a standard Gaussian with mean zero and unit width. Thanks to the central limit theorem, this simple property can be applied in a wide range of situations from hypothesis testing to parameter estimation, where pulls provide evidence for various forms of bias and allow the verification of error coverage.

Section 2 introduces three definitions of pull in the context of parameter estimation and describes a couple of simple applications. These applications boil down to the comparison of a pull distribution with the expectation of a standard Gaussian. In contrast, in hypothesis testing a single pull is used as a test statistic to decide on the consistency of two measurements. This is described in section 3. Section 4 considers non-asymptotic situations and

how to define pulls in the presence of asymmetric errors. The statement that pull distributions are expected to be standard Gaussian implies a properly constructed ensemble of real or simulated measurements on which pulls are defined. The question of how to construct simulated ensembles is studied in section 5, where we also examine the effect of sample size on pull distributions. Finally, we give some general recommendations on the use of pulls in section 6.

2 Pulls in parameter estimation

Two of the most popular methods of parameter estimation are least-squares and maximum-likelihood. In the former, one minimises a weighted sum of squares

$$S \;=\; \sum_i \left(\frac{y_i^{exp} - y_i^{pred}(\tau)}{\sigma_i} \right)^2 \tag{2}$$

where $y_i^{exp} \pm \sigma_i$ are experimental measurements, and y_i^{pred} are the predicted values, which depend on one or more parameters τ. Then τ_m, the best value of the parameter[1], is determined by minimising S with respect to τ, and its error σ_m is given for example by $1/\sqrt{\frac{1}{2}\frac{d^2S}{d\tau^2}}$.

Alternatively τ could be determined by maximising the likelihood

$$\mathcal{L} \;=\; \prod_i f\Big(y_i^{exp}, y_i^{pred}(\tau), \sigma_i\Big) \tag{3}$$

where f is the probability density for observing y_i^{exp} when the predicted value is $y_i^{pred}(\tau)$.

It is also possible to perform a constrained fit, when other information on the parameter(s) is available. Thus if τ has previously been measured as $\tau_c \pm \sigma_c$, equations (2) and (3) would be modified to

$$S \;=\; \left(\frac{\tau - \tau_c}{\sigma_c} \right)^2 \;+\; \sum_i \left(\frac{y_i^{exp} - y_i^{pred}(\tau)}{\sigma_i} \right)^2 \tag{4}$$

and

$$\mathcal{L} \;=\; \frac{e^{-\frac{1}{2}\left(\frac{\tau-\tau_c}{\sigma_c}\right)^2}}{\sqrt{2\pi}\,\sigma_c} \;\prod_i f\Big(y_i^{exp}, y_i^{pred}(\tau), \sigma_i\Big) \tag{5}$$

[1]Although τ is determined by a fit to the data, we denote its fitted value by τ_m (m for 'measured'), to distinguish it from τ_f (f for 'fitted') when we include some constraint in the fit (see for example equation 4). This is consistent with the way we refer to the **measured** momentum of a track as derived from a fit to the hits along its path, as opposed to the **fitted** momentum, from a kinematic fit incorporating energy and momentum conservation.

where the Gaussian factor gives the probability density for observing τ_c if the true value is τ. It is assumed that the previous and the current measurements are uncorrelated. For large samples (or for a linear model with Gaussian uncertainties), the second factor in equation (5) is Gaussian in τ, and $\tau_f \pm \sigma_f$, the fit result that incorporates the constraint $\tau_c \pm \sigma_c$, is given by:

$$\tau_f = \frac{\tau_m/\sigma_m^2 + \tau_c/\sigma_c^2}{1/\sigma_m^2 + 1/\sigma_c^2} \tag{6}$$

$$\sigma_f = \frac{1}{\sqrt{1/\sigma_m^2 + 1/\sigma_c^2}} \tag{7}$$

2.1 Unconstrained fits

Suppose we obtain a set of measurements of a parameter τ, whose "true" or "generated" value is τ_g. The measurements are statistical fluctuations around τ_g and could, for example, follow an exponential time distribution

$$\frac{1}{\tau_g}\, e^{-t/\tau_g}. \tag{8}$$

If a histogram is produced, there would be Poisson fluctuations on the numbers in each bin. A fit to the data would give a value $\tau_m \pm \sigma_m$. Then, for a large number of events in the distribution, we would expect τ_m to be approximately Gaussian distributed about τ_g, even though the distribution (8) is non-Gaussian. For many repetitions of this procedure, the pull

$$g = \frac{\tau_m - \tau_g}{\sigma_m} \tag{9}$$

should be a standard Gaussian. This is still true when the fit involves additional parameters, as long as the error σ_m has been correctly calculated.

The above definition of pull can be used for checking the properties of a fitting algorithm with large numbers of pseudo-experiments. However, when confronted with real data, the "true" value τ_g is not known and definition (9) is useless. Fortunately there exists an alternative definition of pull for cases where an external constraint is applied.

2.2 Constrained fits

Consider again the example of section 2.1, this time incorporating an extra 'constraint' $\tau = \tau_c \pm \sigma_c$ from some external measurement. In other words, in the S expression we are trying to minimise, there is an extra term $(\tau - \tau_c)^2/\sigma_c^2$. Let the fitted value of τ, taking into account the external constraint, be

$\tau_f \pm \sigma_f$. Then the pull

$$g_c \;=\; \frac{\tau_f - \tau_c}{\sqrt{\sigma_c^2 - \sigma_f^2}} \tag{10}$$

is usually a standard Gaussian. The denominator of the expression for g_c may at first sight look a bit surprising, but it is simply the error on the numerator, taking into account the correlation between the errors in the fit result τ_f and the constraint τ_c.

Equivalently, one can define a pull according to:

$$g_m \;=\; \frac{\tau_m - \tau_f}{\sqrt{\sigma_m^2 - \sigma_f^2}}, \tag{11}$$

where $\tau_m \pm \sigma_m$ is the fit result *without* the extra constraint. For large samples, or for a linear model with Gaussian uncertainties, one can use equations (6) and (7) to show that $g_c = g_m$. It should be noted however, that the large-sample limit is not reached at the same rate by g_c and g_m (see section 5.1.)

The definition of g_m allows one to examine the behaviour of pulls in two limiting cases:

1. If the constraint is totally irrelevant (e.g. it refers to a previous measurement of a variable that is completely unrelated to the present analysis), the fit will not improve the measurement and so

   $$\tau_f \pm \sigma_f \;=\; \tau_m \pm \sigma_m. \tag{12}$$

   Then equation (11) reduces to $g_m = 0/0$, which is not wrong.

2. If in contrast the extra constraint is exact, $\tau_f = \tau_c$ and $\sigma_f = \sigma_c = 0$. In this case, τ_m should have been Gaussian distributed about the constraint with variance σ_m^2. The pull definition gives:

   $$g_m \;=\; \frac{\tau_m - \tau_f}{\sqrt{\sigma_m^2 - 0^2}}, \tag{13}$$

   which is thus again a unit Gaussian. An example of this could be the sum of the measured energies of all the final state particles in a reaction, which should equal the (assumed exactly known) initial state energy.

So far we have stated without proof that pull distributions are expected to be standard Gaussian. In order to study this statement more carefully one needs to specify the ensemble on which pulls are defined. We defer a discussion of this topic to section 5.

2.3 Examples

In this section we give two examples of the use of pulls in constrained fits. The first example (section 2.3.1) illustrates definition (10) of constrained pulls, i.e. g_c, whereas in the second example (section 2.3.2) the nature of the constraint is sometimes such that only definition (11), i.e. g_m, can be used.

2.3.1 Lifetime of CP eigenstates of B_s

In CDF, the decay channel $B_s \to \psi\phi$ can be analysed in terms of two different lifetimes τ_s and τ_ℓ of the CP eigenstates of the B_s, which manifest themselves in the different spin states of the ψ and ϕ, which in turn affect the vector meson decay angular distributions [2].

In the fit of experimental data to these two lifetimes (and to other parameters), it is possible to impose a constraint that their suitably weighted average $\bar{\tau_c}$ is given by the measured B_s lifetime of 1.54 ± 0.07 ps [3]. If we generate a whole series of simulated experiments with values τ_s and τ_ℓ (whose weighted average is 1.54 ps) and perform the constrained fit to extract the average lifetime $\bar{\tau_f} \pm \sigma_f$ and the fractional lifetime difference $\Delta\Gamma/\Gamma$, we would then expect $\bar{\tau_f}$ to be distributed such that its pull

$$g_c \;=\; \frac{\bar{\tau_f} - 1.54}{\sqrt{0.07^2 - \sigma_f^2}} \tag{14}$$

is a unit Gaussian.

2.3.2 Kinematic fitting

This is the situation where we minimise

$$S \;=\; \sum_i \left(\frac{x_{fi} - x_{mi}}{\sigma_{mi}} \right)^2 \tag{15}$$

subject to some constraint(s) (such as energy and momentum conservation for a specific assumed reaction) on the fitted kinematic variables x_{fi} of an event, whose measured values before this fitting procedure are $x_{mi} \pm \sigma_{mi}$. Thus x_i could be the 4-momentum components of the tracks at a given vertex in the event. In reality, the four x_i variables of a track are likely to be correlated with each other, which would require expression (15) to be extended to take their correlations into account.

As a result of the fit, we determine the x_{fi} and their errors σ_{fi} (each σ_{fi} can be calculated as the shift in x_{fi} needed to increase S by 1.0 from its minimum value, when S is re-minimised with respect to the other x_{fj},

$j \neq i$.) Then we expect the pulls

$$g_{mi} \;=\; \frac{x_{fi} - x_{mi}}{\sqrt{\sigma_{mi}^2 - \sigma_{fi}^2}} \tag{16}$$

to be distributed like standard Gaussians. This is just equivalent to equation (11).

3 Pulls in hypothesis testing

The previous section described the use of pulls in parameter estimation, where a pull *distribution* is obtained and compared to a standard Gaussian. We now turn to a situation where a *single* pull is calculated and, assuming its parent distribution to be standard Gaussian, an inference is drawn about the validity of a given hypothesis. A slightly more general treatment of the material presented in this section can be found on pages 277-278 of ref. [1].

Suppose we performed a series of measurements of a quantity τ and wish to test the consistency of the latest measurement, $\tau_\ell \pm \sigma_\ell$, with the average of all measurements, $\tau_a \pm \sigma_a$. We write $\tau_p \pm \sigma_p$ for the average of all measurements *prior* to the latest one, and regard τ_p and τ_ℓ as uncorrelated.

For the combined result we have:

$$\tau_a \;=\; \frac{\tau_p\, w_p + \tau_\ell\, w_\ell}{w_p + w_\ell}, \tag{17}$$

$$\sigma_a \;=\; \frac{1}{\sqrt{w_p + w_\ell}}, \tag{18}$$

where $w_p = 1/\sigma_p^2$ and $w_\ell = 1/\sigma_\ell^2$. The difference between the combined result and the latest one is:

$$\tau_a - \tau_\ell \;=\; \frac{\tau_p\, w_p - \tau_\ell\, w_p}{w_p + w_\ell}, \tag{19}$$

and the error $\sigma_{a\ell}$ on $\tau_a - \tau_\ell$ is given by (remember that τ_ℓ and τ_p are uncorrelated):

$$\sigma_{a\ell}^2 \;=\; \sigma_p^2 \left(\frac{w_p}{w_p + w_\ell}\right)^2 + \sigma_\ell^2 \left(\frac{w_p}{w_p + w_\ell}\right)^2 \tag{20}$$

$$\;=\; (\sigma_p^2 + \sigma_\ell^2) \left(\frac{1/\sigma_p^2}{1/\sigma_p^2 + 1/\sigma_\ell^2}\right)^2 \tag{21}$$

$$\;=\; \frac{\sigma_\ell^4}{\sigma_p^2 + \sigma_\ell^2} \tag{22}$$

Rewriting equation (18) in terms of σ_p and σ_ℓ yields:

$$\sigma_a^2 \;=\; \frac{\sigma_p^2\,\sigma_\ell^2}{\sigma_p^2+\sigma_\ell^2}. \tag{23}$$

Comparing equations (22) and (23), one infers that:

$$\sigma_{a\ell}^2 \;=\; \sigma_\ell^2 - \sigma_a^2. \tag{24}$$

The pull of the latest measurement from the average value is therefore given by

$$g_\ell \;=\; \frac{\tau_\ell - \tau_a}{\sqrt{\sigma_\ell^2 - \sigma_a^2}}. \tag{25}$$

If the latest measurement is consistent with the average, g_ℓ should be distributed as a Gaussian with mean 0 and width 1, and can therefore be used as a test statistic. It is identical to definition (11).

Needless to say, the equivalent definition

$$g_p \;=\; \frac{\tau_a - \tau_p}{\sqrt{\sigma_p^2 - \sigma_a^2}} \tag{26}$$

gives identical numerical values.

4 Non-asymptotic and pathological cases

In most cases we expect the pull distribution to tend to a standard Gaussian only asymptotically. For small numbers of events, the likelihood function is usually skewed, resulting in asymmetric error intervals and pull distributions that are significantly non-Gaussian unless special care is taken in defining the pulls. We discuss the definition of pulls from asymmetric errors in section 4.1. Later, in section 5.3, we will return to this definition with an example that demonstrates the corresponding improvement in Gaussian shape of the pull distribution.

It is also possible to encounter ill-defined problems, where pull distributions will never look Gaussian, regardless of the size of the data sample. We present an example of such a pathology in section 4.2.

4.1 Asymmetric errors

Sometimes a fit returns asymmetric errors for a parameter. This happens for example with the MINOS algorithm in the MINUIT package [4]. In this case

the pull g should be defined as follows:

$$\text{if (fit result)} \leq \text{(true value)}: \quad g = \frac{\text{(true value)} - \text{(fit result)}}{\text{(positive MINOS error)}},$$

$$\text{otherwise}: \quad g = \frac{\text{(fit result)} - \text{(true value)}}{\text{(negative MINOS error)}}. \quad (27)$$

This definition guarantees that the percentage of pulls between -1 and $+1$ equals the coverage of the error interval returned by MINOS, which *should* be 68.27% if 1-σ intervals are requested. This can be seen as follows. Suppose τ_g is the true value of the parameter we are trying to determine, and τ_f is the fit result, with σ_f^+ and σ_f^- the absolute values of the positive and negative errors calculated by MINOS. By definition of these MINOS errors, we have:

$$\alpha = \Pr(\tau_f - \sigma_f^- < \tau_g < \tau_f + \sigma_f^+), \quad (28)$$

where α is (close to) 68.27%. This can be rewritten as:

$$\alpha = \Pr(-\sigma_f^- < \tau_g - \tau_f < +\sigma_f^+). \quad (29)$$

Next, we split the probability on the right-hand side into two non-overlapping cases, $\tau_g - \tau_f < 0$ and $\tau_g - \tau_f \geq 0$:

$$\alpha = \Pr(-\sigma_f^- < \tau_g - \tau_f < 0) + \Pr(0 \leq \tau_g - \tau_f < +\sigma_f^+) \quad (30)$$

Finally, dividing by σ_f^- inside the first probability term and by σ_f^+ inside the second one, we obtain:

$$\alpha = \Pr(-1 < \frac{\tau_g - \tau_f}{\sigma_f^-} < 0) + \Pr(0 \leq \frac{\tau_g - \tau_f}{\sigma_f^+} < +1). \quad (31)$$

The interpretation of this equation is straightforward: when $\tau_g < \tau_f$, divide the difference by σ_f^-, otherwise divide it by σ_f^+, and this guarantees that a fraction α of the time the result will be between -1 and $+1$.

There is of course no guarantee that the pull distribution will be Gaussian. However, if it is, and its width is 1, then the coverage will be correct. It is therefore always useful to plot the pull distribution according to the above definition since it provides a good visual indicator of the accuracy of the error estimates.

4.1.1 Example: exponential distribution

To illustrate some features of asymmetric likelihood functions, we investigate a likelihood fit to a small number N of time values from an exponential

distribution, equation (8), with lifetime parameter $\tau_g = 1$. The likelihood estimate of τ_g is simply $\bar{t}$, the mean of the N time values.

The pull is defined as

$$g = \frac{\bar{t} - \tau_g}{\sigma}. \tag{32}$$

Four different pulls result from four different definitions of the error σ:

- $g(1)$ uses $\sigma = \tau_g/\sqrt{N}$, which is the approximate value of the expected error.
- $g(2)$ uses $\sigma = \bar{t}/\sqrt{N}$, which is the parabolic error returned by the likelihood fit.
- $g(3)$ and $g(4)$ make use of the asymmetric errors on the likelihood fit, defined by the changes in $\bar{t}$ required for the logarithm of the likelihood to decrease by 0.5 from its maximum value. Then $g(3)$ uses the upper error σ_u if $\bar{t} \leq \tau_g$, and the lower error σ_l otherwise. Note that, because of the asymmetry in the likelihood, σ_u will tend to be larger than σ_l.
- $g(4)$ tries out the errors the other way around, i.e. σ_u if $\bar{t} > \tau_g$, and σ_l otherwise.

For samples of size $N = 4$ and $N = 30$, Table 1 shows the pull means and standard deviations[2].

	$N = 4$		$N = 30$	
Pull definition	Mean	Width	Mean	Width
$g(1)$	0.00	1.00	0.00	1.00
$g(2)$	-0.67	1.88	-0.19	1.07
$g(3)$	-0.31	1.43	-0.09	1.03
$g(4)$	-1.06	2.44	-0.29	1.12

Table 1: Means and widths of pull distributions for samples of size 4 and 30, for four definitions of pulls (see text).

The result for $g(1)$ is obvious as the estimate $\bar{t}$ has mean value τ_g and variance τ_g/N. Hence the mean pull is zero and its variance is unity for any value of N. However, for small N the distribution of the pull is non-Gaussian. This is clear for the extreme case of $N = 1$, when the pull distribution is

[2]There is no need to perform Monte Carlo calculations, as the sum x of N independent random variables, each exponentially distributed with lifetime parameter τ_g, is known to have a gamma distribution $\frac{1}{\tau_g^N \Gamma(N)}\, e^{-x/\tau_g}\, x^{N-1}$. Therefore the distribution of $\bar{t}$ is $\frac{N^N}{\tau_g^N \Gamma(N)}\, e^{-N\bar{t}/\tau_g}\, \bar{t}^{\,N-1}$.

e^{-g-1} for pull values above -1, and zero otherwise. It becomes approximately Gaussian for large N, because of the Central Limit Theorem.

It is then clear that $g(2)$ will be biassed negatively. This is because a negative pull, corresponding to a low value of $\bar{t}$, will result in a small estimate of the error used in the denominator of the pull definition. Hence, as compared with $g(1)$, the scale is expanded for negative pulls and contracted for positive ones.

The pulls $g(3)$ and $g(4)$ both use errors which vary with $\bar{t}$, and hence share the tendency of $g(2)$ to have a negative bias. Since $g(4)$ uses a smaller error for calculating negative pulls and a larger error for positive pulls, the extent of the bias is increased. For $g(3)$, the opposite is the case. This tends to confirm the 'obvious' fact that when the data has asymmetric errors, it is appropriate to use the upper error when the data is below the expectation.

Also as expected, the deviations from 0.0 ± 1.0 become smaller for larger N.

4.2 Searching for a non-existent resonance

An interesting example [5] is provided by a smooth mass distribution being fitted by a background shape and a resonance peak of arbitrary position and arbitrary amplitude $A \pm \sigma_A$, which can be positive or negative. Since the mass distribution contains no resonance, the pull is simply A/σ_A. Because of fluctuations however, this turns out to have a bimodal distribution, with peaks more or less symmetrically situated above and below zero. It has a minimum at the origin (where a standard Gaussian pull distribution has its maximum). This arises because the fit of a resonace peak with arbitrary position will pick out the mass region which most deviates from the smooth shape. In order for a fit to return $A = 0$, we thus require there to be no significant deviations across the whole mass distribution; this is very unlikely.

As the number of events in the distribution increases, fluctuations become relatively smaller, and the positions of the bimodal peaks move in towards zero pull. However, the minimum at zero is maintained.

5 Pseudo-experiment ensembles for testing pulls

When generating pseudo-experiments to test the properties of a fitting algorithm that includes constraints, it is necessary to understand which parameters to fluctuate, and how to fluctuate them. For example, an event rate which is subjected to a Gaussian constraint is sometimes fluctuated according to a Poisson distribution whose mean is itself fluctuated around the

Gaussian constraint. This method is wrong, as can easily be seen by considering that the probability for a given event rate to occur in the pseudo-experiment ensemble is different from that predicted by the likelihood model. The correct method is to fluctuate the event rate according to a Poisson distribution with fixed mean, and separately to fluctuate the constraining value according to its Gaussian distribution[3]. Once the question of how to run pseudo-experiments is properly resolved, one can check whether the data sample size is large enough for the pull distribution to be standard Gaussian. In this section we start by examining the effect of sample size on the shape of pull distributions (subsection 5.1). We then calculate the expected widths of pull distributions for a very general pseudo-experiment ensemble that includes the "correct" and "wrong" ensembles described above as special cases (subsection 5.2). This provides a demonstration of the importance of using the proper ensemble to study pulls. In the last subsection we argue that the use of MINOS errors in MINUIT fits yields better-behaved pulls than parabolic errors.

To fix ideas, we will be working with the example first introduced in section 2.1, namely the measurement of a time constant with the following likelihood:

$$\mathcal{L}(\tau) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{\tau-\tau_c}{\sigma_c}\right)^2}}{\sqrt{2\pi}\,\sigma_c} \prod_{i=1}^{N}\left(\frac{1}{\tau}\,e^{-t_i/\tau}\right) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{\tau-\tau_c}{\sigma_c}\right)^2}}{\sqrt{2\pi}\,\sigma_c}\;\frac{e^{-\frac{N\bar{t}}{\tau}}}{\tau^N} \tag{33}$$

In the *absence* of the constraint ($\sigma_c \to \infty$), the maximum likelihood estimate τ_m of τ, and its uncertainty σ_m, are given by:

$$\tau_m \;=\; \bar{t} \;\equiv\; \frac{1}{N}\sum_{i=1}^{N} t_i, \tag{34}$$

$$\sigma_m \;=\; \sigma_{\bar{t}} \;=\; \frac{\bar{t}}{\sqrt{N}}. \tag{35}$$

When the constraint is enforced as in section 2.2, the fitted value τ_f is no longer simply equal to $\bar{t}$, although it remains a unique function of $\bar{t}$ and the constraining value τ_c.

5.1 Effect of sample size on pull distributions

We ran sets of pseudo-experiments to study the distributions of the various types of pull defined in this note, and their dependence on the number of measurements N. Each pseudo-experiment was generated as follows:

[3]We can see that this procedure is reasonable for the example of section 3. To test that procedure by Monte Carlo, we would vary both τ_a and τ_ℓ in Gaussian fashion according to their errors. This corresponds in this case to fluctuating the constraint and the Poisson data sample.

1. Generate N random t_i values according to an exponential distribution with fixed time constant τ_g;

2. Generate a constraint τ_c according to a Gaussian with mean $\bar{\tau}_c$ and width σ_c;

3. Fit the t_i to an exponential distribution whose time constant is the fit parameter and is constrained to $\tau_c \pm \sigma_c$.

Unless one is interested in studying the bias introduced by constraining to the wrong time constant, one will usually set $\bar{\tau}_c \equiv \tau_g$.

We generated three sets of pseudo-experiments with $\tau_g = \bar{\tau}_c = 5$ and with $N = 10$, 100 and 1000 respectively. In each case we set the uncertainty σ_c on the constraint to be equal to the expected uncertainty on the corresponding unconstrained result, i.e. $\tau_g/\sqrt{N}$.

The results for $N = 100$ are shown in Figures 1 and 2. Figures 1(a), (b) and (c) show the distributions of the generated constraint τ_c, the fit result without constraint τ_m, and the fit result with constraint τ_f. Because of the large number of measurements per pseudo-experiment, the distribution of τ_m is reasonably Gaussian. So is the distribution of τ_f which, as expected, is narrower than both the distributions of τ_c and τ_m. Plots 1(d), (e) and (f) show distributions of the pulls defined by equations (27), (10) and (11), respectively. The g and g_c pull distributions are Gaussian, but g_m is clearly not. In order to understand this, we plot distributions of the numerators and denominators of the pulls in Figure 2. The numerators all appear to be Gaussian, including the numerator of g_m. In fact, judging by the χ^2/ndf values, the numerator of g_m is even more Gaussian-like than the τ_m distribution, indicating that some cancellation of non-Gaussian effects takes place in the difference $\tau_m - \tau_f$. As expected, the means of the denominator distributions agree with the RMS widths of the corresponding numerator distributions. If one were to divide the pull numerators by these RMS widths, the resulting pull distributions would be perfectly normal (i.e. Gaussian with mean 0 and width 1.) When dividing by the proper denominators however, fluctuations in the latter distort the pull distributions. A measure of the magnitude of these fluctuations is provided by the RMS/mean ratios of the denominator distributions. These are equal to 4%, 5% and 21% for g, g_c and g_m, respectively. The large fluctuations in the denominator of g_m are clearly responsible for the non-Gaussian tail in the corresponding pull distribution.

Figures 3 and 4 show the same plots as Figures 1 and 2 for a set of pseudo-experiments with $N = 10$, i.e. in a regime where the asymptotic limit is no longer a good approximation, as can be seen in the distribution of τ_m (Figure 3(b).) Not only g_m, but now also the g_c pull distribution is beginning to develop a strong non-Gaussian tail.

Finally, Figures 5 and 6 show what happens when N is increased to 1000. Now even the g_m pull is beginning to look quite Gaussian.

We conclude from these studies that different definitions of pulls have different rates of convergence towards the asymptotic limit. Among the three definitions we have considered, g converges the fastest, and g_m the slowest.

5.2 Effect of pseudo-experiment ensembles on pull distributions

To study the behaviour of pulls in various ensembles of pseudo-experiments, we start from a very general ensemble, in which each pseudo-experiment is defined as follows:

1. Generate a random time constant $\tau_\circ$ according to a Gaussian with mean τ_g and width $\sigma_{\tau_\circ}$;
2. Generate N random t_i values according to an exponential distribution with time constant $\tau_\circ$;
3. Generate a constraint τ_c according to a Gaussian with mean τ_g and width σ_{τ_c};
4. Fit the t_i to an exponential distribution whose time constant is the fit parameter and is constrained to $\tau_c \pm \sigma_c$.

This general ensemble depends on five parameters: N, τ_g, $\sigma_{\tau_\circ}$, σ_{τ_c}, and σ_c, and requires the generation of $N+2$ independent random numbers per pseudo-experiment: $\tau_\circ$, τ_c and $t_1 \ldots t_N$. What we called "correct method" in the introduction to section 5 corresponds to $\sigma_{\tau_\circ} = 0$ and $\sigma_{\tau_c} = \sigma_c$, whereas what we called "wrong method" corresponds to $\sigma_{\tau_c} = 0$ and $\sigma_{\tau_\circ} = \sigma_c$.

In the following subsections we calculate analytically the widths of the g and g_c pull distributions in the asymptotic limit, and illustrate the results with Monte Carlo calculations.

5.2.1 Standard deviation of g pulls

The g pull is defined by:

$$g \;=\; \frac{\tau_f - \tau_g}{\sigma_f}. \tag{36}$$

In the asymptotic limit, the fit result $\tau_f \pm \sigma_f$ is given by equations (6) and (7), where $\sigma_m = \tau_\circ/\sqrt{N}$. Since σ_m depends on the random variable $\tau_\circ$ it is itself a random variable, with standard deviation $\sigma_{\tau_\circ}/\sqrt{N}$. For large N we can neglect the fluctuations of σ_m compared to those of $\tau_\circ$, and hence

to those of the numerator of (36). Accordingly we will write $\sigma_m \cong \tau_g/\sqrt{N}$. Thus we have:

$$\tau_f = \frac{N\bar{t}/\tau_g^2 + \tau_c/\sigma_c^2}{N/\tau_g^2 + 1/\sigma_c^2} \tag{37}$$

$$\sigma_f = \frac{1}{\sqrt{N/\tau_g^2 + 1/\sigma_c^2}} \tag{38}$$

We will use these equations to calculate the standard deviation $\sigma_g = \sigma_{\tau_f}/\sigma_f$ of the g pulls, where σ_{τ_f} is the standard deviation of τ_f. Note that in principle σ_{τ_f} could be different from σ_f, because the former depends on how pseudo-experiments are fluctuated, whereas the latter is the result of a fit, and the fitter knows nothing about where the data came from. We have in fact:

$$\sigma_{\tau_f}^2 \equiv E\Big[(\tau_f - \tau_g)^2\Big] \tag{39}$$

$$= E\left[\left(\frac{N(\bar{t} - \tau_g)/\tau_g^2 + (\tau_c - \tau_g)/\sigma_c^2}{N/\tau_g^2 + 1/\sigma_c^2}\right)^2\right] \tag{40}$$

$$= \frac{\frac{N^2}{\tau_g^4} E\Big[(\bar{t} - \tau_g)^2\Big] + \frac{1}{\sigma_c^4} E\Big[(\tau_c - \tau_g)^2\Big] + \frac{2N}{(\tau_g\sigma_c)^2} E[(\bar{t} - \tau_g)(\tau_c - \tau_g)]}{\Big(N/\tau_g^2 + 1/\sigma_c^2\Big)^2} \tag{41}$$

The expectation values depend on the pseudo-experiment ensemble; in this case they are:

$$E\left[(\bar{t} - \tau_g)^2\right] = \sigma_{\tau_\circ}^2 + \frac{\tau_g^2}{N} \tag{42}$$

$$E\left[(\tau_c - \tau_g)^2\right] = \sigma_{\tau_c}^2 \tag{43}$$

$$E\left[(\bar{t} - \tau_g)(\tau_c - \tau_g)\right] = 0 \tag{44}$$

Plugging these expectations back into the expression for $\sigma_{\tau_f}^2$ and taking the square root yields:

$$\sigma_{\tau_f} = \frac{\sqrt{\frac{N}{\tau_g^2}\left(1 + \frac{N}{\tau_g^2}\sigma_{\tau_\circ}^2\right) + \left(\frac{\sigma_{\tau_c}}{\sigma_c^2}\right)^2}}{\frac{N}{\tau_g^2} + \frac{1}{\sigma_c^2}} \tag{45}$$

Dividing by σ_f, we obtain finally:

$$\sigma_g = \sqrt{\frac{\frac{N}{\tau_g^2}\left(1 + \frac{N}{\tau_g^2}\sigma_{\tau_\circ}^2\right) + \left(\frac{\sigma_{\tau_c}}{\sigma_c^2}\right)^2}{\frac{N}{\tau_g^2} + \frac{1}{\sigma_c^2}}} \tag{46}$$

We consider two special cases:

1. $\sigma_{\tau_\circ} = 0$ and $\sigma_{\tau_c} = \sigma_c$.
   This corresponds to the correct way of running pseudo-experiments. In this case, equation (46) gives $\sigma_g = 1$. The distribution of the g-pull will be standard Gaussian.

2. $\sigma_{\tau_c} = 0$ and $\sigma_{\tau_\circ} = \sigma_c$.
   This corresponds to the wrong way of running pseudo-experiments. Equation (46) reduces to $\sigma_g = \sigma_c \sqrt{N}/\tau_g$. The g-pull distribution will not be standard Gaussian, except when $\sigma_c = \tau_g/\sqrt{N}$, i.e. when the uncertainty on the constraint matches the expected uncertainty on the unconstrained result.

5.2.2 Standard deviation of g_c pulls

The g_c pull is defined in equation (10). To calculate σ_{g_c} we will again use the approximation $\sigma_m \cong \tau_g/\sqrt{N}$. The standard deviation of the numerator of the g_c pull, $(\tau_f - \tau_c)$, can be calculated in the same way as σ_{τ_f} in the previous section. We find:

$$\sigma_{(\tau_f - \tau_c)} \;=\; \frac{\frac{N}{\tau_g^2}\sqrt{\frac{\tau_g^2}{N} + \sigma_{\tau_\circ}^2 + \sigma_{\tau_c}^2}}{\frac{N}{\tau_g^2} + \frac{1}{\sigma_c^2}}. \tag{47}$$

On the other hand, the denominator of the g_c pull can be rewritten as:

$$\sqrt{\sigma_c^2 - \sigma_f^2} \;=\; \frac{\frac{\sqrt{N}}{\tau_g}\,\sigma_c}{\sqrt{\frac{N}{\tau_g^2} + \frac{1}{\sigma_c^2}}}, \tag{48}$$

so that:

$$\sigma_{g_c} \;=\; \sqrt{\frac{\frac{\tau_g^2}{N} + \sigma_{\tau_\circ}^2 + \sigma_{\tau_c}^2}{\frac{\tau_g^2}{N} + \sigma_c^2}}. \tag{49}$$

It is easy to see that $\sigma_{g_c} = 1$ in either of the two special cases considered earlier, namely $\sigma_{\tau_\circ} = 0$ and $\sigma_{\tau_c} = \sigma_c$, or $\sigma_{\tau_c} = 0$ and $\sigma_{\tau_\circ} = \sigma_c$. In other words, the g_c pull has a standard Gaussian distribution for both the "correct" and "wrong" ways of running pseudo-experiments. The same conclusion applies to the g_m pull since g_m and g_c are asymptotically equal (section 2.2).

5.2.3 Comparison with Monte Carlo calculations

We illustrate the above results in Figure 7, where we plot the g, g_c and g_m pull distributions for "correct" and "wrong" ensembles of pseudo-experiments with $N = 1000$, $\tau_g = 5$ and $\sigma_c = 0.03162$. As expected, all distributions are standard Gaussian except that of the g pull for the wrong ensemble. Plugging

$N = 1000$, $\tau_g = 5$, $\sigma_{\tau_\circ} = \sigma_c = 0.03162$ and $\sigma_{\tau_c} = 0$ in equations (45) and (38) yields $\sigma_{\tau_f} = 0.0062$ and $\sigma_f = 0.031$, so that $\sigma_{\tau_f}/\sigma_f = 0.2$, in agreement with the width of the distribution in plot (d).

5.3 Pull distributions for MINOS errors

Figure 8 shows distributions of the MINOS error, the parabolic error, and various pulls for an ensemble of "correct" pseudo-experiments with $N = 10$, $\tau_g = 5$ and $\sigma_c = 0.1581$. For this example the magnitudes of the positive and negative MINOS errors differ by about 15% on average. Judging by the χ^2/ndf values, the distribution of the MINOS pull from definition (27) is clearly more Gaussian-like than the g pull using the parabolic error. However, if the MINOS error assignment in equation (27) is reversed, the resulting pull distribution displays a strong non-Gaussian tail. That the assignment of equation (27) is indeed correct can be seen more directly by plotting a combined histogram of the positive and negative errors (plots (c) and (d)).

We conclude that in non-asymptotic situations pulls calculated from MINOS errors are "better behaved" than pulls calculated from parabolic errors, and that equation (27) uses the correct assignment of MINOS errors.

6 General recommendations for the use of pulls in parameter estimation problems

Whenever one is doing a fit, pull distributions should be plotted to check that the fit is giving sensible results. In situations that involve many separate fits (e.g. track fitting for a whole series of events), each fit provides its own pull(s), and the distribution can easily be obtained. If, however, the experiment involves the estimation of just one set of parameters, the pull distribution can be looked at only for a simulated set of repetitions of the experiment. Such pseudo-experiments should always be designed so that the probability of a given pseudo-data sample in the pseudo-experiment ensemble is equal to the probability predicted by the likelihood (or chisquare) model for this sample.

In the majority of cases, one expects the pull distribution to be a standard Gaussian. One thus needs to confirm that it is centered at zero, has unit width, and has no long tails. If this is not the case, one may need to look at the measurement setup, the experimenter's assumptions, etc. We give two simple examples:

1. Suppose we measure the three angles of a triangle as θ_i^m. Improved values θ_i^f can be obtained by imposing the condition that the angles add up to 180°. The pull would be sensitive to effects such as the errors

being incorrectly assigned, the triangle not being closed, the geometry not being flat (e.g. the triangle is drawn on a sphere), etc.

2. In the kinematic fitting example of Section 2.3.2, pulls can be examined to look for effects such as biased momentum measurements, misalignment of the detector, oddities of the kinematic fitting procedure, contamination from other reactions, etc.

It may happen that the pull distribution is approximately Gaussian, but its width is not 1. Assuming that this is understood to be an effect of the non-asymptotic nature of the problem and not a programming error (this can always be tested by running pseudo-experiments closer to the asymptotic limit!), one may want to correct the quoted uncertainties by multiplying them by the width of the pull distribution.

In other cases the non-asymptotic nature of the problem manifests itself by the appearance of tails in the pull distribution. One must then be careful with the interpretation of the uncertainties. If the percentage of pulls between -1 and $+1$ is 68.27%, then "1-σ" errors have the usual meaning. However, since the pull distribution is not Gaussian, "2-σ" errors no longer have a coverage of 95.45%, etc.

Finally, as illustrated in section 5, one should keep in mind that different pull definitions have different rates of convergence towards the asymptotic limit. Thus it may be that the choice of pull definition itself is the cause of non-Gaussian distortions in the pull distribution.

References

[1] W. T. Eadie, D. Drijard, F. James, M. Roos, and B. Sadoulet, "Statistical Methods in Experimental Physics", North Holland (1971).

[2] F. Azfar, private communication; see also F. Azfar, L. Lyons, M. Martin, C. Paus, and J. Tseng, "Prospects for measuring $\left(\frac{\Delta\Gamma}{\Gamma}\right)_{B_0^s}$ using $B_0^s \to J/\psi\phi$, with $J/\psi \to \mu^+\mu^-$, $\phi \to K^+K^-$, in Run-II, an update", CDF/ANAL/BOTTOM/CDFR/5351 (25 June 2000).

[3] D. E. Groom *et al.*, "Review of Particle Physics", Eur. Phys. J. C **15**, 1 (2000).

[4] F. James and M. Roos, "MINUIT, Function Minimization and Error Analysis," CERN D506 (Long Writeup). Available from the CERN Program Library Office, CERN-IT Division, CERN, CH-1211, Geneva 21, Switzerland.

[5] T. Dorigo and M. Schmitt, "On the significance of the Dimuon Mass Bump and the Greedy Bump Bias", CDF/DOC/TOP/CDFR/5239 (26 February 2000).

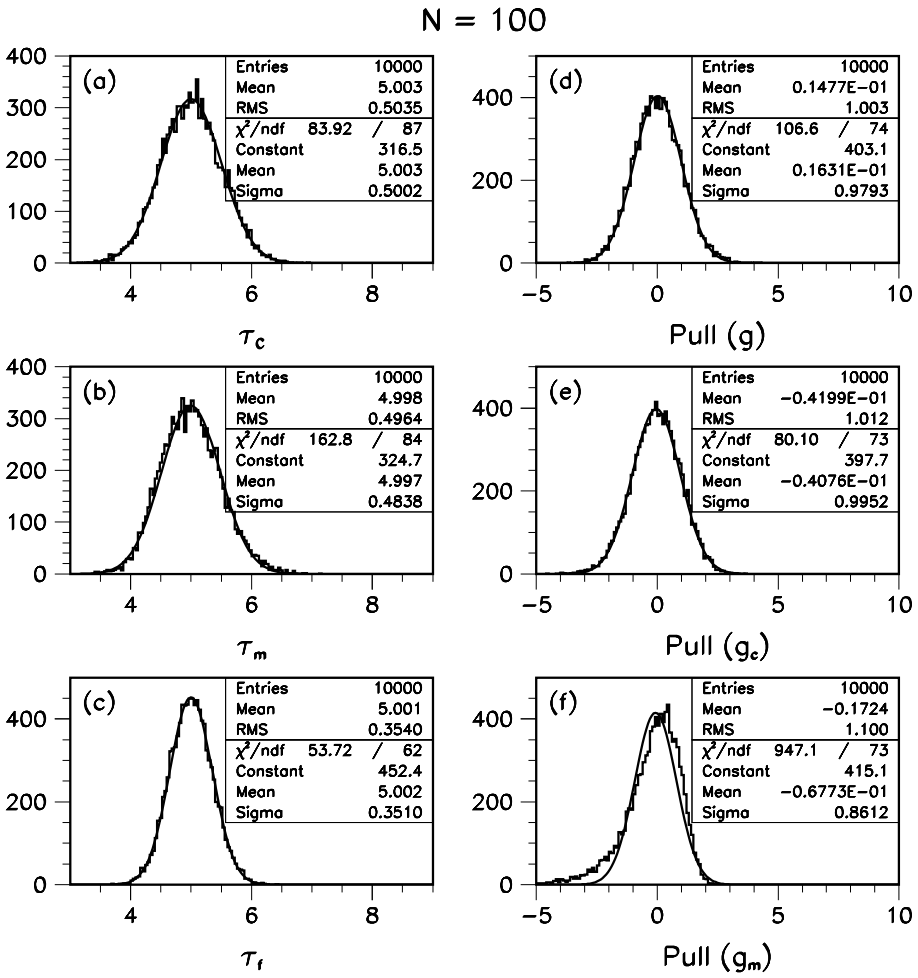

Figure 1: Results of a pseudo-experiment run with $\tau_g = \bar{\tau}_c = 5$, $\sigma_c = 0.5$ and $N = 100$ (see text). Plots (a), (b) and (c) show distributions of the constraint τ_c, the unconstrained fit result τ_m, and the constrained fit result τ_f, respectively. Plots (d), (e) and (f) show pull distributions according to definitions (27), (10) and (11), respectively.

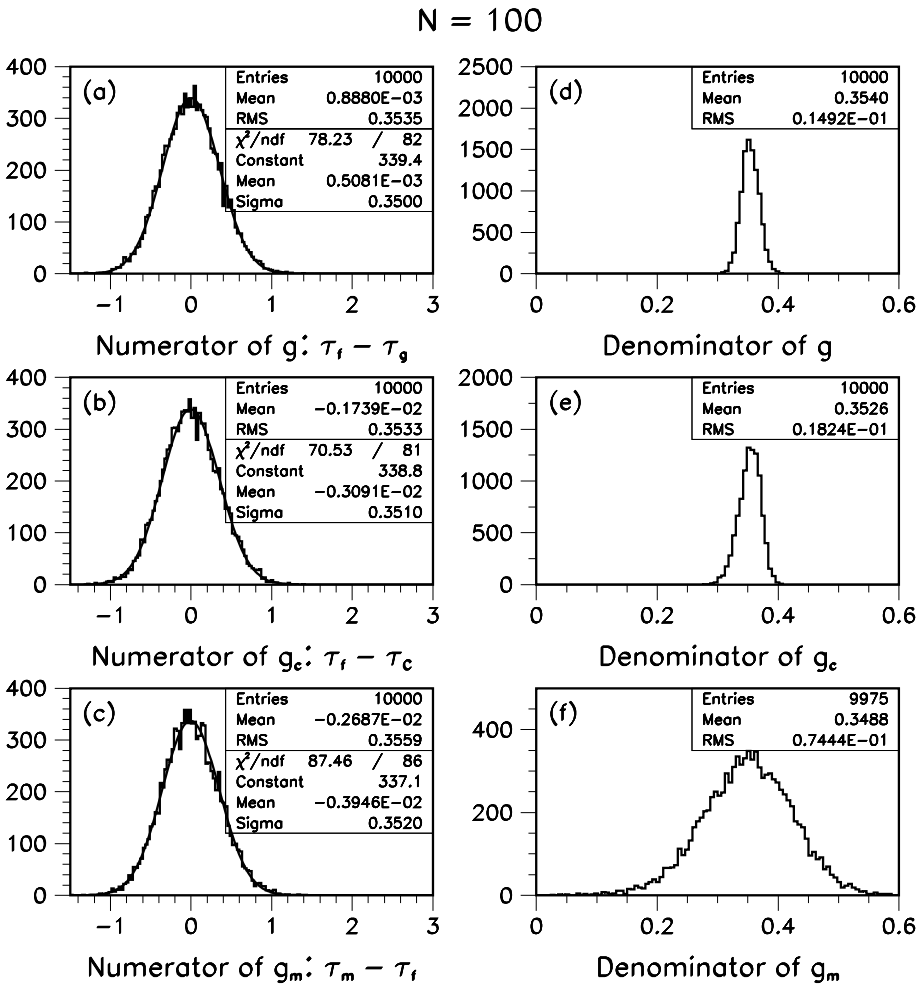

Figure 2: Distributions of the numerators and denominators of the pulls g, g_c and g_m shown in Figure 1.

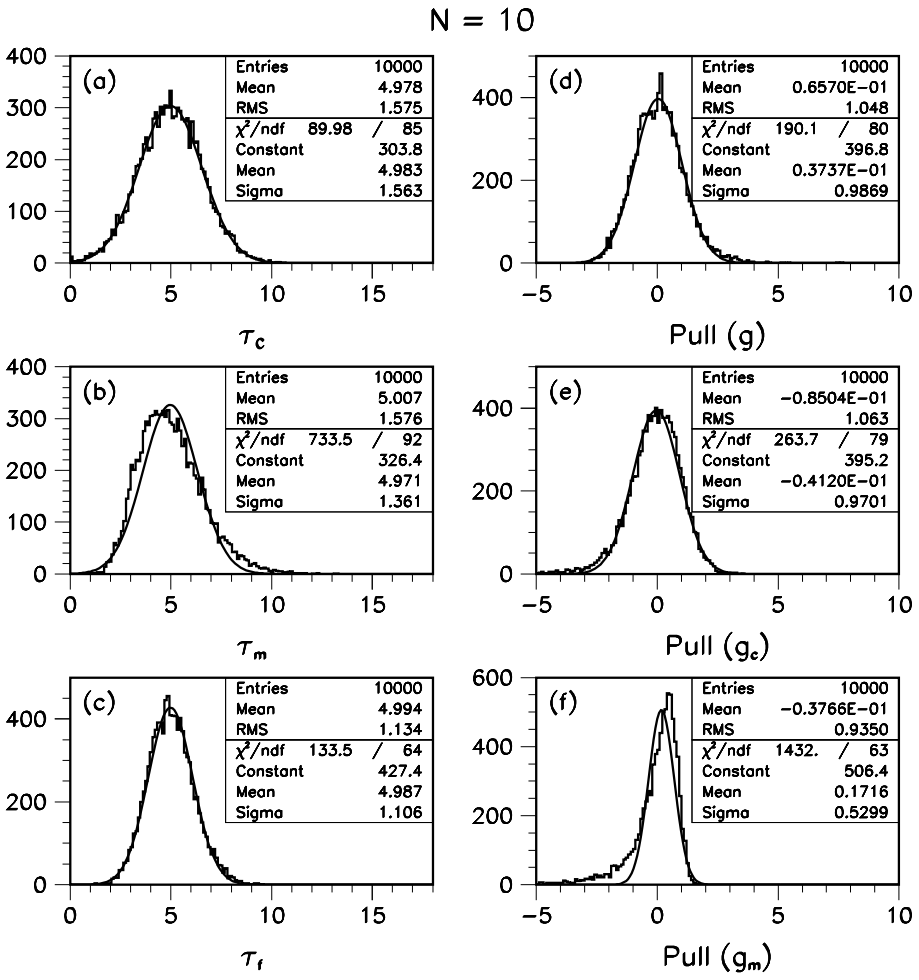

Figure 3: Results of a pseudo-experiment run with $\tau_g = \bar{\tau}_c = 5$, $\sigma_c = 1.581$ and $N = 10$ (see text). Plots (a), (b) and (c) show distributions of the constraint τ_c, the unconstrained fit result τ_m, and the constrained fit result τ_f, respectively. Plots (d), (e) and (f) show pull distributions according to definitions (27), (10) and (11), respectively.

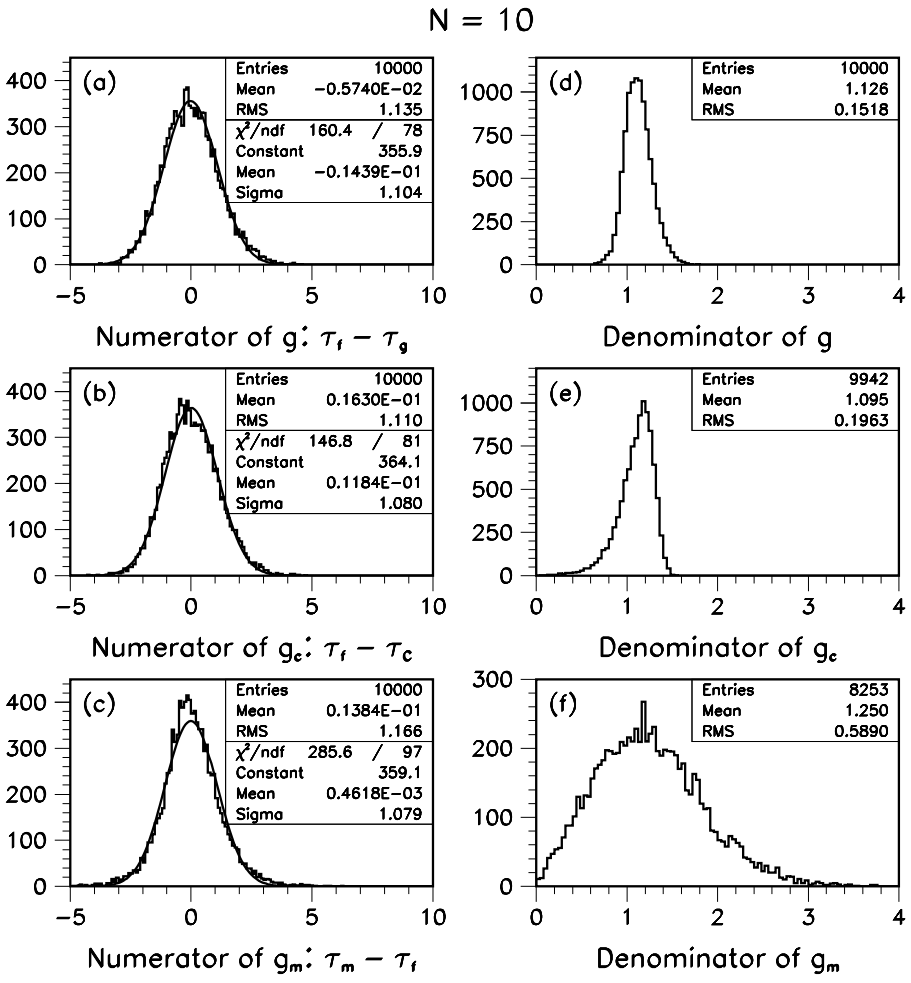

Figure 4: Distributions of the numerators and denominators of the pulls g, g_c and g_m shown in Figure 3.

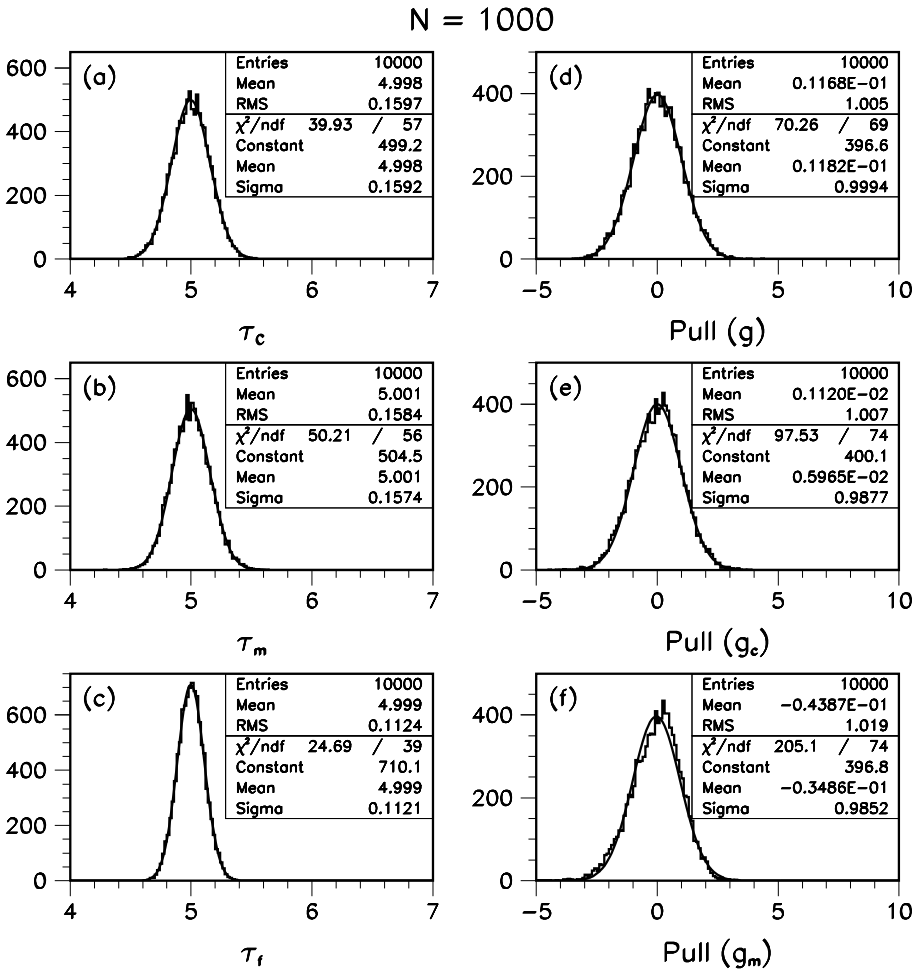

Figure 5: Results of a pseudo-experiment run with $\tau_g = \bar{\tau}_c = 5$, $\sigma_c = 0.1581$ and $N = 1000$ (see text). Plots (a), (b) and (c) show distributions of the constraint τ_c, the unconstrained fit result τ_m, and the constrained fit result τ_f, respectively. Plots (d), (e) and (f) show pull distributions according to definitions (27), (10) and (11), respectively.

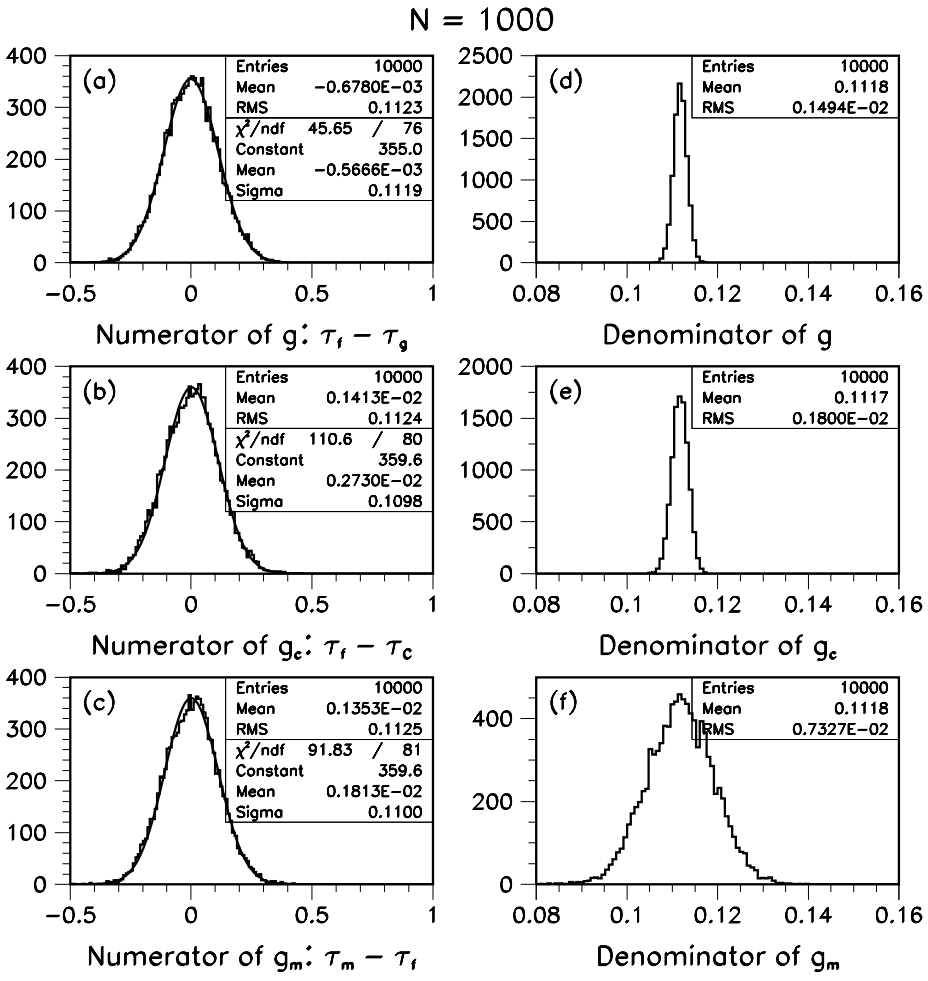

Figure 6: Distributions of the numerators and denominators of the pulls g, g_c and g_m shown in Figure 5.

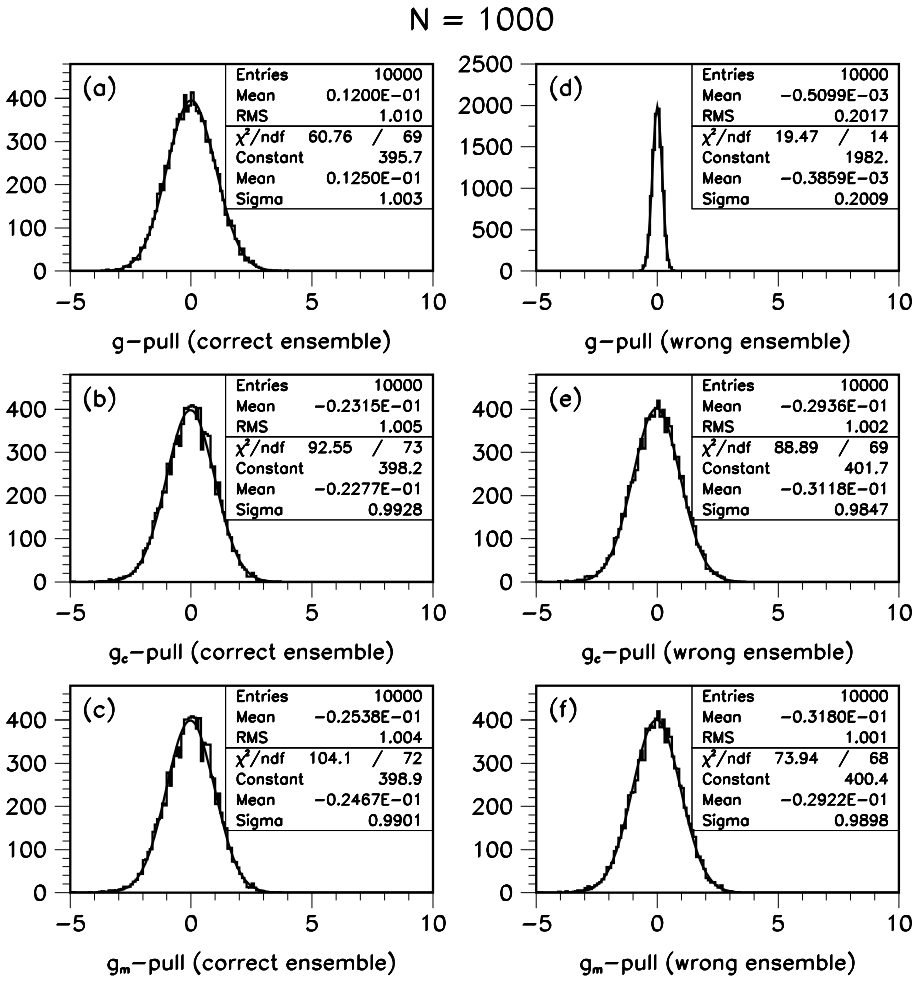

Figure 7: Pull distributions for pseudo-experiments with $N = 1000$, $\tau_g = 5$ and $\sigma_c = 0.03162$. Plots (a), (b) and (c) show the result of using the correct ensemble of pseudo-experiments ($\sigma_{\tau_\circ} = 0$, $\sigma_{\tau_c} = \sigma_c$), whereas plots (d), (e) and (f) show the result of using a wrong ensemble ($\sigma_{\tau_c} = 0$, $\sigma_{\tau_\circ} = \sigma_c$). See text for further details.

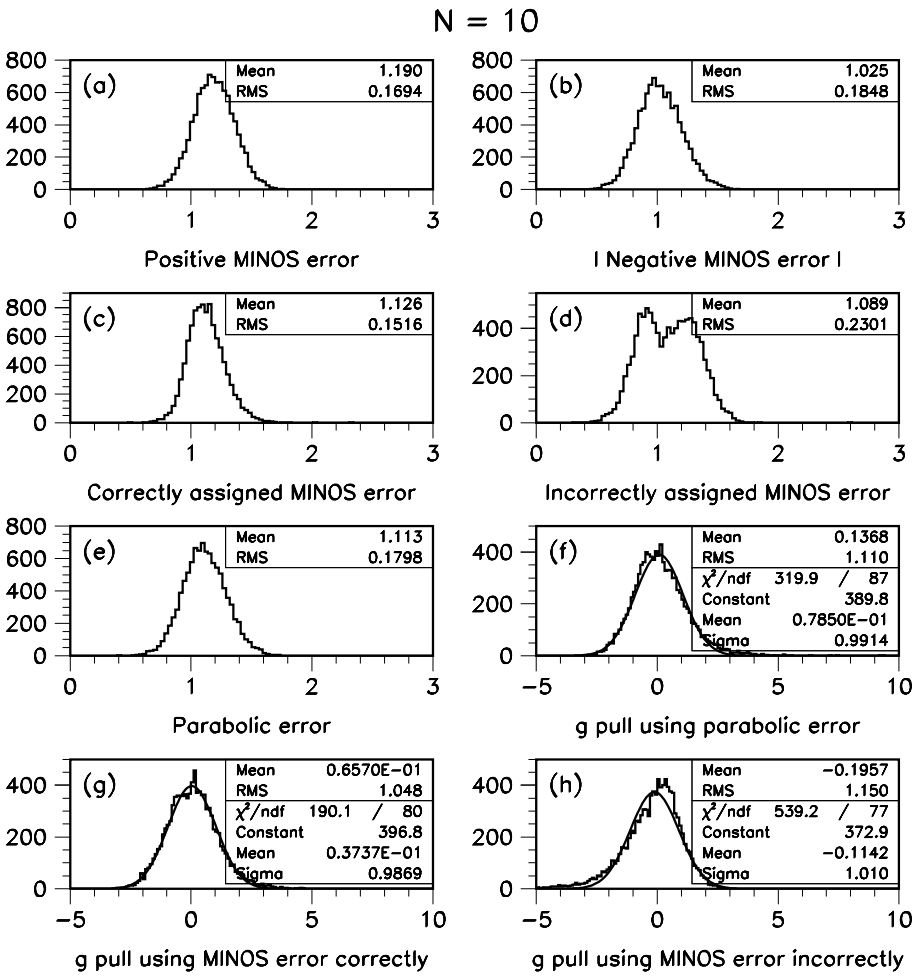

Figure 8: Result of a pseudo-experiment run with $N = 10$, $\tau_g = 5$ and $\sigma_c = 0.1581$. Each pseudo-experiment was generated according to the algorithm described in section 5.1. Plot (c) is a histogram of the positive MINOS error for pseudo-experiments where the fit result τ_f is smaller than the "true" value τ_g, and of minus the negative MINOS error for the remaining pseudo-experiments. Plot (d) shows the opposite MINOS error assignment. Similarly, plot (g) shows the g pull according to equation (27) and plot (h) the g pull with the opposite MINOS error assignment.