Parameter Estimation Part A
The Likelihood Method

Christoph Rosemann

DESY

November 2014
Parameter estimation

Common task

- Determine from measurements with uncertainties the best values of (physical) parameters
- Estimation is a mathematical procedure (!)
- Any parameter makes sense only within a model
- The model is encoded in the pdf of the parameters
- Wrong models deliver wrong answers!
- Uncertainties must be known: Variances and Covariances
- Distinguish between:
  - Statistical uncertainties
  - Systematic uncertainties
Parameter estimation

**Fundamental properties of estimators**

Estimators can be characterized as *good* or *bad*

The characterization classes are:

- **Consistency**: the true value and the estimated value are equivalent
  \[
  \lim_{n \to \infty} \hat{a} = a
  \]

- **Bias**: the expectation value is equivalent to true value
  \[
  \langle \hat{a} \rangle = a
  \]

- **Efficiency**: small variance

The inherent accuracy of an estimator is limited!
Consistency

- Parameters are estimated from limited samples
- Any sample exhibits statistical fluctuations
- For large samples, the effect of fluctuations lessens
- If the difference between the true value and the estimated value vanishes, the estimator is consistent

**Formal definition**

An estimator is consistent, if it tends to the true value as the number of data tends to infinity:

\[
\lim_{{n \to \infty}} \hat{a} = a
\]
For finite amounts of data the estimated parameter is unlikely to have the true value.

A good estimator has the equal chances of over- and underestimation of the true value.

Such an estimator is unbiased.

This can be expressed in terms of the expectation value of the estimator.

**Formal definition**

An estimator is unbiased, if its expectation value is the same as the true value:

\[ \langle \hat{a} \rangle = a \]
Efficiency

- The estimated value depends on the given data sample
- The fluctuations of the sample influence the estimator
- An efficient estimator exhibits a small fluctuation or spread
- The spread is measured in terms of the variance of the estimator

**Formal definition**

An estimator is efficient if its variance is small.
(Without proof) There is a lower bound on the variance of an estimator!

* There are different names for this:
  - Cramér-Rao bound (or inequality), Fréchet inequality, MVB, CRLB
  - It uses the (in the simple/unbiased form) the Likelihood function $\mathcal{L}$:
    $$\sigma_{\hat{a}}^2 \leq \frac{1}{\langle (d\mathcal{L}/da)^2 \rangle}$$

* An estimator is efficient, if its variance is equal to the MVB
Characterization of Maximum Likelihood

Most important parameter estimation method

- Maximum Likelihood estimators are (usually) consistent
- Maximum Likelihood are biased (!) for small $N$
  for large $N$ it becomes unbiased
- It is usually the optimal estimation in terms of the Minimum Variance Bound

Warning

- Maximum Likelihood is (usually) consistent, but biased!
- Maximum Likelihood estimators invariant under parameter transformations!:

\[ \hat{f}(\hat{a}) = f(\hat{a}) \quad \text{e.g.} : \hat{\sigma}^2 = (\hat{\sigma})^2 \]
Bias example

Consider a symmetric pdf around $a_0$, let $\hat{a}$ be an unbiased estimator

**Equal chances that $\hat{a}$ is either 10% too large or too small**

- Equally possible:
  
  $\hat{a} = 1.1a_0$ \quad $\hat{a} = 0.9a_0$

- Now consider (non-linear) transformation $y : x \rightarrow x^2$, then

  $\hat{a}^2 = 1.21a_0^2$ \quad $\hat{a}^2 = 0.81a_0^2$

- Probability content doesn’t change, equal chances that $\hat{a}^2$ is 21% larger or 19% smaller than $a_0^2$

- In short: the pdf becomes asymmetric and therefore biased
The maximum Likelihood method

Requirements

- Data, e.g. $n$ measurements $x_i$
- A model, e.g. a pdf $f(x; a)$
- The function has to be normalized for all $a$:

$$\int f(x; a)dx = 1$$

The formula

Maximize the product of all functions at the given measurements:

$$\mathcal{L}(\bar{x}; a) = f(x_1; a) \cdot f(x_2; a)\ldots f(x_n; a) = \prod_{i}^{n} f(x_i; a)$$

to obtain the best estimator for the parameter(s).
Maximization

Finding the maximum is straightforward

- For a single parameter $a$
  \[
  \frac{d\mathcal{L}(\vec{x}; a)}{da} = 0
  \]

- For multiple parameters $\vec{a} = a_1, \ldots, a_m$:
  \[
  \frac{\partial \mathcal{L}(\vec{a})}{\partial a_k} = 0 \quad \forall k = 1, \ldots, m
  \]
Log Likelihood

Different formulation

- Often: too much data to calculate $L$ accurately
- Take logarithm of $L \implies \ln L$
- Use negative value in order to use only one numerical routine for minimization (like for $\chi^2$ minimization)

Formula

$$\ell(\tilde{x}; a) = -\ln L(\tilde{x}; a)$$
Important reminder:

- One needs to know the underlying pdf
- Wrong pdf will yield a wrong or non-sensical result
- Always check the result:
  - Do the found parameters describe the data (at all!?)
  - Parameter at boundary of parameter space?
    This is always trouble
- There is no consistency check inherent to the method
Consider (once again) a radioactive source; \( n \) measurements are taken under the same conditions, counted are the number of decays \( r_i \) in a given, constant time interval. What's the mean number of decays?

- **Naive (?):** Simply take the arithmetic mean

\[
\mu = \frac{1}{n} \sum_{i}^{n} r_i
\]

- **Wrong (!):** Take the weighted mean

- **Maximum Likelihood**
Estimation via ML

\( r_i \) follows a Poisson distribution:

\[
P(r_i; \mu) = \frac{\mu^{r_i} e^{-\mu}}{r_i!}
\]

The Likelihood function is therefore

\[
\mathcal{L}(\mu) = \prod_{i} P(r_i; \mu) = \prod_{i} \frac{\mu^{r_i} e^{-\mu}}{r_i!}
\]

Negative logarithm:

\[
\ell(\mu) = -\ln \mathcal{L}(\mu) = - \sum_{i} \ln \frac{\mu^{r_i} e^{-\mu}}{r_i!} = \sum_{i} (-r_i \ln \mu + \mu + \ln r_i!)
\]
Likelihood estimation of mean III

Estimation via ML

Differentiate for the parameter $\mu$:

$$\frac{d}{d\mu} \ell(\mu) = \frac{d}{d\mu} \sum_{i}^{n} (-r_i \ln \mu + \mu + \ln r_i!) = \sum_{i}^{n} \left( -r_i \frac{1}{\mu} + 1 \right)$$

set to zero:

$$0 = \sum_{i}^{n} \left( -r_i \frac{1}{\mu} + 1 \right) = n - \frac{1}{\mu} \sum_{i}^{n} r_i$$

$$\implies \mu = \frac{1}{n} \sum_{i}^{n} r_i$$

This yields the same result as the naive expectation.
What is the uncertainty of the estimation?

Consider the following statements (without proof):
- In the limit of $n \to \infty$ the likelihood function $L$ is approximately Gaussian,
- the mean $\mu$ of this distribution is the true mean value of the parameter and
- the variance goes to zero $\sigma \to 0$

(we will formalize this a little later.)

Intuitive explanation:
If you sample from a certain population that follows a certain distribution, the best estimator for a parameter is itself a random variable.

Now evolve the likelihood function around the best estimator.
Series evolution of the likelihood function

With

\[ \frac{d}{da} \ell(a) \bigg|_{a=\hat{a}} = 0 \]

this is

\[ \ell(a) = \ell(\hat{a}) + \frac{1}{2} (a - \hat{a})^2 \frac{d^2 \ell(a)}{da^2} \bigg|_{a=\hat{a}} + \ldots \]

For the likelihood function \( \mathcal{L} \) this is

\[ \mathcal{L} \approx \text{const} \cdot e^{-\frac{1}{2} \left\{ (a-\hat{a})^2 \frac{d^2 \ell(a)}{da^2} \bigg|_{\hat{a}} \right\}} \]

From this expression the variance can be identified:

\[ \sigma_a^2 = \left( \frac{d^2 \ell(a)}{da^2} \bigg|_{\hat{a}} \right)^{-1} \]
What is the uncertainty of the estimation of the mean number of decays?

The best estimator was the arithmetic mean:

\[
\mu = \frac{1}{n} \sum_{i} r_i
\]

Now calculate the variance of \( \mu \), take the second derivative at \( \mu = \hat{\mu} \):

\[
\left. \frac{d^2 \ell(\mu)}{d\mu^2} \right|_{\mu=\hat{\mu}} = \frac{1}{\hat{\mu}} \sum_{i} r_i = \frac{1}{\hat{\mu}^2} \hat{\mu} n = \frac{n}{\hat{\mu}} = \frac{1}{\sigma^2_{\mu}}
\]

\[\implies \sigma^2_{\mu} = \frac{\hat{\mu}}{n}\]

If the true value \( \mu \) is not known, then the variance is calculated from the best estimation.
Numerical example

A set of rate measurements at fixed intervals of a radioactive source yielded

\[ r_i = [1, 1, 5, 4, 2, 0, 3, 2, 4, 1, 2, 1, 0, 1, 1, 2, 1] \]

Assume a Poisson distribution

Better check: histogram the values and compare it with a Poisson. The estimated, best value for the mean is \( \mu = \frac{1}{n} \sum_i^n r_i = 1.78 \) the estimated uncertainty from this is \( \sigma_\mu = \sqrt{\mu/n} = 0.31 \)

Looks OK!
Often the likelihood function $\ell = -\ln L$ can be approximated by a parabola in the direct vicinity of the minimum:

$$\ell(\mu) \approx \ell(\hat{\mu}) + \frac{1}{2} \frac{(\mu - \hat{\mu})^2}{\sigma^2_\mu}$$

From $\mu = \hat{\mu} + \sigma_\mu$ can be then deduced, that the standard deviation can be determined implicitly from the points of intersection of the parabola with the constant

$$\ell_{\text{min}} + \frac{1}{2}$$
In almost all cases, the second derivative of $\ell(a)$ can’t be calculated (accurately) – how is the uncertainty determined then?

The relation still holds:

$$\ell(\hat{\mu} \pm \sigma_\mu) = \ell_{\text{min}} + \frac{1}{2}$$

- In the parabolic approximation is $\mathcal{L}(a) = e^{-\ell(a)}$ a Gaussian distribution around the true value $\hat{\mu}$
- What if the approximation is not very good?
Uncertainty estimation: general solution

If the symmetric Gauss function isn’t a good description, asymmetric errors $\sigma_l$ and $\sigma_r$ can be derived from

$$\ell(\hat{\mu} - \sigma_l) = \ell(\hat{\mu} + \sigma_r) = \ell_{\text{min}} + \frac{1}{2}$$

- In principle it’s always possible to transform the parameter $a$ with $b(a)$, so that $\ell(b(a))$ becomes parabolic.
- One doesn’t even need to know the transformation, the probability content in an interval is always conserved!

$\Rightarrow$ This interval always contains the central 68% probability.

The result can then be written as

$$\mu_{-\sigma_l}^{+\sigma_r}$$
Continue numerical example

- Estimated mean is $\mu = \frac{1}{n} \sum_{i}^{n} r_i = 1.78$
- In the parabolic approximation the uncertainty is $\sigma_{\mu} = \sqrt{\mu/n} = 0.31$
- For finding the true parameter uncertainty, solve the actual Likelihood function for the intersection points with $\ell_{\text{min}} + \frac{1}{2}$:

The result is either

$$\mu = 1.78 \pm 0.31$$

or

$$\mu = 1.78^{+0.33}_{-0.30}$$
The intervals that contain $k$ standard deviations can be determined likewise:

$$\ell(\hat{a} - k\sigma_I) = \ell(\hat{a} + k\sigma_r) = \ell_{\text{min}} + \frac{k^2}{2}$$

The amount of probability is the same as for the Gaussian distribution.

E.g. $2\sigma$ are in $\ell_{\text{min}} + 2$ and corresponds to 95% probability.

$3\sigma$ are defined by $\ell_{\text{min}} + \frac{9}{2}$, corresponding to 99%, etc.
Binned Likelihood

The task

- \( J \) number of bins, each with \( n_j \) entries
- Fit pdf \( f(x; a) \) to the number of entries in each bin
- Obtain the best value for \( a \) using the data

Consider the number of bin entries \( n_j \) as random variables

- Underlying pdf is Poisson with mean value \( \mu_j \):

\[
P(n_j; \mu_j) = \frac{\mu_j^{n_j} e^{-\mu_j}}{n_j!}
\]

- The mean value \( \mu_j \) depends on the fit parameter \( a \): \( \mu_j(a) \)
- The Poissonian describes the distribution of entries in each bin
### How to obtain $\mu_j(a)$?

- Get the probability "amount" by integrating the pdf $f(x; a)$ for the bin $j$

$$p_j = \int_{bin_j} f(x; a) \, dx$$

- This can be approximated (mean value theorem of integration), with $x_c$ the bin center position and $\Delta x$ the interval width

$$p_j \approx f(x_c; a) \Delta x$$

- The expected mean number of entries is obtained by multiplying with the total number of entries $n$, so

$$\mu_j(a) = np_j \approx nf(x_c; a) \Delta x$$
Binned Likelihood function

Master formula for binned Likelihood

\[ F(a) = - \sum_{j}^{J} \ln \left( \frac{\mu_j^{n_j} e^{-\mu_j}}{n_j!} \right) = - \sum_{j}^{J} n_j \ln \mu_j + \sum_{j}^{J} \mu_j + \sum_{j}^{J} \ln(n_j!) \]

- This is the formula to use for Poisson distributed variables (since it’s unbiased)
- It’s also valid if the \( n_j \) are small or even zero (!)
- The last term doesn’t play any role in the minimization, since it’s constant for given data
- It’s directly related to the binned \( \chi^2 \) formula (not shown here)
Multi-dimensional parameters

The generalization to more than parameter $\vec{a} = a_1, \ldots, a_m$ leads to the Likelihood function for $n$ measurements:

$$\mathcal{L}(\vec{a}) = \prod_{i}^{n} f(x_i; \vec{a})$$

- The minimization procedure is the same
- What’s with the uncertainties of the parameters? And Correlations? Answer (as so often): evolve the Likelihood function in a Taylor series
Taylor series evolution of $\ell(\vec{a})$

Evolve $\ell(\vec{a}) = -\ln \mathcal{L}(\vec{a})$ around the true values $\hat{\vec{a}}$:

$$
\ell(\vec{a}) = \ell(\hat{\vec{a}}) + \frac{1}{2} \sum_{i} \sum_{j} (a_i - \hat{a}_i)(a_j - \hat{a}_j) \frac{\partial^2 \ell(\vec{a})}{\partial a_i \partial a_j} + \ldots
$$

$$
= \ell(\hat{\vec{a}}) + \frac{1}{2} \sum_{i} \sum_{j} (a_i - \hat{a}_i)(a_j - \hat{a}_j) G_{ij} + \ldots
$$

The Likelihood function will become Gaussian for $n \to \infty$. Comparing

$$
\mathcal{L}(\vec{a}) = e^{-\ell(\vec{a})}
$$

yields the identification of the inverse covariance matrix

$$
G = V^{-1}
$$

with the Hesse Matrix $G_{ij} = \frac{\partial^2 \ell(\vec{a})}{\partial^2 \vec{a}}$
Probability contents

Also in the case of more than one dimension all results can be taken from the integrated Gaussian distribution.

- The $1\sigma$ contour is defined by $\ell(\hat{a}) + \frac{1}{2}$
- The $2\sigma$ contour is defined by $\ell(\hat{a}) + 2$
- etc.

The probability contents can be calculated with integrating the Gauss function.

Likelihood for two parameters

- The probability to find a pair within the $1\sigma$ contour is 39%
- In the parabolic approximation the contour is an ellipsis in the $a_1, a_2$ plane
- In the general case the curves are asymmetric but contain the same amount of probability
Uncertainty of parameters

The uncertainty of a parameter is determined by minimizing w.r.t. all other parameters. The minimum of this function $\ell'$ serves as reference for $\ell_{\text{min}}$.

Example:

- This is the $1\sigma$ contour for two parameters $a, b$
- Parabolic approximation doesn’t fit
- Still within contour area with 39% probability

Blue curve: to find uncertainty on $a$, $\ell(a, b)$ must be minimized w.r.t $b$ for fixed value of $a$.
Parameter Estimation is a well defined mathematical procedure
- Presented method: Maximum Likelihood
- Uncertainties and Covariances are also extract-able
- No consistency check method – check plausibility of results
- Even more carefully check the pdfs/the model
- The results can still be ill-defined: crap in, crap out