Maximum likelihood method

Olaf Behnke

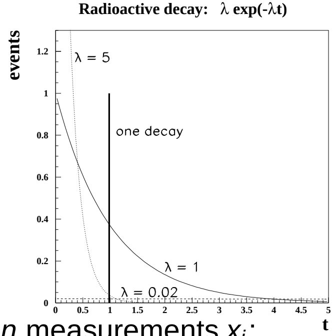
DESY

Outline

- The method
- Weighted averages
- Uncertainties
- Exponential decay
- Two signal processes:
  - fit fractions
  - fit rates (extended Likelihood)
- Binned fits

The Maximum Likelihood (ML) Method

- Single measurements follow PDF $p(x, \vec{a})$ with $\int p(x, \vec{a})\, dx = 1$
- Basic idea: for typical measurement x_i, the $p(x_i, \vec{a}$ should be larger for true $\vec{a}$ then for wrong $\vec{a}$
- Example: radioactive decay $p(t, \lambda) = \lambda e^{-\lambda t}$, one decay at $t = 1$



$\Rightarrow \lambda = 1$ seems a reasonable choice

For n measurements x_i:

> Take for estimator $\hat{\vec{a}}$ the value of $\vec{a}$ for which $L = \prod_{i=1}^{n} p(x_i, \vec{a}) = max$

Practical: Max. of $\omega = lnL = \sum_{i=1}^{n} ln(p(x_i, \vec{a}))$

$\Rightarrow \frac{d\omega}{d\vec{a}}_{|\vec{a}=\hat{\vec{a}}} = 0$

ML example - weighted average

- Likelihood for averaging n measurements y_i with known uncertainties σ_i:

$$L = p(y_1, y_2, ..., y_n|a) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(y_i-a)^2}{2\sigma_i^2}\right\} =$$

$$c \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-a)^2}{\sigma_i^2}\right\}$$

$$= c \exp\left\{-\frac{\chi^2}{2}\right\} \quad \text{with } c = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i} \text{ and } \chi^2 = \sum_{i=1}^{n}\frac{(y_i-a)^2}{\sigma_i^2}$$

$\Rightarrow$ $\omega := \ln L = -\frac{1}{2}\chi^2 + \ln c$

$\Rightarrow$ maximising $\omega \Leftrightarrow$ minimising $\chi^2 \Rightarrow$ Both methods yields same $\hat{a}$

Error estimate	χ^2 method	ML method
$\Rightarrow$ 2nd derivative	$\sigma_{\hat{a}} = \left[\frac{1}{2}\frac{d^2\chi^2}{da^2}\Big|_{a=\hat{a}}\right]^{-1/2}$	$\sigma_{\hat{a}} = \left[-\frac{d^2 \ln L}{da^2}\Big|_{a=\hat{a}}\right]^{-1/2}$
Value change	$\chi^2 = \chi^2_{min} + 1$	$\ln L = \ln L_{max} - 0.5$

$\Rightarrow$ can define: $\tilde{\chi}^2 = -2\ln L$ and use this for fitting

Differences between $\tilde{\chi}^2$ and χ^2

- Likelihood for averaging n measurements y_i with the same but unknown uncertainty σ:

$$L = p(y_1, y_2, ..., y_n|a) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - a)^2}{2\sigma^2}\right\}$$

$$= c \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a)^2}{\sigma^2}\right\}$$

$$= c \exp\left\{-\frac{\chi^2}{2}\right\} \quad \text{with } c = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma} \text{ and } \chi^2 = \sum_{i=1}^{n}\frac{(y_i - a)^2}{\sigma^2}$$

$$\Rightarrow \tilde{\chi}^2 = -2\ln L = \chi^2 - 2\ln c = \chi^2 + 2\sum_{i=1}^{n}\ln\sigma + const.$$

- Find estimate $\hat{\sigma}$ from minimum of

χ^2: $\hat{\sigma} \to \infty$ can you explain why?

$$\tilde{\chi}^2 : \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$\Rightarrow$ “Normal” χ^2 not suitable for this task, but ML method is ok!

ML parameter uncertainties

- Note: L is invariant under a parameter transformation $a \rightarrow b$
- Example weighted average, transform $b = 1/a$:

$$L(a) = p(y_1, y_2, ..., y_n|a) = c \; \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a)^2}{\sigma_i^2}\right\}$$

$$L(b) = p(y_1, y_2, ..., y_n|b) = c \; \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - 1/b)^2}{\sigma_i^2}\right\} = L(a)$$

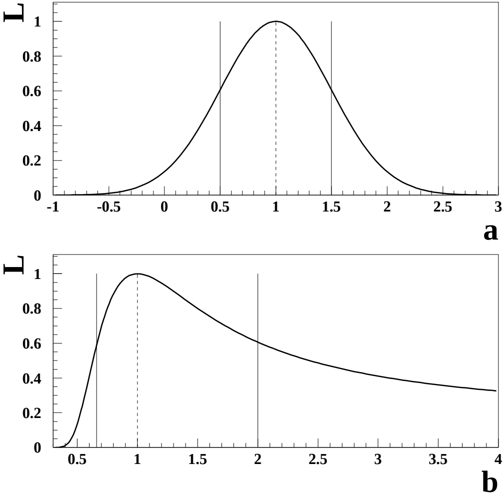
$\Rightarrow$ $\hat{b} = 1/\hat{a}$ Note: for any likelihood and transformation: $\hat{b} = b(\hat{a})$

- Number example: $\hat{a} = 1$, $\sigma_{\hat{a}} = 0.5$
- $L(a) \sim \exp\left\{\frac{(a-1)^2}{0.5}\right\}$
- $L(b) \sim \exp\left\{-\frac{(1/b-1)^2}{0.5}\right\}$

$\Rightarrow$ Assess errors on $\hat{a}$ and $\hat{b}$ from likelihood curves (next slide)

ML parameter uncertainties

Read off uncertainties from points where L drops by 40% (corresponds to $\Delta \ln L = -0.5$)



- Introduce negative and positive uncertainties $\Delta\hat{b}_-$ and $\Delta\hat{b}_+$
- Interval $[\hat{b} - \Delta\hat{b}_-, \hat{b} + \Delta\hat{b}_+]$ is estimated 68% C.L. interval for b
- Quote results as $b = \hat{b}^{+\Delta\hat{b}_+}_{-\Delta\hat{b}_-}$

ML parameter uncertainties

- For any likelihood function $L(a)$: estimating uncertainties from the two points where $\ln L$ drops by 0.5 from maximum is a good method!
- Reasoning: in theory one can always find a parameter transformation $\psi(a)$ which makes the likelihood in ψ gaussian and from the invariance of L we know that the 68% confidence intervals in ψ correspond to 68% confidence intervals in a.
  A small warning: for many/most likelihoods and finite statistics the estimated intervals will not be exact $\Rightarrow$ *"the error has an error"*

Mini Exercise: ML for radioactive decay

Probability density $p(t, \lambda) = \lambda e^{-\lambda t}$

Determine an ML-estimate $\hat{\lambda}$ for case of *one single decay* at time t_1

- Analytically
  - calculate $\omega = lnL = ln\, p(t_1, \lambda)$ and find $\hat{\lambda}$ from $d\omega/d\lambda = 0$
  - Estimate the uncertainty of $\hat{\lambda}$ from the gaussian approximation of L:

    $$\sigma_{\hat{\lambda}} = \left(-\frac{d^2\omega}{d\lambda^2}_{|\lambda=\hat{\lambda}}\right)^{-1/2}$$

- Graphically

  Plot the $\chi^2 = -2lnL$ in ROOT (case $t_1 = 1$):
  - TF1 *f1 = new TF1("f1","-2*log(x)+2*x",0,5); f1->Draw();
  - Determine $\hat{\lambda}$ from the min. χ^2 and an uncertainty estimate from $\chi^2_{min} + 1$

Mini Exercise: ML for radioactive decay solution

Probability density $p(t, \lambda) = \lambda e^{-\lambda t}$

Determine an MLH-estimate $\hat{\lambda}$ for case of *one single decay* at time t_1

- Analytically:
  - calculate $\omega = lnL = ln\, p(t_1, \lambda)$ and find $\hat{\lambda}$ from $d\omega/d\lambda = 0$

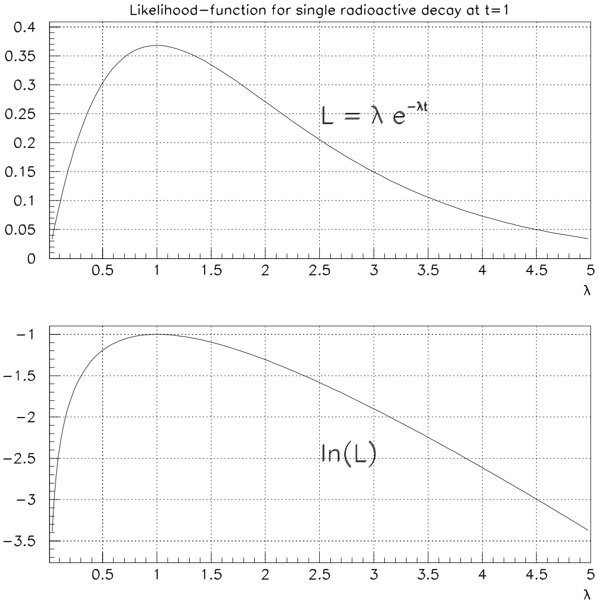
    $d\omega/d\lambda = \frac{1}{\lambda} - t_i$

    $d\omega/d\lambda = 0 \leftrightarrow \hat{\lambda} = \frac{1}{t_i}$
  - Determine an estimate for the uncertainty of $\hat{\lambda}$ from $\sigma_{\hat{\lambda}} = \left(-\frac{d^2\omega}{d\lambda^2}_{|\lambda=\hat{\lambda}}\right)^{-1/2}$ (parabola approximation of *lnL* around the maximum) $\frac{d^2\omega}{d\lambda^2} = \frac{d}{d\lambda}(\frac{1}{\lambda} - t_i) = -\frac{1}{\lambda^2} \Rightarrow \sigma_{\hat{\lambda}} = \hat{\lambda} = 1/t_i$
- graphically:

  Plot the $\chi^2 = -2lnL$ in ROOT (case $t_1 = 1$):
  - TF1 *f1 = new TF1("f1","-2*log(x)+2*x",0,5); f1->Draw();
  - Determine $\hat{\lambda}$ from the min. χ^2 and its uncertainty from $\chi^2_{min} + 1$

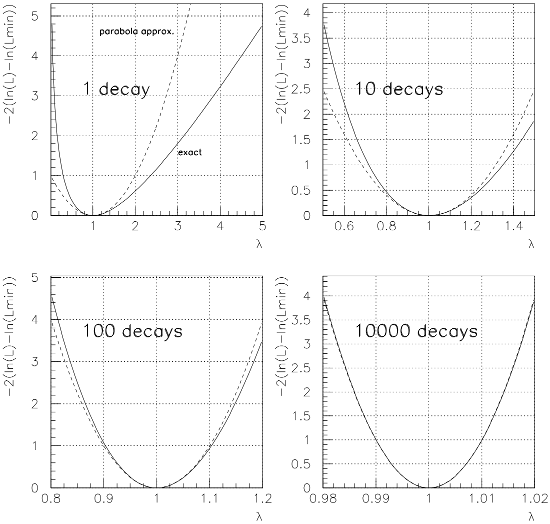
    $\lambda = 1.0^{+1.4}_{-0.6}$

Mini Exercise: ML for radioactive decay solution



Mini Exercise: ML for radioactive decay solution

Maximum Likelihood (MLH): Radioactive decay $L = \prod_{i=1}^{N} \lambda e^{-\lambda t_i}$
Define $\chi^2 = -2ln(L)$ and plot $\chi^2 - \chi^2_{min}$



- For illustration here for all cases: $\hat{\lambda} = 1$
- More decays $\rightarrow$ principal shape of L doesn't change, just zooming in!

ML application - Fractions of two processes

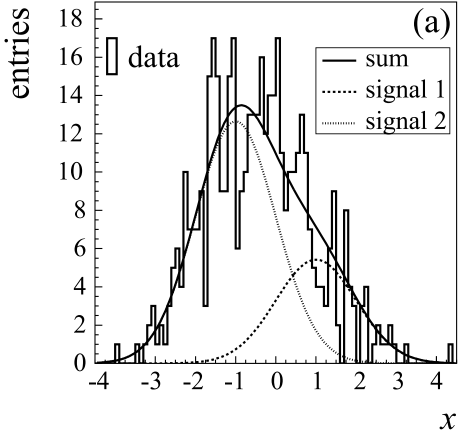
- Often two processes contribute to data (e.g. Higgs production and QCD background) $\Rightarrow$ want to determine fractions f_1 and $f_2 = 1 - f_1$
- Exploit different shapes in variable x (e.g. multivariate discriminator)
- Probability density: $p(x) = f_1 p_1(x) + (1 - f_1) p_2(x)$
- Example:
  - gaussian shapes for p_1 and p_2 with mean values of -1 and $+1$ and unit variance
  - 453 events recorded

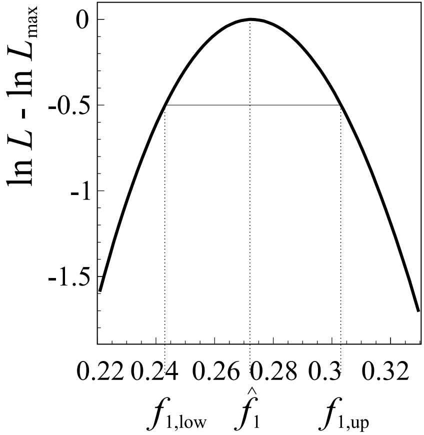
$\Rightarrow$ Likelihood function:

$$L \sim \prod_{i=1}^{453} \left[f_1\, e^{-(x_i - 1)^2/2} + (1 - f_1)\, e^{-(x_i + 1)^2/2} \right]$$

For illustration data are shown binned

unbinned ln L





$\Rightarrow$ fitted fraction $f_1 = 0.273^{+0.030}_{-0.030}$

Extended ML

- Often one wants to determine absolute rates of processes (e.g. Higgs production and QCD background)
- For repeated experiments rates will fluctuate according to Poisson statistics

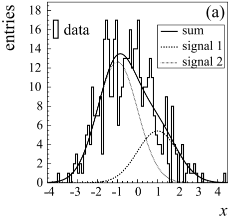
$\Rightarrow$ Introduce multiplicative factor in Likelihood:

$$L(\nu, \vec{a}) = \exp\{-\nu\} \frac{\nu^n}{n!} \prod_{i=1}^{n} p(x_i|\vec{a})$$

$$\ln L = \sum_{i=1}^{n} \ln p(x_i|\vec{a}) + n \ln \nu - \nu + const.$$

- When ν is independent of $\vec{a}$: $\Rightarrow \hat{\nu} = n$ and $\hat{\vec{a}}$ stays unaltered
- When ν is a function of $\vec{a}$: $\Rightarrow$ improved estimates can be obtained, example: $m(top)$ determination from observed $t\bar{t}$ production cross section at CMS, arXiv:1307.1907, needs theory input.

Extended ML example - Rates of two processes



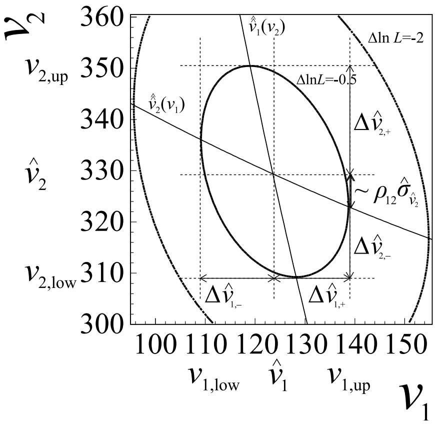
- Extended likelihood for our earlier two process example:

$$L = e^{-\nu}\,\nu^{453}\prod_{i=1}^{453}\left[f_1\,e^{-(x_i-1)^2/2}+(1-f_1)\,e^{-(x_i+1)^2/2}\right]$$

$$= e^{-\nu_1-\nu_2}\prod_{i=1}^{453}\left[\nu_1\,e^{-(x_i-1)^2/2}+\nu_2\,e^{-(x_i+1)^2/2}\right],$$

where we have used the equivalence $\nu_1 = f_1\,\nu$ and $\nu_2 = (1-f_1)\,\nu$

- Plot shows $\ln L$ contours vs ν_1 and ν_2 around $\ln L_{max}$:



- Use Profile Likelihood method to determine $\Delta\hat{\nu}_{1,-}$ and $\Delta\hat{\nu}_{1,+}$
- Profiled curve: $\hat{\hat{\nu}}_2(\nu_1)$ are the the points in ν_2 where $\ln L$ has a maximum for given fixed ν_1
- the two points where $\hat{\hat{\nu}}_2(\nu_1)$ crosses the $\Delta \ln L = -0.5$ contour
  - $\nu_{1,low} = \hat{\nu}_1 - \Delta\hat{\nu}_{1,-}$
  - $\nu_{1,up} = \hat{\nu}_1 + \Delta\hat{\nu}_{1,+}$,

  define a 68% CL interval for ν_1.
- Results: $\nu_1 = 124^{+15}_{-15}$ and $\nu_2 = 329^{+21}_{-21}$.

Profile Likelihood

- The Profile Likelihood method is an generalisation/extension of the $\chi^2_{min} + 1$ ($\equiv \ln L_{max} - 1/2$) method for one parameter a to a parameter vector $\vec{a}$ of dimension j
- $(\hat{\hat{a}}_2, \hat{\hat{a}}_3, ..., \hat{\hat{a}}_j)(a_1)$ denote the "profiled" points in $(a_2, a_3, ..., a_j)$ with maximal $\ln L$ for given fixed a_1
- the two points where $(\hat{\hat{a}}_2, \hat{\hat{a}}_3, ..., \hat{\hat{a}}_j)(a_1)$ crosses the $\Delta \ln L = -0.5$ contour define a 68% CL interval for a_1:
- the uncertainties coincide with the Hesse (= 2nd derivative) approach for multivariate gaussian likelihoods

From unbinned to binned fits (multinomial)

- Unbinned ML: $L = \prod_{i=1}^{n} p(x_i|\vec{a}) = max$

$\Rightarrow$ Can become CPU intensive for large event numbers n

- Binned fits in m bins: provide an alternative
  - Probability for events to appear in bin i:

$$p_i(\vec{a}) = \int_{x_i^{low}}^{x_i^{up}} p(x|\vec{a})\,dx; \qquad \text{note that } \sum_{i=1}^{m} p_i = 1$$

  - k_i = observed number of events in bin i; note that $\sum_{i=1}^{m} k_i = n$

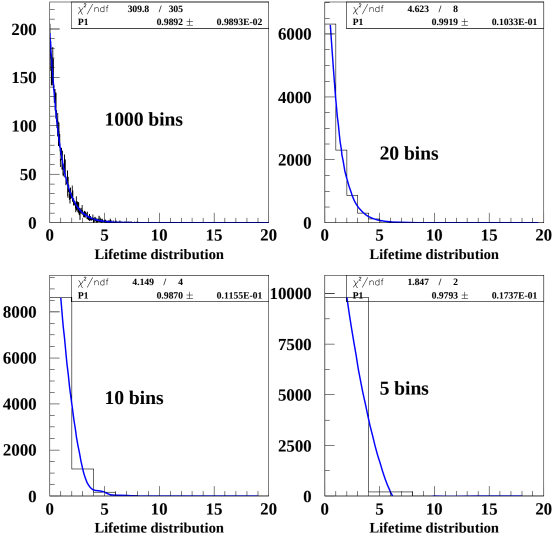
$\Rightarrow$ Multinomial statistics: $\boxed{L = n! \prod_{i=1}^{m} \frac{p_i^{k_i}}{k_i!} = max}$

  - Popular bin-centre approximation: $p_i(\vec{a}) \approx p(x_i^c|\vec{a})\Delta x_i$

    with x_i^c the bin-centre position and Δx_i the bin-width

Example binned multinomial fit: exponential decay

- 10000 decays according to $p(t,\lambda) = \lambda e^{-\lambda t}$ with true $\lambda = 1$:
- Multinomial fit with proper bin-integration

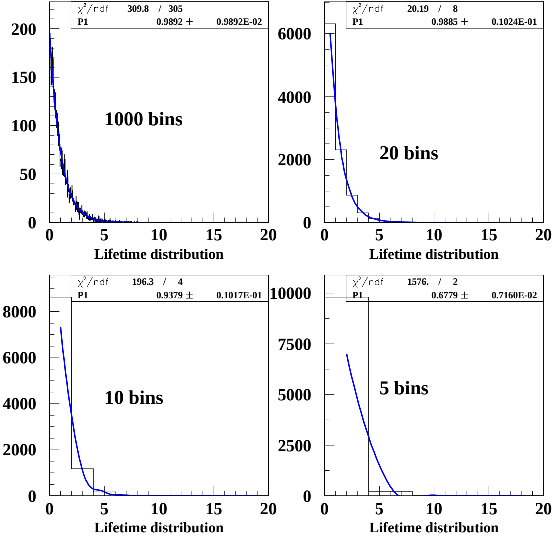


$\Rightarrow$ proper results for any binning

$\Rightarrow$ Information loss (error increase) only for very rough binning

Example binned multinomial fit: exponential decay

- 10000 decays according to $p(t,\lambda) = \lambda\, e^{-\lambda t}$ with true $\lambda = 1$:
- Multinomial fit with bin-centre approximation



$\Rightarrow$ becomes problematic at rough binning

Binned fits: from Multinomial to Poisson (extended ML)

⇒ Multinomial statistics: $L = n! \prod_{i=1}^{m} \frac{p_i^{k_i}}{k_i!} = max$

⇒ Poisson statistics: The total number of expected events ν is a free parameter

$$L = e^{-\nu} \cdot \frac{\nu^N}{N!} \cdot N! \prod_{i=1}^{m} \frac{p_i^{k_i}}{k_i!} = \prod_{i=1}^{m} e^{-\nu_i} \frac{\nu_i^{k_i}}{k_i!} = max\,, \quad \text{with } \nu_i = \nu p_i$$

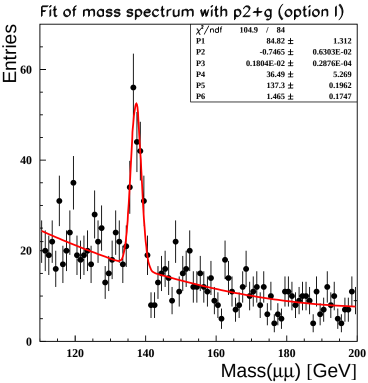
Poisson is usually a good choice for fits to histograms!

Binned fits: estimator choices

Histogram with m bins:

k_i = number of observed events in bins

ν_i = number of expected events (depending on fit parameters)



- Poisson Likelihood:

$$\tilde{\chi}^2 = -2\ln L = 2[\sum_{i=1}^{m} \nu_i - k_i \ln \nu_i]$$

- Neyman χ^2:

$$\chi^2 = \sum_{i=1}^{m} \frac{(k_i - \nu_i)^2}{k_i}$$

- Pearson χ^2:

$$\chi^2 = \sum_{i=1}^{m} \frac{(k_i - \nu_i)^2}{\nu_i}$$

- Both χ^2 estimators have problems: biased results, cannot treat bins with $k_i = 0$, $\Rightarrow$ use Poisson likelihood!

Summary:

- Maximum Likelihood method is a powerful tool to estimate underlying physics parameters from data
- Choose the appropriate likelihood function for your problem:
  - χ^2
  - Unbinned: normal likelihood or extended
  - Binned: multinomial or Poisson,
  - Binomial (not discussed here)
  - etc.
- Estimate 68% CL intervals from parameter points where $\ln L$ drop by 0.5 from maximum (or by 1.0 if you use $\tilde{\chi}^2 = -2 \ln L$)
  For many parameters use profile likelihood