The maximum-likelihood method

Volker Blobel – University of Hamburg

October 2003

1. The maximum likelihood principle
2. Properties of maximum-likelihood estimates
3. Application of the maximum likelihood method
4. Binned distributions

The maximum-likelihood principle

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 Likelihood function

The *maximum-likelihood method* starts from the *joint* probability distribution of the n measured values $x_1, \ldots, x_i, \ldots, x_n$.

For *independent* measurements this 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*.

The likelihood function is a *function*, it is not a probability density of the parameter a (→ Bayes interpretation).

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 is considered,

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

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

Likelihood equation, defining estimate $\hat{a}$: $$\frac{\mathrm{d}F(a)}{\mathrm{d}a} = 0$$

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: $F(a) = -2 \ln \mathcal{L}(a)$.

Combining results of experiments

The combination of results

- from different experiments or
- from different measurements,

depending on the same parameter(s), is straightforward:

$$\begin{array}{ll} \mathcal{L}(a) = \mathcal{L}_2(a) \cdot \mathcal{L}_2(a) & \qquad \text{multiply Likelihood functions} \\ F(a) = F_1(a) + F_2(a) & \qquad \text{add log. Likelihood functions} \end{array}$$

Least Squares and Maximum Likelihood

If the measurements y_i are gaussian distributed around the expected value $f(x_i;a)$ (containing p parameters a to be estimated) with variance σ_i^2, i.e. if they follow a density

$$\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{\left(y_i - f(x_i;a)\right)^2}{2\sigma_i^2}\right] ,$$

then the neg. log. Likehood function $F(a) = -\ln \mathcal{L}(a)$ is

$$F(a) = \frac{1}{2} \sum_{i=1}^{n} \frac{\left(y_i - f(x_i;a)\right)^2}{\sigma^2} + \text{ const.}$$

i.e. the expressions to be minimized are identical in the methods of Least Squares and Maximum Likelihood (except for a factor 1/2).

In case of a correct model the quantity $2F(a)$ at the minimum follows the χ^2-distribution with $(n-p)$ degrees of freedom.

Comparison

Comparison between Maximum Likelihood (ML) method and the Least Squares (LS) method:

ML requires full knowledge on the probability density of data.

LS requires no detailed knowledge on the probability density of data, only the mean and variance (first two moments of distribution), have to be known (data unbiased and variance known).

Efficiency:

ML estimate $\hat{a}$ is usually a nonlinear function of the data, and reaches asympotically the full efficiency, defined by the information I,

LS estimate $\hat{a}$ is a linear function of the data (in linear least squares) and is most efficient among the linear estimates.

In a special case (previous page): ML → LS

Example of angular distribution

The value $x \equiv \cos\vartheta$ is measured in n decays of an elementary particle. According to theory the distribution is

$$p(\cos\vartheta) = \frac{1}{2}\,(1 + a\cos\vartheta)$$

This probability density is normalized for all physical values of the parameter a, if the whole range of $\cos\vartheta$ can be measured.

The aim is to get an estimate of the parameter a.

$$\begin{aligned} \text{minimize} \quad \mathcal{L}(a) &= \prod_{i=1}^{n} \left[\frac{1}{2}\,(1 + a\cos\vartheta_i)\right] \\ \text{maximize} \quad F(a) &= -\sum_{i=1}^{n} \ln\,(1 + a\cos\vartheta_i) + \text{ const.} \end{aligned}$$

Note: The normalization is parameter dependent, if the measured range of $\cos\vartheta$ is limited.

Observations:

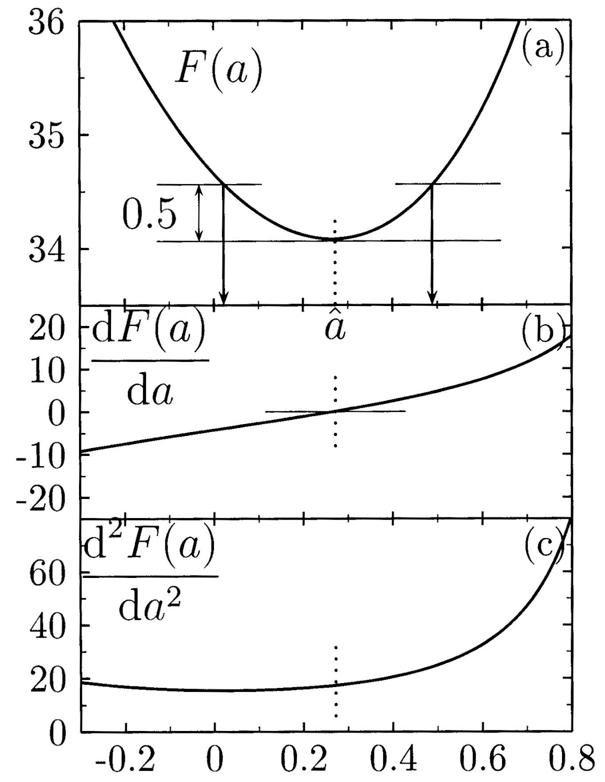
- the value of $F(a)$ at the minimum is fluctuating;
- the shape of $F(a)$ is close to a parabola;
- the value of the curvature increases with increasing number of observations n; the minimum is getting sharper.

value of $F(a)$ at the minimum: provides for binned Maximum likelihood a test of goodness-of-fit (how well are the data described by the model?), but in general **not** for unbinned cases!
In practice this may require repeated MC simulations of the experiment to determine the distribution of $F(a)_{\text{min}}$

inverse curvature: corresponds to the variance of the parameter estimate, at least asymptotically.

. . . contnd.

- shape of $F(a)$ approximately parabolic
- first derivative approximately linear
- second derivative approximately constant



Estimate of variance

Second derivative H of $F(a)$ corresponds to inverse of the variance σ^2 of the parameter estimate: $\sigma^2 = 1/H$ and $\sigma = 1/\sqrt{H}$.

Taylor expansion of $F(a)$:

$$F(a) - F(\hat{a})_{\text{min}} = \frac{1}{2} H \left(a - \hat{a}\right)^2 + \ldots$$

If $|\,(a - \hat{a})\,| = 1\,\sigma$, then

$$F(a) - F(\hat{a})_{\text{min}} = \frac{1}{2}$$

Values a with $\Delta F = F(a) - F(\hat{a})_{\text{min}} = 1/2$ can be used to estimate the standard deviation σ of the parameter estimate.

But distinguish between

- fluctuations ΔF of the minimum value (in binned maximum likelihood → goodness-of-fit), and
- curvature, which can be estimated from ΔF.

Search for special events in two channels

channel	meas. n_i	expected: (total)	signal S	background B
a	6	1.1 ± 0.3	0.9 ± 0.3	0.2 ± 0.1
b	24	28.0 ± 6.0	4.00 ± 0.6	24.00 ± 6.0

Model includes factor f : $\mu_i = f \cdot \; S \; + \; B$

Questions:

- Are the two measurements compatible?
- Are both measurements compatible with $f = 1$, which is the standard expectation from theory?

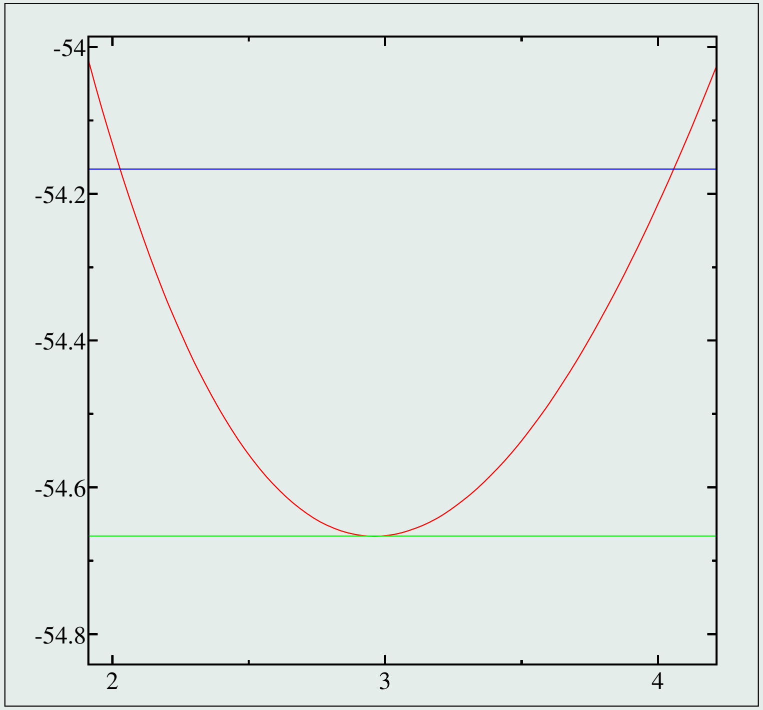
Method to answer the questions

Use Maximum Likelihood method to obtain best estimate for the factor f from all data!

Likelihood function and negative log Likelihood function, based on Poisson distribution of data n_i with mean values given by model:

$$\begin{aligned} \mathcal{L}(f) &= P(n_1|\mu_1) \cdot P(n_2|\mu_2) = \frac{e^{-\mu_1}\mu_1^{n_1}}{n_1!} \cdot \frac{e^{-\mu_2}\mu_2^{n_2}}{n_2!} \\ F = -\ln \mathcal{L}(f) &= \sum_{i=1}^{2} (\mu_i - n_i \ln \mu_i) + \text{ const.} \end{aligned}$$

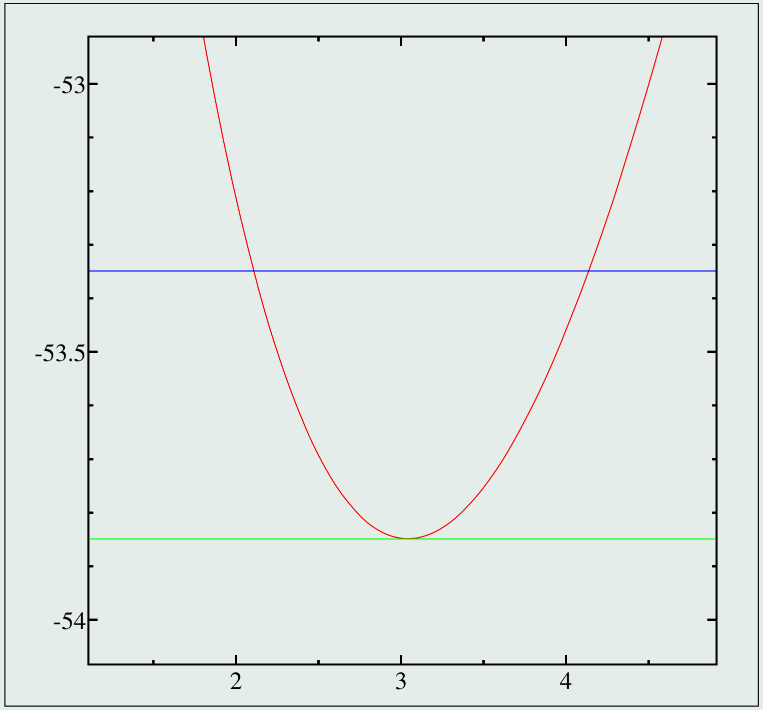
Negative log likelihood function



Result for factor: $f = 2.96 \begin{array}{l} +1.09 \\ -0.94 \end{array}$

Note: statistical fluctuations for SM(...) ignored.

Negative log likelihood function



Result for factor: $f = 3.05 \begin{array}{l} +1.09 \\ -0.94 \end{array}$

Note: statistical fluctuations for SM(...) NOT ignored.

Example: exponential distribution

Measured are n times t_i, which should be distributed according to the density

$$p(t;\tau) = \frac{1}{\tau} \exp\left[-\frac{t}{\tau}\right] \; .$$

Log. Likelihood function for parameter τ, to be estimated from the data:

$$F(\tau) = -\sum_{i=1}^{n} \ln p(t;\tau) = -\sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

By minimization of $F(\tau)$ the resulting estimate is

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i \qquad \text{with} \quad E\left[\hat{\tau}(t_1,\, t_2, \ldots)\right] = \tau$$

i.e. the estimator is unbiased.

Note: in general mean values are unbiased.

. . . contnd.

Instead of parameter τ the parameter λ in the density

$$p(t;\lambda) = \lambda \exp\left[-\lambda\, t\right] \ .$$

has to be estimated. Can the previous result be used?

yes, because of $$\left(\frac{\partial\mathcal{L}}{\partial\tau}\right) = \left(\frac{\partial\mathcal{L}}{\partial\lambda}\right) \cdot \frac{\partial\lambda}{\partial\tau} = 0$$

the Maximum Likelihood estimate for λ is

$$\hat{\lambda} = \frac{1}{\hat{\tau}}$$

(note: $\mathcal{L}(a)$ is a function of a, not a density).

But:

$$E\left[\hat{\lambda}(t_1,\, t_2, \ldots)\right] = \frac{n}{n-1}\lambda = \frac{n}{n-1}\frac{1}{\tau} \qquad \text{biased!}$$

i.e. there is invariance of the Maximum Likelihoid estimates w.r.t. transformations, but only one parametrization can be unbiased.

Properties of the maximum-likelihood estimates

Maximum-likelihood estimates $\widehat{a}$

Consistency: The estimate $\widehat{a}$ of the MLM is asymptotically ($n \rightarrow \infty$) 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.

Information inequality

$$\boxed{\text{Information} \quad I(a) = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial a}\right)^2\right] = \int_\Omega \left(\frac{\partial \ln \mathcal{L}}{\partial a}\right)^2 \mathcal{L}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n}$$

This is the definition of *information*, where $\mathcal{L}$ is the joint density of the n observed values of the random variable x.

$$\boxed{\text{Information inequality} \qquad V[\widehat{a}] \geq \frac{1}{I}}$$

The *inverse* of the information $I_n(a)$, or short I, is the lower limit of the variance of the parameter estimate $\widehat{a}$ – minimum variance bound MVB.

The inequality is also called Rao-Cramér-Frechet inequality, and is valid in this form for any unbiased estimate $\widehat{a} = \widehat{a}(x)$.

Efficiency and bias

Definition of the *efficiency* of an estimator:
For an unbiased estimator $\hat{a}$ one can define the *efficiency* $\text{eff}(\widehat{a})$ by the ratio of the mininal to the actual variance:

$$\boxed{\text{Efficiency} \qquad \text{eff}(\widehat{a}) = \frac{I^{-1}}{V[\widehat{a}]} \qquad 0 \leq \text{eff}(\widehat{a}) \leq 1}$$

The actual efficiency of an estimator depends on the specific problem and method.

Variance limit in case of a bias $B_n(\widehat{a}) = E\left[\widehat{a}\right] - a_{\text{true}} \neq 0$:

$$\boxed{\text{Variance} \qquad V\left[\widehat{a}\right] \geq \frac{\left(1 + \partial B/\partial a\right)^2}{I}}$$

Alternative expression of information I

From the proof of the information inequality in previous chapter:

$$\int_\Omega \left(\frac{\partial \ln \mathcal{L}}{\partial a} \frac{\partial \mathcal{L}}{\partial a} + \frac{\partial^2 \ln \mathcal{L}}{\partial a^2} \mathcal{L} \right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 0 \; ,$$

Rewritten in terms of expectation values:

$$I(a) = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial a} \right)^2 \right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a^2} \right]$$

i.e. either square of first derivative or negative second derivative.

The second derivative is *almost* constant: expectation value is close to value at the minimum

$$I(a) = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a^2} \right] \approx \left. \frac{\partial^2 F(a)}{\partial a^2} \right|_{a=\hat{a}}$$

Case of several variables

Case of m variables $a_1, \ldots, a_j, \ldots, a_m$: information I becomes a m-by-m symmetric matrix $\boldsymbol{I}$ with elements

$$I_{jk} = E\left[\frac{\partial \ln \mathcal{L}}{\partial a_j}\frac{\partial \ln \mathcal{L}}{\partial a_k}\right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a_j \partial a_k}\right]$$

The minimal variance $\boldsymbol{V}[\hat{\boldsymbol{a}}]$ of an estimate $\hat{\boldsymbol{a}}$ is given by the inverse of the information matrix $\boldsymbol{I}$:

$$\boxed{\text{minimal variance} \qquad \boldsymbol{V}[\hat{\boldsymbol{a}}] = \boldsymbol{I}^{-1}}$$

Normality

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

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

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 $\lambda = 1/\tau$.

. . . contnd.

Negative log-likelihood function for given data $t_1, \ldots, t_i, \ldots, t_n$ is

$$F(\tau) = -\sum_{i=1}^{n} \ln p(t_i; \tau) = \sum_{i=1}^{n} \left(\frac{t_i}{\tau} - \ln \frac{1}{\tau} \right) .$$

Minimizing $F(\tau)$ with respect to τ:

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i \qquad E\left[\widehat{\tau}(t_1, \ldots, t_i, \ldots, t_n)\right] = \tau_{\text{true}} \quad \text{unbiased}$$

If decay constant $\lambda = 1/\tau$ is used: $\widehat{\lambda} = 1/\widehat{\tau}$, or

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} \qquad E\left[\widehat{\lambda}\right] = \frac{n}{n-1}\lambda_{\text{true}} = \lambda_{\text{true}} + \frac{1}{n-1}\lambda_{\text{true}} \quad \text{bias} \neq 0$$

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 $\boldsymbol{x}$ expressed by $\mathcal{L}(\boldsymbol{x}|\boldsymbol{a})$

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

to obtain the posterior density $P(\boldsymbol{a}|\boldsymbol{x})$ of the parameter a; i.e. the function $\mathcal{L}(\boldsymbol{x}|\boldsymbol{a})$ of the parameter is transformed to a density of the parameter.

"The information contained in an observation x with respect to the parameter a is summarized by the density $P(\boldsymbol{a}|\boldsymbol{x})$. The density $P(\boldsymbol{a}|\boldsymbol{x})$ of the actual observation is all what matters for the parameter inference."

→ relevant for the calculation of limits.

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

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

"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 $f(x_i, \boldsymbol{a})$ by

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

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.

The correct treatment will require a large effort in computation: during the minimization the normalization has to be repeated for every new value of the parameter.

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).

The parameter value will be biased, if this normalization is neglected.

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 normalized and 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.

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

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, ...

$$\begin{aligned} d_i &= \text{number of events in the real data that fall into bin } i \\ N_D &= \sum_{i=1}^{n} d_i = \text{total number in the data sample} \end{aligned}$$

The real data arise from a number of sources (or physical processes) and the aim is to determine the proportions P_j of the different sources in the data from the statistical data d_i and from models for the sources.

One has to distinguish the case, where analytic forms are available for the distribution of the sources and where no analytic forms are available.

Model with analytic prediction

Calculate, by integration over the bins, numbers a_{ji} proportional to the expected number of events from source j in bin i; these values have no statistical errors.

$$\begin{aligned}
P_j &= \text{proportion of source } j \text{ in the data (sum to unity)} \\
a_{ji} &= \int_{\text{bin } i} f_j(x)\,\mathrm{d}x = \text{numbers, proportional} \\
& \text{to expected number from source } j \text{ expected in bin } i \\
N_j &= \sum_{i=1}^{n} a_{ji} \\
a_{ji}/N_j &= \text{fraction of events from source } j \text{ expected in bin } i
\end{aligned}$$

The predicted number of events in bin i is, for the proportions P_j, given by the sum

$$f_i = N_D \sum_{j=1}^{m} P_j \, \frac{a_{ji}}{N_j} \qquad \text{or} \quad f_i = \sum_{j=1}^{m} p_j \, a_{ji}$$

with strength factors $p_j = N_D P_j / N_j$.

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 $\sigma_i = \sqrt{d_i}$

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

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

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

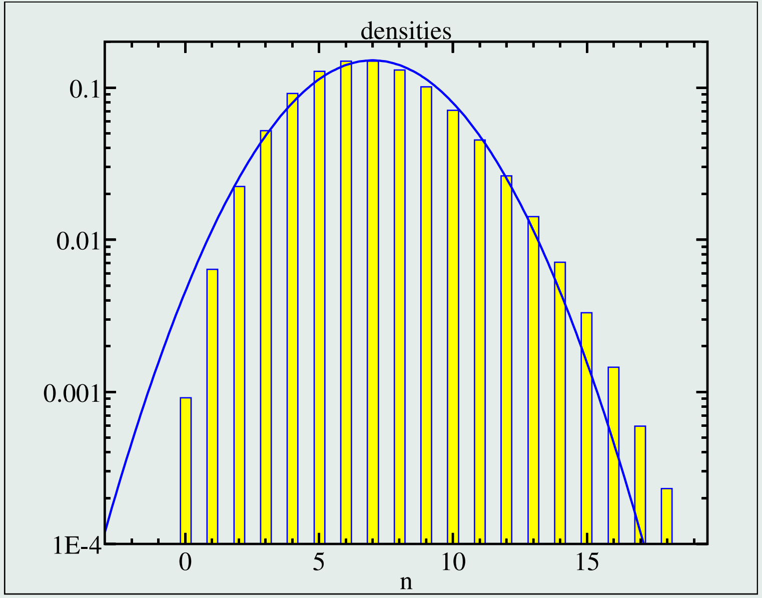
The position of the minimum and the shape of F is not changed by his modification, but now $2F$ is approximately distributed according to the χ^2 distribution (with $(n - m)$ degrees of freedom).

Poisson approximation by Gaussian   Example: observed $y = 7$

Blue curve is Gaussian approximation with $\mu = \sigma^2 = 7$ in both figures.

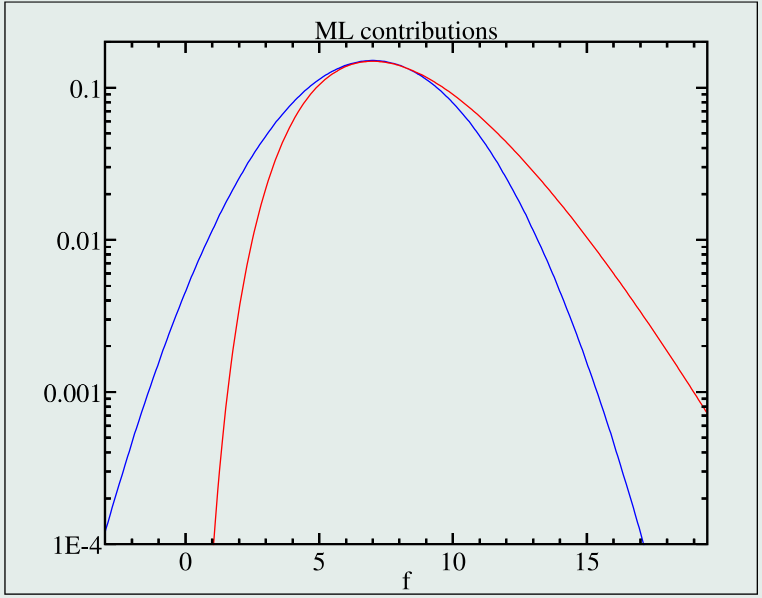
Poisson density for $\mu = 7$:

$$P = \frac{\mu^n e^{-\mu}}{n!}$$



Poisson ML contribution:

$$P = \frac{f^y e^{-f}}{y!}$$



Histogram fits

Should on use a least squares fit (χ^2 minimization) of Poisson maximum likelihood in a fit to histogram data?

> Some people put the requirement as low as $\lambda = 5$, but 10 is probably safer. [**?**]

> It is undesibale to have less than five events in any bin. [**?**]

Just excluding bins with no entries will introduce a bias.

Poisson contribution to objective function

$$
\begin{aligned}
F(\boldsymbol{a}) &= \sum_i f(x_i,\boldsymbol{a}) - y_i \ln f(x_i,\boldsymbol{a}) \\
\text{or better} \quad F(\boldsymbol{a}) &= \sum_i \left(f(x_i,\boldsymbol{a}) - y_i\right) + y_i \ln \frac{y_i}{f(x_i,\boldsymbol{a})} \\
\frac{\partial F}{\partial a_j} &= \sum_i y_i \frac{\frac{\partial f}{\partial a_j}}{f(x_i,\boldsymbol{a})} - \frac{\partial f}{\partial a_j} \\
\frac{\partial^2 F}{\partial a_j \partial a_k} &= \sum_i y_i \frac{\frac{\partial f}{\partial a_j}\frac{\partial f}{\partial a_k} - \frac{\partial^2 f}{\partial a_j \partial a_k} f(x_i,\boldsymbol{a})}{f^2(x_i,\boldsymbol{a})} - \sum_i \frac{\partial^2 f}{\partial a_j \partial a_k}
\end{aligned}
$$

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

- 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; the fluctuations in the a_{ji} are damped by the factor N_D/N_j. 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.

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 $\{A_{ji}\}$:

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

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

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:

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

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

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

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

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

. . . contnd.

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.

ML fit of binned distributions

Expectation value for bin content y_i, $i = 1,\, 2, \ldots n$:

$$y_i = \sum_{J=1}^{m} a_{ij}\, x_j \qquad a_{ij} \text{ are known (without statistical errors)}$$

Distribution of expected number Poisson distribution $P_y(\widehat{y}) = e^{-y}\ \frac{y^{\widehat{y}}}{\widehat{y}!}$ is used to construct (negative log of) likelihood function:

$$F(\boldsymbol{x}) = -\ln \mathcal{L}(\boldsymbol{x}) = -\ln \left[\prod_{i=1}^{n} P_{y_i}(\widehat{y}_i) \right] = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i \cdot \ln y_i \right) + \text{const.}\ ,$$

which has to be minimized – or better ($F_{\text{min}} \sim \chi^2$)

$$F(\boldsymbol{x}) = \sum_{i=1}^{n} g_i \qquad \text{with} \quad g_i = \begin{cases} (y_i - \widehat{y}_i) - \widehat{y}_i \cdot \ln(y_i/\widehat{y}_i) & \text{if } \widehat{y}_i > 0 \\ f_i & \text{if } \widehat{y}_i = 0 \end{cases}$$

Fitting using finite Monte Carlo samples

…but the a_{ij} have itself statistical errors, if determined by Monte Carlo simulation.

Method of R.Barlow and Chr.Beeston (Comp. Phys. Comm. 77 (1993)): redefine expectation

$$y_i = \sum_{j=1}^{m} A_{ij} x_j \quad \text{with (unknown) expected number of events } A_{ij}$$

From each A_{ij} the corresponding a_{ij} is generated by a distribution taken to be Poisson, and included in the definition of the likelihood function.
This introduces a large number $n \times m$ of parameters, which however can be reduced to n parameters t_i, $i = 1, 2, \ldots n$ (see paper above).

Solution

There is still a large number $(n+m)$ of parameters to be determined in the fit:

$$\boldsymbol{x}: \quad m \text{ parameters} \qquad\qquad \boldsymbol{t}: \quad n \text{ parameters}$$

(time-consuming calculation).

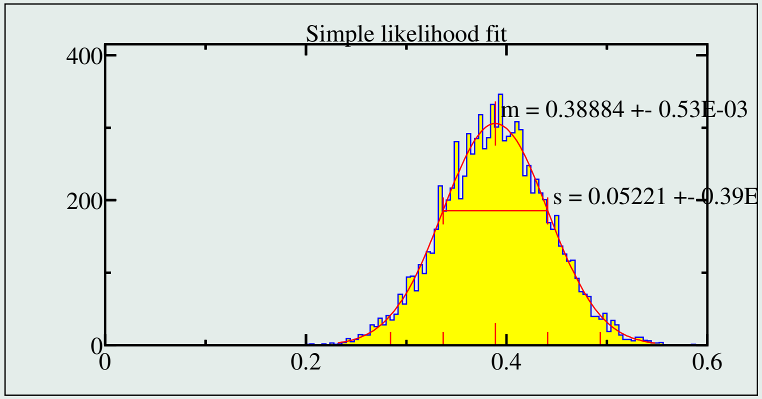
Special fast solution: the $n+m$-by-$n+m$ Hessian has a *diagonal* submatrix of the n-by-n derivatives w.r.t. the t_i and can effectively be reduced by partitioning to a small m-by-m matrix.

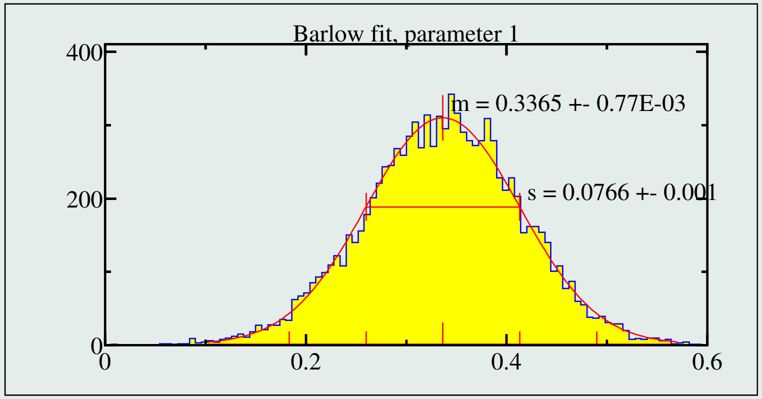
Test with example C of Barlow paper on the next two slides:
100 bins with 1000 entries for "data" and MC with two parameters.

$$1. \text{ parameter } = 1/3 \qquad\qquad 2. \text{ parameter } = 2/3$$

Fit of binned distributions: result on first parameter

Value of first parameter in simulation is 1/3. Shown is the result of 10 000 ”experiments”.





top Simple (Poisson) likelihood fit: result biased

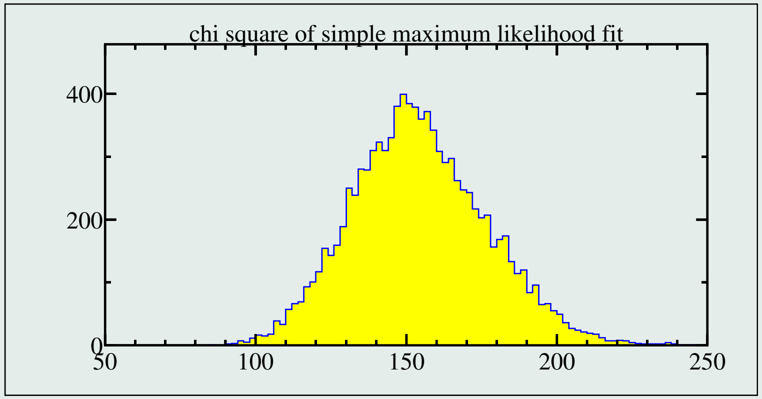
bottom Method of Barlow: result unbiased

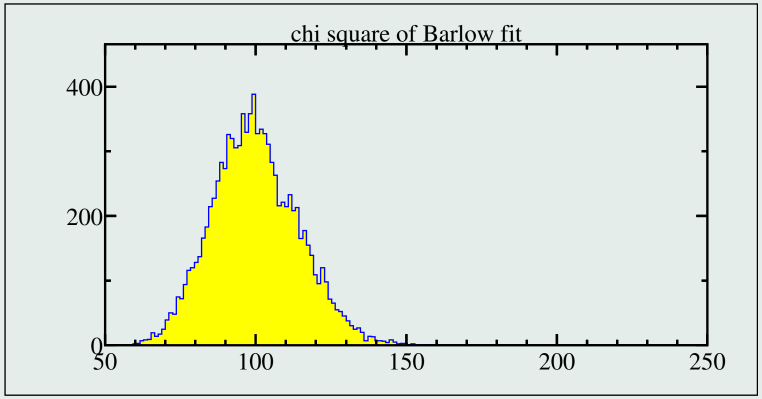
”χ^2” of the two fits methods

Goodness of fit can be checked by (modified) value of (negative log) likelihood function. Expected value is $100 - 2 = 98$.

top Simple (Poisson) likelihood fit: large ”χ^2”, because fluctuation of MC simulation partly neglected.

bottom Method of Barlow: value as expected.





The maximum-likelihood method