Data fitting

Volker Blobel – University of Hamburg
March 2005

1. χ^2 minimization
2. Fits in case of of systematic errors

χ^2 minimisation

Confusion in terminology: A popular method for parameter estimation is χ^2 minimisation $\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta}$ – is this identical to least squares?
The minimum value of the objective function in Least Squares follows often (not always) a χ^2 distribution.

In contrast to the well-defined standard methods

- in χ^2 minimisation a variety of different non-standard concepts is used,
- often apparently motivated by serious problems to handle the experimental data in a consistent way;
- especially for the error estimation there are non-standard concepts and methods.

From publications:

> To determine these parameters one must minimize a χ^2 which compares the measured values . . . to the calculated ones . . .

> Our analysis is based on an effective global chi-squared function that measures the quality of the fit between theory and experiment . . .

Two examples are given, which demonstrate that χ^2 minimisation can give **biased results**:

- Calorimeter calibration
- Averaging data with common normalisation error

Calorimeters for energy measurements in a particle detector require a calibration, usually based on test beam data (measured cell energies y_{ik}) with known energy E. A common method [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**] based on the χ^2 minimisation of

$$\chi^2 = \frac{1}{N}\sum_{k=1}^{N}\left(a_1 y_{1,k} + a_2 y_{2,k} + \ldots + a_n y_{n,k} - E\right)^2$$

for the determination of the a_j can produce biased results, as pointed out by D. Lincoln et al. [**?**].

If there would be one cell only, one would have data y_k with standard deviation σ, with a mean value of $\overline{y} = \sum_k y_k/N$, and the intended result is simply $a = E/\overline{y}$.

A one-cell version of the above χ^2 definition is

$$\chi^2 = \frac{1}{N}\sum_{k=1}^{N}\left(a \cdot y_k - E\right)^2$$

and minimizing this χ^2 has the biased result

$$a = \frac{E \cdot \overline{y}}{\left(\sum_k y_k^2\right)/N} = \frac{E \cdot \overline{y}}{\overline{y}^2 + \sigma^2} \quad \neq E/\overline{y}$$

The bias mimics a non-linear response of the calorimeter.

A known bias in fitted parameters is easily corrected for.

Example: for a hadronic calorimeter one may have

$$\text{Energy resolution } \frac{\sigma}{E} = \frac{0.7}{\sqrt{E}} \qquad \text{which have a result biased by a ratio} = \frac{E}{E + 0.7^2}$$

(at $E = 10$ GeV the resolution is 22 % and the bias is 5 %).

There would be no bias, if the inverse constant a_{inv} would have been determined from

$$\chi^2 = \frac{1}{N} \sum_{k=1}^{N} (y_k - a_{\text{inv}} E)^2$$

General principle: In a χ^2 expression the measured values y_k should not be modified; instead the expectation has to take into account all known effects.

There are N data x_k with different standard deviations σ_k and a **common relative normalisation error** of ε. Apparently the mean value $\overline{y}$ can not be affected by the normalisation error, but its standard deviation is.
One method is to use the full covariance matrix for the correlated data, e.g. in the case $N = 2$:

$$\boldsymbol{V}_a = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} + \varepsilon^2 \cdot \begin{pmatrix} y_1^2 & y_1 y_2 \\ y_1 y_2 & y_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2 y_1^2 & \varepsilon^2 y_1 y_2 \\ \varepsilon^2 y_1 y_2 & \sigma_2^2 + \varepsilon^2 y_2^2 \end{pmatrix}$$

and minimising

$$\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta} \qquad \text{with} \quad \boldsymbol{\Delta} = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \end{pmatrix}$$

Example (from [**?**]): Data are
$y_1 = 8.0 \pm 2\%$ and $y_2 = 8.5 \pm 2\%$, with a common (relative) normalisation error of $\varepsilon = 10\%$.
The mean value resulting from χ^2 minimisation is:

$$7.87 \pm 0.81 \qquad \text{i.e. } < y_1 \text{ and } < y_2$$

- this is apparently wrong.

> . . . that including normalisation errors in the correlation matrix will produce a fit which is biased towards smaller values . . . [**?**]

> . . . the effect is a direct consequence of the hypothesis to estimate the empirical covariance matrix, namely the linearisation on which the usual error propagation relies. [**?**, **?**]

The contribution to $\boldsymbol{V}$ from the normalisation error was calculated from the measured values, which were different; the result is a covariance ellipse with axis different from 45° and this produces a biased mean value.

Distinguish *measured* and *fitted* values.

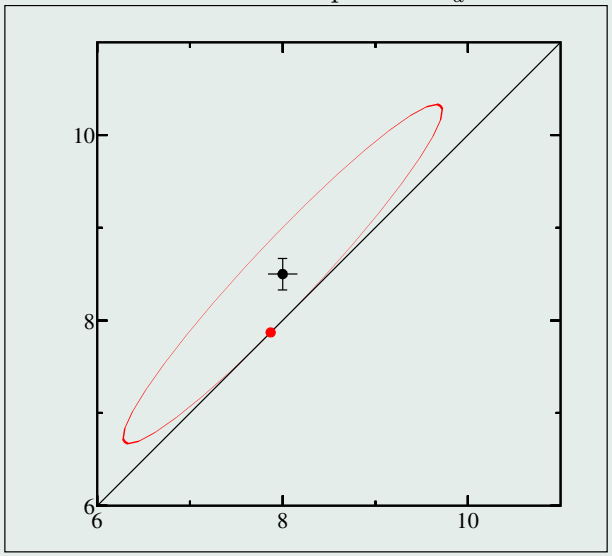
The correct model is: y_1 and y_2 have the same true value, then the normalisation errors $\varepsilon \cdot$ value are identical, with

$$\boldsymbol{V}_b = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} + \varepsilon^2 \cdot \begin{pmatrix} \overline{y}^2 & \overline{y}^2 \\ \overline{y}^2 & \overline{y}^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2 \overline{y}^2 & \varepsilon^2 \overline{y}^2 \\ \varepsilon^2 \overline{y} & \sigma_2^2 + \varepsilon^2 \overline{y}^2 \end{pmatrix}$$

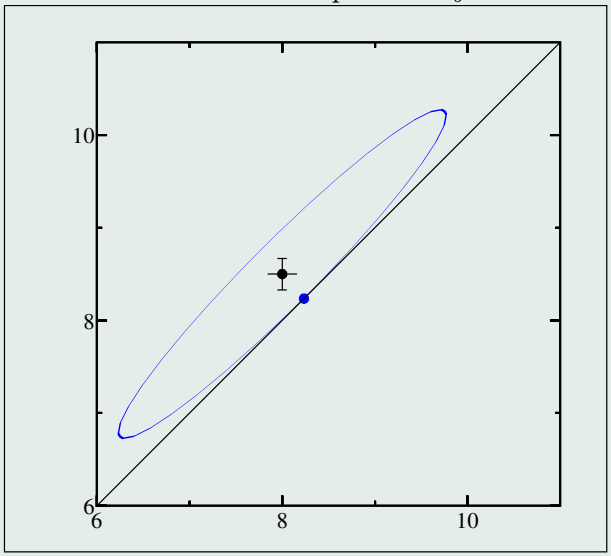
i.e. the covariance matrix depends on the resulting parameter.

Ellipses

Covariance ellipse for $\boldsymbol{V}_a$



Covariance ellipse for $\boldsymbol{V}_b$



Axis of ellipse is tilted w.r.t. the diagonal and ellipse touches the diagonal at a biased point.

Axis of the ellipse is $\approx 45^\circ$ and ellipse touches the diagonal at the correct point.

The result may depend critically on certain details of the model implementation.

The method with one additional parameter ...

Another method often used is to define

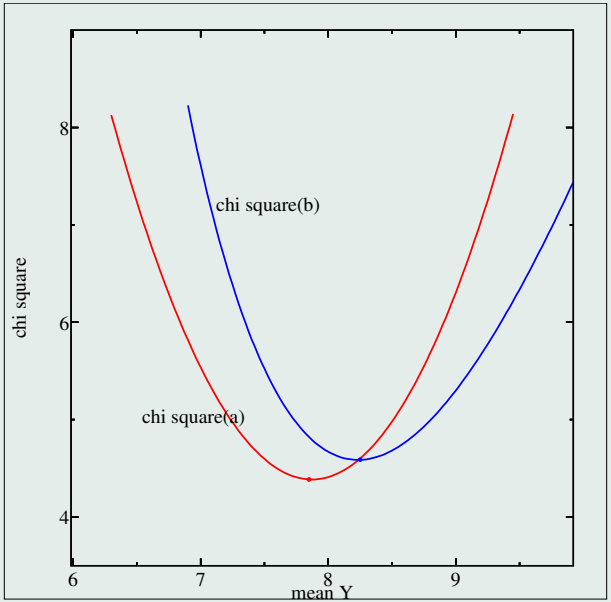
$$\chi_a^2 = \sum_k \frac{(f \cdot y_k - \overline{y})^2}{\sigma_k^2} + \frac{(f-1)^2}{\varepsilon^2} \,,$$

which includes a normalization factor f and which will also produce a biased result.

The χ^2 definition for this problem

$$\chi_b^2 = \sum_k \frac{(y_k - f \cdot \overline{y})^2}{\sigma_k^2} + \frac{(f-1)^2}{\varepsilon^2}$$

will give the correct result (data unchanged and fitted value according to the model), as seen by blue curve.



Standard methods

Standard statistical methods for parameter determination are

- Method of Least Squares $S(\boldsymbol{a})$ • χ^2 minimisation is equivalent: $\chi^2 \equiv S(\boldsymbol{a})$
- Maximum Likelihood method $F(\boldsymbol{a})$
  ... improves the parameter estimation if the detailed probability density is known.

Least squares and Maximum Likelihood can be combined, e.g

$$F_{\text{total}}(\boldsymbol{a}) = \frac{1}{2} S(\boldsymbol{a}) + F_{\text{special}}(\boldsymbol{a})$$

Doubts about justification of χ^2 minimisation from publications:

> The justification for using least squares lies in the assumption that the measurement errors are Gaussian distributed. [**?**]

> However it is doubtful that Gaussian errors are realistic.

> A bad χ^2 ... Finally the data may very well not be Gaussian distributed.

The standard linear least squares method

The model of Linear Least Squares: $\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{a}$

$\boldsymbol{y} =$ measured data $\quad \boldsymbol{A} =$ matrix (fixed) $\quad \boldsymbol{a} =$ parameters $\quad \boldsymbol{V}_y =$ covariance matrix of $\boldsymbol{y}$

Least Squares Principle: minimize the expression $\quad (\boldsymbol{W} = \boldsymbol{V}_y^{-1})$

$$S(\boldsymbol{a}) = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a})^T\ \boldsymbol{W}\ (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a}) \qquad \text{or} \quad F(\boldsymbol{a}) = \frac{1}{2} S(\boldsymbol{a})$$

with respect to $\boldsymbol{a}$.

Derivatives of expression $F(\boldsymbol{a})$:

$$\begin{aligned} \boldsymbol{g} = \frac{\partial F}{\partial \boldsymbol{a}} &= -\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} + (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})\,\boldsymbol{a} \\ \boldsymbol{H} = \frac{\partial^2 F}{\partial a_j a_k} &= (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}) \qquad = \text{constant} \end{aligned}$$

Solution (from $\partial F/\partial \boldsymbol{a} = 0$) is linear transformation of the data vector $\boldsymbol{y}$:

$$\hat{\boldsymbol{a}} = \left[(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\ \boldsymbol{A}^T\boldsymbol{W} \right] \boldsymbol{y} \qquad = \boldsymbol{B}\,\boldsymbol{y}$$

Covariance matrix of $\boldsymbol{a}$ by "error" propagation

$$\boldsymbol{V}[\hat{\boldsymbol{a}}] = \boldsymbol{B}\,\boldsymbol{V}[\boldsymbol{y}]\,\boldsymbol{B}^T = (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1} \qquad = \text{inverse of } \boldsymbol{H}$$

Properties of the solution

Starting from **Principles** properties of the solution are derived, which are valid under certain conditions:

- Data are unbiased: $E[\boldsymbol{y}] = \boldsymbol{A}\,\bar{\boldsymbol{a}}$ ($\bar{\boldsymbol{a}}$ = true parameter vector)
- Covariance matrix $\boldsymbol{V}_y$ of the data is known (and correct).

Distribution-free properties of least squares estimates in linear problems are:

- Estimated parameters are unbiased:

$$E[\hat{\boldsymbol{a}}] = \left(\boldsymbol{A}^T\boldsymbol{W}\,\boldsymbol{A}\right)^{-1}\;\boldsymbol{A}^T\boldsymbol{W}\,E[\boldsymbol{y}] = \bar{\boldsymbol{a}}$$

- In the class of unbiased estimates, which are linear in the data, the Least Squares estimates $\hat{\boldsymbol{a}}$ have the smallest variance (Gauß-Markoff theorem).
- The expectation of the sum of squares of the residuals is $\hat{S} = (n - p)$.

Special case of Gaussian distributed measurement errors:

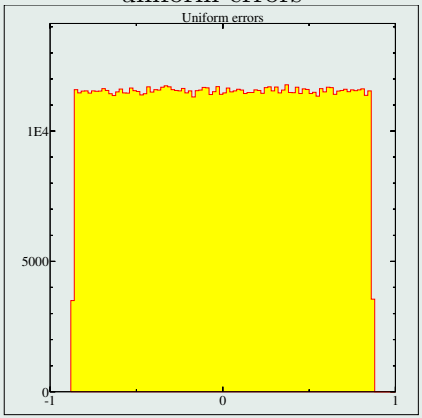
$$\hat{S}/\sigma^2 \quad \text{distributed according to the} \quad \chi^2_{n-p} \quad \text{distribution}$$

to be used for goodness-of-fit test. Properties are not valid, if conditions violated.

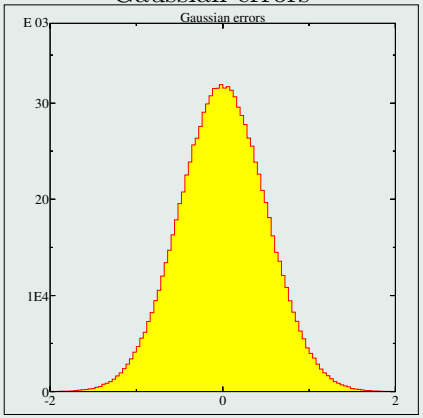
Test of non-Gaussian data

MC test of least squares fit of 20 data points to straight line (two parameters), generated with data errors from different distributions, but always mean = 0 and same standard deviation $\sigma = 0.5$.

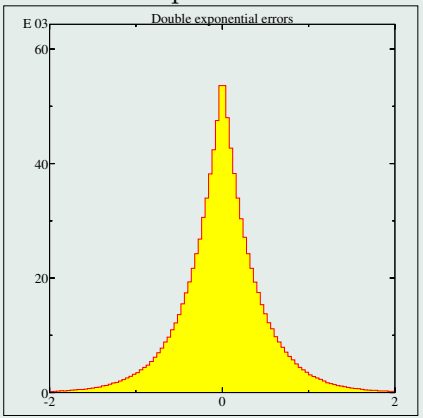
uniform errors



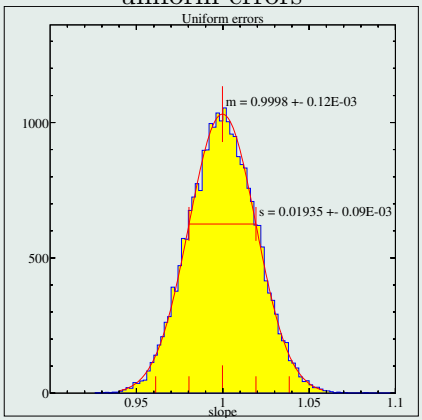
Gaussian errors



double exponential errors

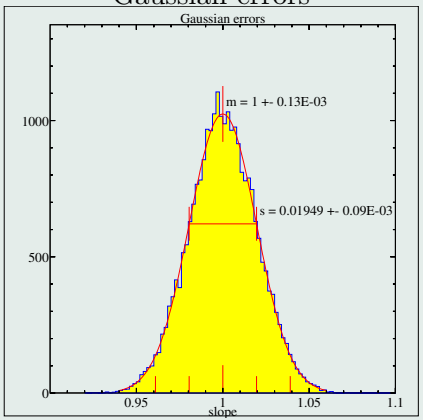


uniform errors



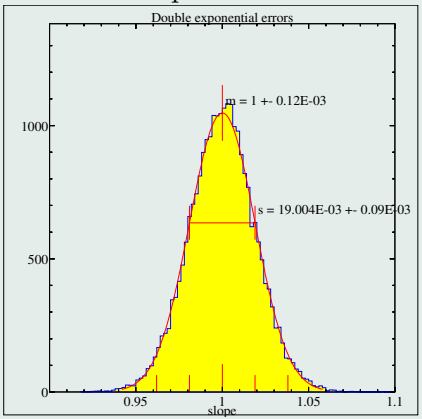
$\sigma = 0.0194$

Gaussian errors



$\sigma = 0.0195$

double exponential errors

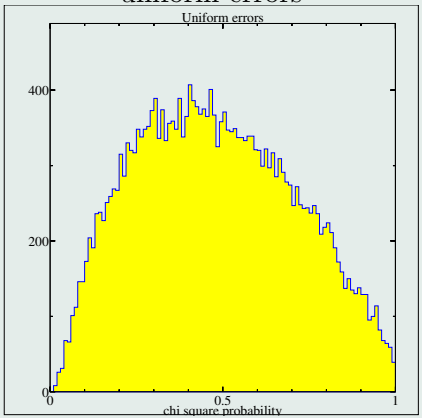


$\sigma = 0.0190$

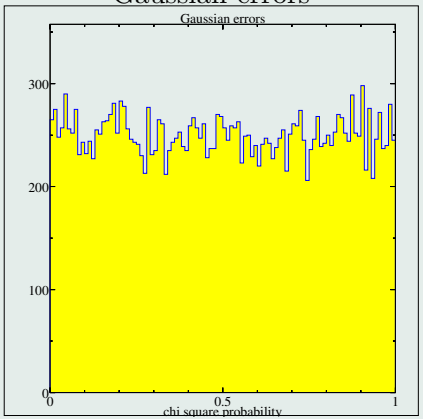
- All parameter distributions are Gaussian, and of the width, expected from the standard error calculation.
- This is valid for both fitted parameters.

- Mean χ^2-values are all equal to $n_{\rm df} = 20 - 2 = 18$, as expected, but
- χ^2-probabilities have different distributions, as expected.

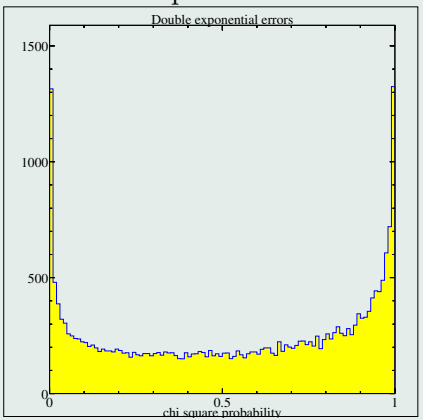
uniform errors



Gaussian errors



double exponential errors



Conclusion: Least squares works fine and as expected, also for non-Gaussian data,
if . . . and only if

- data are unbiased and covariance matrix is complete and correct.

> Everyone believes in the normal law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact. [Poincaré]

Outliers – single unusual large or small values among a sample – are dangerous and will usually introduce a bias in the result.

Modifications of the standard least squares procedure with

- recognition and
- special treatment of outliers

may be useful to reduce the unwanted bias in fitted parameters.

Recipe for robust least square fit

Assume estimate for the standard error of y_i (or of r_i) to be s_i.
Do least square fit on observations y_i, yielding fitted values $\hat{y}_i$, and residuals $r_i = y_i - \hat{y}_i$.

- "Clean" the data by pulling outliers towards their fitted values: winsorize the observations y_i and replace them by pseudo-observations y_i^*:

$$\begin{aligned} y_i^* &= y_i\,, && \text{if} \quad |r_i| \le c\,s_i\,, \\ &= \hat{y}_i - c\,s_i\,, && \text{if} \quad r_i < -c\,s_i\,, \\ &= \hat{y}_i + c\,s_i\,, && \text{if} \quad r_i > +c\,s_i\,. \end{aligned}$$

  The factor c regulates the amount of robustness, a goid choice is $c = 1.5$.

- Refit iteratively: the pseudo-observations y_i^* are used to calculate new parameters and new fitted values $\hat{y}_i$.

Assume a Gaussian density function with 3 parameters N, μ and σ

$$f(x) = N \cdot \Delta x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

is fitted to a histogram (bin size Δx) using the Poisson maximum likelihood method. All three parameters are (almost) uncorrelated. The result for N will be the true value with an error of $\sqrt{N}$ because of the Poisson model (and error propagation).

If however the density is expressed by

$$f(x) = N \cdot \Delta x \quad \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

then N is (negatively) correlated with σ and the relative error of N is enlarged due to the correlation. After a proper full matrix error propagation $\boldsymbol{AVA}^T$ of course the previous error expression is obtained.

An example from parton density fits: the gluon parametrization is

$$xg(x, Q_0^2) = \quad \ldots \quad - A_- \, (1-x)^{\eta_-} \, x^{-\delta_-}$$

where $A_- \sim 0.2$, $\delta_- \sim 0.3$ and η_- fixed at ~ 10. A change of δ_- changes both shape *and* normalisation.

> . . . very small changes in the value of δ_- can be compensated almost exactly by a change in A_- and (to a lesser extent) in the other gluon parameters . . . [**?**]

> . . . we notice that a certain amount of redundancy in parameters leads to potentially disastrous departures . . . For example, in the negative term in the gluon parameterization very small changes in the value of δ_- can be compensated almost exactly by a change in A_- and in the other gluon parameters . . . [**?**]

> We found our input parameterization was sufficiently flexible to accomodate data, and indeed there is a certain redundancy evident. [**?**]

In that case the Hessian will be (almost) singular, inversion is impossible and the convergence of the fit is doubtful.

Fits in case of systematic errors

Data errors: Statistical and systematic uncertainties can only be correctly taken into account in a fit, if there is a clear **model** describing all aspects of the uncertainties.

Statistical data errors: described either

- by ("uncorrelated") errors – standard deviation σ_i for data point y_i (origin is usually counts – Poisson distribution),
- by a covariance matrix $\boldsymbol{V}_y$.

Two alternative models for **systematic errors**:

- **multiplicative effects** – normalisation errors
- **additive effects** – offset errors

that had to be accounted for in *different* ways in a fit.

Data y_i in Particle Physics are often (positive) cross sections, obtained from counts and several factors (Luminosity, detector acceptance, efficiency).

In general there is a normalisation error, given by a relative error ε. If data from > 1 experiment are combined, the normalisation error ε has to be taken into account.

Method: Introduce one additional factor α, which has been measured to be $\alpha = 1 \pm \varepsilon$, modify expectation according to

$$f_i = \alpha \cdot f(x_i, \boldsymbol{a})$$

and make fit with

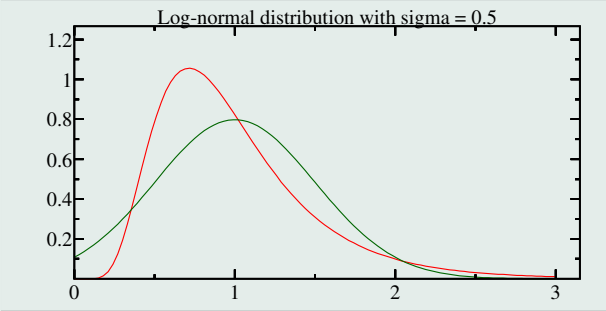
$$S(\boldsymbol{a}) = \sum_i \frac{(y_i - \alpha \cdot f(x_i, \boldsymbol{a}))^2}{\sigma_i^2} + \Delta S^{\text{norm}} \qquad \text{with} \quad \Delta S^{\text{norm}} = \frac{(\alpha - 1)^2}{\varepsilon^2}$$

One factor α_k has to be introduced for each experiment, if data from more than one experiment are fitted.

The normalisation factor determined in an experiment is more the product than the sum of random variables. According to the *multiplicative* central limit theorem the product of positive random variables follows the log-normal distribution, i.e. the logarithm of the normalisation factor follows the normal distribution.

For a log-normal distribution of a random variable α with $E[\alpha] = 1$ and standard deviation of ε the contribution to $S(\boldsymbol{a}, \alpha)$ is

$$\begin{aligned} \Delta S^{\text{norm}} &= \ln\alpha \left(3 + \frac{\ln\alpha}{\ln\left(1+\varepsilon^2\right)}\right) \\ &\to \frac{(\alpha-1)^2}{\varepsilon^2} \quad \text{for small } \varepsilon \end{aligned}$$



The normal and the log-normal distribution, both with mean 1 and standard deviation $\varepsilon = 0.5$.

Example: error of calorimeter constant – a change of the constant will change *all* data values y_i – events are moved between bins.

Determine **shifts** s_i of data values y_i, for a one-standard deviation change of the calorimeter constant – the shifts s_i will carry a relative sign.

1. Method: Modify covariance matrix to include contribution(s) due to systematic errors

$$\boldsymbol{V}_a = \boldsymbol{V}_{\text{stat}} + \boldsymbol{V}_{\text{syst}} \qquad \text{with} \quad \boldsymbol{V}_{\text{syst}} = \boldsymbol{s}\boldsymbol{s}^T \quad \text{(rank=1 matrix)}$$

e.g. $V_{ij}^{\text{stat}} = s_i s_j$, and use modified matrix in fit with $S(\boldsymbol{a}) = \boldsymbol{\Delta}^T \boldsymbol{V}_a^{-1} \boldsymbol{\Delta}$

- Requires inversion (once) of the $n \times n$ matrix of the data.
- Otherwise no change of formalism necessary.
- Used e.g. by LEP Electroweak Heavy Flavour WG.[**?**]

2. Method: Introduce one additional parameter β, which has been measured to be 0 ± 1, for each systematic error source, modify expectation according to

$$f_i = f(x_i, \boldsymbol{a}) + \beta \cdot s_i$$

and make fit with

$$S(\boldsymbol{a}) = \sum_i \frac{(y_i - (f(x_i, \boldsymbol{a}) + \beta s_i))^2}{\sigma_i^2} \; + \; \beta^2$$

Advantage of additional parameter β:

- Allows to test the pull $= \hat{\beta}/\sqrt{1 - V_{\beta\beta}}$ due to the systematic error (should follow a $N(0, 1)$ distribution).
- Allows to test the effect of the fit model on the systematic effect from the global correlation coefficient $\rho_\beta^{\text{global}}$.
- Allows more insight into systematic effect by inspection of the correlation coefficients ρ_{β,a_j} between β and the other parameters.
- First derivative of expectation (for fits) is trivial: $\partial f_i / \partial \beta = s_i$.

Data fitting