Statistical methods for neutrino physics (1/2)

Alessandra Tonazzo
Université Paris-Diderot, Laboratoire APC
tonazzo@in2p3.fr

Why ?

4 Physics Topics

Within the realm of neutrino physics, subjects for which statistical issues seem particularly relevant and which produced interesting discussions included:

- Fitting parameters for 3 neutrino oscillation situations
- Searching for sterile neutrinos
- Determining the neutrino mass hierarchy
- Determining the CP phase
- Searching for rare processes, e.g. ultra high energy cosmic neutrinos, neutrinoless double beta decay*, supernovae neutrinos, etc.
- Neutrino cross-sections
- Reconstruction and classification issues, e.g. for rings in Cerenkov detectors

L. Lyons, arXiv:1705.01874 [hep-ex]

PHYSTAT-ν Workshop Series :

- May 30 - June 1, 2016, IPMU, Japan
- September 19-21, 2016, Fermilab, USA

What we need

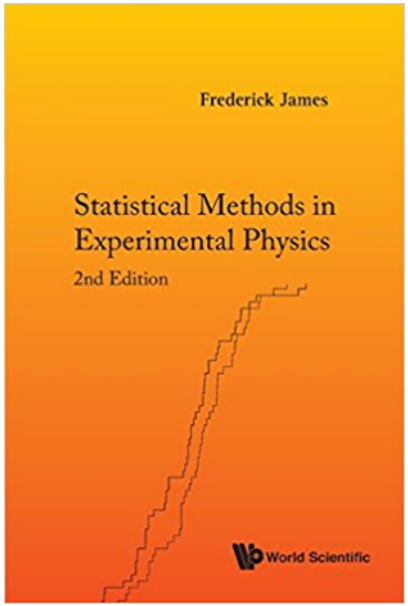
TOPIC 1

- Estimation of parameters
  - the likelihood method
  - the min-chi2 method
- Estimation of confidence intervals
  - Neyman's C.I.and belt method
  - Feldman-Cousins construction
  - Use of the likelihood, 1D and ND => tomorrow…
- *Tutorial n. 1: Feldman-Cousins construction of confidence intervals*

TOPIC 2

- Test of hypotheses
  - the case of neutrino Mass Hierarchy in future experiments
- Goodness-of-fit
- *Tutorial n. 2: Sensitivity of future experiment to Mass Hierarchy*

References



- F. James, “Statistical methods in experimental physics”, World Scientific
- L. Lyons, “Statistics for Nuclear and Particle Physicists”, Cambridge University Press
- R. J.Barlow, “A guide to the use of statistical methods in the physical science”, wiley
- L. Lista, “Statistical Methods for Data Analysis in Particle Physics”, Springer
- CERN Academic Training programme
- … and, once you understand what it’s doing: RooFit/RooStats in Root
  https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome

Reminders

Probability laws for a random variable X :

- **X discrete**: N possible values $x_1 \ldots x_N$, $p_i = P(X = x_i)$ with $0 \le p_i \le 1$ and $\sum_{i=1}^{N} p_i = 1$
- **X continuous**: probability law specified by the cumulative function F(x)

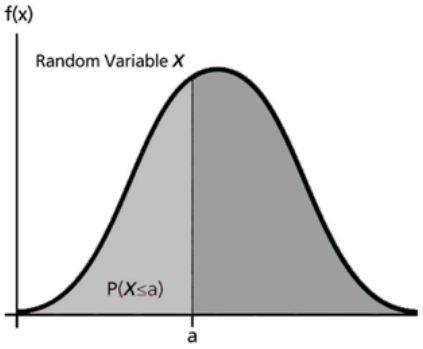
$$F(x_0) = P(x \le x_0) \quad ; \quad F(-\infty) = 0, F(+\infty) = 1 \qquad F \text{ monotonic}$$

the Probability Density Function (PDF) is *f(x)* such that

$$f(x)dx = F(x+dx) - F(x) = P(x \in [x, x+dx])$$

$$\text{so } f(x) = \frac{dF}{dx}$$

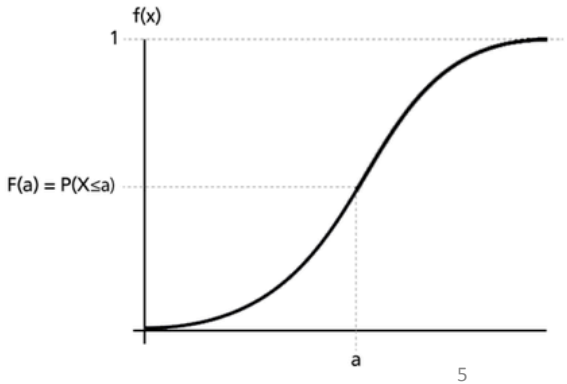
Multidimensional PDF for random variables X,Y,Z…: f(x,y,z,…)



The Central Limit Theorem

N independent random variables $X_1,\ldots,X_N$
each having a PDF $f_i(x_i)$ with mean μ_i and variance σ_i^2

$$S = \sum X_i \qquad PDF(S) \underset{N \to \infty}{\to} Gaus\left(\mu = \sum \mu_i, \sigma = \sqrt{\sum \sigma_i^2}\right)$$



Reminders

Properties of distributions

given a random variable X with PDF *f(x)*

- Expectation of a function *g(x)* : $E[g(x)] \equiv \langle g(x) \rangle = \int g(x) f(x) dx$
- Mean = expectation of X $E[X] = \int x f(x) dx$
- Variance = expectation of $(x-\mu)^2$ $V[X] = E[(x - E[X])^2] = E[x^2] - (E[X])^2$
  - standard deviation $\sigma = \sqrt{V}$
- Covariance (multi-dimensional case) $C_{XY} = E\left[\left(x - E[X]\right)\left(y - E[Y]\right)\right]$
  - Correlation coefficient $\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y}$
- Variance-Covariance matrix of N random variables

$$V = \begin{pmatrix} \sigma_1^2 & C_{12} & \dots & C_{1N} \\ C_{12} & \sigma_2^2 & & \dots \\ \dots & & \dots & \dots \\ C_{1N} & \dots & \dots & \sigma_N^2 \end{pmatrix}$$

Reminders

- Bayes' theorem (on conditional probabilities)
  - A and B are sets of <u>events</u> for random variables X_i

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

frequentist !

- Bayesian use of Bayes' theorem
  - not events, but <u>hypotheses</u>
  - $P(\theta_i)$ = "degree of belief" in hypothesis θ_i
  - X = observed data

$$P(\theta_i \mid X) = \frac{P(X \mid \theta_i)P(\theta_i)}{P(X)}$$

Posterior probability

Prior probability

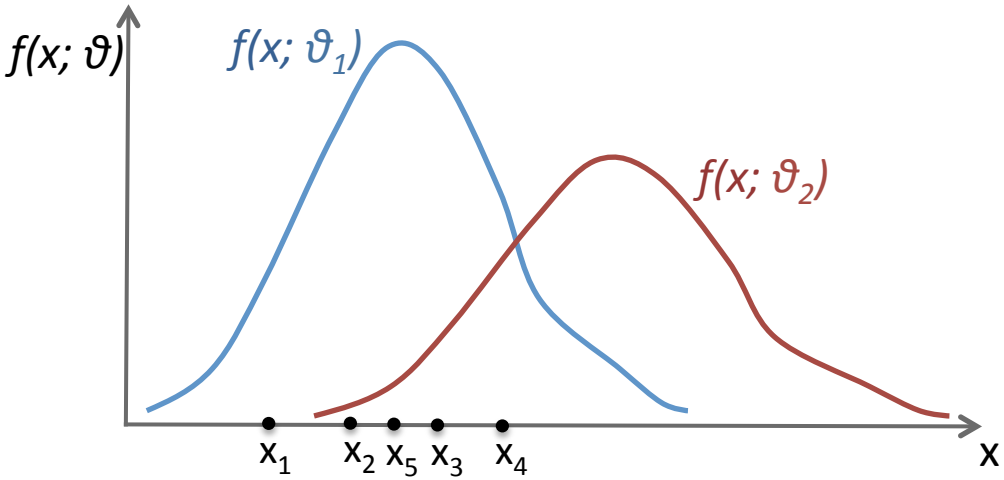
ESTIMATION OF PARAMETERS

Parameter estimation

- X: random variable with p.d.f. $f(x;\vartheta_0)$, where ϑ_0 is the true value of an unknown parameter ϑ
- N independent trials of X: $x_1, \ldots, x_N$

What can we say about the value of ϑ_0 ?

Example (N=5) :



Would you say $\vartheta_0=\vartheta_1$ or $\vartheta_0=\vartheta_2$?

Parameter estimation

- X: random variable with p.d.f. $f(x;\vartheta_0)$, where ϑ_0 is the true value of an unknown parameter ϑ
- N independent trials of X: $x_1, \dots, x_N$

What can we say about the value of ϑ_0 ?

Aim: construct a random variable, function of the x_i, whose expectation value is ϑ_0 (at least asymptotically) and with variance as small as possible

$t_N = h(x_1, \dots, x_N)$: $E[t_N] \to \vartheta_0$

t_N is an estimator or statistics for ϑ_0

Properties of estimators

$$t_N = h(x_1, \ldots, x_N)$$

- t_N is an unbiased estimator of ϑ_0 if $E[t_N] = \vartheta_0$
  - def.: "Bias" $b_N = E[t_N] - \vartheta_0$
- t_N is a consistent or convergent estimator of ϑ_0 if, as N->∞,

  b_N->0 like 1/N and $V(t_N)$ ->0 like 1/N

- an estimator which is unbiased and has smaller variance than any other is optimal
- t_N is an efficient estimator if it is unbiased and its variance reaches the theoretical lower bound, the "Minimum Variance Bound" (Information)

$$I_N(\vartheta) = -NE\left[\frac{\partial^2 \ln L}{\partial \vartheta^2}\right] = MVB^{-1}$$

- t_N is robust if it is independent of the assumptions on the p.d.f. for ϑ

ESTIMATION OF PARAMETERS
The Maximum Likelihood method

The Maximum Likelihood method

Compute the **Likelihood** (="joint probability") of the set of N independent trials:

$$L(x_1,\ldots,x_N;\vartheta)=\prod_{i=1}^{N} f(x_i;\vartheta)$$

The **Maximim Likelihood Estimator (MLE)** of the parameter ϑ_0 is the value $\hat{\vartheta}_{ML}$ for which $L(x;\vartheta)$ has its maximum, given the particular set of observations $x_1 .. x_N$

It is easier to compute sums than products: take the log

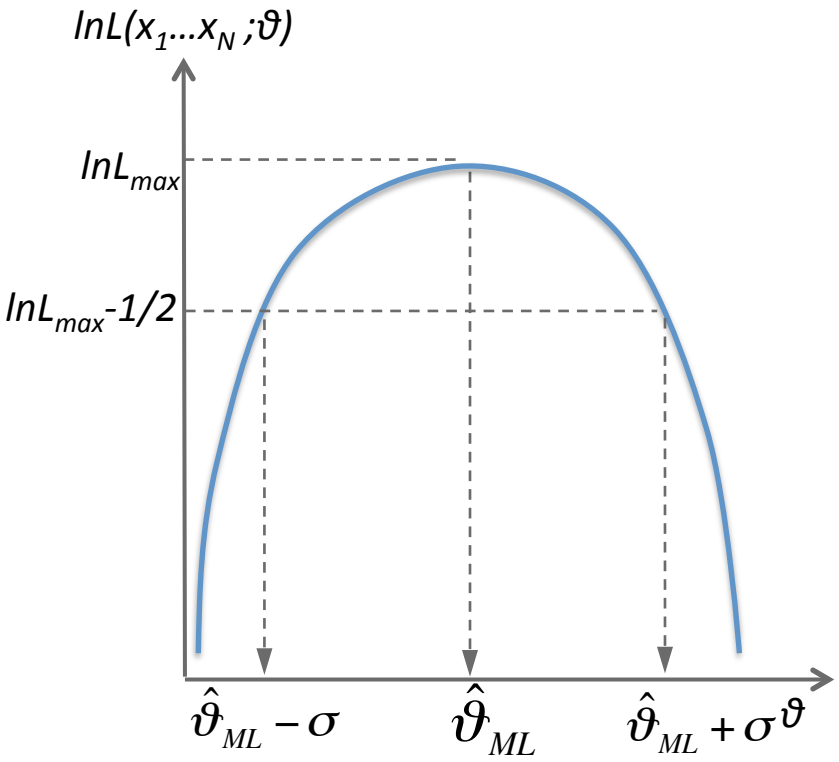
$$\ln L(x_1,\ldots,x_N;\vartheta)=\sum_{i=1}^{N} \ln f(x_i;\vartheta)$$

Likelihood equation :

$$\frac{\partial \ln L(x_1,\ldots,x_N;\vartheta)}{\partial \vartheta}=\sum_{i=1}^{N}\frac{\partial \ln f(x_i;\vartheta)}{\partial \vartheta}=0$$

$\hat{\vartheta}_{ML}$ is a root of the likelihood equation (if it exists)

The Maximum Likelihood method



Error on the ML estimator

$$V\left(\hat{\vartheta}_{ML}\right) \underset{N\to\infty}{\to} -E\left[\left.\frac{\partial^2 \ln L}{\partial \vartheta^2}\right|_{\vartheta=\vartheta_0}\right]^{-1} \sim -E\left[\left.\frac{\partial^2 \ln L}{\partial \vartheta^2}\right|_{\vartheta=\hat{\vartheta}}\right]^{-1}$$

the MVB !

Since *L* is (asymptotically) Gaussian because of the CLT,
⇒ L_{max}-1/2
gives the "1-sigma" error

$$\frac{\partial \ln L(x_1,\ldots,x_N;\theta)}{\partial\theta} = \sum_{i=1}^{N} \frac{\partial \ln f(x_i;\theta)}{\partial\theta} = 0$$

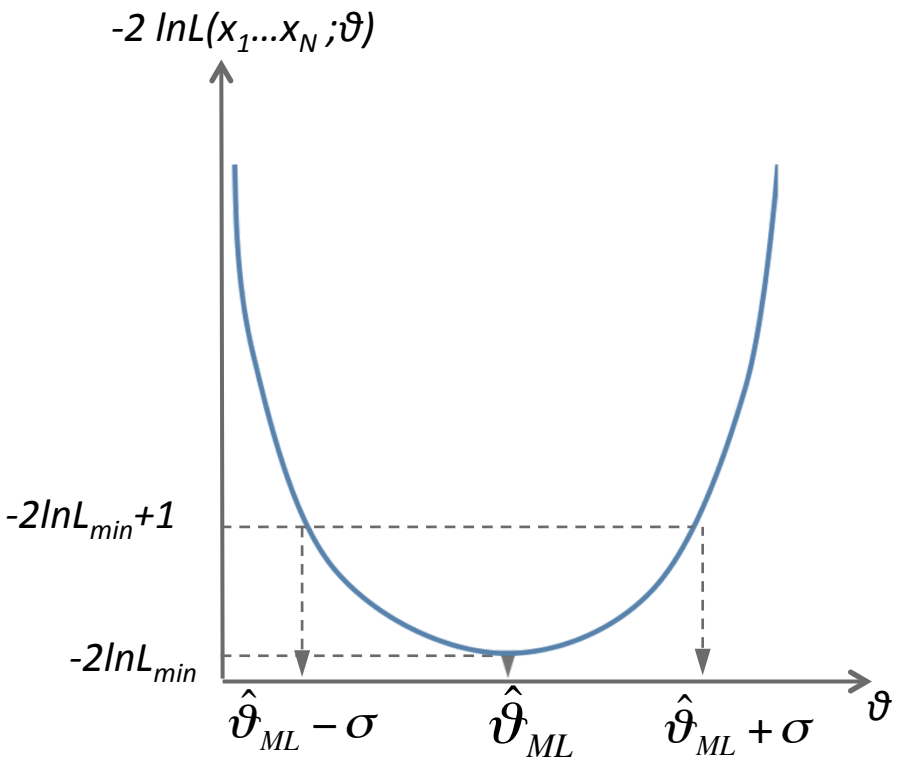
$\hat{\vartheta}_{ML}$ is a root of the likelihood equation (if it exists)

MLE in practice

It is easier for computer algorithms to find a minimum than a maximum, and we like integers better…

=> minimize *-2 ln L*

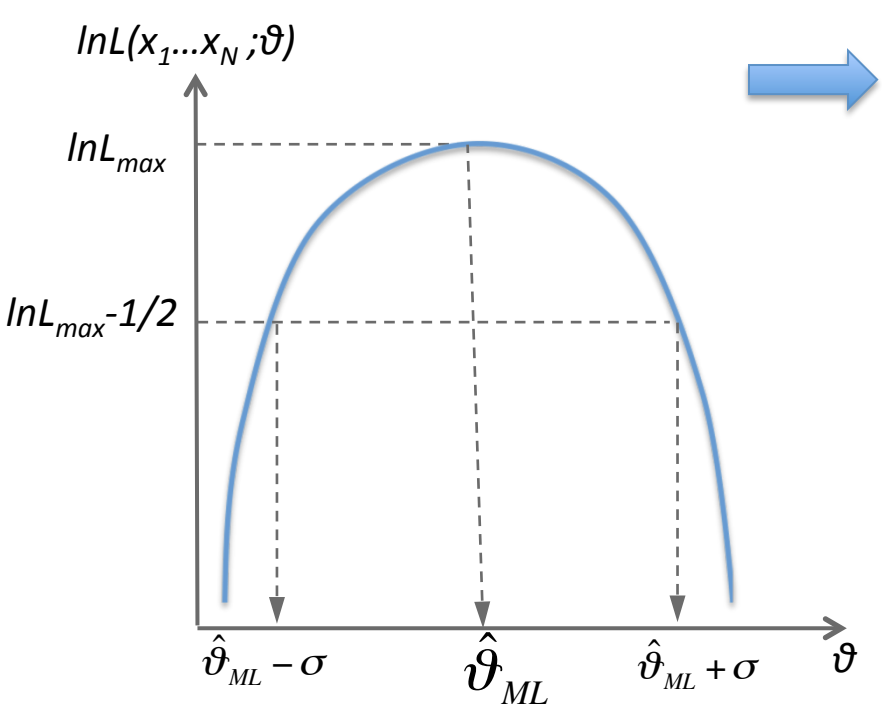
MINUIT package (in Root)

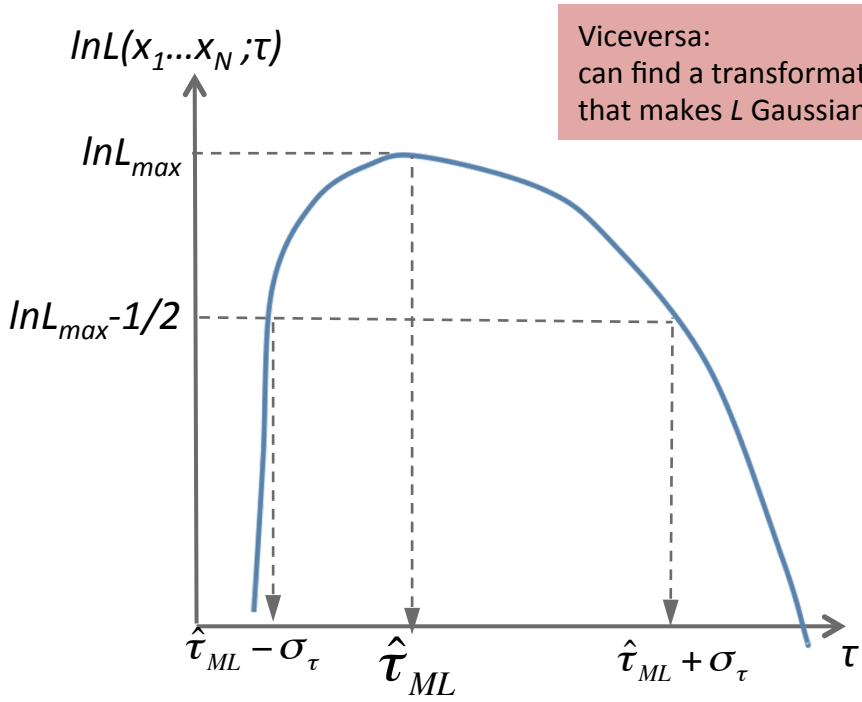


… but we will stick to maximising lnL

The MLE : properties

- consistent
- asymptotically normally distributed, with minimum variance (but may have more than one max for finite N)
- for finite N, optimal only under Darmois theorem (exp. family): $f(x;\vartheta) = \exp\left(a(x)\cdot\alpha(\vartheta) + b(x) + \beta(\vartheta)\right)$
- <u>invariance</u>: the MLE $\hat{\tau}$ of a function $\tau(\vartheta)$ is $\hat{\tau} = \tau\left(\hat{\vartheta}\right)$





Viceversa:
can find a transformation that makes *L* Gaussian !

Simple examples

- MLE of the mean of a Gaussian $f(x;\mu)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ σ known

$$L(x_1,..,x_N;\mu)=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^N}\prod_{i=1}^{n}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \qquad \ln L(x_1,..,x_N;\mu)=-\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}-N\ln\left(\sqrt{2\pi}\sigma\right)$$

$$\frac{\partial \ln L(x_1,..,x_N;\mu)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)=0 \quad\Leftrightarrow\quad \mu=\frac{1}{N}\sum_{i=1}^{N}x_i\equiv\hat{\mu}_{ML}$$

The MLE of the mean is the sample average: $\hat{\mu}_{ML}=\bar{x}$

it is easy to verify that $\hat{\mu}_{ML}\to\mu$ for $N\to\infty$ (unbiased) and $V(\hat{\mu}_{ML})=\frac{\sigma^2}{N}$ (efficient)

- Estimate of σ, or rather of the Variance $V=\sigma^2$ (not strictly with ML)

it is natural to use the sample variance $\hat{V}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$, which is "good" if μ is known.

If μ is not known it needs to be replaced by $\hat{\mu}_{ML}=\bar{x}$ and…

$$\sum_{i=1}^{N}(x_i-\bar{x})^2=\sum_{i=1}^{N}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)=\sum_{i=1}^{N}x_i^2-2\bar{x}\sum_{i=1}^{N}x_i+N\bar{x}=\sum_{i=1}^{N}x_i^2-2N\bar{x}+N\bar{x}=\sum_{i=1}^{N}\left(x_i^2-\bar{x}^2\right)$$

$$E\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i^2-\bar{x}^2\right)\right]=E\left[\frac{1}{N}\sum_{i=1}^{N}\left((x_i-\mu)-(\bar{x}-\mu)\right)^2\right]=\frac{1}{N}\left(E\left[\sum_{i=1}^{N}(x_i-\mu)^2\right]-E\left[(\bar{x}-\mu)^2\right]\right)=V(x)-V(\bar{x})=V(x)-\frac{1}{N}V(x)$$

so $\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right]=V(x)\left(1-\frac{1}{N}\right)\neq V(x)$ => use "Bessel's correction" to obtain an unbiased estimator $\hat{V}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$

Simple examples

- ML and weighted average

measurements with the same mean μ and different variances $V_i=\sigma_i^2$, which are known: estimate μ

$$f(x_i;\mu)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right)$$

$$L(x_1,..,x_N;\mu)=\frac{1}{\left(\sqrt{2\pi}\sigma_i\right)^N}\prod_{i=1}^{n}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \qquad \ln L(x_1,..,x_N;\mu)=-\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma_i^2}-N\ln\left(\sqrt{2\pi}\sigma_i\right)$$

$$\frac{\partial \ln L(x_1,..,x_N;\mu)}{\partial\mu}=\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma_i^2}=0 \quad \Leftrightarrow \quad \mu=\frac{\sum_{i=1}^{N}(x_i/\sigma_i^2)}{\sum_{i=1}^{N}(1/\sigma_i^2)}\equiv\hat{\mu}_{ML}$$

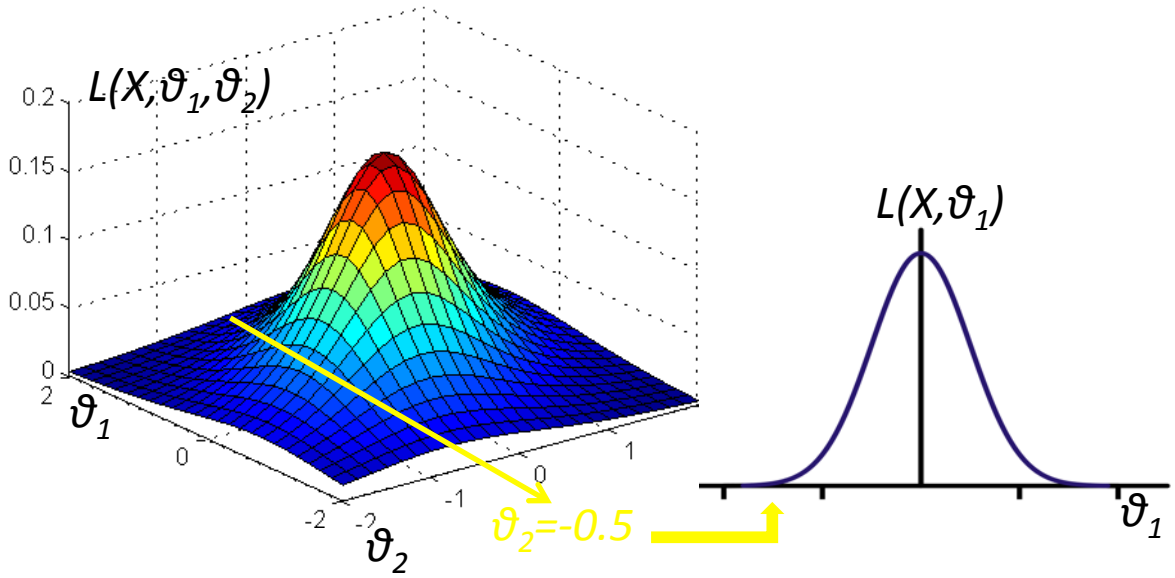
The MLE of the mean is the weighted average

Nuisance parameters

Parameters whose value are not of interest
but which need to be taken into account to estimate the parameter of interest

One option :
- fix the nuisance parameters to a given value
- maximise L with respect to interesting parameters

=> "PROFILE LIKELIHOOD"



A more general approach :
- assume the Likelihood to factorize
- assume normal distributions with known mean and sigma for the nuisance parameters

$$\ln L\left(X \mid \vartheta_1, \vartheta_2\right) = \ln L\left(X \mid \vartheta_1\right) - \frac{\left(\vartheta_2 - \mu\right)^2}{2\sigma^2}$$

Nuisance term

ESTIMATION OF PARAMETERS
The Least Squares method

The Least Squares method

N observations

$\vec{X} = \{x_1, \ldots, x_N\}$

Expectation values depending on k parameters $\vec{\vartheta} = \{\vartheta_1, \ldots, \vartheta_K\}$

$\vec{M}(\vec{\vartheta}) = \{m_1(\vec{\vartheta}), \ldots, m_N(\vec{\vartheta})\} = \{E[X_1](\vec{\vartheta}), \ldots, E[X_N](\vec{\vartheta})\}$

V: covariance matrix of the data (NxN)

Consider the quadratic form

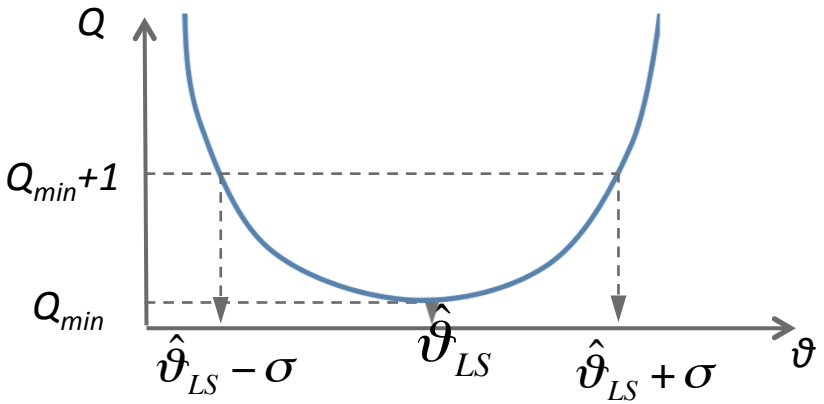
$$Q(\vec{X}, \vec{\vartheta}) = \left[\vec{X} - \vec{M}(\vec{\vartheta})\right]^T V^{-1} \left[\vec{X} - \vec{M}(\vec{\vartheta})\right] = \sum_{i=1}^{N} \sum_{j=1}^{N} \left[x_i - m_i(\vec{\vartheta})\right] \left(V^{-1}\right)_{ij} \left[x_j - m_j(\vec{\vartheta})\right]$$

The Least Squares Estimator of $\vec{\vartheta}$, $\hat{\vec{\vartheta}}_{LS}$ is the value minimising Q

Particular case: independent observations, V diagonal: $V_{ii} = \sigma_i^2 \Rightarrow Q = \sum_{i=1}^{N} \frac{\left[x_i - m_i(\vec{\vartheta})\right]^2}{\sigma_i^2(\vec{\vartheta})}$

Case k=1 (1 parameter):

$$V(\hat{\vartheta}_{LS}) \underset{N \to \infty}{\to} 2\left[E\left[\frac{\partial^2 Q}{\partial \vartheta^2}\right]_{\vartheta=\vartheta_0}\right]^{-1} \sim 2D_2^{-1}\Big|_{\vartheta=\hat{\vartheta}_{LS}}$$



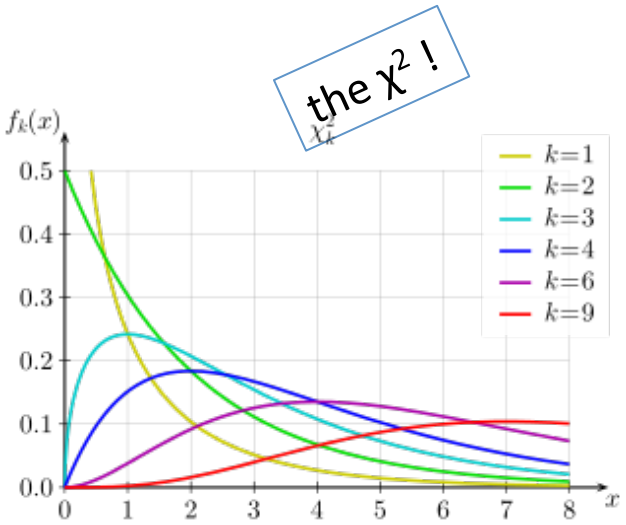
The Least Squares estimator: properties

- The LSE can be shown to be consistent, in general biased, non-optimal
- Case of the linear model
  - if the σ_i are independent of ϑ and the $\mu_i(\vartheta)$ are linear functions of ϑ, then the minimisation can be done analytically and the estimator is optimal and convergent

- What about the asymptotic distribution ?
  - case of x_i's normally distributed: $f\left(x_i;\vec{\vartheta}\right)=G\left(x_i;\mu_i\left(\vec{\vartheta}\right),\sigma_i\left(\vec{\vartheta}\right)\right)$

then $Q_o=\sum_{i=1}^{N}\frac{\left[x_i-\mu_i\left(\vec{\vartheta}\right)\right]^2}{\sigma_i^2\left(\vec{\vartheta}\right)}$ is distributed according to a χ^2 law with N degrees of freedom

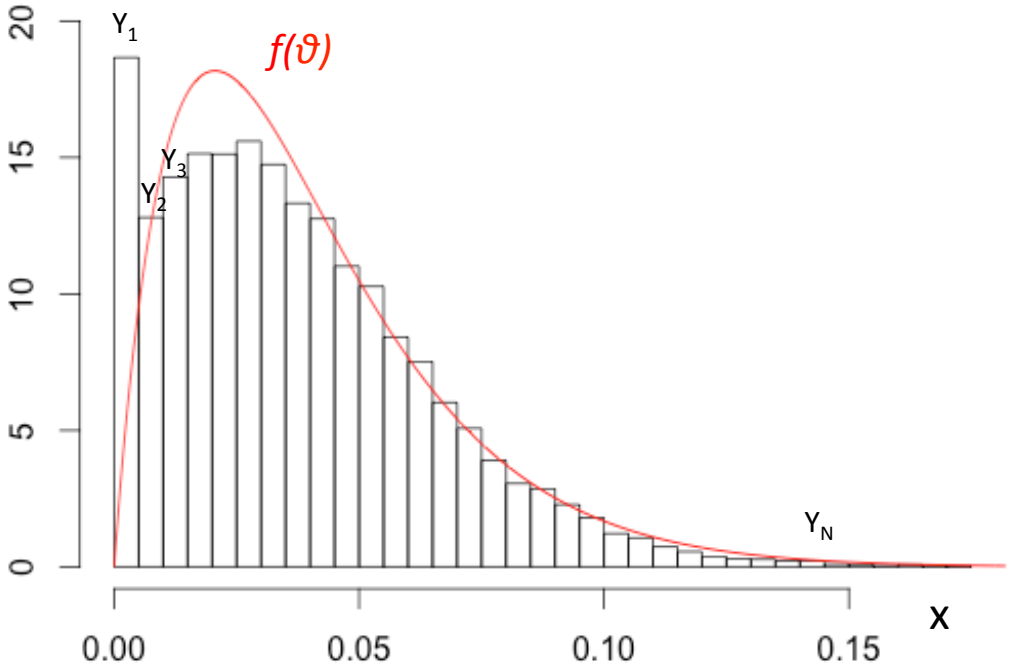
$$f\left(Q_0\right)=\frac{Q_0^{N/2-1}}{2\Gamma\left(N/2\right)}e^{-Q_0^2/2} \qquad \left\langle Q_0\right\rangle=N, V\left(Q_0\right)=2N$$



When the model is linear, $Q_{\min}=Q_0\left(\hat{\vartheta}\right)$ is distributed according to a χ^2 law with N-K degrees of freedom
and $\hat{\vec{\vartheta}}$ according to a N-dim normal law with mean $\vec{\vartheta}_0$ and variance $2D_2^{-1}$

Applications: fitting a histogram

Histograms with N bins (N large), Y_i events in each bin, $E[Y_i]=f_i(\vartheta)$
the number of events per bin follows a multinominal distribution with $\sigma_i^2=E[Y_i]=f_i(\vartheta)$



Find an estimator for ϑ

Applications: fitting a histogram

Histograms with N bins (N large), Y_i events in each bin, $E[Y_i]=f_i(\vartheta)$
the number of events per bin follows a multinominal distribution with $\sigma_i^2=E[Y_i]=f_i(\vartheta)$

Least square estimator of ϑ : minimize $Q^2 = \sum_{i=1}^{N} \frac{\left[Y_i - f_i(\vartheta)\right]^2}{f_i(\vartheta)}$: the usual "minimum chi-square method"

In general no predictions on its properties if there are few events in some bins.
However, it can be proven to have optimal asymptotic properties: consistent, asymptotically Normal, efficient

Reasonable to use $\langle Y_i \rangle = f_i(\vartheta)$ and minimize $Q'^2 = \sum_{i=1}^{N} \frac{\left[Y_i - f_i(\vartheta)\right]^2}{Y_i}$: "modified min. chi-square method"

Same properties as Q^2 asymptotically (for large N)

Maximum Likelihood method (Multinomial) $\ln L = \sum_{i=1}^{N} Y_i \ln f_i(\vartheta)$: "Binned Max. Likelihood method"

Asymptotically equivalent to the previous two, but converges faster. And no problem with low-content or empty bins. Recommended !
P.S.: converges to "unbinned M.L." when N->∞

Extended Maximum Likelihood

- The total number of events does not intervene in the maximisation of *lnL* if it is independent of the parameter
- If N_{tot} depends on the parameter => Extended Maximum Likekihood (EML)

Case of a histogram:
in each bin, Poisson distribution with mean $f_i(\theta)$ and number of events per bin Y_i

$$f_i(Y_i,\vartheta) = f_i(\vartheta)^{Y_i} e^{-f_i(\vartheta)} / Y_i! \qquad \sum_{i=1}^{N} Y_i = N_{tot}(\vartheta)$$

$$\ln L = \sum_{i=1}^{N} Y_i \ln f_i(\vartheta) - f_i(\vartheta) - \ln Y_i! = -N_{tot}(\vartheta) + \sum_{i=1}^{N} Y_i \ln f_i(\vartheta) + cste$$

identical to ML if N_{tot} does not depend on ϑ

⇒ ML (normalisation independent of the parameter) uses shape

⇒ EML (normalisation dependent on the parameter) uses shape + normalization

- Same results when N_{tot} is independent of ϑ
- Needs care in the interpretation of errors (e.g. when size and shape are not indep.)

ESTIMATION OF CONFIDENCE INTERVALS

Interval estimation

Find the range $[\vartheta_L, \vartheta_U]$ which contains the true value ϑ_0 with a given probability β :

$$P(\vartheta_L < \vartheta_0 < \vartheta_U) = \beta$$

$[\vartheta_L, \vartheta_U]$ is the Confidence Interval for ϑ with probability content β

In physics, often used for "errors": 1σ ⇔ β=68.3%, 2σ ⇔ β = 95.5% etc.

Strictly true only for Normal distributions !

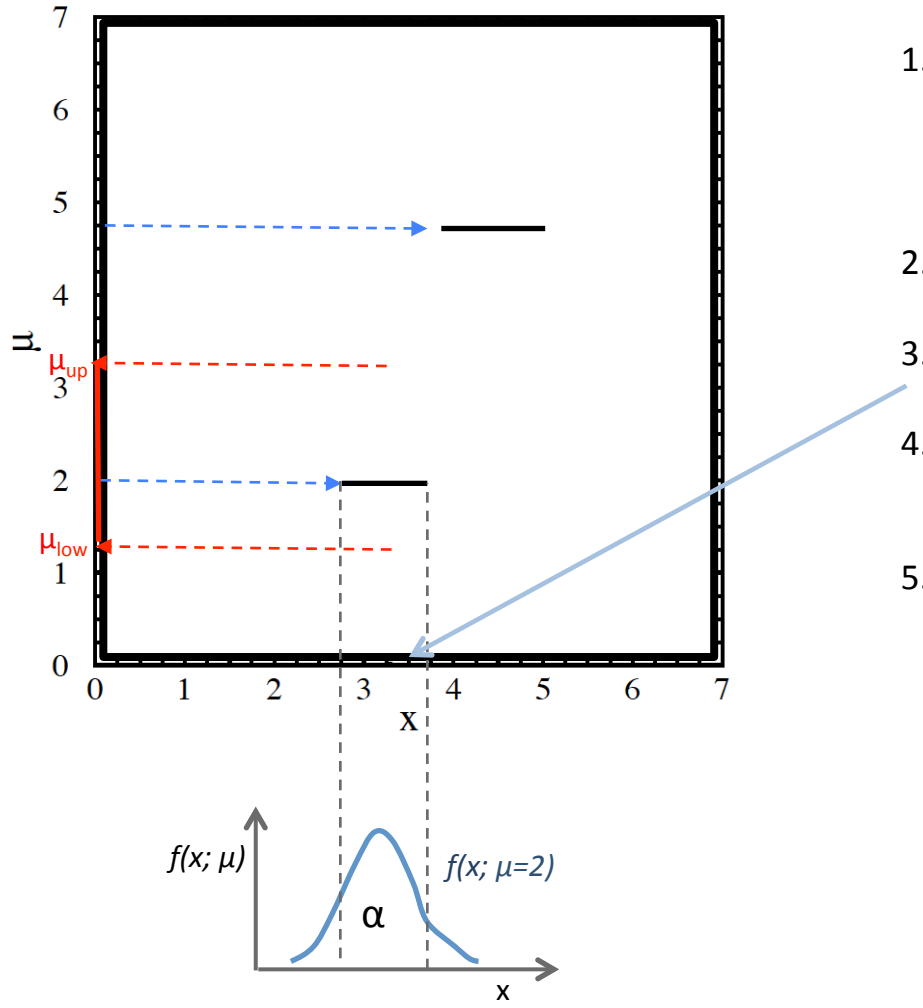
Probability content of a region [a,b] in the space of the variable X, given the PDF of X and if the parameter ϑ is known: $\beta = P(a < X < b) = \int_a^b PDF(X \mid \vartheta) dX$

If ϑ is unknown: find a new variable such that its PDF is independent of ϑ => find the range of ϑ such that $P(\vartheta_a < \vartheta_0 < \vartheta_b) = \beta$

- *property of COVERAGE*

Interval estimation: general case 1-D

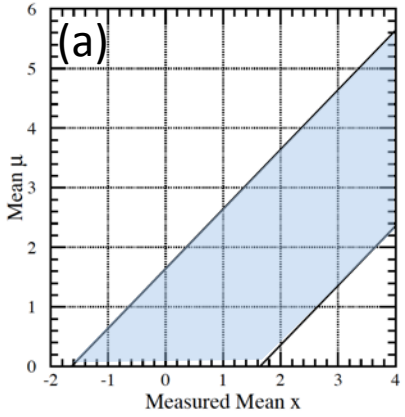
Construction of Neyman's confidence belt



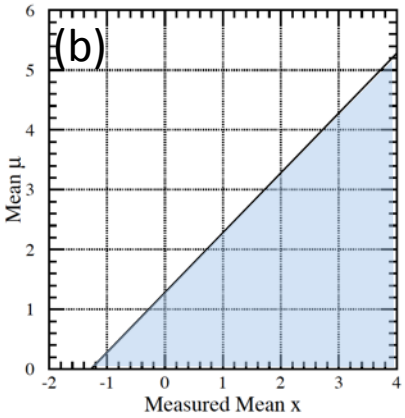
1. For a given value of the parameter μ, draw the interval with probability content α (horizontal line)
   - arbitrary positioning… here we choose a symmetric one, leaving out (1-α)/2 on each side
2. Repeat for all values of μ => you have the Confidence Belt
3. Consider the result x of your experiment and draw a vertical line
4. Take the values μ_{low} , μ_{up} where the vertical lines intersects the confidence belt
5. The interval $[\mu_{low}, \mu_{up}]$ is the Confidence Interval with probability content α for the true value μ_0 of the parameter μ

Interval estimation: problems

90%CL central interval for the mean of a Gaussian



90%CL upper limit for the mean of a Gaussian

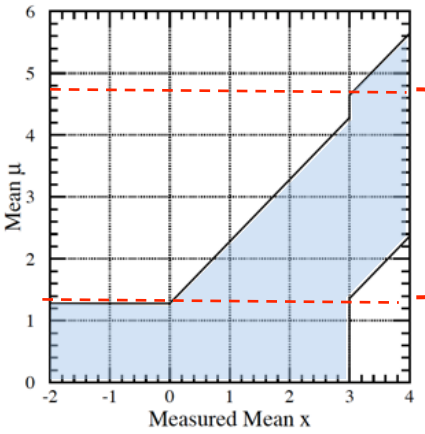


Both ensure correct coverage if the choice is made beforehand

However, suppose that …

1) you decide to use (a) if your result is x>3 and (b) if x<3
   - **FLIP-FLOPPING**
2) you don't want negative values, because allowed physical values for your parameter can only be positive (e.g. mass)
   - **UNPHYSICAL VALUES**

… so you use this confidence belt



G.J. Feldman and R.D. Cousins, Phys.Rev.D57:3873-3889,1998

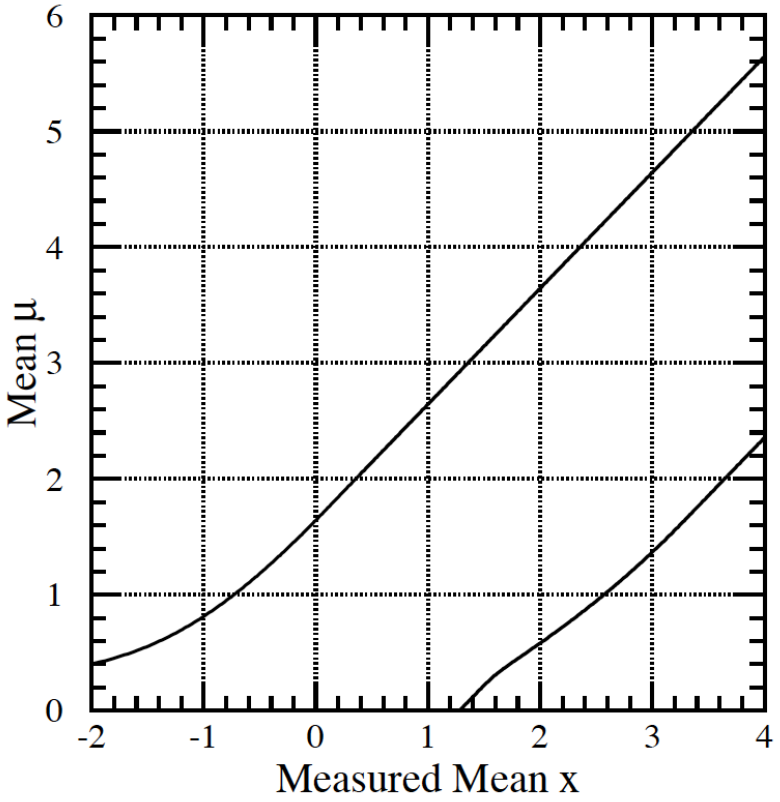
Interval estimation: Feldman-Cousins

Feldman-Cousins "unified approach" (originally for neutrino physics):
likelihood ratio ordering principle

- in the interval for $\mu=\mu_0$, include the elements of probability $P(x|\mu_0)$ giving the largest value of the likelihood ratio

$$R(x) = \frac{P(x \mid \mu_0)}{P(x \mid \hat{\mu})}$$

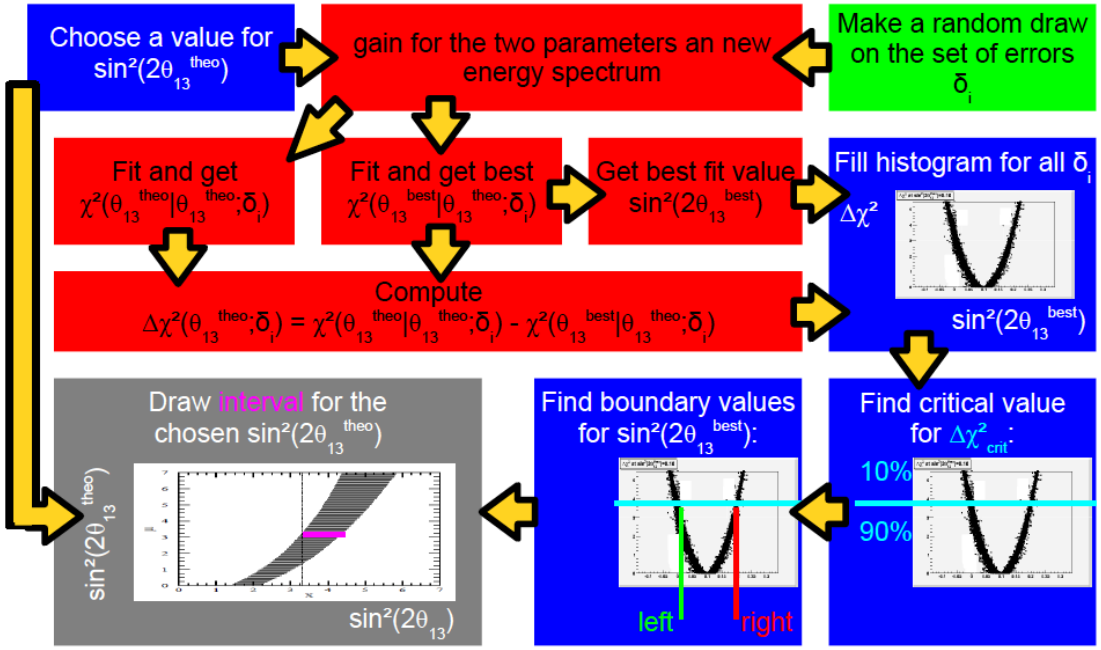
where $\hat{\mu}$ is the value of μ for which $P(x|\mu)$ is maximal within the physical region.



➔ Hands-on session today

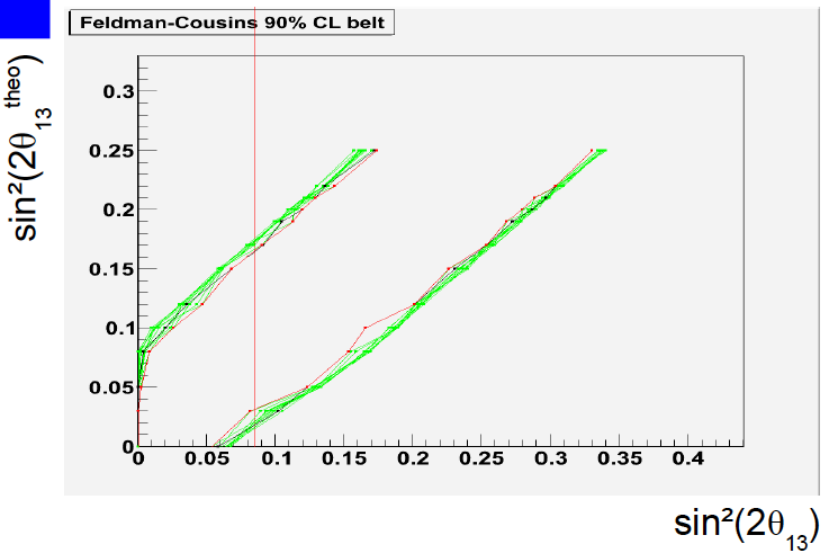
G.J. Feldman and R.D. Cousins, Phys.Rev.D57:3873-3889,1998

Example: measuring θ_{13} in Double Chooz



S. Schoppmann, RWTH Aachen
DC internal meeting, 2012
(confidential ??? !!!)

"Alternative method" to the published analysis



SUMMARY

- Parameter estimation
  - The Maximum Likelihood Estimator: construction and properties
  - The Least Squares estimator: construction and properties

- Estimation of confidence intervals
  - general case 1D: Neyman belt construction
  - the Feldman-Cousins approach
    - hands-on: the Feldman-Cousins approach
  - use of the Likelihood function
  - case of multiple parameters
  - Bayesian credibility intervals