Modern Methods of Data Analysis

Lecture VII (26.11.07)

Contents:

- Maximum Likelihood (II)

Exercise: Quality of Estimators

- Assume hight of students is Gaussian distributed. You measure the size of N students. Which of the following estimators is a) consistent? b) unbiased? c) efficient?

- 1) Add all measurements, divide by N
- 2) Add the first 10 measurements, divide by 10
- 3) Add all measurements, divide by N-1
- 4) Assume 1.8 m
- 5) Add smallest & largest meas., divide by 2
- 6) Add every second measurement, divide by N/2

Re: Maximum Likelihood (I)

- N independent measurements of a random variable x_i distributed according to f(x|a), with unknown parameter a
- Want to get the best estimate $\hat{a}$ for the true parameter a.
- Likelihood = joint probability: $$\boxed{L(a) = \prod_{i=1}^{n} f(x_i|a)}$$
- According to the ML principle the best estimation of a is the value $\hat{a}$ which maximizes L(a), i.e., which maximizes the probability to obtain the observed data
- The maximum is computed by $dL(a)/da = 0$
- $\hat{a}$ is an efficient (often biased but consistent) estimator

Exercise

- If you have two independent measurements of equal accuracy, one of $\sin\Theta$ and one of $\cos\Theta$, find the ML estimate of Θ.

Error on Estimate (I)

- Evolve (negative) Log-Likelihood around $a = \hat{a}$

$$-\ln L(a) = -\ln L(\hat{a}) - \left.\frac{d\ln L}{a}\right|_{a=\hat{a}}(a-\hat{a}) - \frac{1}{2}\left.\frac{d^2\ln L}{da^2}\right|_{a=\hat{a}}(a-\hat{a})^2 + \ldots$$

$$-\ln L(a) \approx -\ln L(\hat{a}) - \frac{1}{2}\left.\frac{d^2\ln L}{da^2}\right|_{a=\hat{a}}(a-\hat{a})^2$$

$$L(a) \approx const * e^{\frac{1}{2}\frac{d^2\ln L}{da^2}(a-\hat{a})^2}$$ L(a) is Gaussian distributed!

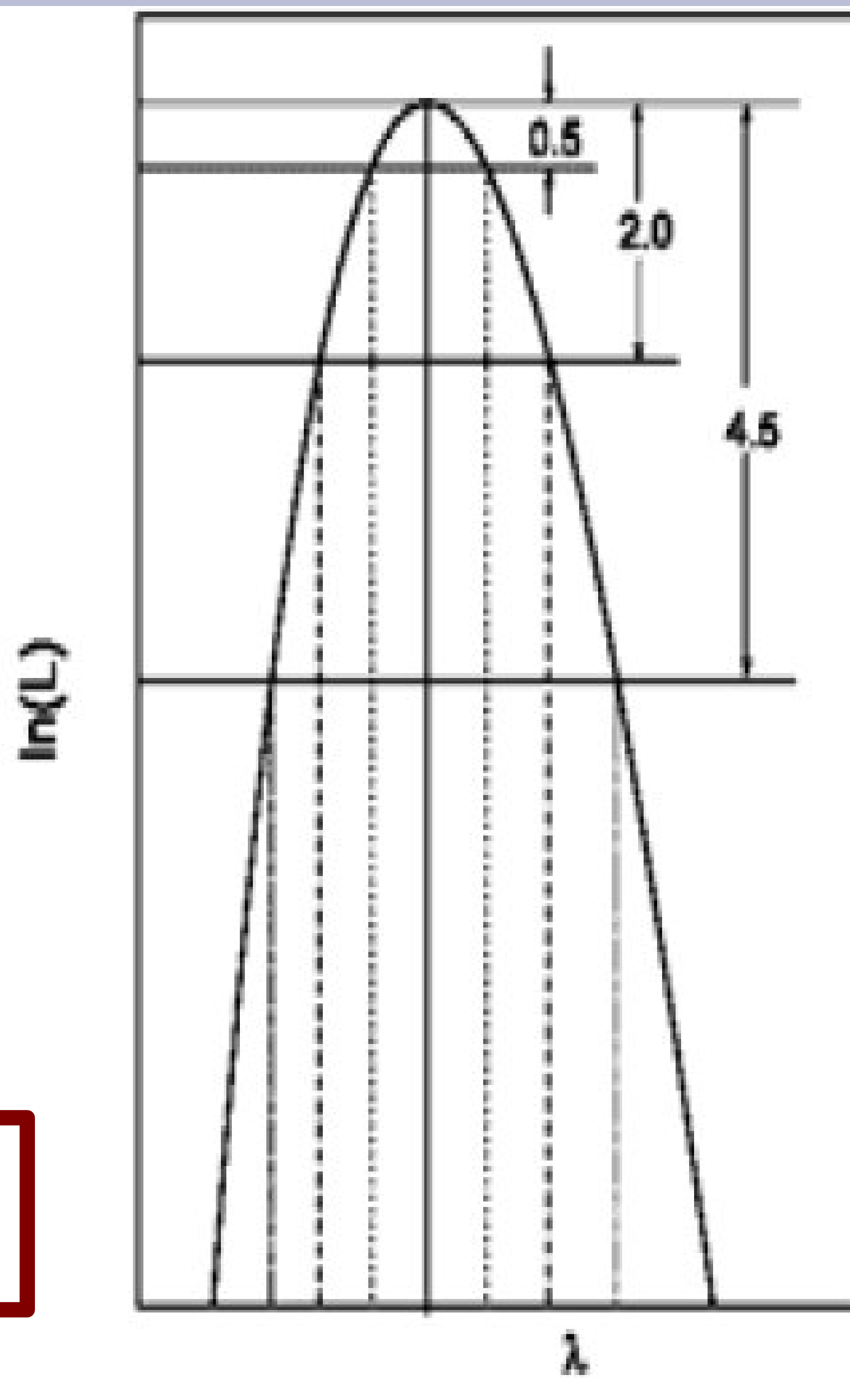
$$\sigma^2 = \left.\frac{d^2\ln L}{da^2}\right|_{a=\hat{a}}$$

$$\boxed{-\ln L(\hat{a}\pm n\sigma) = -\ln L(\hat{a}) + \frac{1}{2}n^2}$$

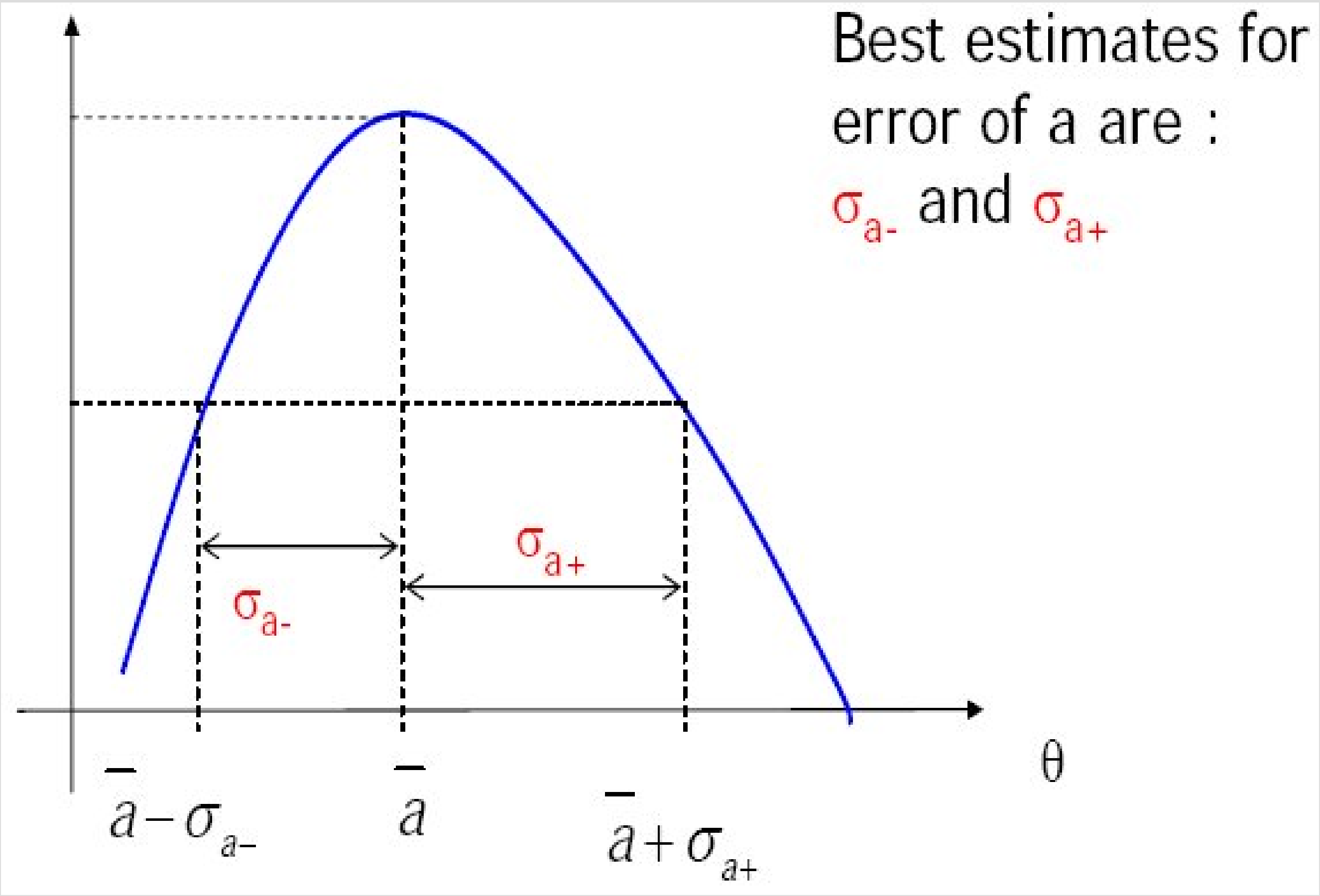
Error on Estimate (II)

- This means Log-Likelihood decreases
  - for ±1σ by ±0.5
  - for ±2σ by ±2.0
  - for ±3σ by ±4.5
- in case of too small n, Log-Likelihood is not parabola, but rather asymmetric
  - quote asymmetric uncertainties

$$\boxed{-\ln L(\hat{a} \pm n\sigma) = -\ln L(\hat{a}) + \frac{1}{2}n^2}$$



Error on Estimate (III)

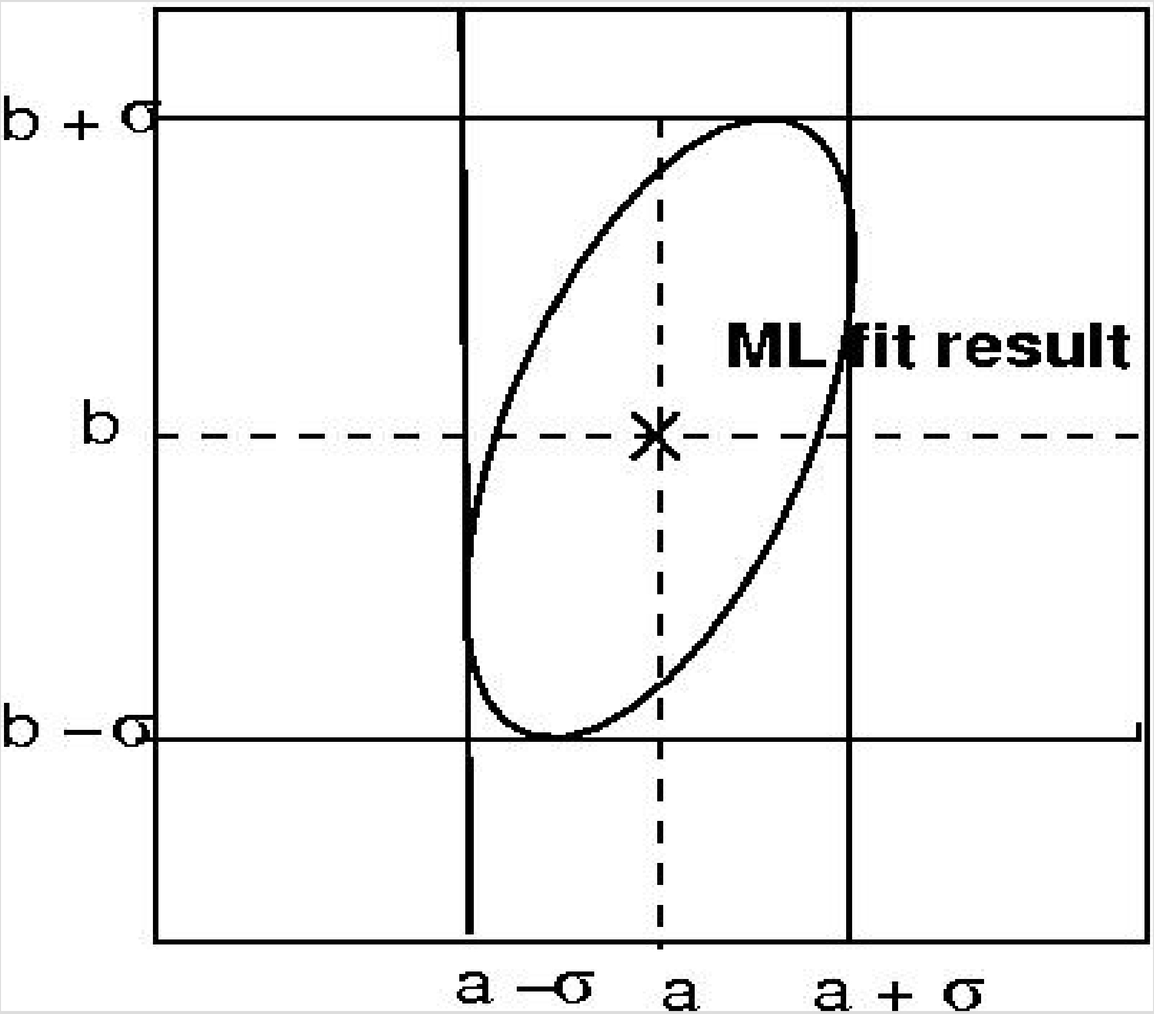


Idea:
There is some transformation to make log L parabolic. Due to Invariance of log L definition for 1,2,3 σ are still valid.

For non parabolic distribution the 2σ interval is not necessarily twice as large as the 1σ interval!

Error Estimate for Multiple Parameters

$$cov(a,b)^{-1} = -\frac{\delta^2 \ln L}{\delta a \delta b}\Big|_{a=\hat{a},b=\hat{b}}$$



Error on Estimate (IV)

- In many not-so-easy cases, a Monte Carlo based Method is used:
  - simulate with a toy-MC the experimental measurements many times (use realistic resolutions etc.)
  - As a true value for parameter a in the MCs, $\hat{a}$ obtained from data is a sensible choice.
  - determine in every pseudo-experiment i the estimator $\hat{a}_i$
  - determine the sample variance of $\hat{a}$ from the many pseudo-experiments
  - this also can be used to check/correct for bias

Example: Signal Enhancement (I)

- Search for special events in two independent channels, uncertainties on expected numbers due to limited MC sample for study (pure statistical)

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	4.0±0.6	24.0±6.0

Model includes factor f: $\mu_i = f S_i + B_i$

Question: Are both measurements compatible with f=1, which is standard expectation from theory?

Signal Enhancement (II)

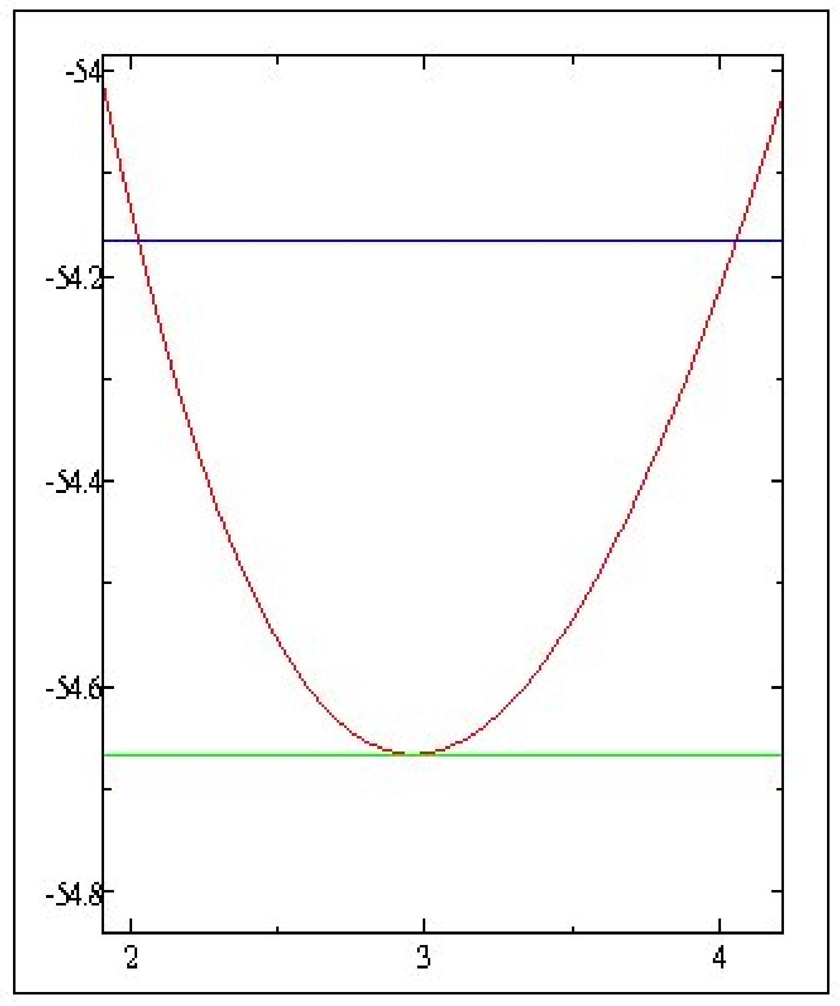
- Use ML method to obtain best estimate for factor f from all data!
- Assume Poisson distribution of data n_i with mean values given by theory model:

$$L(f) = P(n_1|\mu_1)P(n_2|\mu_2) = \frac{e^{-\mu_1}\mu_1^{n_1}}{n_1!}\frac{e^{-\mu_2}\mu_2^{n_2}}{n_2!}$$

$$F = -\ln(L(f)) = \sum_{i=1}^{2}(\mu_i - n_i \ln \mu_i) + const$$

Signal Enhancement (III)

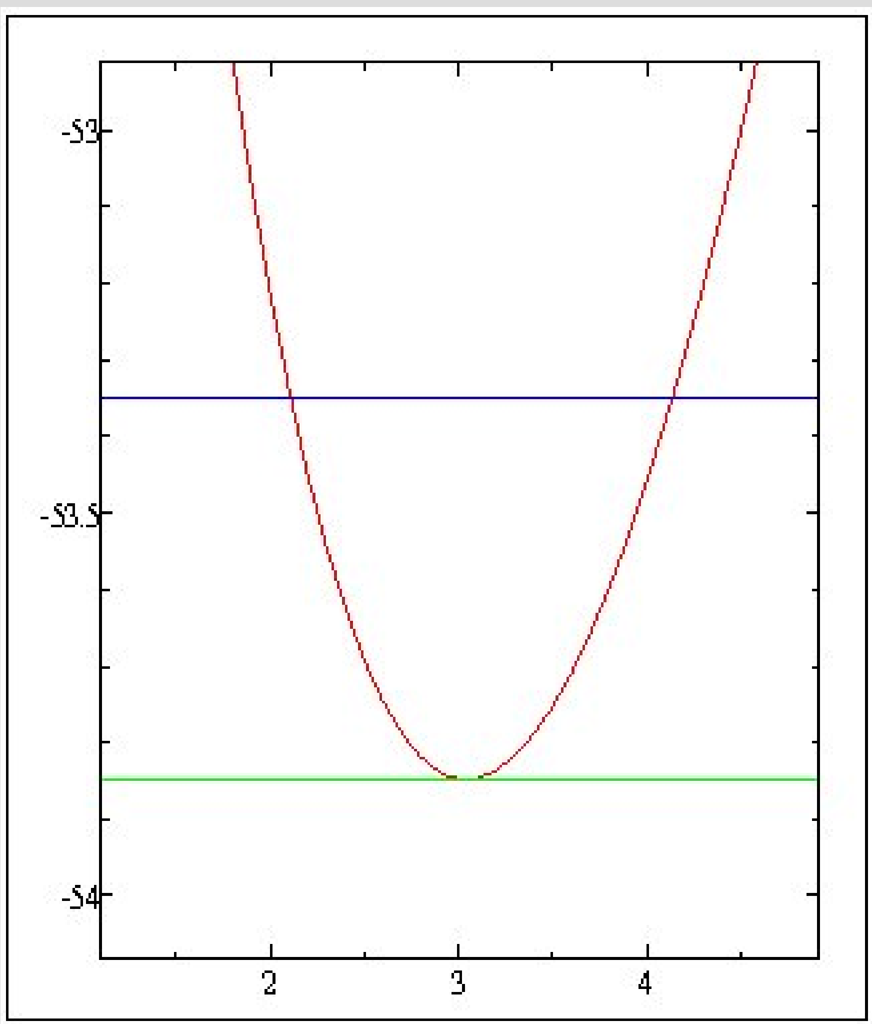
- Result for factor: $f = \mathbf{2.96^{+1.09}_{-0.94}}$



Note: statistical fluctuations for model predictions ignored

Signal Enhancement (III)

$$P(n_1|f) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}} \frac{e^{-x} x^{n_1}}{n_1!} dx$$



$$\sigma_1 = \sqrt{\sigma_{S_1}^2 + \sigma_{B_1}^2}$$

$$L(f) = P(n_1|\mu_1)P(n_2|\mu_2)$$

$$f = 3.05^{+1.09}_{-0.94}$$

Combination of Measurements with LH

- first experiment measure x_i with pdf f(x|a), second experiment measures y_i with pdf g(y|a). Functions f and g can be different but have to depend on same true parameter a:

$$L(a) = \prod_{i=1}^{n} f(x_i|a) \prod_{i=1}^{m} g(y_i|a) = L_x(a) * L_y(a)$$

- combined likelihood is product of single likelihoods.
- alternatively: $\ln L(a) = \ln L_x(a) + \ln L_y(a)$
- often used to combined complex analysis from two different experiments (example later)

Extended Maximum Likelihood (I)

- random variable x distributed according to f(x,Θ), $\theta = (\theta_1, ..., \theta_m)$. Often number of observed events n is itself a Poisson random variable with mean value v.

$$L(\nu, \theta) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i, \theta) = \frac{e^{-\nu}}{n!} \prod_{i=1}^{n} \nu f(x_i, \theta)$$

This is called extended Likelihood function.

$$\ln L(\nu, \theta) = -\nu(\theta) + \sum_i \ln[v(\theta) f(x_i, \theta)] + const$$

1) v is independent of Θ
2) v is a function of Θ

Extended Maximum Likelihood (II)

- v is independent of Θ:

$$\frac{d\ln L}{d\nu} = -1 + \sum_i \frac{1}{\nu} \quad \rightarrow \quad \hat{\nu} = n$$

$$\frac{d\ln L}{d\theta}: \quad \text{same as normal LH}$$

- v depend on Θ: E.g. measurement of angular distribution, which depend on mass of particle. Number of observed events is function of cross section which depend as well on mass of particle. Adding v as measurement to LH improves resolution on Θ (on mass), additional information is used!

Binned Maximum Likelihood (I)

- For very large data samples, the log-likelihood function becomes difficult to compute
- Compute the number of expected entries in a bin

$$\nu_i(\theta) = n_{tot} \int_{x_{min,i}}^{x_{max,i}} f(x,\theta)dx$$

$$f(n,\nu) = \frac{n_{tot}!}{n_1!...n_N!} \left(\frac{\nu_1}{n_{tot}}\right)^{n_1} .... \left(\frac{\nu_N}{n_{tot}}\right)^{n_N}$$

$$\ln(L(\theta)) = \sum_{i=1}^{N} n_i \ln \nu_i(\theta) + const$$

- Uncertainties are slightly larger than in unbinned fit
- limit of very small bins -> unbinned fit -> no problems with low number of entries

Binned Maximum Likelihood (II)

- One may regard the total number of entries n_{tot} as random variable from a Poisson distribution with mean ν_{tot} .

$$f(n,\nu) = \frac{\nu_{tot}^{n} e^{-\nu_{tot}}}{n_{tot}!} \frac{n_{tot}!}{n_1!...n_N!} \left(\frac{\nu_1}{\nu_{tot}}\right)^{n_1} ... \left(\frac{\nu_N}{\nu_{tot}}\right)^{n_N}$$

mit $$\nu_{tot} = \sum_{i=1}^{N} \nu_i, \; n_{tot} = \sum_{i=1}^{N} n_i$$

$$f(n,\nu) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

$$\nu_i(\nu_{tot},\theta) = \nu_{tot} \int_{x_{min}}^{x_{max}} f(x,\theta)dx$$

Binned Maximum Likelihood (III)

- Independent Poisson distribution of each bin or Poisson distribution of overall number of entries plus multinomial distribution!

$$\ln L(\nu_{tot}, \theta) = -\nu_{tot} + \sum_{i=1}^{N} n_i \ln \nu_i(\nu_{tot}, \theta)$$

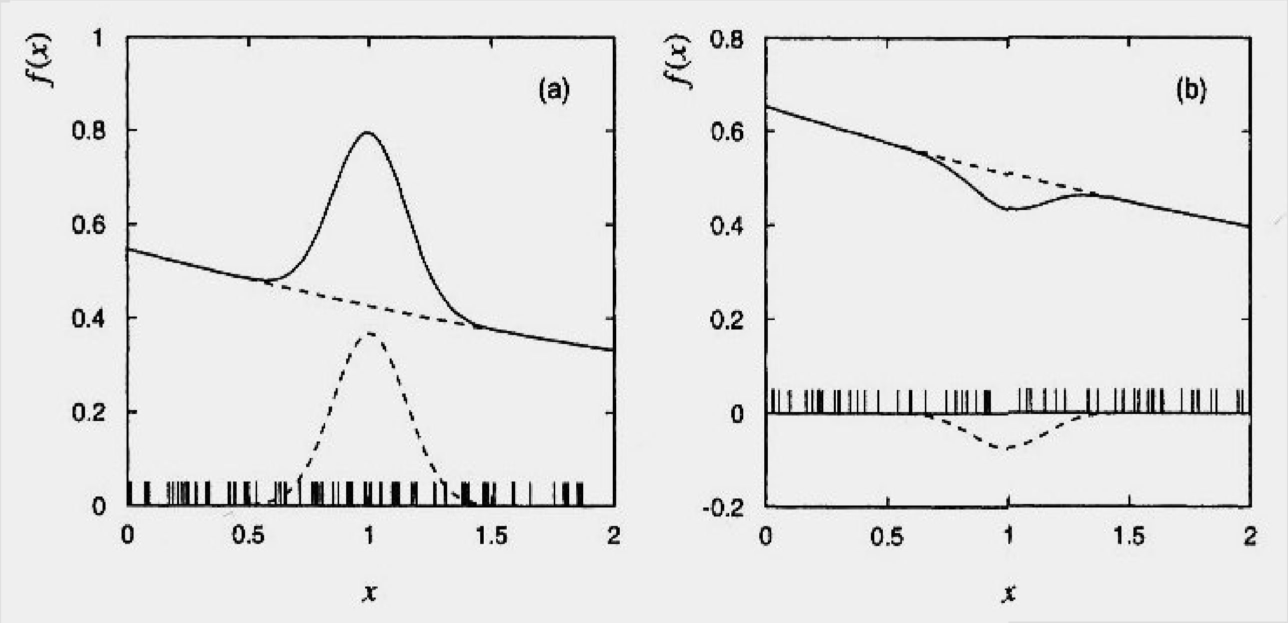
- This is extended LH for binned case. As before: if there is any relation between ν_{tot} & θ uncertainties on θ get smaller, otherwise best estimator for $\nu_{tot} = n_{tot}$. Uncertainties on θ stay the same.

Signal to Background

- Likelihood often sum of two or more components (signal + background)

$$L(\theta) = \theta * f_S(x) + (1-\theta) * f_B(x)$$

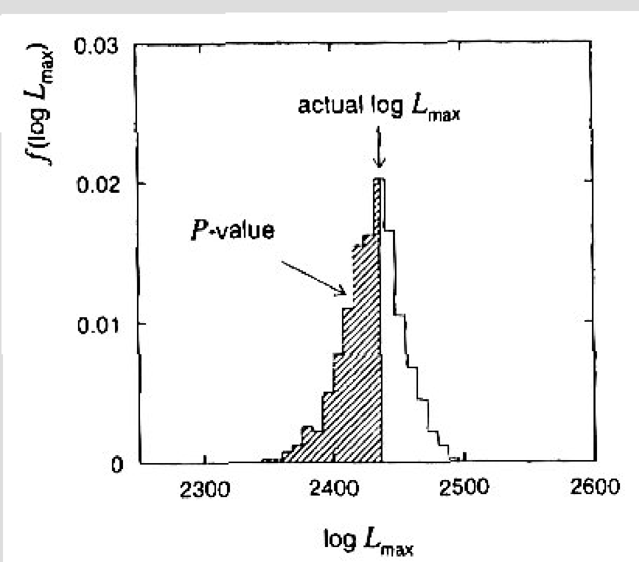
- e.g. toy MC with 6 signal & 60 background events

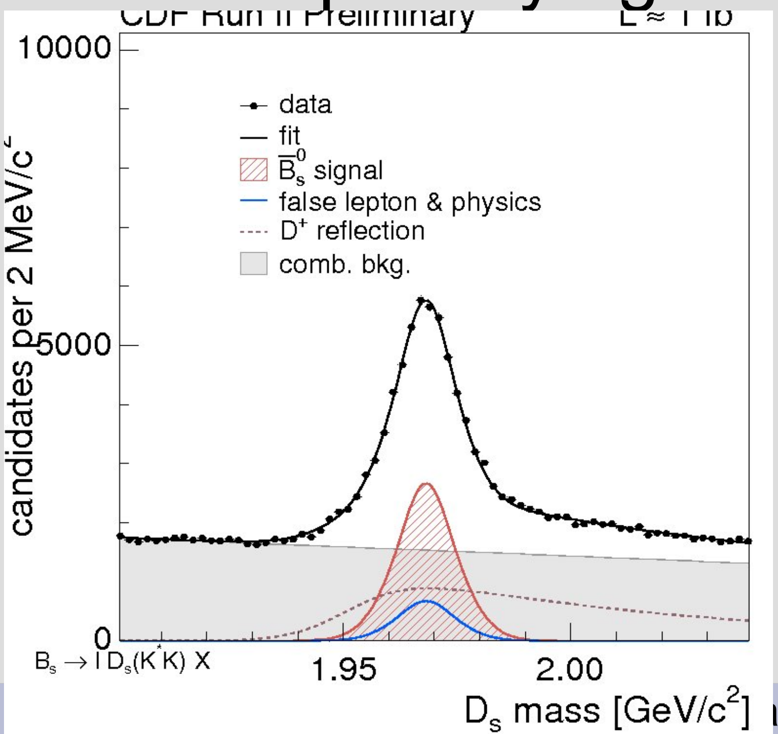


Although negative number of # signal unphysical, need to use them, when combining with other experiments otherwise bias.

Goodness of Fit

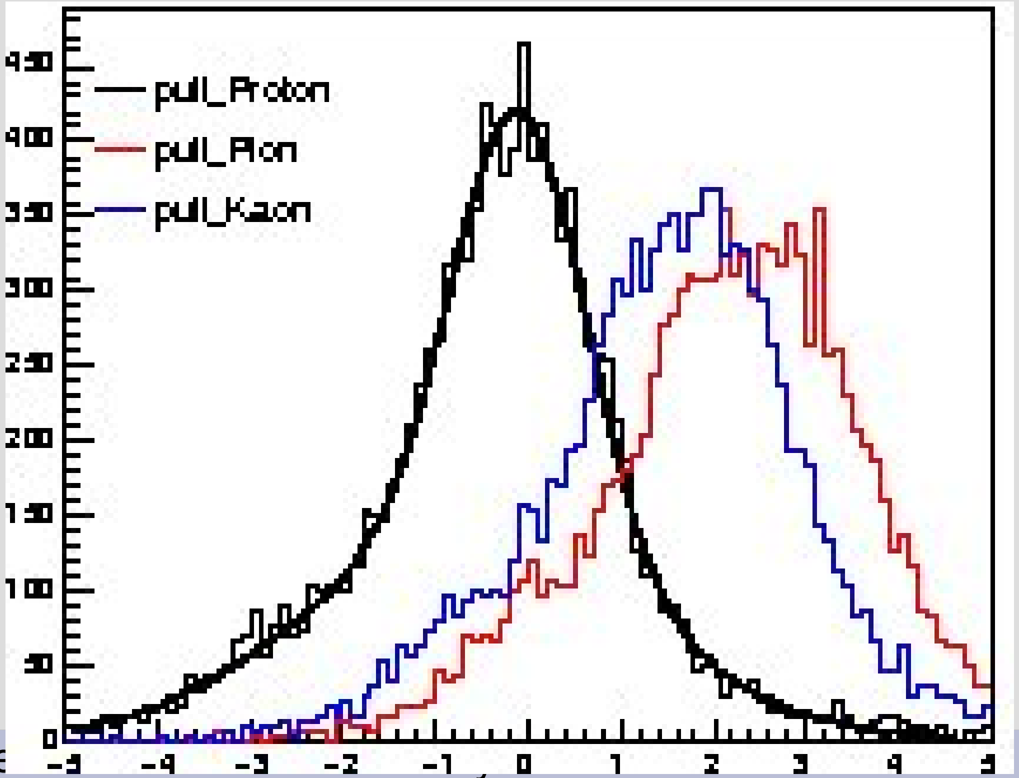
- LH does not provide any information on the Goodness of the fit.
- This has to be checked separately.
  - e.g. simulate toy MC according to estimated pdf (using fit results from data as “true” parameter values) compare max Likelihood value in toy to the one in data
  - draw data in (binned) histogram, “compare” distribution with result of LH fit. [Methods to quantify agreement in later lecture]





Likelihood Ratio

- Observed stable particle has to be kaon, pion or proton. One of the hypothesis has to be true -> which one is the most likely one.
- The relative probability for proton is given by:
  - L(data| proton)/L(data/proton or pion or kaon)
  - be aware of a priori probabilities



It is crucial to well describe tails in the distribution!

Exercise: Throwing a Coin

- There are two type of coins, which are not distinguishable from looking at them
  - Type 1: p(head) = 0.9, p(number) = 0.1
  - Type 2: p(head) = 0.1, p(number) = 0.9
- Throwing 10 times the coin give 6 times head and 4 times number
- If you don't know anything about the properties of the coin before, what is the best estimator for p using ML?
- Is the result of the experiment consistent with any of the two type of coins?
- Compute the Likelihood ratio of L(Type1)/L(Type2).

Comments (I)

- Advantages
  - no binning needed, retains full information
  - also possible with binned data, no problem with zero entries
  - good method to combine results from different experiments (simply add the log-likelihood functions)

- ML estimates are
  - Gaussian for large N
  - consistent (asymptotically unbiased), i.e. bias disappears for large n
  - efficient, reaching the minimal variance bound

- this is why ML is very popular!

Comments (II)

- Disadvantages:
  - can be extremely CPU-time consuming for large sample
  - Need to know pdf f(x|a), but often pdf very complicated or actually not known
  - no general way to estimate "goodness of fit"
    - compare simply fitted pdf with data distributions
    - perform MC experiments to get distrib. of L(max)
  - for smaller n, there is generally a bias. Important to study ML behavior in toy-MC and correct for bias

  - ML method requires normalization of f(x|a). This has to be done at every step in the minimization/maximization. Programs like MINUIT are doing this numerically (CPU intense).