



Probability and Statistics

Basic concepts II

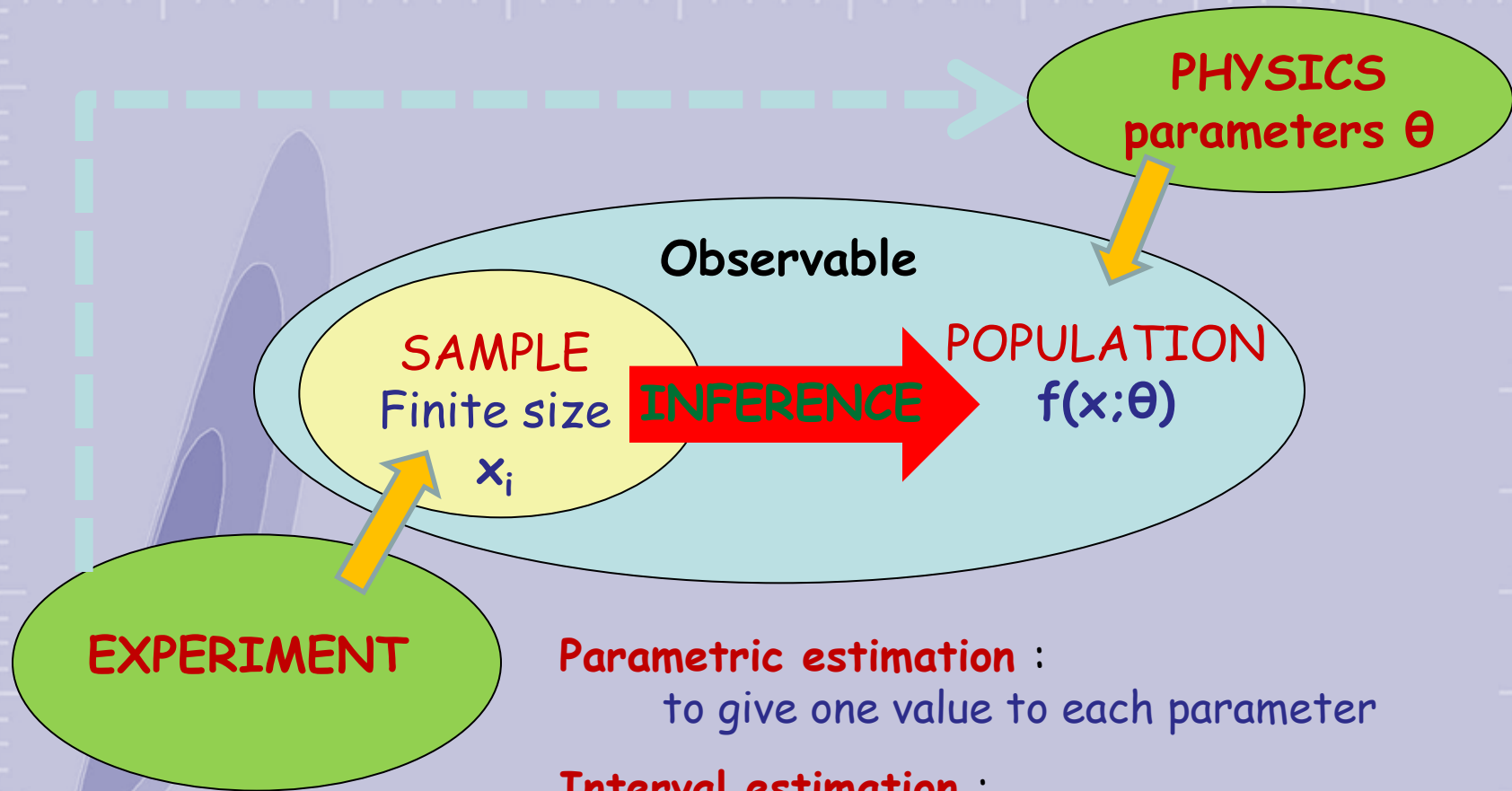
(from a physicist point of view)

Benoit CLEMENT - Université J. Fourier / LPSC

bclement@lpsc.in2p3.fr



Statistics



Parametric estimation :
to give one value to each parameter

Interval estimation :
to derive a interval that probably contains the true value

Non-parametric estimation :
to estimate the full pdf of the population

Parametric estimation

From a finite sample $\{x_i\}$ \rightarrow estimating a parameter θ

Statistic = a function $S = f(\{x_i\})$

Any statistic can be considered as an **estimator** of θ

To be a good estimator it needs to satisfy :

- **Consistency** : limit of the estimator for a infinite sample.
- **Bias** : difference between the estimator and the true value
- **Efficiency** : speed of convergence
- **Robustness** : sensitivity to statistical fluctuations

A good estimator should at least be **consistent** and **asymptotically unbiased**

Efficient / Unbiased / Robust often contradict each other
 \Rightarrow different choices for different applications

Bias and consistency

As the sample is a set of realization of random variables (or one vector variable), so is the estimator :

$\hat{\theta}$ is a realization of $\hat{\Theta}$

it has a mean, a variance,... and a probability density function

Bias : Mean value of the estimator $b(\hat{\theta}) = E[\hat{\Theta} - \theta_0] = \mu_{\hat{\Theta}} - \theta_0$

unbiased estimator : $b(\hat{\theta}) = 0$

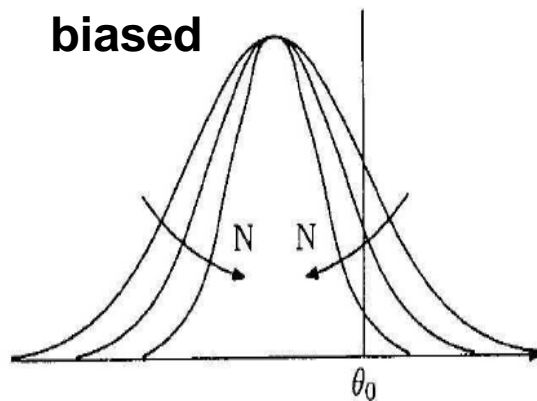
asymptotically unbiased : $b(\hat{\theta}) \xrightarrow{n \rightarrow +\infty} 0$

Consistency: formally $P(|\hat{\theta} - \theta| > \varepsilon) \xrightarrow{n \rightarrow +\infty} 0, \forall \varepsilon$

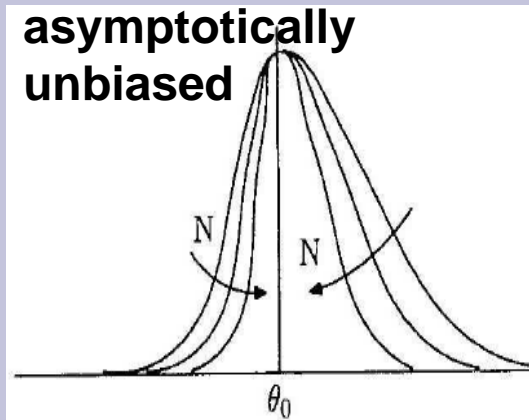
in practice, if asymptotically unbiased

$$\sigma_{\hat{\theta}} \xrightarrow{n \rightarrow +\infty} 0$$

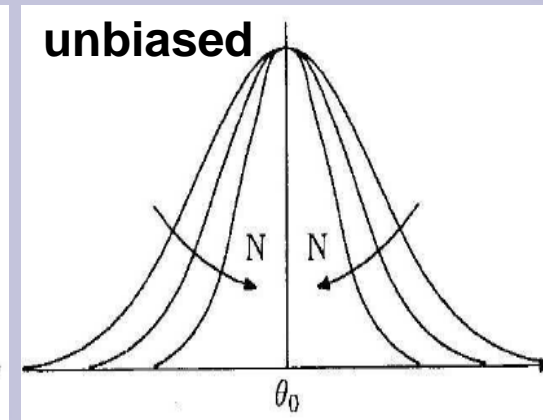
biased



asymptotically unbiased



unbiased



Empirical estimator

Sample mean is a good estimator of the **population mean**
 -> weak law of large numbers : convergent, unbiased

$$\hat{\mu} = \frac{1}{n} \sum x_i, \quad \mu_{\hat{\mu}} = E[\hat{\mu}] = \mu, \quad \sigma_{\hat{\mu}}^2 = E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$$

Sample variance as an estimator of the population variance :

$$\hat{s}^2 = \frac{1}{n} \sum_i (x_i - \hat{\mu})^2 = \left(\frac{1}{n} \sum_i (x_i - \mu)^2 \right) - (\mu - \hat{\mu})^2$$

$$E[\hat{s}^2] = \left(\frac{1}{n} \sum_i \sigma^2 \right) - \sigma_{\hat{\mu}}^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2$$

biased,
asymptotically unbiased

unbiased variance estimator :

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

variance of the estimator
(convergence)

$$\sigma_{\hat{\sigma}^2}^2 = \frac{\sigma^4}{n-1} \left(\frac{n-1}{n} \gamma_2 + 2 \right) \rightarrow \frac{2\sigma^4}{n}$$

Errors on these estimator

Uncertainty \Leftrightarrow Estimator standard deviation

Use an estimator of standard deviation : $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ (!!! Biased)

Mean : $\hat{\mu} = \frac{1}{n} \sum x_i,$

$$\sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n}$$

$$\Delta\hat{\mu} = \sqrt{\frac{\hat{\sigma}^2}{n}}$$

Variance : $\hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \hat{\mu})^2, \sigma_{\hat{\sigma}^2}^2 \approx \frac{2\sigma^4}{n}$

$$\Delta\hat{\sigma}^2 = \sqrt{\frac{2}{n}} \hat{\sigma}^2$$

Central-Limit theorem \rightarrow empirical estimators of mean and variance are **normally distributed**, for large enough samples

$\hat{\mu} \pm \Delta\hat{\mu}$; $\hat{\sigma}^2 \pm \Delta\hat{\sigma}^2$ define 68% confidence intervals

Likelihood function

Generic function $k(x, \theta)$

x : random variable(s)

θ : parameter(s)

fix $\theta = \theta_0$ (true value)

fix $x = u$ (one realization of the random variable)

Probability density function

$$f(x; \theta) = k(x, \theta_0)$$

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

for Bayesian $f(x|\theta) = f(x; \theta)$

Likelihood function

$$\mathcal{L}(\theta) = k(u, \theta)$$

$$\int \mathcal{L}(\theta) d\theta = ???$$

for Bayesian $f(\theta|x) = \mathcal{L}(\theta) / \int \mathcal{L}(\theta) d\theta$

For a **sample** : n independent realizations of the same variable X

$$\mathcal{L}(\theta) = \prod_i k(x_i, \theta) = \prod_i f(x_i; \theta)$$

Estimator variance

Start from the generic k function, differentiate twice, with respect to θ , the pdf normalization condition: $1 = \int k(x, \theta) dx$

$$0 = \int \frac{\partial k}{\partial \theta} dx = \int k \frac{\partial \ln k}{\partial \theta} dx = E \left[\frac{\partial \ln k}{\partial \theta} \right] \Rightarrow (b + \theta) E \left[\frac{\partial \ln k}{\partial \theta} \right] = 0$$

$$0 = \int \frac{\partial^2 k}{\partial \theta^2} dx = \int k \frac{\partial^2 \ln k}{\partial \theta^2} dx + \int k \left(\frac{\partial \ln k}{\partial \theta} \right)^2 dx \Rightarrow E \left[\frac{\partial^2 \ln k}{\partial \theta^2} \right] = -E \left[\left(\frac{\partial \ln k}{\partial \theta} \right)^2 \right]$$

Now differentiating the estimator bias: $\theta + b = \int \hat{\theta}(x) k(x, \theta) dx$

$$1 + \frac{\partial b}{\partial \theta} = \frac{\partial}{\partial \theta} \int \hat{\theta}(x) k(x, \theta) dx = \int \hat{\theta} \frac{\partial k}{\partial \theta} dx = \int \hat{\theta} k \frac{\partial \ln k}{\partial \theta} dx = \int (\hat{\theta} - b - \theta) k \frac{\partial \ln k}{\partial \theta} dx$$

Finally, using Cauchy-Schwartz inequality

$$\left(1 + \frac{\partial b}{\partial \theta} \right)^2 \leq \int (\hat{\theta} - b - \theta)^2 k dx \int k \left(\frac{\partial \ln k}{\partial \theta} \right)^2 dx \Rightarrow \sigma_{\hat{\theta}}^2 \geq - \frac{(1 + b')^2}{E \left[\left(\frac{\partial \ln k}{\partial \theta} \right)^2 \right]}$$

Cramer-Rao bound

Efficiency

For any unbiased estimator of θ , the variance cannot exceed :

$$\sigma_{\hat{\theta}}^2 \geq \frac{-1}{\mathbb{E} \left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta} \right)^2 \right]}$$

The **efficiency** of a convergent estimator, is given by its **variance**.

An **efficient estimator** reaches the Cramer-Rao bound (at least asymptotically) : Minimal variance estimator

MVE will often be biased, asymptotically unbiased

Maximum likelihood

For a sample of measurements, $\{x_i\}$

The analytical form of the density is known

It depends on several unknown parameters θ

eg. event counting : Follow a Poisson distribution, with a parameter that depends on the physics : $\lambda_i(\theta)$

$$\mathcal{L}(\theta) = \prod_i \frac{e^{-\lambda_i(\theta)} \lambda_i(\theta)^{x_i}}{x_i!}$$

An estimator of the parameters of θ , are the ones that **maximize of observing the observed result.**

-> Maximum of the likelihood function

$$\left. \frac{\partial \mathcal{L}}{\partial \theta} \right|_{\theta = \hat{\theta}} = 0$$

rem : system of equations for several parameters

rem : **often minimize** $-\ln \mathcal{L}$: simplify expressions

Properties of MLE

Mostly **asymptotic properties** : valid for large sample, often assumed in any case for lack of better information

Asymptotically **unbiased**

Asymptotically **efficient** (reaches the CR bound)

Asymptotically **normally distributed**

-> Multinormal law, with covariance given by generalization of CR Bound :

$$f(\hat{\theta}; \vec{\theta}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(\hat{\theta}-\vec{\theta})^T \Sigma^{-1}(\hat{\theta}-\vec{\theta})} \quad \Sigma_{ij}^{-1} = -E \left[\frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \ln \mathcal{L}}{\partial \theta_j} \right]$$

Goodness of fit = The value of $-2\ln L(\hat{\theta})$ is Khi-2 distributed, with $ndf = \text{sample size} - \text{number of parameters}$

$$p\text{-value} = \int_{-2\ln L(\hat{\theta})}^{+\infty} f_{\chi^2}(x; ndf) dx$$

Probability of getting a worse agreement

Errors on MLE

$$f(\hat{\theta}; \bar{\theta}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(\hat{\theta}-\bar{\theta})^T \Sigma^{-1}(\hat{\theta}-\bar{\theta})}$$

$$\Sigma_{ij}^{-1} = -E \left[\frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \ln \mathcal{L}}{\partial \theta_j} \right]$$

Errors on parameter -> from the covariance matrix

For **one parameter**, 68% interval $\Delta\theta = \hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{-1}{\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}}}$ only one realization of the estimator -> empirical mean of 1 value...

More generally :

$$\Delta \ln \mathcal{L} = \ln \mathcal{L}(\hat{\theta}) - \ln \mathcal{L}(\theta) = \frac{1}{2} \sum_{i,j} \Sigma_{ij}^{-1} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) + O(\theta^3)$$

Confidence contour are defined by the equation : $\Delta \ln \mathcal{L} = \beta(n_{\theta}, \alpha)$ with $\alpha = \int_0^{2\beta} f_{x^2}(x; n_{\theta}) dx$

Values of β for different number of parameters n_{θ} and confidence levels α

$n_{\theta} \rightarrow$ $\alpha \downarrow$	1 (0.5 * n_{θ}^2)	2	3
68.3	0.5	1.15	1.76
95.4	2	3.09	4.01
99.7	4.5	5.92	7.08

Least squares

Set of measurements (x_i, y_i) with uncertainties on y_i

Theoretical law : $y = f(x, \theta)$

Naïve approach : use **regression**

$$w(\theta) = \sum_i (y_i - f(x_i, \theta))^2, \quad \frac{\partial w}{\partial \theta_i} = 0$$

Reweight each term by the error

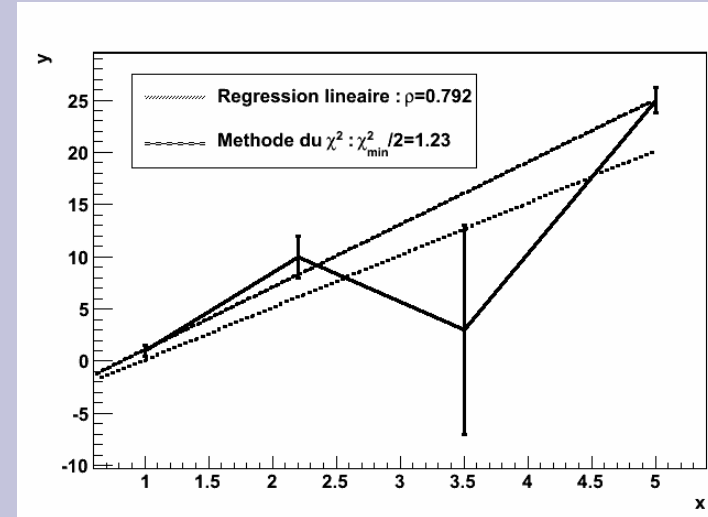
$$K^2(\theta) = \sum_i \left(\frac{y_i - f(x_i, \theta)}{\Delta y_i} \right)^2, \quad \frac{\partial K^2}{\partial \theta_i} = 0$$

Maximum likelihood : assume each y_i is normally distributed with a mean equal to $f(x_i, \theta)$ and a variance equal to Δy_i

Then the **likelihood** is : $\mathcal{L}(\theta) = \prod_i \frac{1}{\sqrt{2\pi\Delta y_i}} e^{-\frac{1}{2} \left(\frac{y_i - f(x_i, \theta)}{\Delta y_i} \right)^2}$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \Leftrightarrow -2 \frac{\partial \ln \mathcal{L}}{\partial \theta} = \frac{\partial K^2}{\partial \theta} = 0$$

Least squares or **Khi-2 fit** is the MLE, for Gaussian errors



Generic case with correlations: $K^2(\vec{\theta}) = \frac{1}{2} (\vec{y} - \vec{f}(x, \vec{\theta}))^T \Sigma^{-1} (\vec{y} - \vec{f}(x, \vec{\theta}))$

Example : fitting a line

For $f(x)=ax+b$

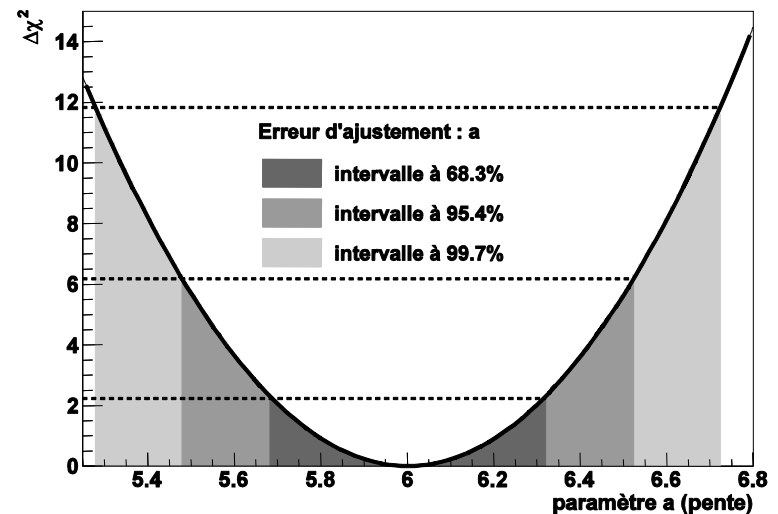
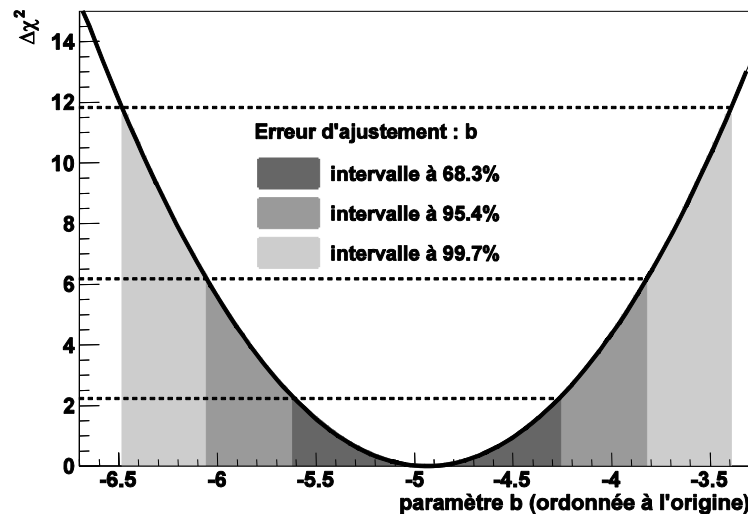
$$A = \sum_i \frac{x_i y_i}{\Delta y_i^2}, \quad B = \sum_i \frac{x_i^2}{\Delta y_i^2}, \quad C = \sum_i \frac{x_i}{\Delta y_i^2}, \quad D = \sum_i \frac{y_i}{\Delta y_i^2}, \quad E = \sum_i \frac{1}{\Delta y_i^2}$$

$$\hat{a} = \frac{AE - DC}{BE - C^2},$$

$$\hat{b} = \frac{DB - AC}{BE - C^2}$$

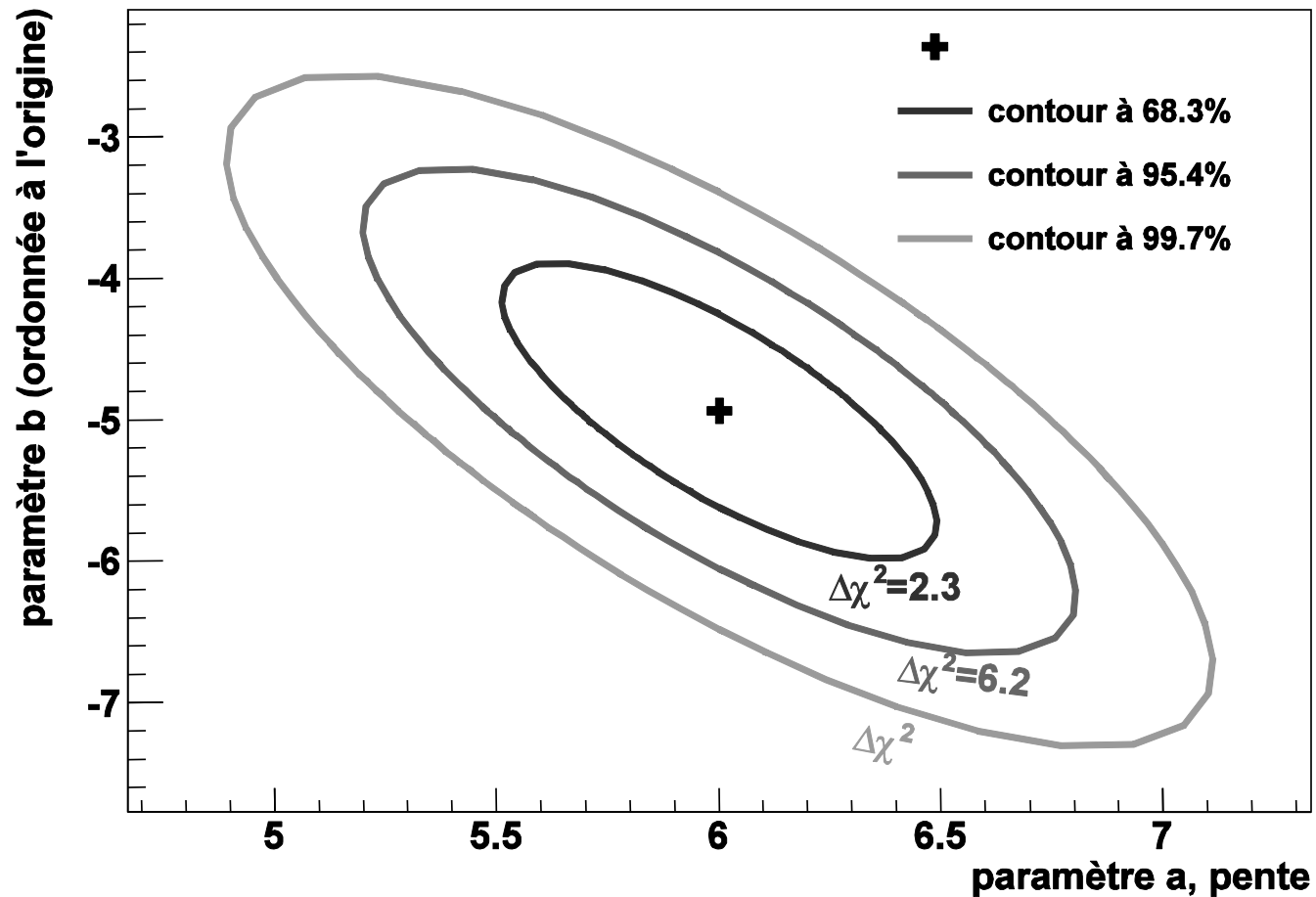
$$\Delta \hat{a} = \frac{1.52}{\sqrt{B}},$$

$$\Delta \hat{b} = \frac{1.52}{\sqrt{E}}$$



Example : fitting a line

2 dimensional error contours on a and b



Confidence interval

For a random variable, a **confidence interval** with **confidence level α** , is any interval $[a, b]$ such as :

$$P(X \in [a, b]) = \int_a^b f_X(x) dx = \alpha$$

Probability of finding a realization inside the interval

Generalization of the concept of uncertainty:

interval that contains the **true value** with a given probability

-> slightly different concepts

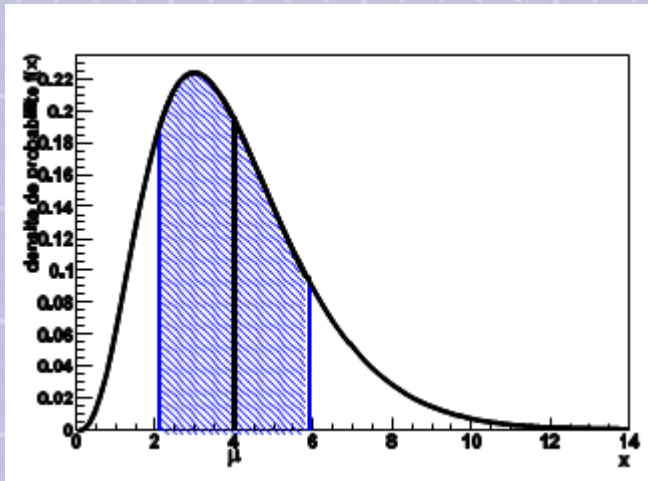
For **Bayesians** : the posterior density is the probability density of the true value. It can be used to derive interval :

$$P(\theta \in [a, b]) = \alpha$$

No such thing for a **Frequentist** : The interval itself becomes the random variable $[a, b]$ is a realization of $[A, B]$

$$P(A < \theta \text{ and } B > \theta) = \alpha \text{ Independently of } \theta$$

Confidence interval



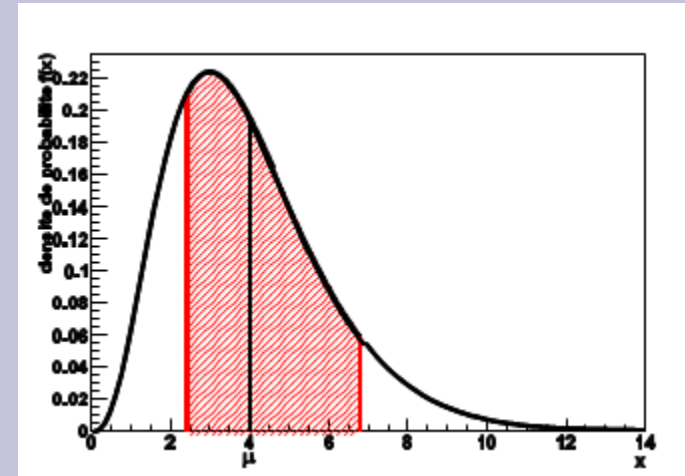
Mean centered, symmetric interval

$$[\mu - a, \mu + a]$$

$$\int_{\mu - a}^{\mu + a} f(x) dx = a$$

Mean centered, probability symmetric interval : $[a, b]$

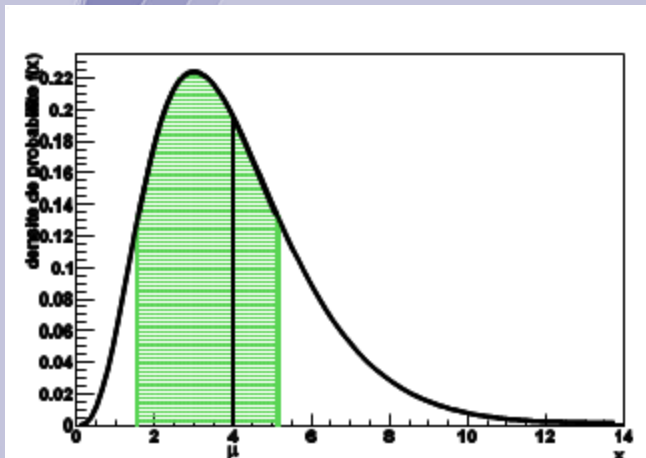
$$\int_a^{\mu} f(x) dx = \int_{\mu}^b f(x) dx = \frac{a}{2}$$



Highest Probability Density (HDP)

$$[a, b] \int_a^b f(x) dx = a$$

$f(x) > f(y)$ for $x \in [a, b]$ and $y \notin [a, b]$



Confidence Belt

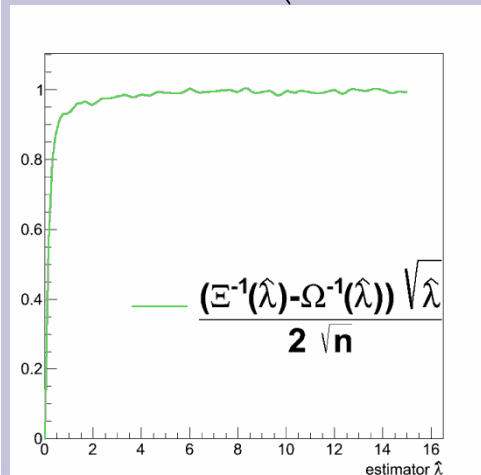
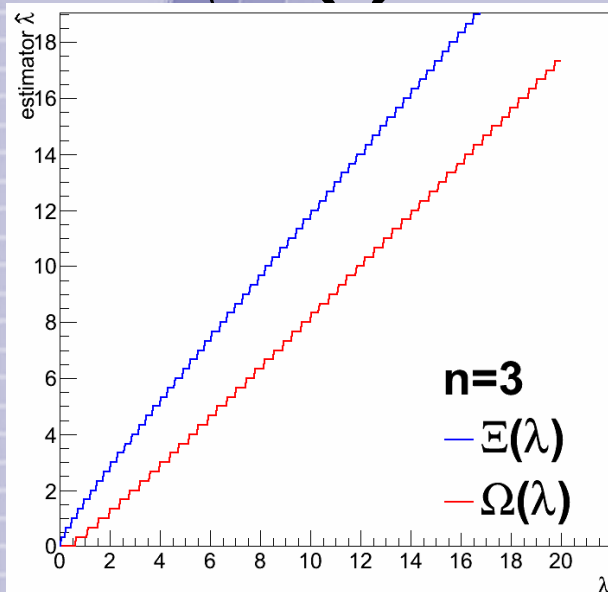
To build a frequentist interval for an estimator $\hat{\theta}$ of θ

1. Make pseudo-experiments for several values of θ and compute the estimator $\hat{\theta}$ for each (MC sampling of the estimator pdf)
2. For each θ , determine $A(\theta)$ and $B(\theta)$ such as :
 $\hat{\theta} < \Xi(\theta)$ for a fraction $(1-\alpha)/2$ of the pseudo-experiments
 $\hat{\theta} > \Omega(\theta)$ for a fraction $(1-\alpha)/2$ of the pseudo-experiments

These 2 curves are the **confidence belt**, for a **CL α** .

3. Inverse these functions. The interval $[\Omega^{-1}(\hat{\theta}), \Xi^{-1}(\hat{\theta})]$ satisfy:

$$\begin{aligned} P(\Omega^{-1}(\hat{\theta}) < \theta < \Xi^{-1}(\hat{\theta})) &= 1 - P(\Xi^{-1}(\hat{\theta}) < \theta) - P(\Omega^{-1}(\hat{\theta}) > \theta) \\ &= 1 - P(\hat{\theta} < \Xi(\theta)) - P(\hat{\theta} > \Omega(\theta)) = \alpha \end{aligned}$$



Confidence Belt for Poisson parameter λ estimated with the empirical mean of 3 realizations (68%CL)

Dealing with systematics

The variance of the estimator only measure the statistical uncertainty.

Often, we will have to deal with some **parameters** whose **value is known with limit precision**.

Systematic uncertainties

The likelihood function becomes :

$$L(\theta, \mathbf{v}) \quad \mathbf{v} = \mathbf{v}_0 \pm \Delta \mathbf{v} \text{ or } \mathbf{v}_{0-\Delta\Delta}^{+\Delta\mathbf{v}_+}$$

The known parameters \mathbf{v} are **nuisance parameters**

Bayesian inference

In **Bayesian statistics**, nuisance parameters are dealt with by assigning them a prior $\pi(\mathbf{v})$.

Usually a multinormal law is used with mean \mathbf{v}_0 and covariance matrix estimated from $\Delta\mathbf{v}_0$ (+correlation, if needed)

$$f(\boldsymbol{\theta}, \mathbf{v} | \mathbf{x}) = \frac{f(\mathbf{x} | \boldsymbol{\theta}, \mathbf{v})\pi(\boldsymbol{\theta})\pi(\mathbf{v})}{\iint f(\mathbf{x} | \boldsymbol{\theta}, \mathbf{v})\pi(\boldsymbol{\theta})\pi(\mathbf{v})d\boldsymbol{\theta} d\mathbf{v}}$$

The final prior is obtained by **marginalization** over the nuisance parameters

$$f(\boldsymbol{\theta} | \mathbf{x}) = \int f(\boldsymbol{\theta}, \mathbf{v} | \mathbf{x})d\mathbf{v} = \frac{\int f(\mathbf{x} | \boldsymbol{\theta}, \mathbf{v})\pi(\boldsymbol{\theta})\pi(\mathbf{v})d\mathbf{v}}{\iint f(\mathbf{x} | \boldsymbol{\theta}, \mathbf{v})\pi(\boldsymbol{\theta})\pi(\mathbf{v})d\boldsymbol{\theta} d\mathbf{v}}$$

Profile Likelihood

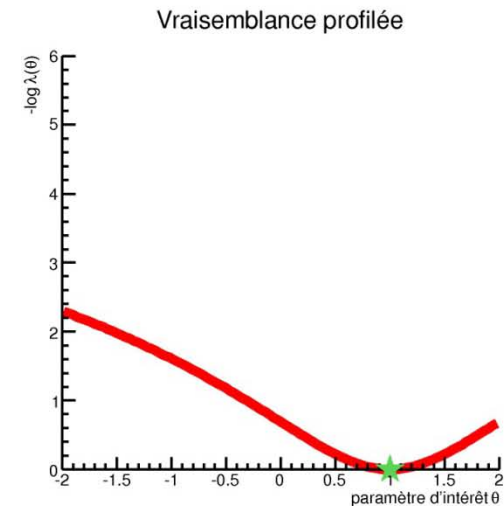
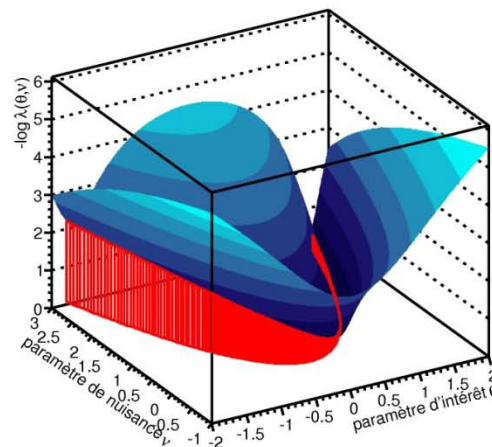
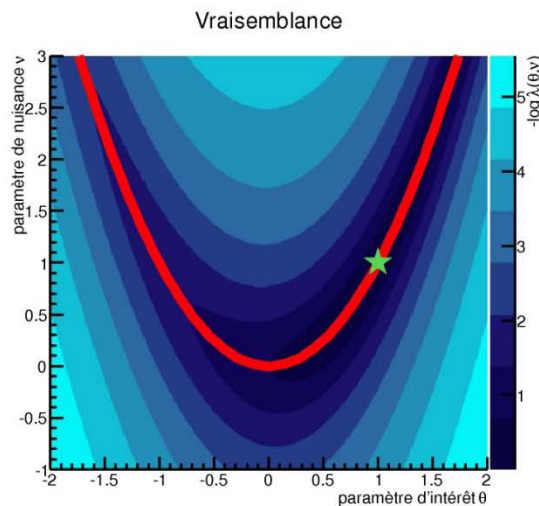
No true frequentist way to add systematic effects. Popular method of the day : **profiling**

Deal with nuisance parameters as realization if random variables :
extend the likelihood : $L(\theta, \nu) \rightarrow L'(\theta, \nu)G(\nu)$

$G(\nu)$ is the likelihood of the new parameters (identical to prior)

For each value of θ , minimize the likelihood with respect to nuisance : **profile likelihood** $PL(\theta)$.

$PL(\theta)$ has the same statistical asymptotical properties than the regular likelihood



Non parametric estimation

Directly estimating the probability density function

- Likelihood ratio discriminant
- Separating power of variables
- Data/MC agreement
- ...

Frequency Table : For a sample $\{x_i\}$, $i=1..n$

1. Define successive intervals (bins) $C_k = [a_k, a_{k+1}[$

2. Count the number of events n_k in C_k

Histogram : Graphical representation of the frequency table

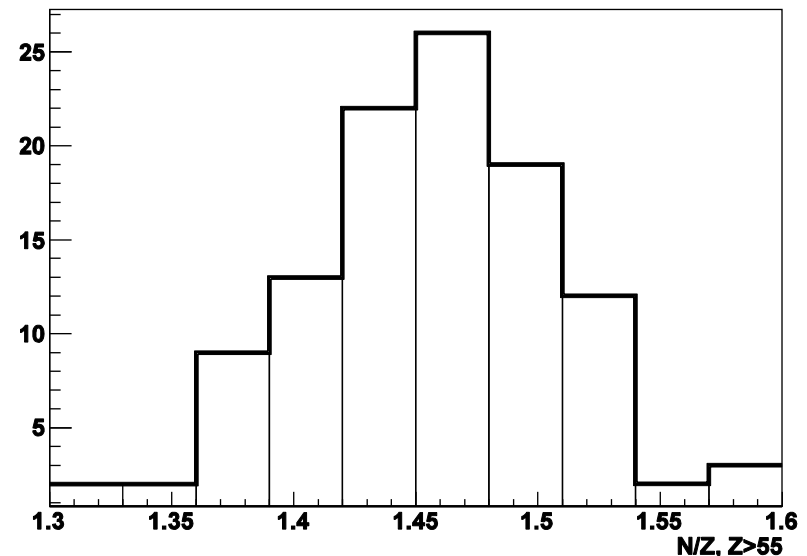
$$h(x) = n_k \text{ if } x \in C_k$$

Histogram

Classe	Nombre de N/Z	Fréquence	Classe	Nombre de N/Z	Fréquence
< 1.30	0	0	1.45 - 1.48	26	0.2363
1.30 - 1.33	2	0.0182	1.48 - 1.51	19	0.1727
1.33 - 1.36	2	0.0182	1.51 - 1.54	12	0.1091
1.36 - 1.39	9	0.0818	1.54 - 1.57	2	0.0182
1.39 - 1.42	13	0.1182	1.57 - 1.60	3	0.0273
1.42 - 1.45	22	0.2	≥ 1.60	0	0

N/Z for stable heavy nuclei

1.321, 1.357, 1.392, 1.410, 1.428, 1.446,
 1.464, 1.421, 1.438, 1.344, 1.379, 1.413,
 1.448, 1.389, 1.366, 1.383, 1.400, 1.416,
 1.433, 1.466, 1.500, 1.322, 1.370, 1.387,
 1.403, 1.419, 1.451, 1.483, 1.396, 1.428,
 1.375, 1.406, 1.421, 1.437, 1.453, 1.468,
 1.500, 1.446, 1.363, 1.393, 1.424, 1.439,
 1.454, 1.469, 1.484, 1.462, 1.382, 1.411,
 1.441, 1.455, 1.470, 1.500, 1.449, 1.400,
 1.428, 1.442, 1.457, 1.471, 1.485, 1.514,
 1.464, 1.478, 1.416, 1.444, 1.458, 1.472,
 1.486, 1.500, 1.465, 1.479, 1.432, 1.459,
 1.472, 1.486, 1.513, 1.466, 1.493, 1.421,
 1.447, 1.460, 1.473, 1.486, 1.500, 1.526,
 1.480, 1.506, 1.435, 1.461, 1.487, 1.500,
 1.512, 1.538, 1.493, 1.450, 1.475, 1.500,
 1.512, 1.525, 1.550, 1.506, 1.530, 1.487,
 1.512, 1.524, 1.536, 1.518, 1.577, 1.554,
 1.586, 1.586



Histogram

Statistical description : n_k are **multinomial** random variables.
with parameters :

$$n = \sum_k n_k$$

$$p_k = P(x \in C_k) = \int_{C_k} f_X(x) dx$$

$$\mu_{n_k} = np_k$$

$$\sigma_{n_k}^2 = np_k(1-p_k) \underset{p_k \ll 1}{\approx} \mu_{n_k}$$

$$\text{Cov}(n_k, n_r) = -np_k p_r \underset{p_k \ll 1}{\approx} 0$$

For a large sample :

For small classes (width δ):

$$\lim_{n \rightarrow +\infty} \frac{n_k}{n} = \frac{\mu_k}{n} = p_k$$

$$p_k = \int_{C_k} f_X(x) dx \approx \delta f(x_c) \Rightarrow \lim_{\delta \rightarrow 0} \frac{p_k}{\delta} = f(x)$$

So finally :

$$f(x) = \lim_{\substack{n \rightarrow +\infty \\ \delta \rightarrow 0}} \frac{1}{n\delta} h(x)$$

The **histogram** is an **estimator** of the **probability density**

Each bin can be described by a **Poisson density**.

The 1σ error on n_k is then : $\Delta n_k = \sqrt{\hat{\sigma}_{n_k}^2} = \sqrt{\hat{\mu}_{n_k}} = \sqrt{n_k}$

Kernel density estimators

Histogram is a step function -> sometime need smoother estimator

On possible solution : **Kernel Density Estimator**

Attribute to each point of the sample a "kernel" function $k(u)$

$$u = \frac{x - x_i}{w}, \quad k(u) = k(-u), \quad \int k(u) du = 1$$

Triangle kernel : $k(u) = 1 - |u|$, for $-1 < u < 1$

Parabolic kernel : $k(u) = \frac{3}{4}(1 - u^2)$, for $-1 < u < 1$

Gaussian kernel : $k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$

...

w = **kernel width**, similar to bin width of the histogram

The a pdf estimator is :

$$K(x) = \frac{1}{n} \sum_i k(u_i) = \frac{1}{n} \sum_i k\left(\frac{x - x_i}{w}\right)$$

Rem : for multidimensional pdf : $u^2 = \sum_k \left(\frac{x^{(k)} - x_i^{(k)}}{w^{(k)}}\right)^2$

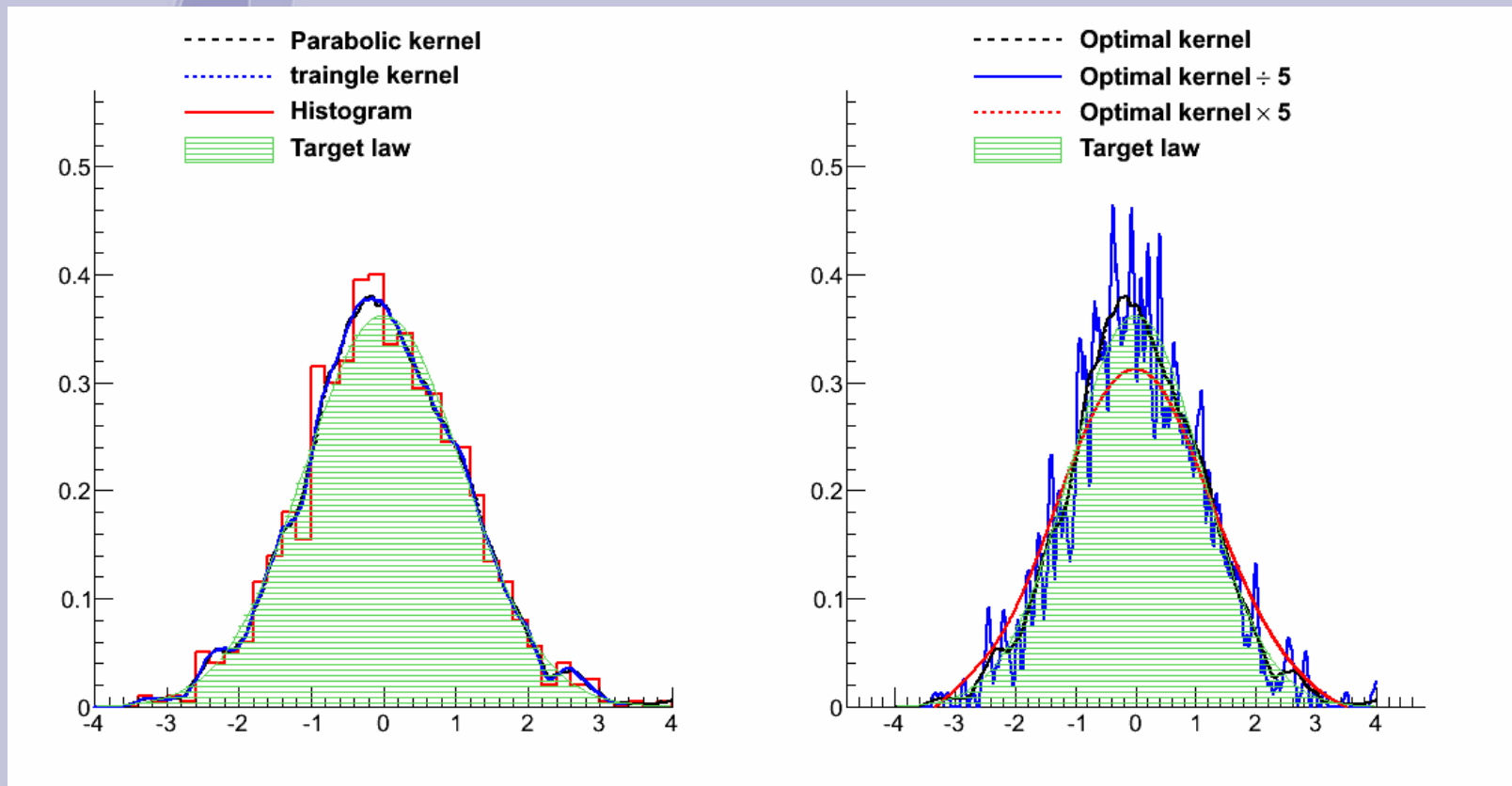
Kernel density estimators

If the estimated density is normal, the **optimal width** is :

$$w = \sigma \left(\frac{3}{(d+2)n} \right)^{\frac{1}{d+4}}$$

with n that sample size and d the dimension

As for the histogram binning, no generic result : try and see



Statistical Tests

Statistical tests aim at:

- Checking the **compatibility** of a dataset $\{x_i\}$ **with a given distribution**
- Checking the **compatibility of two datasets** $\{x_i\}$, $\{y_i\}$: are they issued from the same distribution.
- **Comparing different hypothesis** : background vs signal+background

In every case :

- build a statistic that quantify the agreement with the hypothesis
- **convert it into a probability** of compatibility/incompatibility : **p-value**

Pearson test

Test for **binned data** : use the Poisson limit of the histogram

- Sort the sample into k bins $C_i : n_i$
- Compute the probability of this class : $p_i = \int_{C_i} f(x) dx$
- The test statistics compare, for each bin the deviation of the observation from the expected mean to the theoretical standard deviation.

$$\chi^2 = \sum_{\text{bins } i} \frac{(n_i - np_i)^2}{np_i}$$

Data → $(n_i - np_i)^2$ Poisson mean → np_i
Poisson variance → np_i

Then χ^2 follow (asymptotically) a Khi-2 law with $k-1$ degrees of freedom (1 constraint $\sum n_i = n$)

P-value : probability of doing worse, $p\text{-value} = \int_{\chi^2}^{+\infty} f_{\chi^2}(x; k-1) dx$

For a "good" agreement $\chi^2 / (k-1) \sim 1$,

More precisely $\chi^2 \in (k-1) \pm \sqrt{2(k-1)}$ (1 σ interval $\sim 68\%CL$)

Kolmogorov-Smirnov test

Test for **unbinned data** : compare the sample cumulative density function to the tested one

Sample Pdf (ordered sample)

$$f_s(x) = \frac{1}{n} \sum_i \delta(x - i) \Rightarrow F_s(x) = \begin{cases} 0 & x < x_0 \\ \frac{k}{n} & x_k \leq x < x_{k+1} \\ 1 & x > x_n \end{cases}$$

The the Kolmogorov statistic is the largest deviation :

$$D_n = \sup_x |F_s(x) - F(x)|$$

The test distribution has been computed by Kolmogorov:

$$P(D_n > \beta\sqrt{n}) = 2 \sum_r (-1)^{r-1} e^{-2r^2\beta^2}$$

$[0;\beta]$ define a confidence interval for D_n

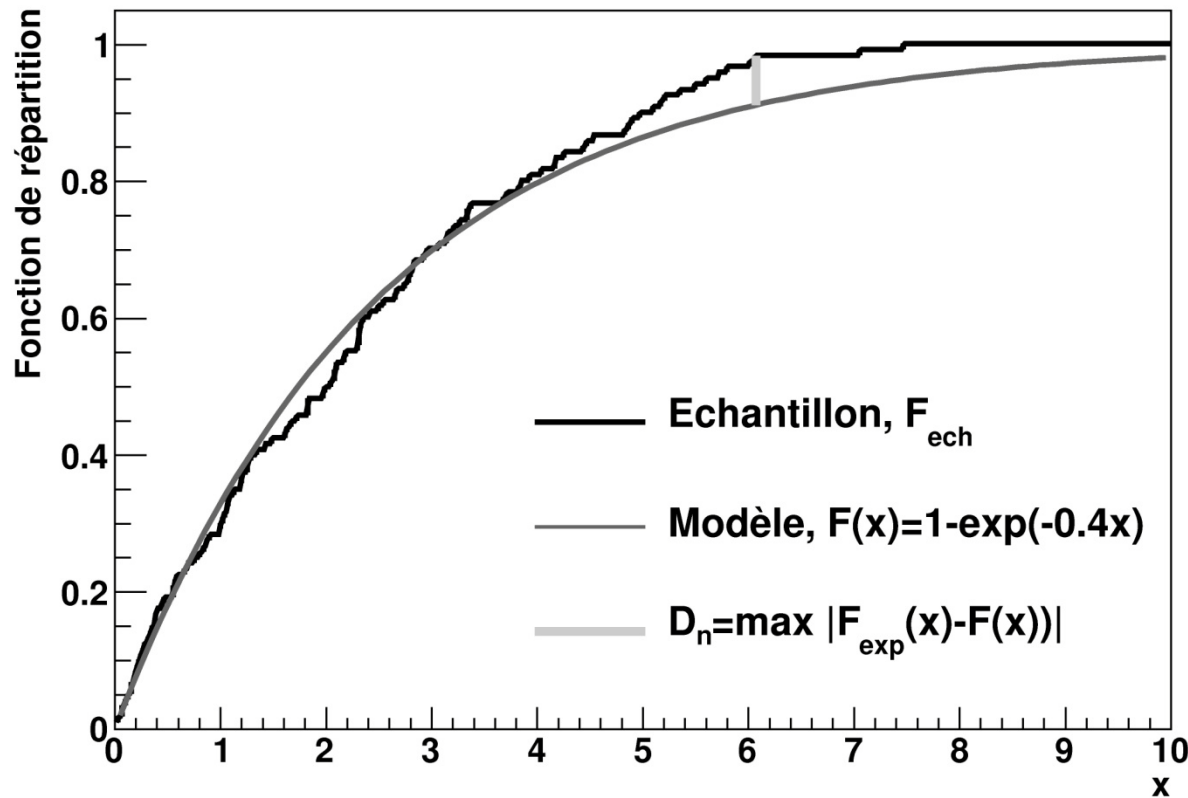
$\beta=0.9584/\sqrt{n}$ for 68.3% CL

$\beta=1.3754/\sqrt{n}$ for 95.4% CL

Example

Test compatibility with an exponential law : $f(x) = \lambda e^{-\lambda x}$, $\lambda = 0.4$

0.008, 0.036, 0.112, 0.115, 0.133, 0.178, 0.189, 0.238, 0.274, 0.323, 0.364, 0.386, 0.406, 0.409, 0.418, 0.421, 0.423, 0.455, 0.459, 0.496, 0.519, 0.522, 0.534, 0.582, 0.606, 0.624, 0.649, 0.687, 0.689, 0.764, 0.768, 0.774, 0.825, 0.843, 0.921, 0.987, 0.992, 1.003, 1.004, 1.015, 1.034, 1.064, 1.112, 1.159, 1.163, 1.208, 1.253, 1.287, 1.317, 1.320, 1.333, 1.412, 1.421, 1.438, 1.574, 1.719, 1.769, 1.830, 1.853, 1.930, 2.041, 2.053, 2.119, 2.146, 2.167, 2.237, 2.243, 2.249, 2.318, 2.325, 2.349, 2.372, 2.465, 2.497, 2.553, 2.562, 2.616, 2.739, 2.851, 3.029, 3.327, 3.335, 3.390, 3.447, 3.473, 3.568, 3.627, 3.718, 3.720, 3.814, 3.854, 3.929, 4.038, 4.065, 4.089, 4.177, 4.357, 4.403, 4.514, 4.771, 4.809, 4.827, 5.086, 5.191, 5.928, 5.952, 5.968, 6.222, 6.556, 6.670, 7.673, 8.071, 8.165, 8.181, 8.383, 8.557, 8.606, 9.032, 10.482, 14.174



$$D_n = 0.069$$

$$p\text{-value} = 0.617$$

$$1\sigma : [0, 0.0875]$$

Hypothesis testing

- Two exclusive hypotheses H_0 and H_1
- which one is the most compatible with data
 - how incompatible is the other one

$$P(\text{data}|H_0) \quad \text{vs} \quad P(\text{data}|H_1)$$

Build a **statistic**, define an interval w

- if the observation falls into w : accept H_1
- else accept H_0

Size of the test : how often did you get it right

$$\alpha = \int_w \mathcal{L}(x | H_0) dx$$

Power of the test : how often do you get it wrong !

$$1 - \beta = \int_w \mathcal{L}(x | H_1) dx$$

Neyman-Pearson lemma : optimal statistic for testing hypothesis is the **Likelihood ratio**

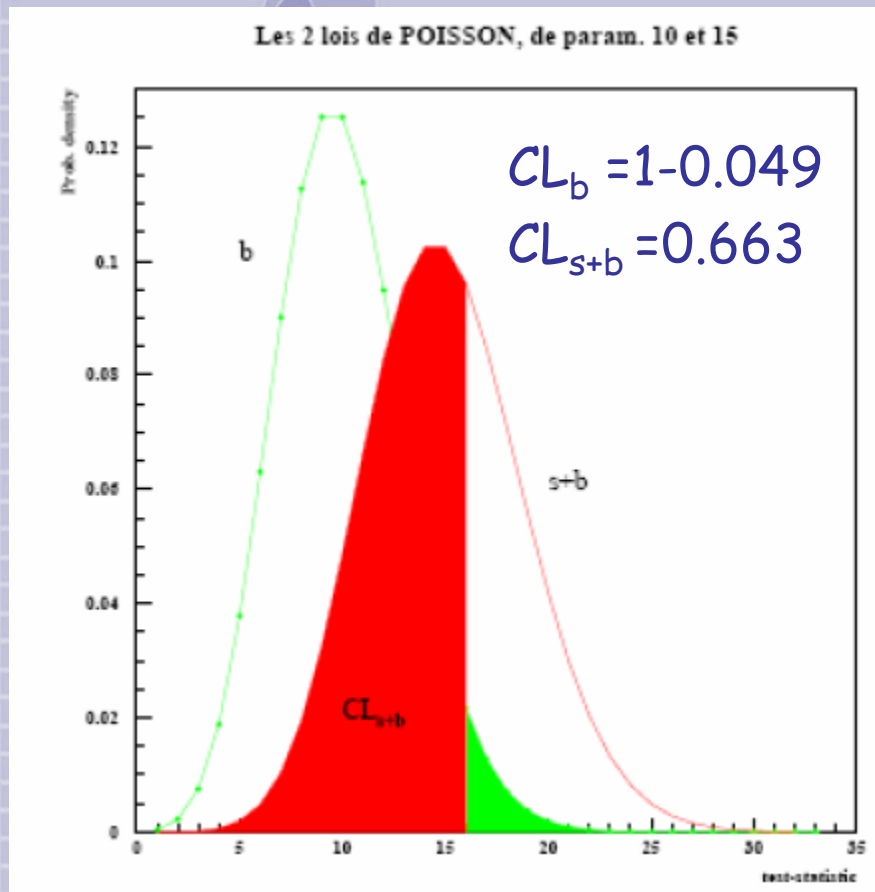
$$\lambda = \frac{\mathcal{L}(x | H_0)}{\mathcal{L}(x | H_1)} < k_\alpha$$

CL_b and CL_s

Two hypothesis, for counting experiment

- background only : expect 10 events
- signal+ background : expect 15 events

You observe 16 events



CL_b = confidence in the background hypothesis (power of the test)

Discovery : $1 - CL_b < 5.7 \times 10^{-7}$

CL_{s+b} = confidence in the signal+background hypothesis (size of the test)

Rejection : $CL_{s+b} < 5 \times 10^{-2}$

Test for signal (non standard)

$CL_s = CL_{s+b} / CL_b$