Learning goals of the week:

- Maximum Likelihood Fit
  - perform a likelihood fit
    - estimate the uncertainty on the best fit value
    - fit binned data
    - fit constrained parameters
- Understand similarities/differences wrt to Least Squares

# Week 6

# Likelihood

# Maximum Likelihood

The maximum likelihood method is intuitively the best you can do to setup a generic estimator:

- take N independent measurements of random variable $x_i$, distributed according to a known pdf f(x|a), with an unknown parameter a.

- the probability to measure $x_1$, $x_2$, ... $x_N$ (in any order) is:

  $P(x_1 \text{ in } [x_1, x_1+dx_1]) = f(x_1|a)dx_1$

  $P(x_2 \text{ in } [x_2, x_2+dx_2]) = f(x_2|a)dx_2$

  ...

  $P(x_N \text{ in } [x_N, x_N+dx_N]) = f(x_N|a)dx_N$

  simultaneously for each i

- the joint probability to obtain these measurements is the product of the probabilities

  $f(x_1|a)dx_1\ f(x_2|a)dx_2 \ldots f(x_N|a)dx_N$

We define Likelihood L(a) the joint probability

$$L(a) = \prod_{i=1}^{N} f(x_i|a)$$

The best estimate of a is the value that maximizes the probability of each measurement, hence that maximizes the likelihood:

L(a) = maximum

$$\frac{dL(a)}{da} = 0$$

# Maximum likelihood

Remarks:
The Likelihood is a sampling function, ie. a random variable, not the pdf of the true parameter a. If that was the case we would just compute expectation value of a (and all its moments)

The Likelihood has to be normalized for every value of a

The likelihood is not a probability: it's a product of probability density functions:
—>the probability would be = L $\Pi dx_i$

For a given sample or a given set of measurements $x_i$ the Likelihood function L=L(a) $\Pi dx_i$ is the probability to observe a certain set of data points {xi} given the parameter "a".

In L=L(a), "a" is the "parameter of interest" (p.o.i.), but L could depend on several other parameters $\vartheta$: L(a| $\vartheta$)

Notation $\hat{a}$ is the maximum likelihood estimator (MLE) of a

Notation: in general we will neglect the explicit dependence on the data points

We can be interested in several p.o.i. at the same time: L(**a**).
In this case we will maximize simultaneously:

$$\frac{\partial L(a_i)}{\partial a_i} = 0$$

# Maximum likelihood

Practicalities:

For numerical reasons we will work with -ln L(a). L(a) is the product of numbers in [0,1] (pdf). Multiplying several of those will reach the numerical precision of the computer. Instead of multiplying $f(x_i)$ we will sum ln $f(x_i)$

$$l(a) = \ln L(a) = \sum_i \ln f(x_i|a)$$

We will minimize -L (negative L) instead of maximizing L, because the numerical libraries are typically written to find minima instead of maxima

# Example: lifetime/exponential fit

Compute the maximum Likelihood estimator for an exponential distribution (use ln L)

$$f(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

t = proper decay time
τ = lifetime

Write the log-likelihood

$$\ln L(\tau) = l(\tau) = \sum_i \ln f(t_i;\tau) = \sum_i (\ln \frac{1}{\tau} - \frac{t_i}{\tau})$$

Find the derivative:

$$\frac{\partial l}{\partial \tau} = -\frac{n}{\tau} + \frac{1}{\tau^2} \sum t_i$$

Set it to zero and solve for the lifetime:

$$\hat{\tau} = \frac{1}{n} \sum_i t_i$$

which is simply the average of the measured proper decay times

NB: the bias decreases as 1/n. The ML estimator is asymptotically unbiased.
General property: likelihood estimators are unbiased.

# Example: Gaussian pdf

Take a Gaussian pdf: $$f(x_i;\mu) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_i}\cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma_i}\right)^2}$$

The likelihood function for the estimator of the mean is:

$$l(\mu) = \ln L(\mu) = \ln \prod_i f(t_i;\mu) = \sum_i \ln f(t_i;\mu) = \sum_i \left( \ln \frac{1}{\sqrt{2\pi}} - \ln \sigma_i - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma_i}\right)^2 \right)$$

which maximized gives:

$$\frac{dl(\mu)}{d\mu} = \frac{d}{d\mu}\sum_i -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma_i}\right)^2 = \sum_i \frac{x_i-\mu}{\sigma_i^2} = 0$$

$$\hat{\mu} = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$

The estimator of the mean is the weighted mean of the sample with weights $1/\sigma_i^2$

# Example: Poisson pdf

Take a Poisson distribution with expected mean λ. The likelihood estimator of λ is given by:

$$l(\lambda) = \sum_i \ln \frac{\lambda^{r_i}}{r_i!} e^{-\lambda} = \sum_i \ln \lambda^{r_i} - n\lambda - \sum_i \ln r_i! = \ln \lambda \cdot \sum_i r_i - n\lambda - \sum_i \ln r_i!$$

Taking the derivative and setting it to zero provides the MLE for λ:

$$\hat{\lambda} = \frac{1}{n} \sum_i r_i$$

Which is the mean of the measurements

# MLE uncertainty

Expand around the maximum of the Likelihood

$$F(\theta) = -\ln L(\theta) = F(\hat{\theta}) + \frac{1}{2}\frac{d^2F}{d\theta^2}|_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2 + \cdots$$

$$L(\theta) \sim const \cdot \exp\left(-\frac{1}{2}\cdot\frac{d^2F}{d\theta^2}|_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2\right) := const \cdot \exp\left(-\frac{(\theta-\hat{\theta})^2}{2\sigma^2}\right)$$

In the neighbour of the maximum we can approximate the function with a gaussian.
Comparing the exponents:

$$\sigma^2(\hat{\theta}) = \frac{1}{d^2F/d\theta^2|_{\theta=\hat{\theta}}}$$

ie: Variance(â)=inverse of the 2nd derivative of the Log-Likelihood

If L is Gaussian, then Log-Likelihood is a parabola (in general true for N—>∞).
The value of F(a)= –ln L around minimum at a = â ± n·σ is: (replace in the expansion)

$$F(\hat{a} \pm n\cdot\sigma) = F(\hat{a}) + \frac{1}{2}n^2$$

# MLE uncertainty

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

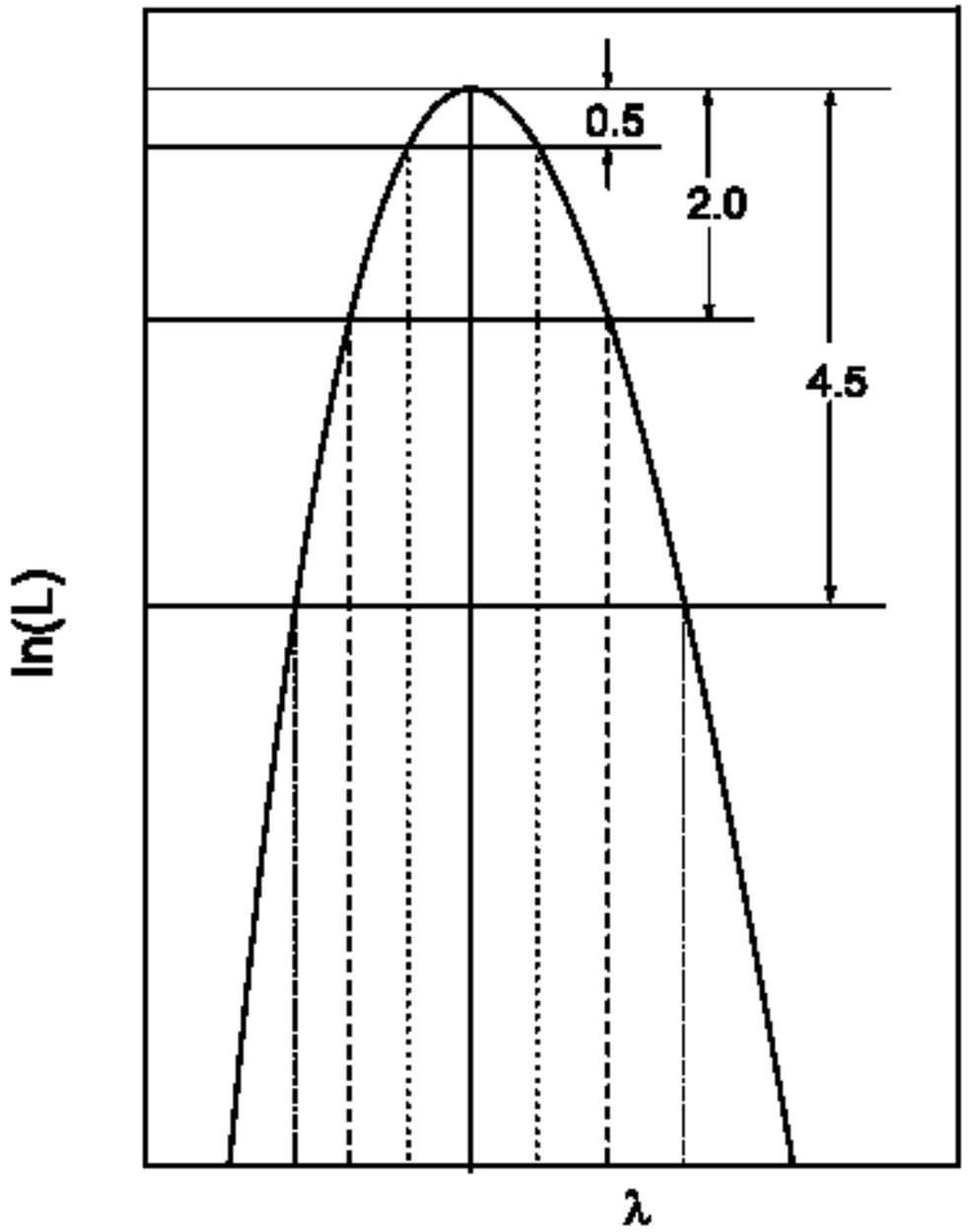
This means that Log-Likelihood decreases

for ±1σ by ±0.5

for ±2σ by ±2.0

for ±3σ by ±4.5

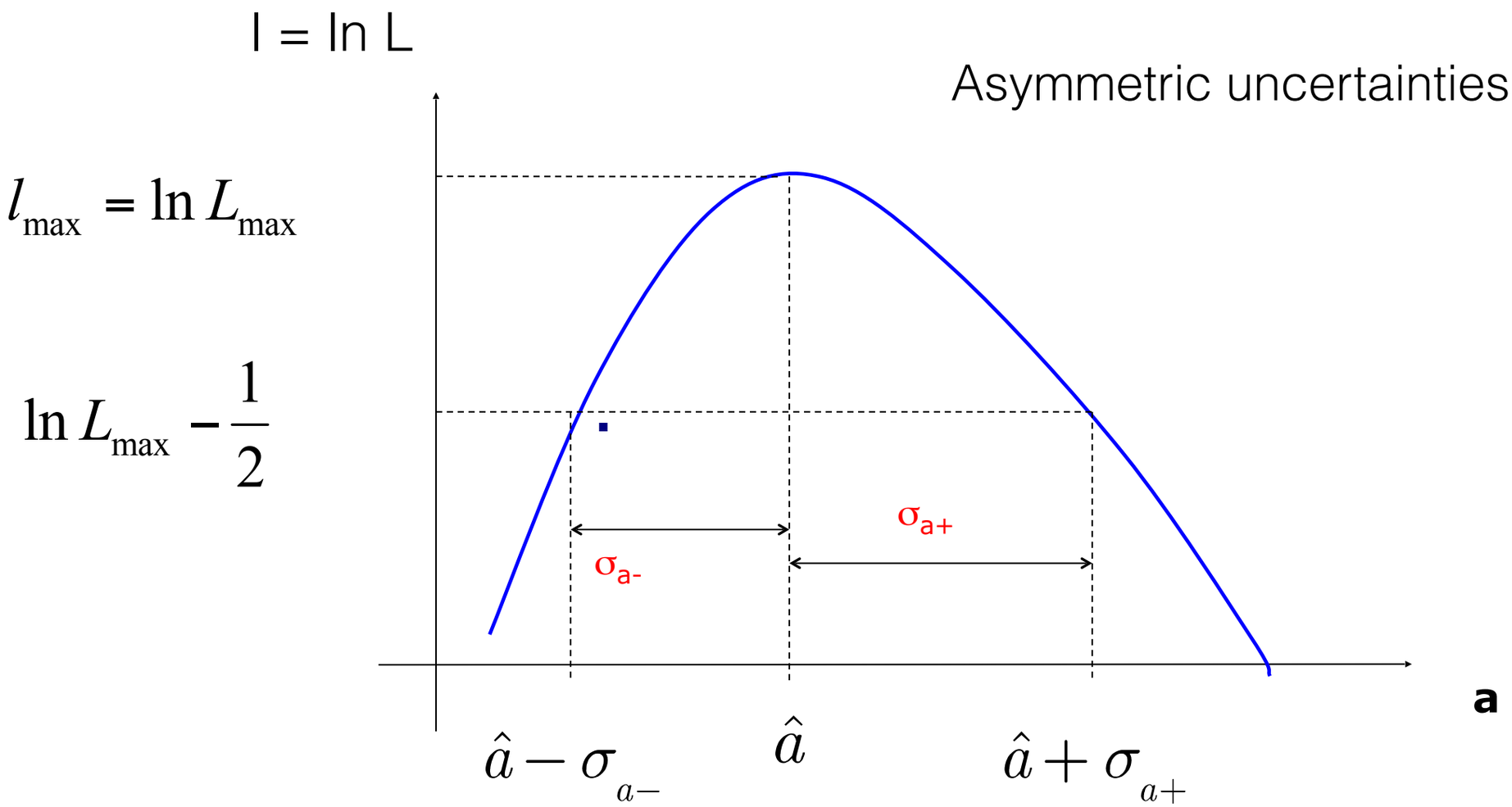
from its maximum



# MLE uncertainty

If the log-Likelihood is not parabolic, typical case for low statistics samples where CLT does not apply, we can compute the uncertainty numerically and obtain asymmetric uncertainties

(change a by nσ from its ML-estimate until $\ln(L_{max})$ decreases by $n^2/2$ )




Best estimates for error of a are : $\sigma_{a-}$ and $\sigma_{a+}$

# Fisher information

"The precision of the estimation should be greater if we have more information."

The definition of Fisher information

$$\sigma^2(\hat{\theta}) = \frac{1}{d^2F/d\theta^2|_{\theta=\hat{\theta}}} = \frac{1}{I(\hat{\theta})}$$

The variance is the inverse of the second derivative of the likelihood,
i.e. the inverse of the information.

To reduce the uncertainty of a MLE we need to increase the information,
but that is limited by the "minimum variance bound" (—> see backup slides)

# Extended MLE

When the number of events is unknown, we can fit for it.

We need to multiply "extend" the Likelihood by a Poisson term modelling the probability to obtain n observed events when v are expected:

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

and the log-Likelihood becomes:

$$l(x;\theta) = \sum_{i=1}^{n} \ln f(x;\theta) \rightarrow l_E(x;\theta,\nu) = \sum_{i=1}^{n} \ln \nu f(x;\theta) - \nu + const$$

# Binned MLE

The likelihood considered so fa is called “unbinned”: we use the maximum information by using each single point in the dataset.

This can be time consuming for large datasets (each time you compute the NLL for a value of $\vartheta$ you need to loop over all points)

The expected number of events in a bin $[x_i^{min}, x_i^{max}]$ is given by the integral of the pdf

$$\nu_i = \int_{x_i^{min}}^{x_i^{max}} f(x_i;\theta)dx \qquad \nu = (\nu_1, \ldots, \nu_N)$$

the histogram is an N-dimensional random vector described by a multinomial pdf. (generalization of the binomial distribution with k-types of outcomes instead of only 0 / 1)

$$P(r_1, ..., r_k; n, p_1, ..., p_k) = \left(\frac{n!}{r_1!...r_k!}\right) p_1^{r_1}...p_k^{r_k}$$

Assuming the total number of events is given $N_{tot}$ :

$$f_{comb}(\mathbf{n};\nu) = \frac{n_{tot}!}{n_1!\cdots n_N!}\left(\frac{\nu_1}{n_{tot}}\right)^{n_1}\cdots\left(\frac{\nu_N}{n_{tot}}\right)^{n_N}$$

NB: the dependence on $\vartheta$ is in the model of the expected number of events $\nu_i=\nu_i(\vartheta)$ ($n_{tot}$ and the observed number of events per bin do not depend on $\vartheta$ !)

# Binned MLE

The binned log-Likelihood is written as:

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

Remarks:

- the binned MLE reduces to the unbinned case for large number of bins (when each bin contains one entry)
- contrary to the least square method, there is no problem with bins with zero entries ($n_i = 0$) —> provided the expected number of events $v_i > 0$ (zero or negative value of the pdf will produce infinities)

# Combining measurements

Different measurements of the same parameter $\vartheta$ can be combined using the ML principle. The idea is to maximize the likelihood of each of the measurement for both experiments, i.e. simply multiply the two likelihoods:

$$L(\theta) = \underbrace{\prod_{i=1}^{n} f(x_i;\theta)}_{\text{experiment 1}} \cdot \underbrace{\prod_{i=1}^{m} g(y_i;\theta)}_{\text{experiment 2}} = L_x(\theta) \cdot L_y(\theta)$$

When combining experiments in this way it is important to look for possible correlations among the parameters of the likelihoods

# Fit with constraints

Often the value of the parameter of interest is limited by physical constraints (e.g. mass > 0) or by information gathered by other measurements (e.g. uncertainty on calibrations)

In general the constraints can be implemented as Lagrange multipliers
(you have encountered this when studying implicit equations $\vec{g}(\vec{\theta}) = 0$ )

The MLE with the constraint $\vec{g}(\vec{\theta}) = 0$ can be found by maximizing

$$F(\vec{x}; \vec{\theta}, \vec{\alpha}) = \ln L(\vec{x}; \vec{\theta}) + \vec{\alpha}\vec{g}(\vec{\theta})$$

with respect to both $\vec{\theta}$ and $\vec{\alpha}$

Basically you add to the ln L (or multiply L by) the constrain on the parameter of interest.

**Example** Take the likelihood $L(x; \theta_1; \theta_2)$ and say we want to estimate $\theta_1$ but we know from a different measurement that $\theta_2$ has a a value $\bar{\theta}_2 \pm \sigma_{\theta_2}$. We can introduce the constraint on $\theta_2$ by simply multiplying the likelihood by an gaussian function centred at $\bar{\theta}_2$ with width $\sigma_{\theta_2}$ (or adding the equivalent parabolic term to the log-likelihood). □

# Comments

- The ML is the most widely used estimation method because in the large statistics limit the MLE is:
  - unbiased (MLE normally distributed around the true value $\vartheta$)
  - efficient (minimum variance)
  - consistent

But remember:

- for small samples it has no optimal properties. In particular it can be biased ! other estimators may have greater concentration around the true parameter-value.

- can be extremely CPU-time consuming for large samples (use binned likelihoods)

- no general way how to estimate “goodness of fit”
  compare simply fitted pdf with data distribution, or perform Monte Carlo experiments to obtain distribution of $L_{max}$

# From MLS to LSE

Take the usual example of the dataset (xi, yi) where the uncertainty on $x_i$ is negligible. Under the assumption that the $y_i$ are gaussian distributed around the true value we can write:

$$p(y_i|a) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-(y_i - f(x_i|a))^2/2\sigma_i^2}$$

We can write the likelihood as:

$$L(a,y) = \prod_i p(y_i|a)$$

$$\ln L(a,y) = -\frac{1}{2}\sum_i \left(\frac{y_i - f(x_i|a)}{\sigma_i}\right)^2 - \sum_i \ln \sigma_i\sqrt{2\pi},$$

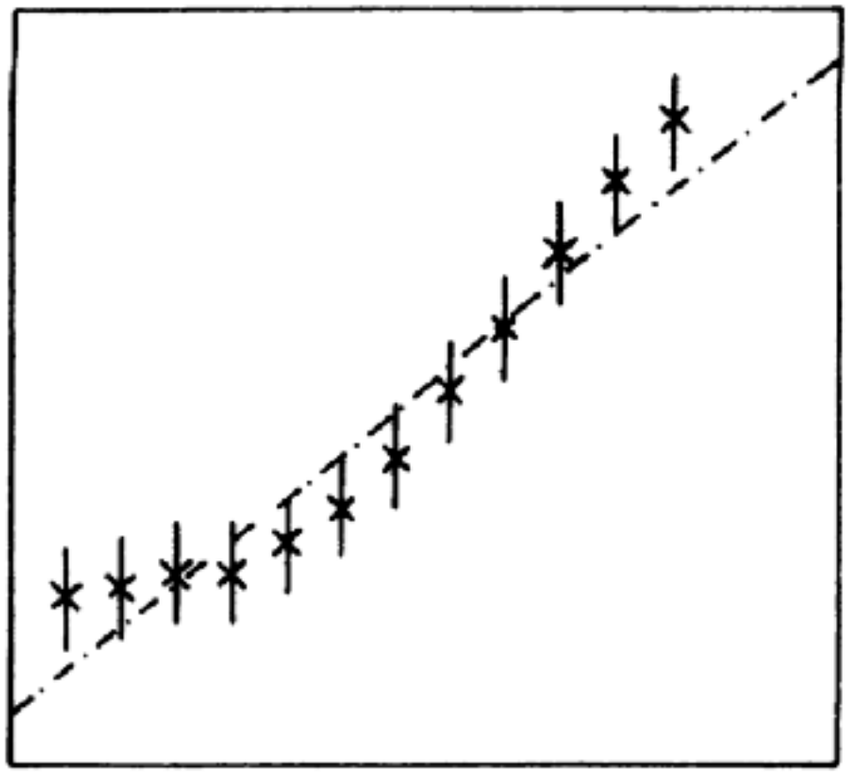
does not depend on the p.o.i. “a”

Minimizing the negative log likelihood correspond to maximize

Which is what we called $$\chi^2 = \sum_i \left(\frac{y_i - f(x_i|a)}{\sigma_i}\right)^2$$

# Run test

χ2/ndof = 1: Is this a good fit ?



A=above  B=below
NA+NB = N

Number of  combinations of $N_A$ in N

$$\binom{N}{N_A} = \frac{N!}{N_A!N_B!}$$

Define: r = run or sequence
Compute the average number of runs and its variance

$$< r >= 1 + \frac{2N_AN_B}{N}$$

$$V(r) = \frac{2N_AN_B(2N_AN_B - N)}{N^2(N-1)}$$

The χ2 ignores the sign of the deviation; the run test looks only at the signs !

# The pull knows everything

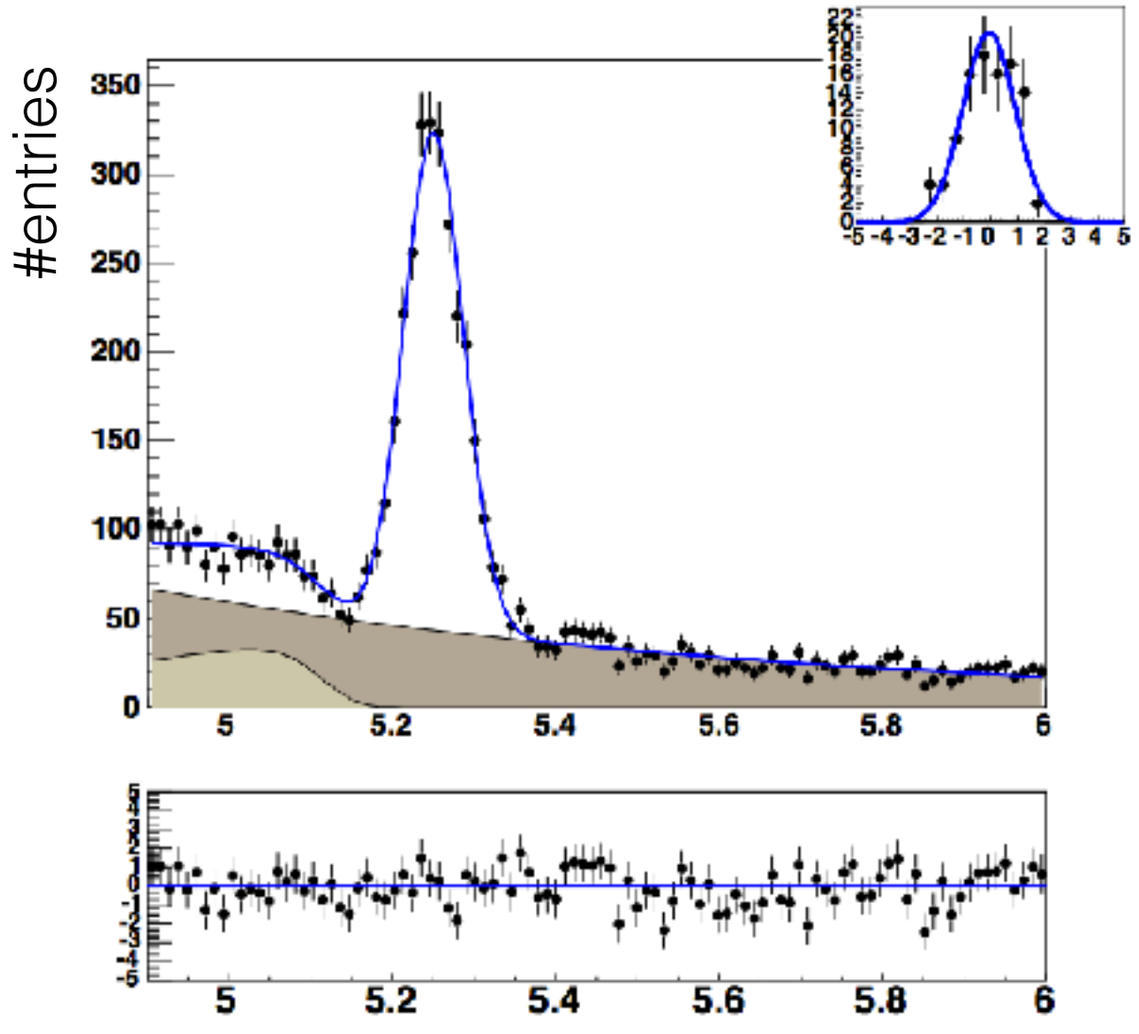
A simple way to assess the quality of a fit is to overlap the fitted curve on data and plot the pull distribution below it.
(when not divided by the uncertainty it's called residual)

$$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

Even better you can:

- Fill the "pull-distribution", i.e. the histogram filled with the value
- The distribution should be centred at zero and have width 1

If not centred at zero —> bias
If larger than 1 —> underestimate uncertainties
If narrower than 1 —> overestimate uncertainties



# Bibliography

Likelihood:

Lyons ch 4.6

Cowan Ch 6

# Backup

# Fisher information

The variance of a MLE is related to the amount of information carried by the dataset on the parameter of interest $\vartheta$.

Intuitively the information should have the properties:

1 information should increase if we make more observations

2 data, which are irrelevant to the estimation of the parameters we wish to estimate or to the hypothesis we wish to test, should contain no information

3 the precision of the estimation should be greater if we have more information

The Fisher information carried by a dataset $\{x_i\}$ on the parameter $\vartheta$ is defined as:

$$I_{\vec{x}}(\theta) = \langle \left( \frac{\partial \ln L(\vec{x};\theta)}{\partial \theta} \right)^2 \rangle =$$

$$= \langle \left( \frac{\partial l}{\partial \theta} \right)^2 \rangle = \int \left( \frac{\partial \ln L(\vec{x};\theta)}{\partial \theta} \right)^2 L(\vec{x};\theta) d\vec{x}$$

$$= -\langle \frac{\partial^2}{\partial \theta^2} \log L \rangle$$

← compute the second derivative

# Fisher information

To simplify notation we define the "score" of one measurement:

$$S_1 = \frac{\partial}{\partial\theta} \ln f(x;\theta)$$

The score of a set of measurements is simply the sum of the scores of each measurement and equal to the derivative of the Likelihood wrt the p.o.i.:

$$S(\vec{x},\theta) = \sum_{i=1}^{n} S_1(x_i;\theta) \;\; = \frac{\partial \ln L(\vec{x},\theta)}{\partial\theta}$$

With this notation the information becomes:

$$I_{\vec{x}}(\theta) = \langle S^2(\vec{x};\theta)\rangle$$

# Fisher information

The Fisher definition satisfies the properties to be called information:

1 information should increase if we make more observations

For n measurements using $V(a) = \langle a^2 \rangle - \langle a \rangle^2$

$$I(\theta) = \langle \left( \sum_{i=1}^{n} S_1(x_i;\theta) \right)^2 \rangle = V\left( \sum_{i=1}^{n} S_1(x_i;\theta) \right) + \langle \sum_{i=1}^{n} S_1(x_i;\theta) \rangle^2$$

Assuming independent measurements, the variance of the sum is the sum of the variances, and being sampled from the same pdf they are all equal (same for the expectation value of the sum of scores)

$$I(\theta) = nV\left(S_1(x;\theta)\right) + n^2 \langle S_1(x;\theta) \rangle^2$$

Proving that the information increases with the number of measurements

2 irrelevant data carry no information

for irrelevant data the p.d.f. will not depend on θ and the score by definition will be 0

adding no information

# MLE uncertainty

To appreciate the third point ("the precision of the estimation should be greater if we have more information") let's first look again at the uncertainty of the MLE.

With the Fisher information definition we see that

$$\sigma^2(\hat{\theta}) = \frac{1}{d^2F/d\theta^2|_{\theta=\hat{\theta}}} = \frac{1}{I(\hat{\theta})}$$

The variance is the inverse of the second derivative of the likelihood, i.e. the inverse of the information.

To reduce the uncertainty of a MLE we need to increase the information, but there is a bound: "minimum variance bound"

# Efficient estimator

Looking back at last week definition

Efficiency
if it has the smallest possible variance (see later Rao-Cramer-Frechet inequality / variance bound)

$$\epsilon = \frac{\text{minimal Variance of } \hat{\theta}}{\text{Variance of } \hat{\theta}}$$

$$\epsilon(\hat{\theta}) = \frac{V_{min}(\theta)}{V(\hat{\theta})} \leq 1$$

When ε = 1 the estimator is called efficient

It is not always possible to find an efficient estimator, but it can be shown that:

- if an efficient estimator for a given problem exist, it will be found using the ML method
- ML estimators are efficient in the large sample limit.

Th: An efficient estimator can be found if and only if it belongs to the exponential family:

$$f(x;\theta) = \exp\left[A(\theta)\,\hat{\theta}(x) + B(\theta) + K(x)\right]$$

# Rao-Cramer-Frechet inequality

Both the score and the MLE are random variables.
Let's compute the covariance among them

$$cov[S(\vec{x},\hat{\theta}),\hat{\theta}(\vec{x})] = \langle S(\vec{x},\hat{\theta})\hat{\theta}(\vec{x})\rangle - \langle S(\vec{x},\hat{\theta}(\vec{x}))\rangle\langle\hat{\theta}(\vec{x})\rangle$$

The first term is

$$\begin{aligned}
\langle\hat{\theta}S(\vec{x},\theta)\rangle &= \int\dots\int\hat{\theta}\left(\frac{\partial}{\partial\theta}\ln L(\vec{x},\theta)\right)L(\vec{x},\theta)dx_1\dots dx_n \\
&= \int\dots\int\hat{\theta}\left(\frac{1}{L(\vec{x},\theta)}\frac{\partial}{\partial\theta}L(\vec{x},\theta)\right)L(\vec{x},\theta)dx_1\dots dx_n \\
&= \int\dots\int\hat{\theta}\left(\frac{\partial}{\partial\theta}L(\vec{x},\theta)\right)dx_1\dots dx_n \\
&= \int\dots\int\hat{\theta}\frac{\partial}{\partial\theta}\left(\prod_{i=1}^{n}f(x_i;\theta)dx_i\right) \\
&= \int\dots\int\frac{\partial}{\partial\theta}\left(\hat{\theta}\prod_{i=1}^{n}f(x_i;\theta)dx_i\right) \\
&= \frac{\partial}{\partial\theta}\int\dots\int\hat{\theta}\prod_{i=1}^{n}f(x_i;\theta)dx_i \\
&= \frac{\partial}{\partial\theta}\langle\hat{\theta}\rangle = \frac{\partial}{\partial\theta}(\theta+b_n(\hat{\theta})) \\
&= 1+\frac{\partial}{\partial\theta}b_n(\hat{\theta})
\end{aligned}$$

# Rao-Cramer-Frechet inequality

Both the score and the MLE are random variables.
Let's compute the covariance among them

$$cov[S(\vec{x},\hat{\theta}),\hat{\theta}(\vec{x})] = \langle S(\vec{x},\hat{\theta})\hat{\theta}(\vec{x})\rangle - \langle S(\vec{x},\hat{\theta}(\vec{x}))\rangle\langle\hat{\theta}(\vec{x})\rangle$$

The second term is zero:

$$\langle S(\vec{x};\theta)\rangle = \sum \langle S_1(x_i;\theta)\rangle$$

$$\begin{aligned}
\langle S_1(\vec{x};\theta)\rangle &= \langle \frac{\partial}{\partial\theta}\ln f(x;\theta)\rangle \\
&= \int \left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right) f(x;\theta)dx \\
&= \int \frac{1}{f(x;\theta)}\left(\frac{\partial}{\partial\theta} f(x;\theta)\right) f(x;\theta)dx \\
&= \int \frac{\partial}{\partial\theta} f(x;\theta)dx
\end{aligned}$$

$$= \frac{\partial}{\partial\theta}\int f(x;\theta)dx = \frac{\partial}{\partial\theta}1 = 0$$

# Rao-Cramer-Frechet inequality

so we have $$cov[S(\vec{x},\hat{\theta}),\hat{\theta}(\vec{x})] = 1 + \frac{\partial}{\partial\theta} b_n(\hat{\theta})$$

And the linear correlation:

$$\rho^2 = \frac{(cov[S,\hat{\theta}])^2}{V(S)V(\hat{\theta})} = \frac{\left(1 + \frac{\partial}{\partial\theta} b_n(\hat{\theta})\right)^2}{I(\theta)V(\hat{\theta})}$$

since $$\rho^2 \leq 1$$

$$V(\hat{\theta}) \geq \frac{\left(1 + \frac{\partial}{\partial\theta} b_n(\hat{\theta})\right)^2}{I(\theta)}$$

Rao-Cramer-Frechet inequality

which for an unbiased estimator is

$$V(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

This is called minimum variance bound

# Comments on the extended MLE

The total number of events is unknown

Case 1): v=v(ϑ) the parameter v depends on ϑ.

$$\begin{aligned} \ln L(\theta) &= n \ln \nu(\theta) - \nu(\theta) + \sum_{i=1}^{n} \ln f(x_i;\theta) \\ &= -\nu(\theta) + \sum_{i=1}^{n} \ln(\nu(\theta) f(x_i;\theta)) \end{aligned}$$

the resulting variance is usually smaller, because when estimating ϑ, we use the extra information brought in by n.

Case 2): the parameter v does not dependent on $\hat{\vartheta}$. Maximizing the likelihood we find

$$\hat{\nu} = n$$

We also obtain as estimators the same ϑ the standard ML, but with larger variance