Parameter Estimation

Arun Nayak[1]

ML4HEP-2025, IISER, Kolkata

16 - 19th June 2024

[1]Institute of Physics, Bhubaneswar

Reference Books

1. G. Cowan, Statistical Data Analysis, Oxford.
2. Luca Lista, Statistical Methods for Data Analysis in Particle Physics, Springer, 2017
3. L. Lyons, Statistics for Nuclear and Particle Physics

(Some of the Figures in these slides are taken from ref. 1)

Parameter estimation

Suppose x is a random variable described by pdf $f(x)$

Sample space: Set of all possible values of x.

sample: A set of n independent observations of x is a smaple of size n.

Assuming all x_i are independent, we can write the joint pdf $f_{sample}(x_1, ..., x_n)$ as

$$f_{sample}(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i)$$

Parameter estimation – contd.

- Suppose we have n-measurements of x, whose pdf $f(x)$ is not known. **Problem:** Infer properties of $f(x)$ based on the observations.
- Suppose we have a hypothesis that describes the pdf $f(x, \theta)$, where θ is unknown parameter(s).
  Example: Suppose we have a radio active source whose lifetime is not known. We know that the decay rate is distributed according to exponential distribution, with parameter τ (lifetime).
- **parameter fitting:** Estimate the parameter value(s) given n measurements $x_1, ..., x_n$ (data).
  Goal is to construct a function of x_i to estimate the parameter(s).

Estimator(s)

Estimator: A function of observed measurements $x_1, ..., x_n$, which is used to estimate some property of a pdf (e.g. mean, variance or some other parameters)

An estimator for θ is usually written as $\hat{\theta}$. The numerical value of the estimator evaluated with a particular sample is called an **estimate**.

If $\hat{\theta}$ **converges** to θ in the limit of large n, the estimator is said to be **consistent**.
Limit of large n is typically referred as '*large sample*' or '*asymptotic*' limit.

Mean value of Estimator

A function of random variables is also a random variable
$\Longrightarrow$ $\hat{\theta}$ is a random variable, with some pdf $g(\hat{\theta};\theta)$.
The prob. distribution of $\hat{\theta}$ is called a **sampling distribution**.

The expectation value of $\hat{\theta}$

$$\begin{aligned} <\hat{\theta}(\vec{x})> &= \int \hat{\theta} g(\hat{\theta};\theta) d\hat{\theta} \\ &= \int ... \int \hat{\theta}(\vec{x}) f_{sample}(x_1, ..., x_n; \theta) dx_1 ... dx_n \\ &= \int ... \int \hat{\theta}(\vec{x}) f(x_1;\theta) ... f(x_n;\theta) dx_1 ... dx_n \end{aligned}$$

This is the expected mean value of $\hat{\theta}$ from an infinite number of similar experiments, each with a sample of size n.

Quality of Estimators

Define **bias**,

$$b = < \hat{\theta}(\mathbf{x}) > \; - \; \theta$$

b depends on

- sample size,
- functional form of the estimator, and
- the true properties of the pdf $f(x, \theta)$.

If $b = 0$, irrespective of sample size n, $\hat{\theta}$ is **unbiased**.

If $b \to 0$, in the limit $n \to \infty$, $\hat{\theta}$ is **asymptotically unbiased**.

In most practical cases, the bias is small compared to the statistical error (i.e. the standard deviation).

Quality of Estimators – contd.

The mean squared error,

$$\begin{aligned} MSE &= <(\hat{\theta} - \theta)^2> = <\hat{\theta}^2 + \theta^2 - 2\hat{\theta}\theta> \\ &= <\hat{\theta}^2> + \theta^2 - 2<\hat{\theta}>\theta \\ &= <(\hat{\theta} - <\hat{\theta}>)^2> + (<\hat{\theta} - \theta>)^2 \\ &= V[\hat{\theta}] + b^2 \end{aligned}$$

The MSE is the sum of the variance and the bias squared

An estimator is considered **optimal** if $b = 0$ and $V[\hat{\theta}]$ is minimum, though MSE could also be considered.

Estimator for Mean

Consider a sample of a r.v. x, of size n: $(x_1, ..., x_n)$.
The pdf $f(x)$ is not known.
Aim: Construct a function $t(x_1, ..., x_n)$ to be an estimator for population mean $< x >= \mu$.

The arithmetic mean or **sample mean** is

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$\overline{x}$ can be considered to be an estimator for $< x >$.

Weak law of large numbers: If the variance of x exists, then $\overline{x}$ is a consistent estimator for the population mean $< x >$.,
i.e. for $n \to \infty$, $\overline{x}$ converges to μ.
Note that the law holds irrespective of the form of pdf $f(x)$.

Estimator for Mean – contd.

The expectation value of the sample mean,

$$< \bar{x} > \;=\; \left\langle \frac{1}{n}\sum_{i=1}^{n} x_i \right\rangle \;=\; \frac{1}{n}\sum_{i=1}^{n} < x_i > \;=\; \frac{1}{n}\sum_{i=1}^{n} \mu \;=\; \mu$$

since

$$< x_i > \;=\; \int \ldots \int x_i f(x_1) \ldots f(x_n) dx_1 \ldots dx_n \;=\; \mu$$

for all i.
Hence, the sample mean $\bar{x}$ is an **unbiased estimator** for the population mean μ.

Estimator for Variance

The **sample variance** s^2, defined by

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n}x_i\right) \\
&= \frac{1}{n-1}\left(n\overline{x^2} + n\bar{x}^2 - 2n\bar{x}^2\right) \\
&= \frac{n}{n-1}(\overline{x^2} - \bar{x}^2)
\end{aligned}
$$

Exercise: Show that $< s^2 > \; = \; \sigma^2$.

So, s^2 is an unbiased estimator for the population variance.

Estimator for Variance – contd.

In case the population mean, μ, is known, define

$$S^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i \;-\; \mu)^2 \;=\; \overline{x^2} \;-\; \mu^2$$

In this case, $< S^2 > \;=\; \sigma^2, \;\Longrightarrow\; S^2$ is an unbiased estimator of the variance σ^2.

Similarly,

$$\hat{V}_{xy} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \;=\; \frac{n}{n-1}(\overline{xy} \;-\; \bar{x}\bar{y})$$

is an unbiased estimator for the covariance V_{xy} of two random variables x and y of unknown mean.

Estimator for correlation coefficient

The estimator r for the correlation coefficient, ρ,

$$
\begin{aligned}
r &= \frac{\hat{V}_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{j=1}^{n}(x_j - \bar{x})^2 \cdot \sum_{k=1}^{n}(y_k - \bar{y})^2\right)^{1/2}} \\
&= \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}
\end{aligned}
$$

The expectation value of r depend on higher moments of the joint pdf $f(x, y)$. For 2d Gaussian pdf,

$$
< r > \; = \; \rho \; - \; \frac{\rho(1 - \rho^2)}{2n} \; + \; \mathcal{O}(n^{-2})
$$

Thus, the estimator r is only asymptotically unbiased. Still it is widely used because of its simplicity.

Error on mean

Given an estimator $\hat{\theta}$, we can compute its variance

$$V[\hat{\theta}] \;=\; \left\langle \hat{\theta}^2 \right\rangle \;-\; \left(< \hat{\theta} > \right)^2 .$$

Note: $V[\hat{\theta}]$ is a measure of the variation of $\hat{\theta}$ about its mean in a large number of similar experiments each with sample size n

$\Longrightarrow$ statistical error of $\hat{\theta}$

e.g., the variance of sample mean $\bar{x}$,

$$V[\hat{x}] \;=\; \left\langle \hat{x}^2 \right\rangle \;-\; (< \hat{x} >)^2 \;=\; \left\langle \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) \left(\frac{1}{n} \sum_{j=1}^{n} x_j \right) \right\rangle \;-\; \mu^2$$

$$= \frac{1}{n^2} \sum_{i,j=1}^{n} < x_i x_j > \;-\; \mu^2$$

$$= \frac{1}{n^2} [(n^2 \;-\; n)\mu^2 \;+\; n(\mu^2 \;+\; \sigma^2)] \;-\; \mu^2 \;=\; \frac{\sigma^2}{n}$$

where, σ^2 is the variance of $f(x)$.

We used $< x_i x_j > \;=\; \mu^2$ for $i \neq j$ and $< x_i^2 > \;=\; \mu^2 + \sigma^2$.

Error on variance

The variance of estimator s^2 is

$$V[s^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$$

where μ_k is the kth central moment, e.g. $\mu_2 = \sigma^2$.
Using simple generalization of definition of s^2,

$$\mu_k = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Similarly, the variance of r, considering 2d Gaussian pdf,

$$V[r] = \frac{1}{n}(1 - \rho^2)^2 + \mathcal{O}(n^{-2})$$

Note that although $\hat{V}_{xy}$, s_x^2, and s_y^2 are unbiased estimators of V_{xy}, σ_x^2, andσ_y^2, the nonlinear function $\hat{V}_{xy}/(s_x s_y)$ is not an unbiased estimator of $V_{xy}/(\sigma_x \sigma_y)$.

Method of maximum likelihood

Likelihood function

Let a r.v. x, measured n times, giving values $(x_1, ..., x_n)$. PDF of x is $f(x)$.
Then, prob. of x to be in $[x, x + dx] = f(x)dx$.
Assuming x_i are independent,

$$Prob.(x_i \text{ in } [x_i, x_i + dx_i] \text{ for all i}) = \prod_{i=1}^{n} f(x_i)dx_i$$

Defining,

$$L = \prod_{i=1}^{n} f(x_i) \qquad \text{(likelihood function)}$$

If $f(x)$ depends on some parameter θ,

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

In case, x_i are not independent, L is the joint probability $f(\vec{x}, \theta)$.
Note: $L(\theta)$ is not same as probability. It is a function of θ, given a sample of data x_i.

Maximum likelihood method

It is a technique for estimating the values of parameters given a finite data sample.

ML estimators: The estimator $\hat{\theta}$ for the parameter θ is the one that maximizes the likelihood function.

i.e., the estimators are given by the solutions to the equations,

$$\frac{\partial L}{\partial \theta_i} = 0, \quad i = 1, ..., m.$$

The estimators for $\vec{\theta} = (\theta_1, ..., \theta_m)$ is usually written as $\hat{\vec{\theta}} = (\hat{\theta_1}, ..., \hat{\theta_m})$.

Advantages of ML method: Ease of use, no binning is necessary.

Example: Exponential distribution

Consider an exponential pdf, with mean τ,

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

(PDF for proper decay times)
Suppose we have n measurements of t, $(t_1, ..., t_n)$ (i.e., n decays)
Task: Estimate the value of the parameter τ.
Construct likelihood

$$L(\tau) \;=\; \prod_{i=1}^{n} f(t_i;\tau)$$

Convenient to use **log-likelihood function** instead of likelihood function.
Since the logarithm is a monotonically increasing function, the parameter value which maximizes L will also maximize $logL$.

Example: Exponential distribution – contd.

The log-likelihood function:

$$logL(\tau) \;=\; \sum_{i=1}^{n} logf(t_i;\tau) \;=\; \sum_{i=1}^{n}\left(log\frac{1}{\tau} \;-\; \frac{t}{\tau}\right)$$

Advantage: The product in L is converted into a sum.

Maximizing $logL$ wrt τ, gives ML estimator

$$\frac{\partial logL(\tau)}{\partial\tau} \;=\; 0 \;\Rightarrow\; \hat{\tau} \;=\; \frac{1}{n}\sum_{i=1}^{n} t_i$$

In this case the ML estimator is simply the sample mean of the measured time values.

Example: Exponential distribution – contd.

The expectation value of $\hat{\tau}$ is

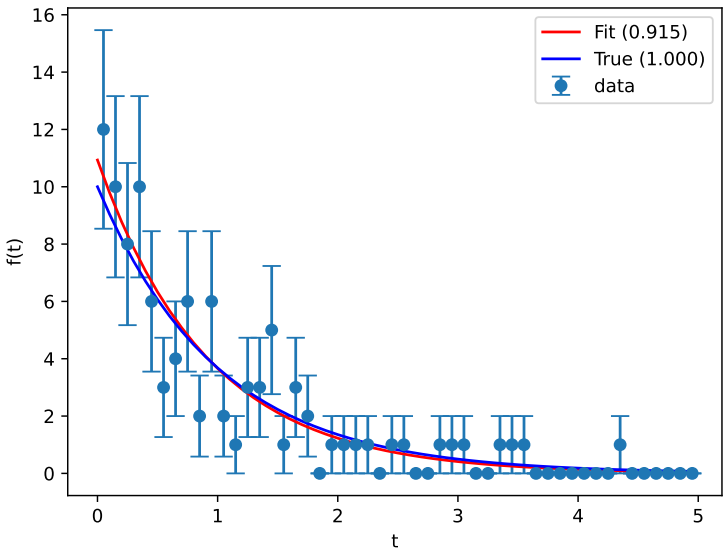
$$\begin{aligned}
\langle\hat{\tau}\rangle &= \int \ldots \int \hat{\tau} f_{joint}(t_1, \ldots, t_n; \tau) dt_1 \ldots dt_n \\
&= \int \ldots \int \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \ldots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \ldots dt_n \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\int \ldots \int t_i \frac{1}{\tau} e^{-t_i/\tau} dt_i \prod_{j \neq i} \frac{1}{\tau} e^{-t_j/\tau} dt_j \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \tau = \tau.
\end{aligned}$$

Thus, $\hat{\tau}$ is an unbiased estimator for τ.

It was also shown previously that the sample mean is an unbiased estimator of the population mean for any pdf.

Example: Exponential distribution – contd.

Example: Consider a sample of 100 MC generated decay times using a true lifetime $\tau = 1.0$.



$\hat{\tau} = 0.915$ from ML fit.
(Note: you may get different value based on your generated sample)

Exercise: Perform this MC experiment and estimate mean lifetime from your dataset. Increase your MC statitics by factor of two and check the result.

Example: Exponential distribution – contd.

If $a(\theta)$ is a function of some parameter θ,

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial a}\frac{\partial a}{\partial \theta} = 0$$

This implies,

$$\frac{\partial L}{\partial a} = 0, \qquad \text{if } \frac{\partial a}{\partial \theta} \neq 0$$

Thus, the ML estimator of a function can be obtained simply by evaluating the function with the original ML estimator, $\hat{a} = a(\hat{\theta})$.
So, the ML estimator for decay constant $\lambda = 1/\tau$ is
$\hat{\lambda} = 1/\hat{\tau} = n/\sum_{i=1}^{n} t_i$.
One can show that, the expectation value of $\hat{\lambda}$ is

$$\left\langle \hat{\lambda} \right\rangle = \lambda \frac{n}{n-1} = \frac{1}{\tau}\frac{n}{n-1}.$$

$\hat{\lambda}$ is only asymptotically unbiased.

Example: Gaussian distribution

Suppose we have n measurements of a r.v. x, assumed to be distributed according to a Gaussian pdf of unknown μ and σ. The log-likelihood function is

$$\begin{aligned} logL(\mu, \sigma^2) &= \sum_{i=1}^{n} log(x_i; \mu, \sigma^2) \\ &= \sum_{i=1}^{n} \left(log \frac{1}{\sqrt{2\pi}} + \frac{1}{2} log \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \end{aligned}$$

Maximizing $logL$ with respect to μ gives,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

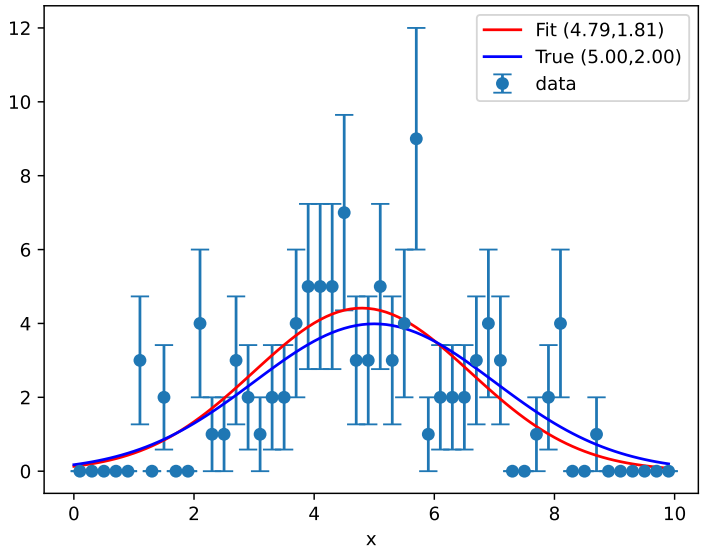
Repeating the procedure for σ^2 and using the result for $\hat{\mu}$ gives

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Example: Gaussian distribution

Exercise: Show that $\langle \hat{\mu} \rangle = \mu$ and $\left\langle \hat{\sigma^2} \right\rangle = \frac{n-1}{n}\sigma^2$.

Thus, $\hat{\mu}$ is an unbiased estimator while $\hat{\sigma^2}$ is only asymptotically unbiased.



Recall from previous lecture: the sample variance, defined by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

is an unbiased estimator for variance for any pdf.
So, it is also an unbiased estimator for the parameter σ^2 of the Gaussian. But it is not the ML estimator.

Variance of ML estimators

What is the statistical uncertainty on the estimates?

One way of estimating it is by considering the **variance (or standard deviation) of the estimator**.

Different ways of estimating the variance:

1. Analytic method
2. Monte Carlo method
3. RCF bound
4. graphical method

Analytic method

In certain cases it is possible to compute the variance using analytic method

e.g., consider the exponential distribution with mean τ estimated by $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$

$$
\begin{aligned}
V[\hat{\tau}] &= \left\langle \hat{\tau}^2 \right\rangle - \left\langle \hat{\tau} \right\rangle^2 \\
&= \int \ldots \int \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right)^2 \frac{1}{\tau} e^{-t_1/\tau} \ldots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \ldots dt_n \\
&- \left(\int \ldots \int \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \ldots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \ldots dt_n \right)^2 \\
&= \frac{\tau^2}{n}
\end{aligned}
$$

Analytic method – contd.

Note that $V[\hat{\tau}]$ is a function of true parameter τ, which is unknown.

How to report the statistical error of the experiment?

Using transformation invariance of ML estimators, we can obtain ML estimate for the variance $\sigma^2_{\hat{\tau}} = \tau^2/n$, simply by replacing τ with its own ML estimator $\hat{\tau}$,
Thus,

$$\widehat{\sigma^2_{\hat{\tau}}} = \frac{\hat{\tau}^2}{n}$$

Monte Carlo Method

Useful when analytic method is not possible.

Procedure:

Simulate a large number of experiments and look at the distribution of ML estimates from MC experiments.

In MC program, the estimated value of the parameter from the real experiment can be used in place of the *true* parameter.

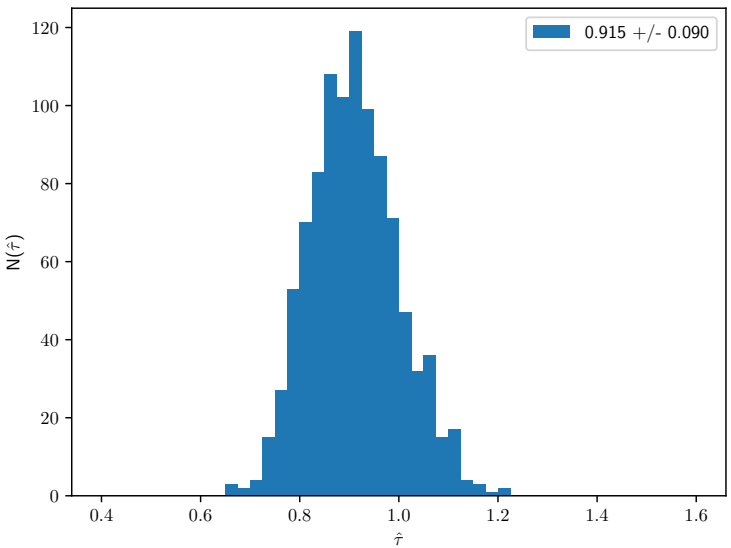
Example:

Consider again mean lifetime measurement with the exponential distribution.

For true lifetime $\tau = 1.0$, a sample $n = 100$ measurement gave the ML estimate $\hat{\tau} = 0.915$.

Considering this measurement as the real one, 1000 further experiments are simulated with 100 measurements each (with $\tau = 0.915$).

Monte Carlo Method – contd.



The sample mean of the estimates $\overline{\hat{\tau}} = 0.911$.
This is close to the input value, as expected, since ML estimator $\hat{\tau}$ is **unbiased**.
The sample standard deviation $\sigma = 0.09$,
essentially same as
$\widehat{\sigma_{\hat{\tau}}} = \hat{\tau}/\sqrt{n} = 0.915/\sqrt{100} = 0.091$

Note: the distribution is approximately Gaussian in shape $\rightarrow$ a general property of ML estimators for the large sample limit, known as **asymptotic normality**.

RCF bound

Rao-Cramer-Frechet (RCF) inequality, also called the **information inequality**
Provides a lower bound on an estimator's variance.

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Bigg/ \left\langle -\frac{\partial^2 log L}{\partial \theta^2} \right\rangle$$

where b is the bias and L is the likelihood function.
In case of equality (i.e. minimum variance) the estimator is said to be **efficient**.
It can be shown that ML estimators are efficient in large sample limit.
In practice $\rightarrow$ Assume efficiency and zero bias.

RCF bound – contd.

Consider exponential distribution with mean τ

$$\frac{\partial^2 logL}{\partial \tau^2} = \frac{n}{\tau^2}\left(1 - \frac{2}{\tau}\frac{1}{n}\sum_{i=1}^{n} t_i\right) = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

Since b = 0, the RCF bound is

$$V[\hat{\tau}] \geq \frac{1}{\left\langle -\frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)\right\rangle} = \frac{1}{-\frac{n}{\tau^2}\left(1 - \frac{2\langle\hat{\tau}\rangle}{\tau}\right)} = \frac{\tau^2}{n}$$

This is same as what we got from exact calculation. In this case equality holds, since $\hat{\tau}$ is an efficient estimator for τ.

RCF bound – contd.

In case of more than one parameter, the corresponding formula for inverse of the covariance matrix $V_{ij} = cov\left[\hat{\theta}_i, \hat{\theta}_j\right]$ is

$$
\begin{aligned}
\left(V^{-1}\right)_{ij} &= \left\langle -\frac{\partial^2 logL}{\partial\theta_i \partial\theta_j} \right\rangle \\
&= \int ... \int \frac{\partial^2}{\partial\theta_i \partial\theta_j} \left(\sum_{k=1}^{n} logf(x_k;\theta) \right) \prod_{l=1}^{n} f(x_l;\theta) dx_l \\
&= n \int -f(x;\theta) \frac{\partial^2}{\partial\theta_i \partial\theta_j} logf(x;\theta) dx
\end{aligned}
$$

$f(x;\theta)$ is pdf for r.v. x, for which there are n measurements.
Note: $V^{-1} \propto n$ or $V \propto 1/n$
$\Longrightarrow$ a well known result that stat. errors decrease in $\propto 1/\sqrt{n}$

RCF bound – contd.

In many situations it is impractical to compute RCF bound analytically.
For sufficiently large data sample, V^{-1} can be estimated by evaluating 2nd derivative with the measuremed data and the ML estimates $\hat{\theta}$

$$\left(\widehat{V^{-1}}\right)_{ij} = -\left.\frac{\partial^2 logL}{\partial\theta_i \partial\theta_j}\right|_{\theta=\hat{\theta}}$$

For single parameter,

$$\widehat{\sigma^2_{\hat{\theta}}} = \left.\left(-1\Big/\frac{\partial^2 logL}{\partial\theta^2}\right)\right|_{\theta=\hat{\theta}}$$

Usual method for estimating cov. matrix when likelihood function is maximized numerically.

Graphical method

An extension of RCF bound

Expanding log-likelihood function about the ML estimate $\hat{\theta}$,

$$logL(\theta) = logL(\hat{\theta}) + \left[\frac{\partial logL}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2}\left[\frac{\partial^2 logL}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + ...$$

By definition $logL(\hat{\theta}) = logL_{max}$ and 2nd term is zero. Thus,

$$logL(\theta) = logL_{max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2_{\hat{\theta}}}}$$

or

$$logL(\hat{\theta} \pm \widehat{\sigma_{\hat{\theta}}}) = logL_{max} - \frac{1}{2}$$

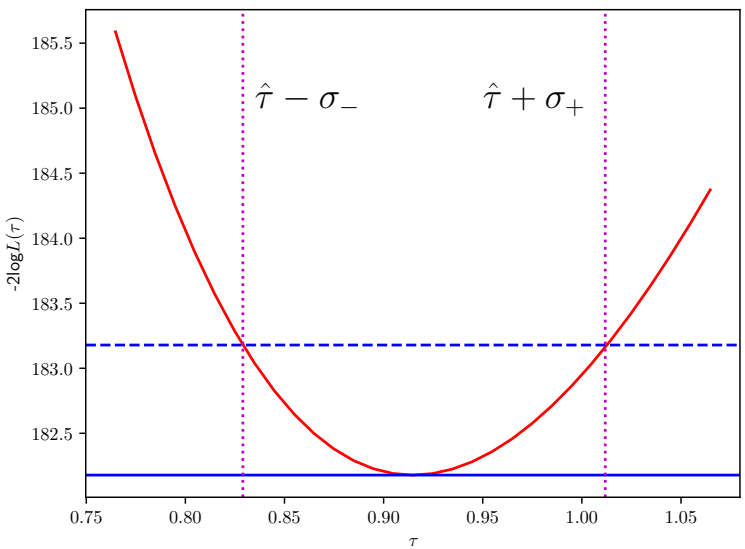
i.e. a change in the parameter θ of one standard deviation from its ML estimate leads to a decrease in the log-likelihood of $1/2$ from its maximum value.

It is also possible to show that the log-likelihood function becomes a parabola (i.e. the likelihood function becomes a Gaussian curve) in the large sample limit.

Graphical method – Example

More convenient to use $-2logL(\hat{\theta} \pm \widehat{\sigma_{\hat{\theta}}}) = -2logL_{max} + 1.$

Example: Consider again the example of exponential distribution. The log-likelihood function $-2log(\tau)$ as a function of the parameter τ for a Monte Carlo experiment consisting of 100 measurements.



The obtained one standard deviations in this case are $\Delta\hat{\tau}_- = 0.086$ and $\Delta\hat{\tau}_+ = 0.095$. Approximately same as $\widehat{\sigma_{\hat{\tau}}} = \hat{\tau}/\sqrt{n} = 0.091$

In this case $-2log(\tau)$ is reasonably close to a parabola.

Extended maximum likelihood

If the n (no. of observations) is itself a Poisson r.v with a mean $= \nu$, then the likelihood function becomes

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

called the **extended likelihood function**.

Two possible situtaions:
(1) when ν is a function of θ, and
(2) when ν is an independent parameter.

Extended maximum likelihood – contd.

When ν is a function of θ, the extended log-likelihood function is

$$
\begin{aligned}
logL(\theta) &= nlog\nu(\theta) - \nu(\theta) + \sum_{i=1}^{n} logf(x_i;\theta) + const. \\
&= -\nu(\theta) + \sum_{i=1}^{n} log(\nu(\theta)) + \sum_{i=1}^{n} logf(x_i;\theta) + const. \\
&= -\nu(\theta) + \sum_{i=1}^{n} log(\nu(\theta)f(x_i;\theta))
\end{aligned}
$$

Including the Poisson term the resulting estimators $\hat{\theta}$ exploits the information from n as well as from the variable x $\Rightarrow$ smaller variances for $\hat{\theta}$.

Extended maximum likelihood – contd.

If ν does not depend on $\theta \Rightarrow \hat{\nu} = n$, and $\hat{\theta}_i$ are same as the usual ML case.
However, still helpful in cases, e.g. when the pdf is the superposition of several components,

$$f(x;\theta) = \sum_{i=1}^{m} \theta_i f_i(x),$$

where, $f_i(x)$ are all known and θ_i are not all independent, but $\sum_{i=1}^{m} \theta_i = 1$.
Then the $logL$ becomes

$$logL(\nu,\theta) = -\nu + \sum_{i=1}^{n} log\left(\sum_{j=1}^{m} \nu\theta_j f_j(x_i)\right)$$

Extended maximum likelihood – contd.

Defining $\mu_i = \theta_i \nu$,

$$logL(\mu) = -\nu \sum_{j=1}^{m} \theta_j + \sum_{i=1}^{n} log \left(\sum_{j=1}^{m} \nu \theta_j f_j(x_i) \right)$$

$$= \sum_{j=i}^{m} \mu_j + \sum_{i=1}^{n} log \left(\sum_{j=1}^{m} \mu_j f_j(x_i) \right)$$

- Parameters $\mu = (\mu_1, ..., \mu_m)$ are not subject to a constraint and n is a sum of independent Poisson variables with means μ_j
- Estimators $\hat{\mu}_j$ give directly the estimated mean numbers of events of different types, which is equivalent to $\hat{\mu}_j = \hat{\theta}_j \hat{\nu} = \hat{\theta}_j n$

Extended maximum likelihood - Example

Let data sample consisting of two types of events: **signal** and **background**
$f_s(x)$ is **Gaussian** and $f_b(x)$ is **Exponential**

Number of signal events: n_s (Poisson distributed with mean μ_s),
Number of bkg events: n_b (Poisson distributed with a mean μ_b)

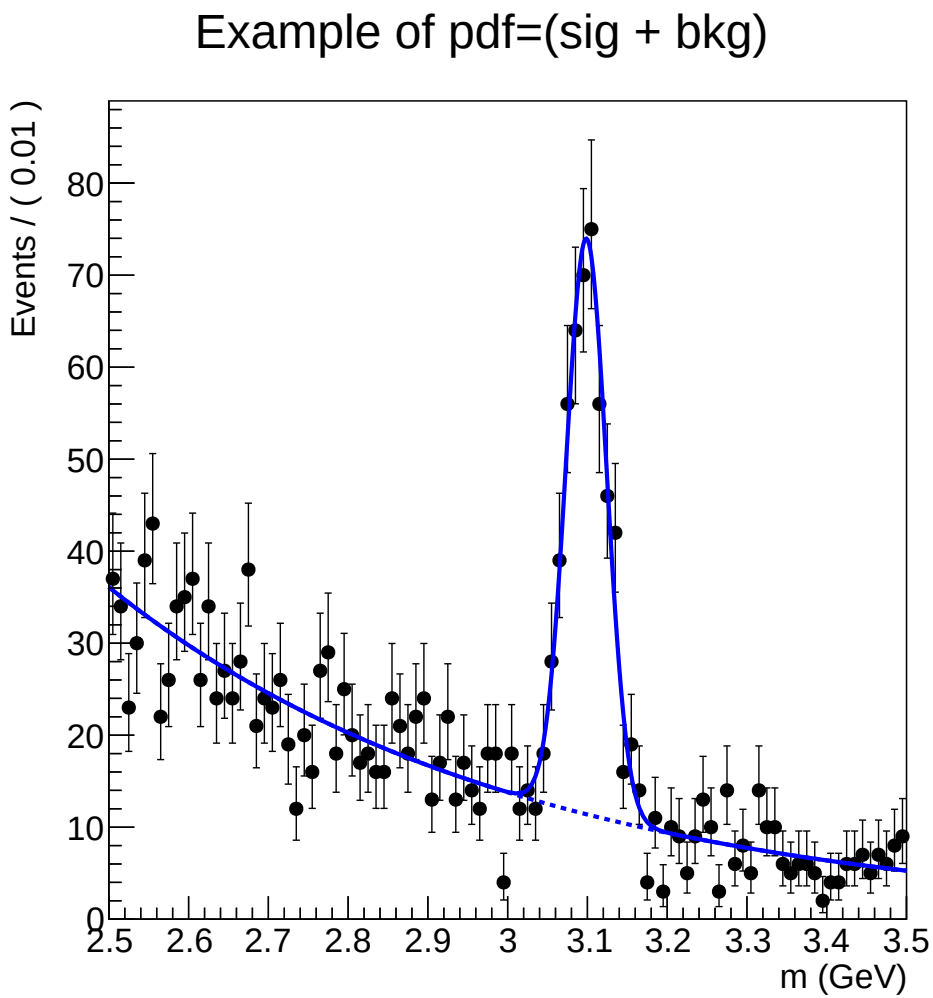
The pdf of x:

$$f(x) = \frac{\mu_s}{\mu_s + \mu_b} f_s(x) + \frac{\mu_b}{\mu_s + \mu_b} f_b(x)$$

Suppose we observed $n = n_s + n_b$ events. Fit to estimate μ_s and μ_b.

Extended maximum likelihood - Example

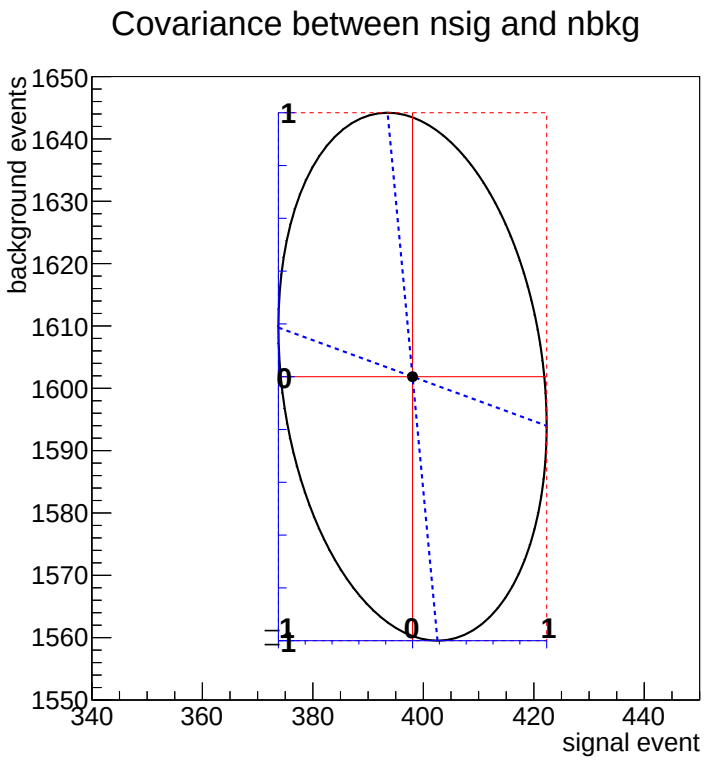
MC samples generated using using $\mu_s = 400$ and $\mu_b = 1600$
Extended ML fit for both μ_s and μ_b.

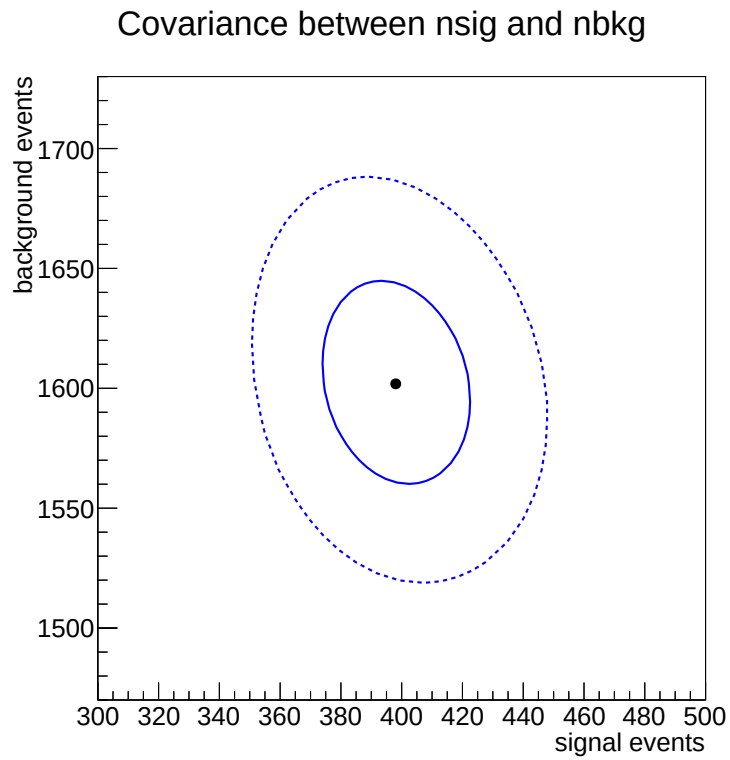


The estimated values from the fit:
$\hat{\mu}_s = 8.7$ and
$\hat{\sigma}_s = 5.5$

Extended maximum likelihood - Example

For more than one parameters the covariance matrix can be computed.





Countour of constant likelihood for one and two standard deviations.

The tangets to the curve correspond to $\hat{n_s} \pm \hat{\sigma_{n_s}}$ and $\hat{n_b} \pm \hat{\sigma_{n_b}}$.

ML with binned data

For very large data sample one can make histograms instead of recording each measurement separately.
Number of entries $\mathbf{n} = n_1, ...n_N$ in N bins, with $\sum_{i=1}^{N} n_i = n_{tot}$.
The expectation values $\nu = \nu_1, ..., \nu_N$ of the numbers of entries are

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

Considering the histogram as a single measurement of an N-dimensional random vector, joint pdf is

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

ν_i/n_{tot} = probability for the event to be in bin i.
And, the $logL$ function

$$logL(\theta) = \sum_{i=1}^{N} n_i log\nu_i(\theta) + const.$$

ML with binned data – contd.

Suppose n_{tot} is a r.v. from a Poisson distribution with mean ν_{tot}. The joint pdf will be

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

'where, $\nu_{tot} \;=\; \sum_{i=1}^{N} \nu_i$ and $n_{tot} \;=\; \sum_{i=1}^{N} n_i$. Implies,

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

where the expected number of entries in each bin ν_i now depends on the parameters θ and ν_{tot},

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

This is equivalent to treating the number of entries in each bin as an independent Poisson r.v. n_i with mean value ν_i.

ML with binned data – contd.

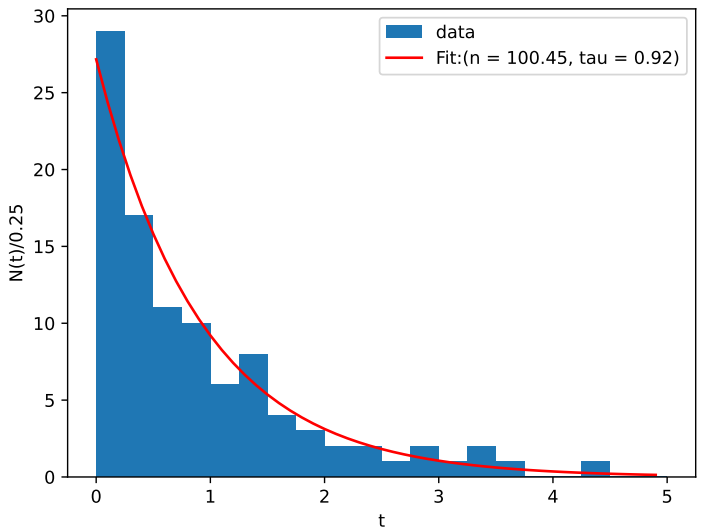
The $logL$ (dropping the constant terms) becomes,

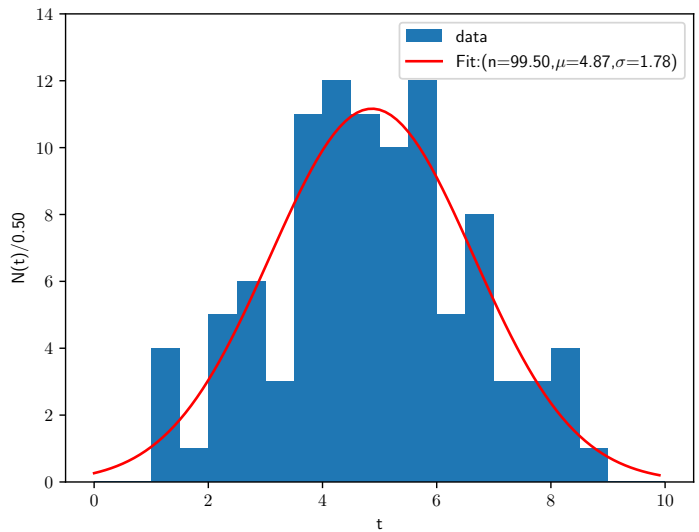
$$logL(\nu_{tot}, \theta) \;=\; -\nu_{tot} \;+\; \sum_{i=1}^{N} n_i log\nu_i(\nu_{tot}, \theta)$$

This is the extended log-likelihood function for the case of binned data.

ML with binned data - Example

Consider our previous exercises of 100 measurements
Histograms with a bin width of $\Delta t = 0.25$ along with the results of the ML fit.





Good agreements with unbinned fit results.

Testing goodness-of-fit

Principle of ML does not directly suggest a method of testing goodness-of-fit.

Possible in some cases to obtain a g.o.f. measurement by finding a proper ratio of likelihood functions

e.g. Consider the ratio

$$\lambda = \frac{L(\mathbf{n}|\nu)}{L(\mathbf{n}|\mathbf{n})} = \frac{f_{joint}(\mathbf{n};\nu)}{f_{joint}(\mathbf{n};\mathbf{n})}$$

For Poisson distributed data

$$\lambda_P = e^{n_{tot}-\nu_{tot}} \prod_{i=1}^{N} \left(\frac{\nu_i}{n_i}\right)^{n_i}$$

If the hypothesis is correct, in the large sample limit, the statistic

$$\chi^2_P = -2log\lambda_P = 2\sum_{i=1}^{N} \left(n_i log\frac{n_i}{\hat{\nu}_i} + \hat{\nu}_i - n_i\right)$$

follows a χ^2 distribution for $N - m$ degrees of freedom.

Method of least squares

Relation to Maximum Likelihood

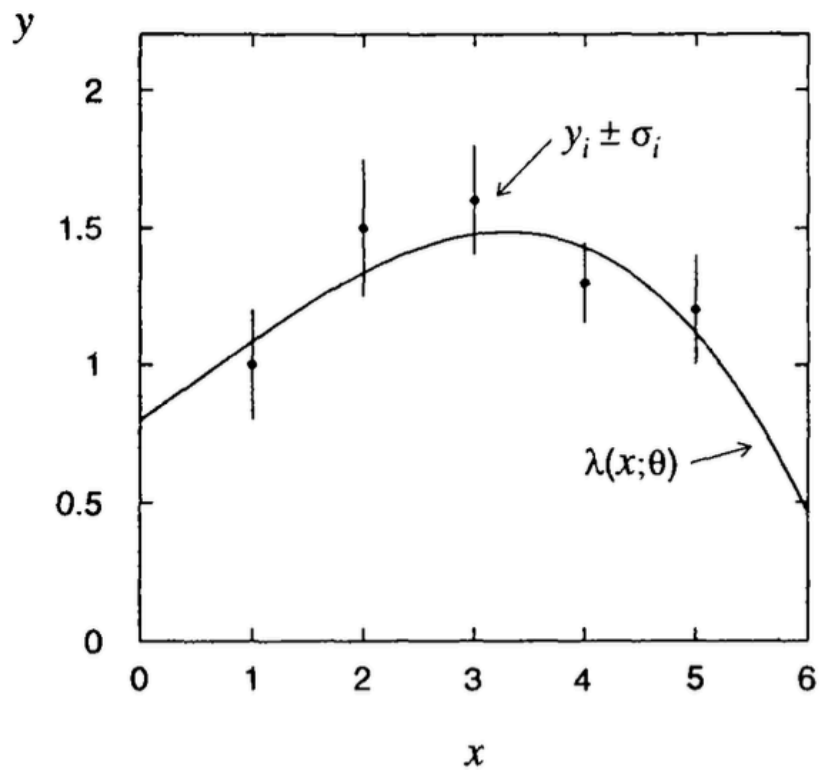
Usually a measured value y can be regared as Gaussian r.v. centered around true value λ (Follows from the CLT).

Consider N independent Gaussian r.v. y_i, related to another variable x_i, e.g. some measurements at positions x_i.

Assume λ_i (unknown) are mean of y_i and σ_i^2 (known) are variances.

y_is can be regarded as a single measurement of N-d random vector, the joint pdf will be

$$g(y_1, ..., y_N; \lambda_1, ..., \lambda_N, \sigma_1^2, ..., \sigma_N^2)$$
$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} exp\left(\frac{-(y_i - \lambda_i)^2}{2\sigma_i^2}\right)$$



Relation to Maximum Likelihood – contd.

Suppose $\lambda = \lambda(x;\theta)$, estimate θ, where $\theta = (\theta_1, ..., \theta_m)$ are unknown parameters.

The logarithm of the joint pdf (or the likelihood) [dropping additive terms]

$$logL(\theta) = -\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

Maximizing $logL(\theta)$ is same as minimizing

$$\chi^2(\theta) = -2logL(\theta) = \sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

This is the basis of the *method of least squares (LS)*. Also used when y_i are not Gaussian, as long as they are independent.

The parameters that minimize the χ^2 are called the LS estimators, $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)$. The resulting mimimum χ^2 follows, under certian circumstances, the Chi-square distribution.

χ^2 for non-independent measurements

If y_i are described by N-d Gaussian PDF with known covariance matrix V, the log-likelihood from joint PDF is

$$logL(\theta) = -\frac{1}{2}\sum_{i,j=1}^{N}(y_i - \lambda(x_i;\theta))(V^{-1})_{ij}(y_j - \lambda(x_j;\theta))$$

and, therefore

$$\chi^2(\theta) = \sum_{i,j=1}^{N}(y_i - \lambda(x_i;\theta))(V^{-1})_{ij}(y_j - \lambda(x_j;\theta))$$

Will reduce to previous expression if V is diagonal.

Linear least-square fit

If λ is a linear function of θ

$$\lambda(x;\theta) = \sum_{j=1}^{m} a_j(x)\theta_j$$

$a_j(x)$ are any linearly independent functions of x.
In this case, the estimators and their variances can be found analytically. Also, the estimators have zero bias and minimum variance.
At x_i,

$$\lambda(x_i;\theta) = \sum_{j=1}^{m} a_j(x_i)\theta_j = \sum_{j=1}^{m} A_{ij}\theta_j$$

Then, in matrix notation

$$\begin{aligned}\chi^2 &= (\mathbf{y}-\lambda)^T V^{-1}(\mathbf{y}-\lambda) \\ &= (\mathbf{y}-A\theta)^T V^{-1}(\mathbf{y}-A\theta)\end{aligned}$$

where $\mathbf{y} = (y_1, ..., y_N)$ and $\lambda = (\lambda_1, ..., \lambda_N)$.

Linear least-square fit

Minimizing χ^2 w.r.t θ_i,

$$\Delta\chi^2 = -2(A^T V^{-1}\mathbf{y} - A^T V^{-1} A\theta) = 0$$

If $A^T V^{-1} A$ is not singular,

$$\hat{\theta} = (A^T V^{-1} A)^{-1} A^T V^{-1}\mathbf{y} \equiv B\mathbf{y}$$

i.e. $\hat{\theta}$ are linear functions of the original measurements $\mathbf{y}$.

Variance of LS estimators

Using error propagation to find the covariance matrix $U_{ij} = cov[\hat{\theta}_i, \hat{\theta}_j]$,

$$U = BVB^T = (A^T V^{-1} A)^{-1}$$

Equivalently,

$$(U^{-1})_{ij} = \frac{1}{2}\left[\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right]\Bigg|_{\theta=\hat{\theta}}$$

Note: it coincides with RCF bound when y_i are Gaussian distributed.

Variance of LS estimators

For $\lambda(x;\theta)$ linear in θ, χ^2 is quadratic in θ

$$\chi^2(\theta) = \chi^2(\hat{\theta}) + \frac{1}{2}\sum_{i,j=1}^{m}\left[\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right]\Bigg|_{\theta=\hat{\theta}}(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

Combining with expression for variance yields 1σ contours in parameter space. i.e. for $\theta_i = \hat{\theta}_i \pm \hat{\sigma}_i$

$$\chi^2(\theta) = \chi^2(\hat{\theta}) + 1 = \chi^2_{min} + 1$$

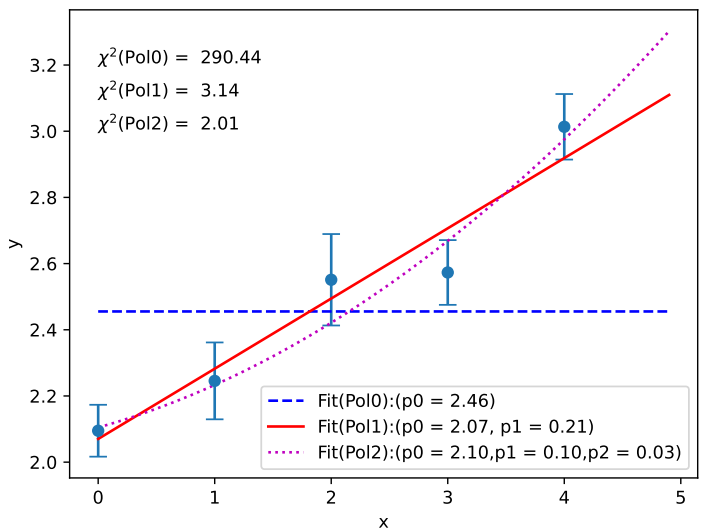
Similar to the contour of constant likelihood in the ML method.

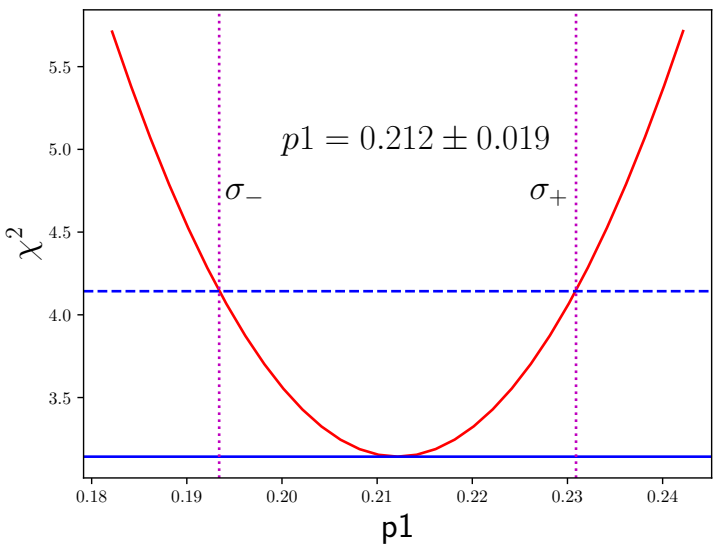
Note: If λ is not linear in θ, then the contour is not in general elliptical. So, the tangents do not correspond to one standard deviations, but defines a region in parameter space which can be interpreted as a *confidence region*.

Example: Least squares fit of a polynomial

Consider λ a polynomial of order m (i.e. $m+1$ parameters)

$$\lambda(x;\theta_0,...,\theta_m) = \sum_{j=0}^{m} x^j \theta_j$$





Example of fit to 0th, 1st, and 2nd order polynomials for five measured points. χ^2 function as a function of the parameter $p1$ (slope) is shown for 1st order polynomial.

LS with binned data

Consider a binned data of n observations of x, filled into a histogram with N bins.
If y_i = number of entries in bin i, $f(x;\theta)$ is a hypothesized PDF, then, the number of entries predicted in bin i, $\lambda_i = \langle y_i \rangle$,

$$\lambda_i(\theta) = n \int_{x_i^{min}}^{x_i^{max}} f(x;\theta)dx = np_i(\theta)$$

where, $p_i(\theta)$ is the probability to have an entry in bin i.
Then, χ^2 function becomes

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\theta))^2}{\sigma_i^2}$$

σ_i^2 is the variance of the no. of entries in bin i.

LS with binned data

If $\langle y_i \rangle = \lambda_i$ are small compared to the total number of entries, then, y_i are approximately Poisson distributed. i.e. variance = mean. Thus,

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\theta))^2}{\lambda_i(\theta)} = \sum_{i=1}^{N} \frac{(y_i - np_i(\theta))^2}{np_i(\theta)}$$

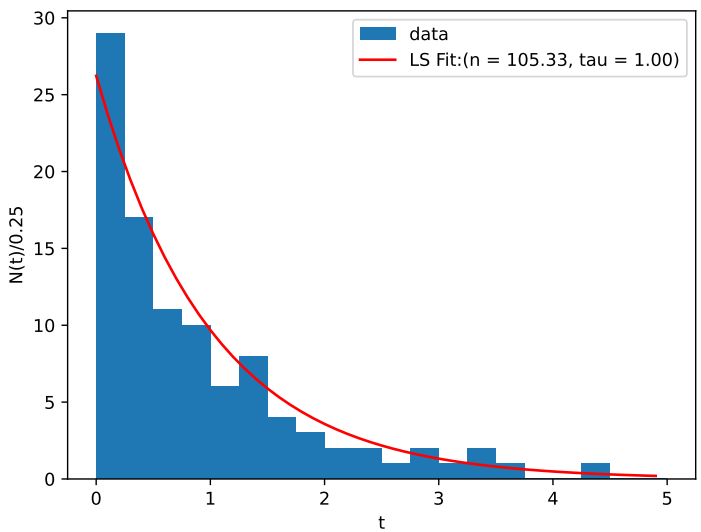
Alternatively, one can approximate $\sigma_i^2 = y_i$ (entries actually observed instead of predicted). In that case

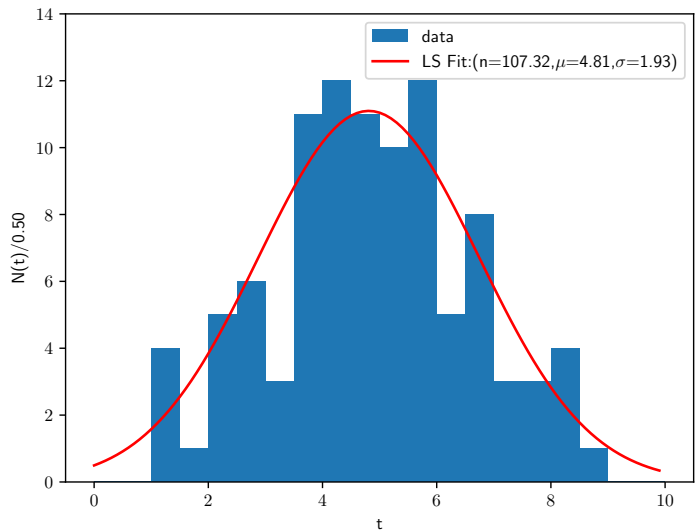
$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\theta))^2}{y_i} = \sum_{i=1}^{N} \frac{(y_i - np_i(\theta))^2}{y_i}$$

So-called **modified least-squares method** (MLS method). Advantage: computationally easier, Disadvantage: errors may be poorly estimated (or χ^2 may even be undefined) if any of the bins contain few or no entries.

LS with binned data – Example

Consider our previous exercises of 100 measurements
Histograms with a bin width of $\Delta t = 0.25$ along with the results of the LS fit.





Good agreements with ML fit results.

Testing goodness-of-fit

$(y_i - \lambda(x_i;\theta))/\sigma_i$ is a measure of the deviation between y_i and the function $\lambda(x_i;\theta) \Rightarrow \chi^2$ is a measure of total agreement between observed data and hypothesis.

If,

1. y_i ($i = 1$ to N) are independent Gaussian r.v. with known variances (or N-d Gaussian with known cov. matrix),
2. the hypothesis λ is linear in θ_j ($j = 1$ to m),
3. the functional form of the hypothesis is correct,

then, χ^2_{min} is distributed according the χ^2 distribution with $N - m$ d.o.f.

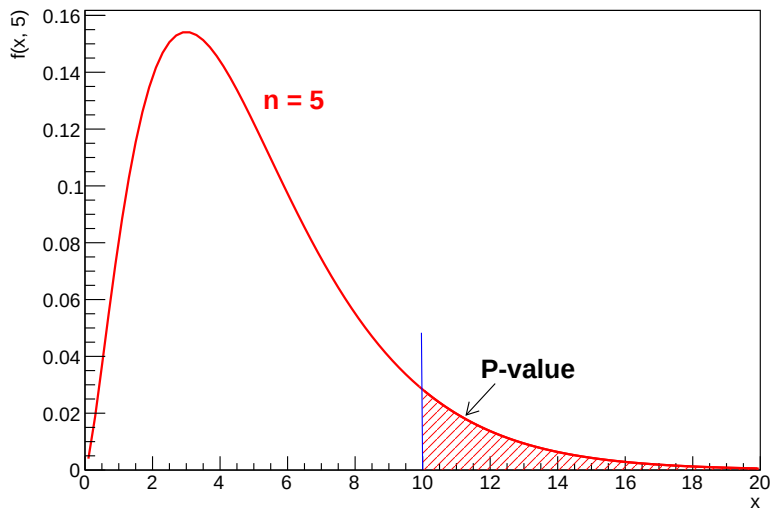
Testing goodness-of-fit

We know $\langle z \rangle = n_d$ for χ^2 distribution.
Thus, χ^2/n_d is a measure of the g.o.f.

If

- χ^2/n_d is similar to 1: Fit is good.
- χ^2/n_d is much less than 1: Fit is too good. Should check whether errors are overestimated or correlated.
- χ^2/n_d is much larger than 1: Decide whether the hypothesis can be rejected based on the P-value.

$$P = \int_{\chi^2}^{\infty} f(z; n_d) dz$$



END