Statistical and numerical methods of data analysis III

V. Blobel
Universität Hamburg

Desy, **June** 28, (+14, 21), **2001**

- The maximum likelihood method
- Confidence intervals and limits

The maximum likelihood method

- The maximum likelihood principle
- Properties of maximum-likelihood estimates
- Application of the maximum likelihood method
- Binned distributions
- Function minimization

A data analysis problem

A standard data analysis problem:

A measurement is performed in the space of the random variable x. The distribution of the measured values x is assumed to be known to follow the (normalized) *probability density* $p(x;a)$

$$p(x;a) \geq 0 \qquad \text{with} \qquad \int_\Omega p(x;a)\,\mathrm{d}x = 1$$

in the x-space, which depends on a single parameter a.

From a given set of n measured values $x_1, \ldots, x_i, \ldots, x_n$ the optimal value of the parameter a has to be estimated.

The maximum likelihood method

The *maximum-likelihood method* starts from the *joint* probability distribution of the n measured values $x_1, \ldots, x_i, \ldots, x_n$, which for *independent* measurements is given by the product of the individual densities $p(x|a)$, which is

$$\mathcal{L}(a) = p(x_1|a) \cdot p(x_2|a) \cdots p(x_n|a) = \prod_{i=1}^{n} p(x_i|a) \; .$$

The function $\mathcal{L}(a)$, for a given set $\{x_i\}$ of measurements considered as a function of the parameter a, is called the *likelihood function* and is a *function* (not a probability density) of the parameter a.

Principle of Maximum Likelihood

The estimate $\widehat{a}$ for the parameters a is the value, which *maximizes* the likelihood function $\mathcal{L}(x|a)$.

For technical and also for theoretical reasons it is easier to work with the logarithm (a monotonically increasing function of its argument) of the likelihood function $\mathcal{L}(\boldsymbol{a})$, or with the *negative* logarithm. In the following the *negative* log-likelihood function[a] is considered,

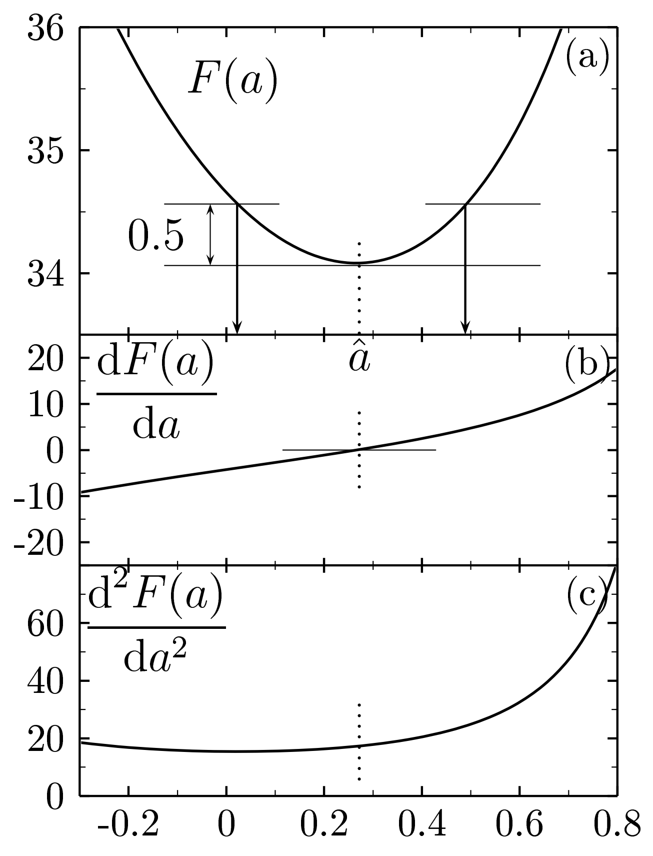
$$\boxed{F(a) = -\sum_{i=1}^{n} \ln p(x_i|a)}$$

and the maximum likelihood estimate $\widehat{a}$ is the value that *minimizes* this function.

$$\text{Likelihood equation} \qquad \frac{\mathrm{d}F(a)}{\mathrm{d}a} = 0$$

[a] Sometimes a factor of 2 is included in the definition of the negative log-likehood function; this factor makes it similar to the χ^2-expression of the method of least squares in certain applications.

Example



Properties of the maximum-likelihood estimates

Maximum-likelihood estimates $\widehat{a}$

Consistency: The estimate $\widehat{a}$ of the MLM is asymptotically ($n \to \infty$) consistent. For finite values of n there may be a bias $B(\widehat{a}) \propto 1/n$.

Normality: The estimate $\widehat{a}$ is, under very general conditions, asymptotical normally distributed with minimal variance $V(\widehat{a})$.

Invariance: The maximum likelihood solution is invariant under change of parameter – the estimate $\widehat{b}$ of a function $b = b(a)$ is given by $\widehat{b} = b(\widehat{a})$. The bias $B(\widehat{a})$ for finite n may be different for different functions of the parameter.

Efficiency: If efficient estimators exist for a given problem the maximum likelihood method will find them.

Normality

Normality: The estimate $\widehat{a}$ is, under very general conditions, asymptotical normally distributed with minimal variance $V(\widehat{a})$, i.e.

$$\lim_{n\to\infty} V\left[\widehat{a}\right] = \frac{1}{n}\left\{ E\left[\frac{\partial \ln p}{\partial a}\right]^2 \right\}^{-1} .$$

Asymptically the likelihood equation becomes a function, which is *linear* in the parameter a (constant second derivative).

Calculation of variance and covariance matrix in practice:

$$V\left[\widehat{a}\right] = \left(\left.\frac{\mathrm{d}^2 F}{\mathrm{d}a^2}\right|_{a=\widehat{a}} \right)^{-1} \qquad \boldsymbol{V}\left[\widehat{\boldsymbol{a}}\right] = \boldsymbol{H} \quad \text{with} \quad H_{jk} = \frac{\partial^2 F}{\partial a_j \partial a_k}$$

Invariance

The Likelihood function is a function of $\boldsymbol{a}$, and is not a probability density of $\boldsymbol{a}$.

Invariance: The maximum likelihood solution is invariant under change of parameter – the estimate $\widehat{b}$ of a function $b = b(a)$ is given by $\widehat{b} = b(\widehat{a})$. The bias $B(\widehat{a})$ for finite n may be different for different functions of the parameter.

Example: Parameter of an exponential distribution

Exponential probability density distribution for the decay of unstable particles

$$p(t;\tau) = \frac{1}{\tau} e^{-t/\tau} \qquad \text{or} \qquad p(t;\lambda) = \lambda e^{-\lambda\, t}$$

dependent on the mean lifetime τ or the decay constant $\lambda = 1/\tau$.

Bayesian parameter estimation

Both the data $\boldsymbol{x}$ and the parameters $\boldsymbol{a}$ are random variables.

- Before the experiment the knowledge about $\boldsymbol{a}$ is summarized by $\pi(\boldsymbol{a})$ – called prior density, for example $\pi(\boldsymbol{a}) =$ const.. $\pi(\boldsymbol{a}) = 0$ outside the physical region.
- Bayes theorem is used to update the prior using the data $\boldsymbol{x}$

$$q(\boldsymbol{a}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{a} = \frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{a})\,\pi(\boldsymbol{a})\,\mathrm{d}\boldsymbol{a}}{\int \mathcal{L}(\boldsymbol{x}|\boldsymbol{a}')\,\pi(\boldsymbol{a}')\,\mathrm{d}\boldsymbol{a}'}$$

to obtain the posterior density.

Application of the maximum-likelihood method

Application of the maximum-likelihood method may be complicated due to imperfections of the measurement:

- limited acceptance,
- finite resolution
- non-negligible background contribution.

In principle there seem to be two possibilities, to take these conditions into account:

- try to *correct* the data, or
- modify the theoretical distribution according to the real properties of the measurement - the correct method.

Limited acceptance

Limited acceptance is described by an acceptance function

$$A(x) \quad = \text{probability of observation} \qquad \text{with} \quad 0 \le A(x) \le 1$$

"Acceptance"-correction: assignment of a weight $w_i = 1/A(x_i)$ to the data element x_i.

If for example $A(x_i) = 0.5$, the measurement x_i would get a weight $w_i = 2$. This method may be acceptable for a histogram of the measured distribution; the weighted histogram is then called "acceptance"-corrected.

Weighting method: replace in the likelihood function $p(x_i, \boldsymbol{a})$ by

$$p(x_i, \boldsymbol{a})^{w_i} \qquad \text{with the result} \qquad F(\mathbf{a}) = -\sum_{i=1}^{n} w_i \ln p(x_i|\mathbf{a}).$$

Resulting errors will be wrong. Especially problems with large weights with acceptance $A(x_i) \ll 1$.

Limited acceptance – correct treatment

The correct way is to modify the expectation and to replace $p(x_i|\boldsymbol{a})$ by the properly normalized probability density

$$\mathcal{N}(\boldsymbol{a})^{-1} \cdot A(x_i) \cdot p(x_i, \boldsymbol{a})$$

with the normalization factor $\mathcal{N}$ defined by

$$\mathcal{N} = \int_\Omega A(x) p(x, \boldsymbol{a}) \, \mathrm{d}x$$

In general the normalizing factor will depend on the actual parameter value.

Limited acceptance - example

According to the exponential decay law the decay time distribution is proportional to $e^{-t/\tau}$, where τ is the mean lifetime. For a measurement, which is sensitive only in the time region $t_1 \ldots t_2$, the p.d.f is correctly normalized by the condition $\int_{t_1}^{t_2} p(t) \, \mathrm{d}t = 1$, resulting in the expression

$$p(t) \, \mathrm{d}t = \frac{e^{-t/\tau}}{\tau \left(e^{-t_1/\tau} - e^{-t_2/\tau} \right)} \, \mathrm{d}t \, ,$$

valid in the rest system of the particle (the mean decay time τ is defined in the rest system of the particle).

Finite resolution

The measured data are ”smeared” by a certain resolution function.

The expected distribution is

$$p(x, a) \quad \text{folded with resolution function} \quad A(x_{meas}, x)$$

Result of folding is new distribution

$$q(x, a) = \int A(x_{meas}, x) p(x, a) \, \mathrm{d}x$$

which has to be used in the Likelihood function.

Background contribution

The measured distribution may contain, in addition to the theoretical distribution $p(x, a)$, an additional contribution due to background. For a M.L. fit the background distribution has to be known, either by a measurement of by a simulation.

$$\begin{array}{rl} p(x, a) \, \mathrm{d}x & \text{signal distribution} \\ q(x) \, \mathrm{d}x & \text{background distribution} \\ (1 - \alpha) p(x, a) \, \mathrm{d}x + \alpha q(x) \, \mathrm{d}x & \text{fit distribution} \end{array}$$

The parameter α has either to be known before or has to be fitted.

Binned maximum-likelihood

Measurement: sample of real data, each element of which consists of a set of values $\{\boldsymbol{x}\}$.

Binning: dividing the one- or multi-dimensional space of the $\{\boldsymbol{x}\}$ into n bins. This subdivision gives a set of numbers $\{d_1, d_2, \ldots, d_n\}$, where d_i is the number of events in the real data that fall into bin i. The values d_i are integers 0, 1, ...

$$
\begin{aligned}
d_i &= \text{number of events in the real data that fall into bin } i \\
f_i &= \text{prediction for bin } i = \sum_{j=1}^{m} p_j \, a_{ji} \text{ (in case of m processes)} \\
p_j &= \text{strength factors}
\end{aligned}
$$

Approximate solution by least squares

Measured numbers d_i follow Poisson distribution.

"Standard" method (χ^2 minimization): approximate Poisson distribution by normal distribution with standard deviation $\sigma_i = \sqrt{d_i}$

adjust p_j to minimize $\boxed{S(\boldsymbol{p}) = \sum_{i=1}^{n} \frac{(d_i - f_i)^2}{d_i}}$

This approximation will lead to biased results for small d_i.

Better use correct Poisson distribution.

Poisson maximum likelihood

If the mean value in a bin is f, then the observed values $d = 0, 1, \ldots$ follow the Poisson distribution

$$P_f(d) = e^{-f}\,\frac{f^d}{d!} \qquad \text{with normalization} \quad \sum_{d=0}^{\infty} P_f(d) = 1 \,.$$

Strength factors are found by maximizing the total likelihood

$$\mathcal{L}(\boldsymbol{p}) = \prod_{i=1}^{n} P_{f_i}(d_i) = \prod_{i=1}^{n} e^{-f_i}\,\frac{f^{d_i}}{d_i!}$$

with respect to the parameters $\boldsymbol{p}$ or equivalently, by minimizing the negative logarithm of the likelihood $F(\boldsymbol{p}) = -\ln \mathcal{L}(\boldsymbol{p})$:

$$\boxed{F(\boldsymbol{p}) = \sum_{i=1}^{n} \left(f_i - d_i \cdot \ln f_i\right)}$$

omitting constant factors in the product like $1/d_i!$.

Poisson maximum likelihood contnd.

This expression correctly accounts for small numbers of data events d_i in a bin or even zero data events in some bins. The method is called *binned maximum likelihood* fit.

There is some advantage in redefining the function F to be minimized by adding some constants as

$$\boxed{F(\boldsymbol{p}) = \sum_{i=1}^{n} g_i \qquad \text{with} \quad g_i = \begin{cases} (f_i - d_i) - d_i \cdot \ln(f_i/d_i) & \text{if } d_i > 0 \\ f_i & \text{if } d_i = 0 \end{cases}} \,.$$

The position of the minimum and the shape of F is not changed by his modification, but now $2F$ is approximately distributed according to the χ^2 distribution (with $(n-m)$ degrees of freedom).

Model without analytic prediction

Often no analytical calculation possible for the distributions of the sources.

Instead a Monte Carlo simulation is used to generate data according to the model of the source.

These MC samples can be binned in the same way as the real data, giving a set of integer numbers $\{a_{j1}, a_{j2}, \ldots a_{jn}\}$ for source j. Now both the real data d_i and the data a_{ji} are of statistical nature with integer values $0, 1, \ldots$.

$$\begin{aligned} a_{ji} &= \text{number of Monte Carlo events from source } j \text{ in bin } i \\ N_j &= \sum_{i=1}^{n} a_{ji} = \text{total number in the MC sample for source } j \end{aligned}$$

The Monte Carlo samples are finite, leading to statistical fluctuations in the numbers a_{ji}.

Model without analytic prediction – What to do?

The Monte Carlo samples are finite, leading to statistical fluctuations in the numbers a_{ji}.

- If the Monte Carlo samples are much larger than the data sample, one may ignore these fluctuations and use

$$f_i = \sum_{j=1}^{m} p_j \, a_{ji}$$

  as before. Usually it is assumed that a factor of 10 in the size of the Monte Carlo samples compared to the date sample should be large enough.

- However often the statistical fluctuations in the a_{ij} of the Monte Carlo sample can not be ignored, and one has to consider them together with the statistical fluctuations of the data d_i. A method to treat the problem within the maximum-likelihood method has been developed by R. BARLOW (Comp. Phys. Comm. **77** (1993) 219-228) .

Model without analytic prediction

Barlow method: there is for each source, in each bin, some (unknown) expected number of events A_{ji}

$$f_i = \sum_{j=1}^{m} p_j A_{ji} \; .$$

From each A_{ji} the corresponding a_{ji} is generated by a distribution which can be taken as Poisson.

The total Likelihood is the combined probability of the observed $\{d_i\}$ and of the observed $\{A_{ji}\}$:

$$F = -\ln \mathcal{L} = \sum_{i=1}^{n} [f_i - d_i \cdot \ln f_i] + \sum_{i=1}^{n} \sum_{j=1}^{m} [A_{ji} - a_{ji} \cdot \ln A_{ji}]$$

$$F = -\ln \mathcal{L} = \sum_{i=1}^{n} \left[\sum_{j=1}^{m} p_j A_{ji} - d_i \cdot \ln \sum_{j=1}^{m} p_j A_{ji} \right] + \sum_{i=1}^{n} \sum_{j=1}^{m} [A_{ji} - a_{ji} \cdot \ln A_{ji}]$$

Large number of unknowns

Large number of unknowns:

- m unknown strength factors p_j plus $m \times n$ unknowns A_{ji},
- compared to n bin data $\{d_1, d_2, \ldots, d_n\}$ and $m \times n$ MC bin data.

Set the derivatives of F with respect to the strength factors p_j and the event numbers A_{ji} to zero:

$$\begin{aligned} \sum_{i=1}^{n} \left[A_{ji} - \frac{d_i A_{ji}}{f_i} \right] &= 0 \qquad j = 1, 2, \ldots m \\ 1 - \frac{a_{ji}}{A_{ji}} + p_j - \frac{d_i\, p_j}{f_i} &= 0 \qquad j = 1, 2, \ldots m \qquad i = 1, 2, \ldots n \end{aligned}$$

Thus one has to solve a system of $m + m \times n$ simultaneous, nonlinear equations for the $m + m \times n$ unknowns.

Large number of unknowns

Original paper by BARLOW: last set of equations rewritten in the form

$$1 - \frac{d_i}{f_i} = \frac{1}{p_j}\left(\frac{a_{ji}}{A_{ji}} - 1\right) \qquad j = 1, 2, \ldots m \qquad i = 1, 2, \ldots n$$

Left hand side depends on index i only, so it can be written as $t_i = 1 - d_i/f_i$ and one can express A_{ji} in the form

$$A_{ji} = \frac{a_{ji}}{1 + p_j\, t_i} \,,$$

This simplifies the problem! For a given set of p_j the $m \times n$ unknowns A_{ji} are given by the n unknowns t_i (defined above). If $d_i = 0$ then $t_i = 1$, and if not, then

$$\frac{d_i}{1 - t_i} = f_i = \sum_{j=1}^{m} p_j A_{ji} = \sum_{j=1}^{m} p_j \frac{a_{ji}}{1 + p_j\, t_i}$$

Large number of unknowns

The equations are *not coupled* and the values t_i are determined independently for each bin i. This is a great simplification: instead of $m \times n$ unknowns A_{ji} the are only n unknowns t_i. Since, for a given bin i, all the m values A_{ji} are combined to one value f_i, which has to be compared to one measured value d_i, it is clear that this reduction of the number of parameters per bin from m to 1 has to be possible.

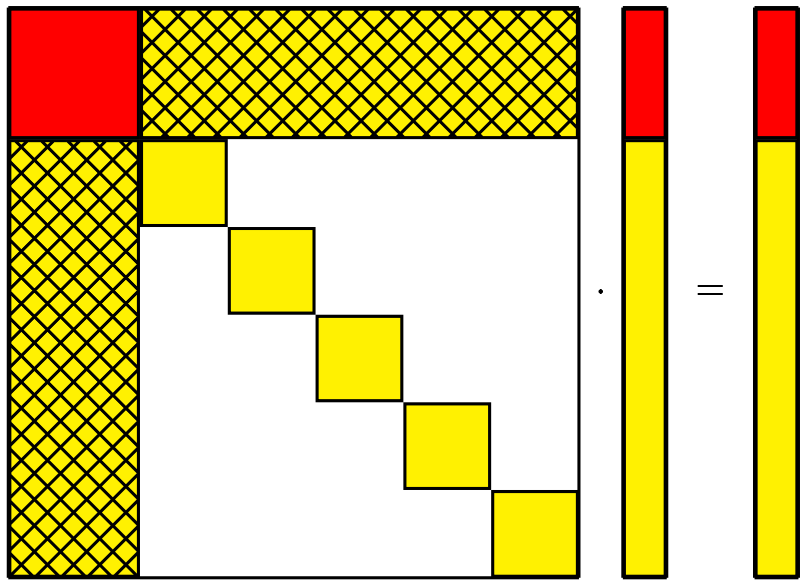
Strategy: the t_i are considered as the essential parameters, not the A_{ji}. The A_{ji} are eliminated and expressed by the t_i. The function to be minimised is written in the form

$$F = -\ln \mathcal{L} = \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{m} \frac{p_j a_{ji}}{1 + p_j t_i} \right) - d_i \cdot \ln \left(\sum_{j=1}^{m} \frac{p_j a_{ji}}{1 + p_j t_i} \right) \right] + \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{a_{ji}}{1 + p_j t_i} - a_{ji} \cdot \ln \frac{a_{ji}}{1 + p_j t_i} \right]$$

with the m unknowns p_j and the n unknowns t_i.

Iterative solution

Iteration procedure with very large number of parameters (p_j and t_i), but solution can be very fast because submatrix for the t_i is diagonal.



Function minimization

Optimization: define objective function $F(x)$ to express optimality e.g for a fit of a parametrized function to measured data.

- Maximum likelihood method: negative log-likelihod function $F(x)$
- Method of least squares: (weighted) sum of squares, chisquare functions

Function minimization and error analysis

- e.g. MINUIT for difficult problems which may require guidance to find the correct solution
- Special code for repeated solution of identically parametrized problems, using general tools

Unconstrained minimization

$$\boxed{\text{minimize } F(\boldsymbol{x}) \qquad \boldsymbol{x} \in R^n}$$

Taylor expansion:

$$\begin{array}{lrcl} \text{function} & F(\boldsymbol{x}_k + \boldsymbol{\Delta x}) & \approx & F_k + \boldsymbol{g}_k^T \boldsymbol{\Delta x} + \frac{1}{2} \boldsymbol{\Delta x}^T \boldsymbol{H}_k \boldsymbol{\Delta x} \\ \text{derivative} & \boldsymbol{g}(\boldsymbol{x}_k + \Delta \boldsymbol{x}) & \approx & \boldsymbol{g}_k + \boldsymbol{H}_k \boldsymbol{\Delta x} \, , \end{array}$$

Function value and derivatives are evaluated at $\boldsymbol{x} = \boldsymbol{x}_k$:

$$\begin{array}{lrcll} \text{function} & F_k & = & F(\boldsymbol{x}_k) & \\ \text{gradient} & (g_k)_i & = & \left. \frac{\partial F}{\partial x_i} \right|_{\boldsymbol{x}=\boldsymbol{x}_k} & i = 1, \, 2, \, \ldots n \\ \text{Hesse matrix} & (H_k)_{ij} & = & \left. \frac{\partial^2 F}{\partial x_i \partial x_j} \right|_{\boldsymbol{x}=\boldsymbol{x}_k} & i, j = 1, \, 2, \, \ldots n \end{array}$$

For negative log-likelihood function: $V = \boldsymbol{H}^{-1}$

Newton step

Step $\boldsymbol{\Delta x}_N$ determined from $\boldsymbol{g}_k + \boldsymbol{H}_k \boldsymbol{\Delta x} = 0$:

$$\boldsymbol{\Delta x}_N = -\boldsymbol{H}_k^{-1} \boldsymbol{g}_k \, .$$

For a quadratic function $F(\boldsymbol{x})$ the Newton step is in length and direction a step to the minimum of the function.

Sometimes large angle ($\approx 90°$) between Newton direction $\boldsymbol{\Delta x}_N$ and $\boldsymbol{g}_k$ (directions of steepest descent).

$$d = -\boldsymbol{g}_k^T \cdot \Delta \boldsymbol{x}_N = \boldsymbol{g}_k^T \boldsymbol{H}_k^{-1} \boldsymbol{g}_k = \boldsymbol{\Delta x}_N^T \boldsymbol{H}_k \boldsymbol{\Delta x}_N > 0$$

if Hessian positive-definite. For a quadratic function the distance to the minimum is $d/2$.

General iteration scheme

Define initial value $\boldsymbol{x}_0$, compute $F(\boldsymbol{x}_0)$ and set $k = 0$

1. **Compute a search vector.** A vector $\boldsymbol{\Delta x}$ is computed as the new search vector. The newton search vector is determined from $\boldsymbol{H}_k \boldsymbol{\Delta x}_N = -\boldsymbol{g}_k$.
2. **Line Search** A one-dimensional minimization is done for the function $f(z) = F(\boldsymbol{x}_k + z \cdot \boldsymbol{\Delta x}_N)$. and z_{min} is determined.
3. **Update.** The point $\boldsymbol{x}_{k+1}$ is defined by $\boldsymbol{x}_k + z_{\text{min}} \cdot \boldsymbol{\Delta x}$, and k is increased by 1.
4. **Test for convergence** If the conditions for convergence are satisfied, the algorithm terminates with $\boldsymbol{x}_k$ as the solution. The difference $F_{k-1} - F_k$ and d are used in the test.

Line search – simple method without derivatives

Find initial interval. Start with inital values x_1, x_2 with $f(x_1) > f(x_2)$ and evaluate function at

$$x_i = x_{i-1} + \alpha(x_{i-1} - x_{i-2}) \qquad i = 3, 4, \ldots \qquad \alpha > 1$$

until function value

$$f(x_i) > f(x_{i-1})$$

is found. Then a minimum is included in interval $[x_{i-2}, x_i]$.

Reduction of the interval. The interval $[x_{i-2}, x_i]$ is divided by the inner point x_{i-1} with

$$f(x_{i-1}) < f(x_{i-2}) \qquad f(x_{i-1}) < f(x_i)$$

Based on a function value an another inner point x_t the interval can be reduced,

- either to the interval $[x_i, x_{i-1}]$
- or to the interval $[x_{i-1}, x_t]$

Method of steepest descent

The search vector $\boldsymbol{\Delta x}$ is equal to the negative gradienten

$$\boldsymbol{\Delta x}_{\mathrm{SD}} = -\boldsymbol{g}_k$$

- Step seems to be natural choice;
- only gradient required (no Hesse matrix);
- no step size defined (in contrast to the Newton step);
- rate of convergence only linear:

$$\frac{F(\boldsymbol{x}_{k+1}) - F(\boldsymbol{x}^*)}{F(\boldsymbol{x}_k) - F(\boldsymbol{x}^*)} = c \approx \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \quad ,$$

($\lambda_{\max}$ and $\lambda_{\min}$ are largest and smallest eigenvalue and κ condition number of Hesse matrix $\boldsymbol{H}$. For a large value of κ c close to one and slow convergence.

Optimal step size, if Hesse matrix known:

$$\boldsymbol{\Delta x}_{\mathrm{SD}} = -\left(\frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \boldsymbol{H}_k \boldsymbol{g}_k}\right) \boldsymbol{g}_k \quad .$$

Problem: Hessian not positive definit

Diagonalization:

$$\boldsymbol{H} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T \; ,$$

(vectors $\boldsymbol{u}_i$are the eigen vectors of $\boldsymbol{H}$, and $\boldsymbol{D}$ is diagonal, with eigenvalues λ_i $\boldsymbol{H}$ in diagonal. At least of the eigenvalues λ_i is zero or less than zero.

1. Strategy: define modified diagonal matrix $\overline{\boldsymbol{D}}$ with elements

$$\overline{\lambda}_i = \max\left(|\lambda_i|, \delta\right) \qquad \delta > 0$$

(which requires diagonalization) and modified positive-definite Hessian

$$\overline{\boldsymbol{H}} = \boldsymbol{U}\overline{\boldsymbol{D}}\boldsymbol{U}^T$$

Problem: Hessian not positive definit

2. Strategy: define modified Hessian by

$$\overline{\boldsymbol{H}} = \boldsymbol{H} + \alpha \boldsymbol{I}_n \qquad (\boldsymbol{I}_n = n \times n \,\text{unit matrix})\ .$$

(simpler)

- large value of α means search direction close to (negative) gradient (safe)
- small value of α means search direction close to Newton direction

factor α allows to interpolate between method of steepeset descent and Newton method.

What is large and what is small: scale defined by

$$\alpha \approx \frac{|\boldsymbol{g}^T \boldsymbol{H} \boldsymbol{g}|}{\boldsymbol{g}^T \boldsymbol{g}}$$

suggested by step of steepest descent method.

Confidence intervals and limits

- Classical confidence intervals – Neyman construction
- Bayesian estimates and intervals
- Poisson with background
- Gaussian with boundary at origin

Probability and the normal distribution

Deviation of a single value x from the mean μ in case of a *normal* distribution:

$$\begin{array}{lrcl} \text{2–sided} & P(|x-\mu| < 1.960\,\sigma) & = & 95\,\% \\ \text{1–sided} & P(x-\mu < 1.645\,\sigma) & = & 95\,\% \end{array}$$

This relation between probability P and standard deviations σ is valid only for the *normal* distribution!

Poisson process – upper limit

Assume $n_{\text{obs}} = 5$ events are observed.

What is the upper limit μ_b of the mean value at the 95 % C.L.?

Naive approach: use $1.645\,\sigma$ with $\sigma = \sqrt{n_{\text{obs}}}$.

$$\mu_b = 5 + 1.645 \cdot \sqrt{5} = 8.68$$

Is this correct?

Assume Poisson distribution with $\mu = 8.68$. What is the probability to observe $k = 5, 4, 3, 2, 1, 0$ events? $P = 13.6\,\%$ i.e. an upper limit of $\mu_b = 8.68$ is not correct (undercoverage).

Assume Poisson distribution with $\mu = 10.51$. What is the probability to observe $k = 5, 4, 3, 2, 1, 0$ events? $P = 5.00\,\%$

The limit $\mu_b = 10.51$ is a 95 % C.L. upper limit for the observed number $n_{\text{obs}} = 5$.

Poisson process – lower limit

Naive approach:

$$\mu_a = 5 - 1.645 \cdot \sqrt{5} = 1.32$$

Assume Poisson distribution with $\mu = 1.32$. What is the probability to observe $k = 5, 6, 7, \ldots$ events? $P = 1.1\,\%$ i.e. overcoverage.

Assume Poisson distribution with $\mu = 1.97$. What is the probability to observe $k = 5, 6, 3, \ldots$ events? $P = 5.000\,\%$

Zero observed events

Important special case: $n_{\text{obs}} = 0$

Upper limit for the confidence level $1 - \beta$:

$$\begin{aligned} \beta &= P(k = 0; b) = P_b(0) = e^{-\mu_b} \\ \rightarrow \mu_b &= -\ln \beta \end{aligned}$$

Numerical example:	95 % C.L.	$\mu_b = 2.996$
	99 % C.L.	$\mu_b = 4.6$
	99.9 % C.L.	$\mu_b = 6.9$

If the true value is $\mu = 6.9$ or > 6.9, then the probability to observe $n = 0$ events is 0.1% or less.

Classical confidence intervals

A confidence interval is a member of a set, such that the set has the property that

$$P(\mu \in [\mu_a, \mu_b]) = 1 - \varepsilon$$

The limits μ_a and μ_b are functions of the measured n_{obs}, and the equation refers to the varying confidence intervals $[\mu_a, \mu_b]$ from a ensemble of experiments with fixed μ.

The intervals $[\mu_a, \mu_b]$ contain the fixed unknown μ_t in a fraction of $1 - \varepsilon$ experiments.

The set of confidence intervals has the correct *coverage*, if the equation is true for every allowed μ.

Basic principle of classical confidence intervals is *coverage*.

Example: if $n_{\text{obs}} = 5$ is observed, the confidence interval at the 90 % C.L. is [1.97, 10.51].

Poisson limits

The number n of events has been observed. What are limits for the underlying mean $\mu = E[n]$, while the observed values k follow the Poisson distribution? The maximum likelihood estimate for μ is $\widehat{\mu} = n$.

$$\begin{aligned}
\alpha &= P\left(k \geq n; \mu_a\right) = \sum_{k=n}^{\infty} P_a\left(k\right) = 1 - \sum_{k=0}^{n-1} P_a\left(k\right) = 1 - \sum_{k=0}^{n-1} \frac{\mu_a^k}{k!}\, e^{-\mu_a} \\
\beta &= P\left(k \leq n; \mu_b\right) = \sum_{k=0}^{n} P_b\left(k\right) = \sum_{k=0}^{n} \frac{\mu_b^k}{k!}\, e^{-\mu_b}
\end{aligned}$$

Lower and upper limits μ_a and μ_b, using a relation between the Poisson and the χ^2 distribution:

$$\mu_a = \frac{1}{2} F_{\chi^2}^{-1}\left(\alpha; n_d = 2n\right) \qquad \mu_b = \frac{1}{2} F_{\chi^2}^{-1}\left(1 - \beta; n_d = 2(n+1)\right)$$

Neyman construction

...for repeatable experiments based on the limiting frequency.

Basis: $p\,(x;\,a)$ = probability density of observing $\boldsymbol{x}$ if the true value of the parameter is a.

Classical construction first proposed by J. Neyman[a].

- Assume fixed value of parameter a and probability ε
- Construct from fixed value of a the limits $x_1(a)$ and $x_2(a)$ with

$$P\,(x_1 < x < x_2) = \int_{x_1}^{x_2} p\,(x;\,a)\;\mathrm{d}x = 1 - \varepsilon$$

  Repeated experiments would produce results x in $x_1 < x < x_2$ with frequency $1 - \varepsilon$.
- Repeat for many assumed values of the parameter a and draw curves $x_1(a)$ and $x_2(a)$ as a function of a.

---

[a] J. Neyman, Outline of a theory of statistical estimation based on the classical theory of probability, Phil. Trans. Royal Soc. London, Series A, **236** (1937) 333

Bayesian estimators

Knowledge about a parameter is summarized by a probability density function (pdf), which expresses the degree of belief for the parameter to take on a given value.

$$\begin{aligned}
\mu &= \text{parameter with true value } \mu_t \\
x &= \text{observable; } x_0 \text{ is a measurement of } x \\
P(x|\mu) &= \text{pdf for obtaining the value } x \\
\mathcal{L}(x_0|\mu) &= P(x_0|\mu) = \text{ likelihood function} \\
P(\mu_t) &= \text{"prior" pdf for the true value} \\
P(\mu_t|x_0) &= \text{"posterior" pdf for the true value } \mu_t
\end{aligned}$$

$P(\mu_t|x_0)$ is the conditional pdf for the unknown μ_t, *given* the result x_0 of the measurement, and represents *the degree of belief* about μ_t after the measurement, and is obtained by applying Bayes' theorem to pdfs for μ_t:

$$\boxed{P(\mu_t|x_0) = \mathcal{L}(x_0|\mu)\,P(\mu_t)/P(x_0)}$$

Prior and posterior probability

$$\begin{aligned} \text{prior pdf} &= \text{state of knowledge before consideration of the data} \\ \text{posterior pdf} &= \text{state of knowledge given the data} \end{aligned}$$

What to use for the prior? How to represent prior ignorance?

Naive choice is a uniform prior, constant in the physically allowed region e.g. restricted to $[0, \infty]$ ("non-informative" prior).

$$\text{Uniform in } \mu_t \text{ or in } 1/\mu_t \text{ or in } \ln \mu_t \text{ ?}$$

A prior uniform in μ_t transforms a likelihood function (function, not density) into a probability density about μ_t, which is non-zero only in the physically allowed region.

The Bayesian estimator is in general not invariant under a transformation of the parameter.

Bayesian intervals

A Bayesian interval $[\mu_1, \mu_2]$ (or "credible interval") corresponding to a certain confidence level can be constructed from the posterior pdf by

$$\int_{\mu_1}^{\mu_2} P(\mu_t|x_0)\, \mathrm{d}\mu_t = 1 - \varepsilon$$

For *making decision* the Bayesian description of inference is probably the way many scientists think: one person's subjective prior incorporates all personal beliefs as well as the results from previous experiments.

Poisson with background – Karmen

Karmen 2 experiment (1998): search for neutrino oscillation signal. For a signal like that reported by the LSND experiment 1 – 1.5 signal events were expected.

Result of Karmen 2:

expected background	$b = 2.88 \pm 0.13$
observed	$n = 0$

Can the effect claimed by LSND be excluded?

Poisson with background

Number n of observed events includes signal events with *unkwown* mean μ and background events with *known* mean B.

The number k of events observed in an experiment follows the Poisson distribution with mean $\mu + b$

$$P_\mu(k) = (\mu + b)^k \, \frac{\exp(-(\mu + b))}{k!}$$

Two problems:

- If the expected is for example $b = 3.0$ and no events are observed, the confidence intervall may be empty.
- If the observed number n is small, and if the data are used to decide either for an upper limit (one-sided) or a confidence interval (two-sided), then the resulting set can undercover.

Proposal to solve the problems by an ordering principle:

Gary J. Feldmann and Robert D. Cousins, A unified approach to the classical statistical analysis of small signals, HUTP-97/A096

Ordering principle based on likelihood ratios

Example: Poisson with *known* background $b = 3.0$

Assume in a Neyman construction a particular value of μ: $\mu = 0.5$

	$P_{0.5}(k)$	μ_{best}	$P_{\mu_{\text{best}}}(k)$	ratio R	rank
$k = 0$	0.030	0	0.050	0.607	6
$k = 1$	0.106	0	0.149	0.708	5
$k = 2$	0.185	0	0.224	0.826	3
$k = 3$	0.216	0	0.224	0.963	2
$k = 4$	0.189	1	0.195	0.966	1
$k = 5$	0.132	2	0.175	0.753	4
$k = 6$	0.077	3	0.161	0.480	7
$k = 7$	0.039	4	0.149	0.259	
...					

Value $k = 0$ has $P(0) = 0.030$, but what is the value of μ which maximizes $P_\mu(0)$? This is called μ_{best} and is given by $\mu_{\text{best}} = \max(0, k - b)$.

Values k are added to the acceptance region in *decreasing* order of the ratio R, until the sum meets the desired C.L. (here $k = 0 \ldots 6$ with sum 0.935).

Gaussian with boundary at origin

Assume only non-negative values for parameter μ are allowed (e.g. μ is a mass).

Observable x is the measured value of μ in an experiment with Gaussian resolution function with fixed standard deviation σ:

$$P(x; \mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/(2\sigma^2))$$

The same two problems as in the Poisson case can happen for small or negative x/σ) (undercoverage and empty confidence interval).

Ordering principle for Gaussian with boundary at origin

Assume μ physically bounded to non-negative values.

Let μ_{best} be the physically allowed value for which $P(x;\mu)$ is maximum. Then $\mu_{\text{best}} = max(0,x)$, and

$$P(x;\mu_{\text{best}}) = \begin{cases} 1/(\sqrt{2\pi}\sigma) & \text{if } x \geq 0 \\ 1/(\sqrt{2\pi}\sigma)\exp(-x^2/(2\sigma^2)) & \text{if } x < 0 \end{cases}$$

Neyman construction of confidence intervals with μ from grid with narrow spacing: the ratio R, given by

$$R(x) = \frac{P(x;\mu)}{P(x;\mu_{\text{best}})}$$

determines the order in which value of x are added to the acceptance region at a particular value of μ.

Goodness-of-fit and confidence interval

Standard classical analysis of a parameter measurement:

Goodness-of-fit C.L. Test reliability of the experiment: Parameter obtained by a poor fit? Measured value far in the unphysical region?

Confidence interval C.L. What is the confidence interval of the parameter at a certain confidence level? (e.g. determined from the curvature of a likelihood function)

These two questions may be coupled!

An advantage of the proposed ordering is a decoupling of the two questions.

Summary

There is no clear and unique solution to the problem of Confidence Limits in practice.

See opinions and discussion in:
Workshop on Confidence Limits, CERN 17-18 January 2000, ed. F. James, L. Lyons, Y. Perrin, CERN 2000-005 (yellow report)

"You see, a question has arisen,
about which we cannot come to an agreement,
probably because we have read too many books"
(Brecht: Leben des Galilei)