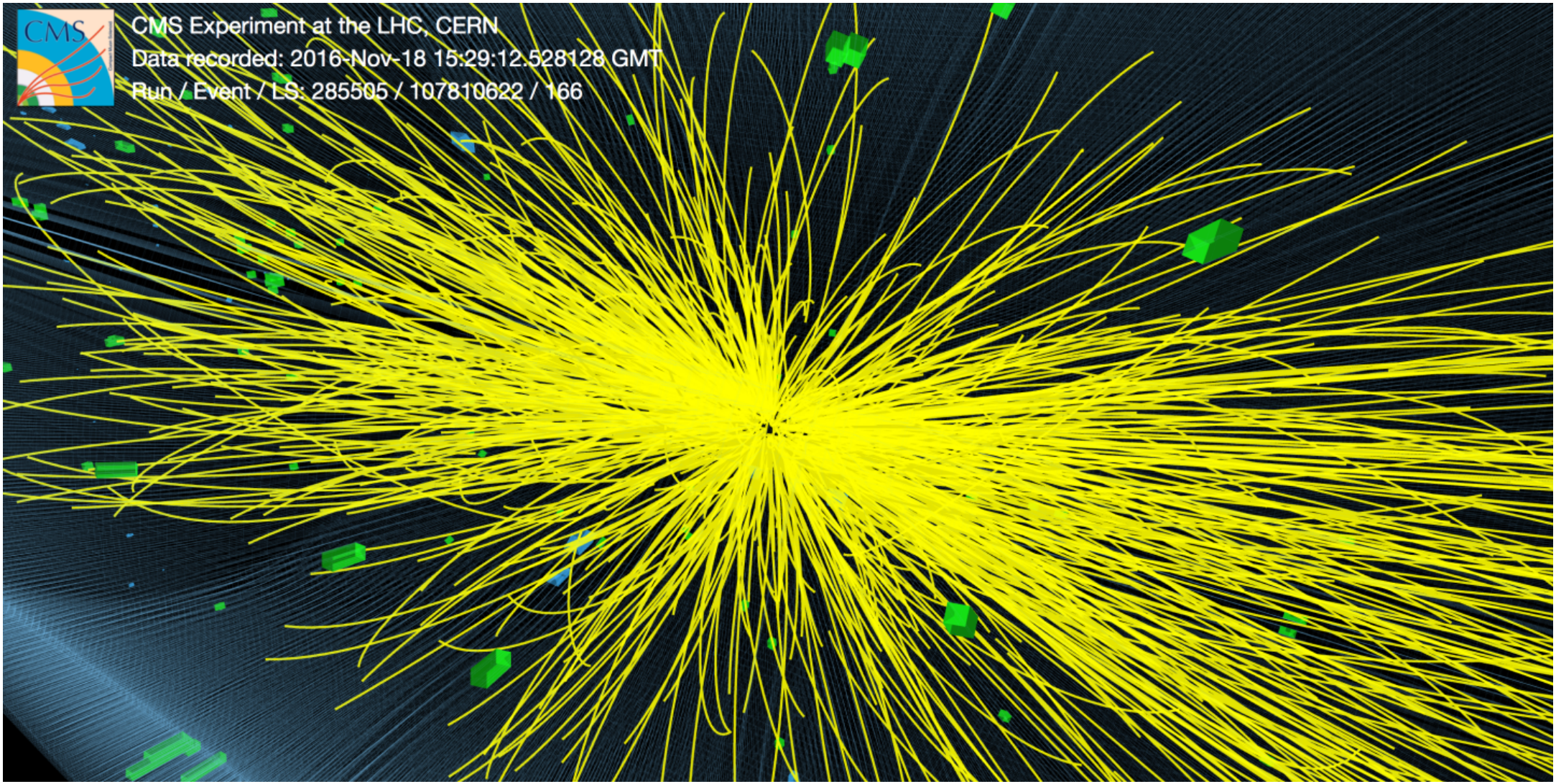
Statistical Data Analysis

Getting physics from HEP data



Giacomo Graziani (INFN Firenze)

Third WISHEPP School - Nablus, 2018

INFN Istituto Nazionale di Fisica Nucleare

Getting physics from LHC data may look like this

But we want even more:

- not only “find a needle” but all the needles hidden there, or be sure that none can be found
- using a reasonable amount of resources

In short, the goal is to use all the available information in your dataset about the physics you’re looking after

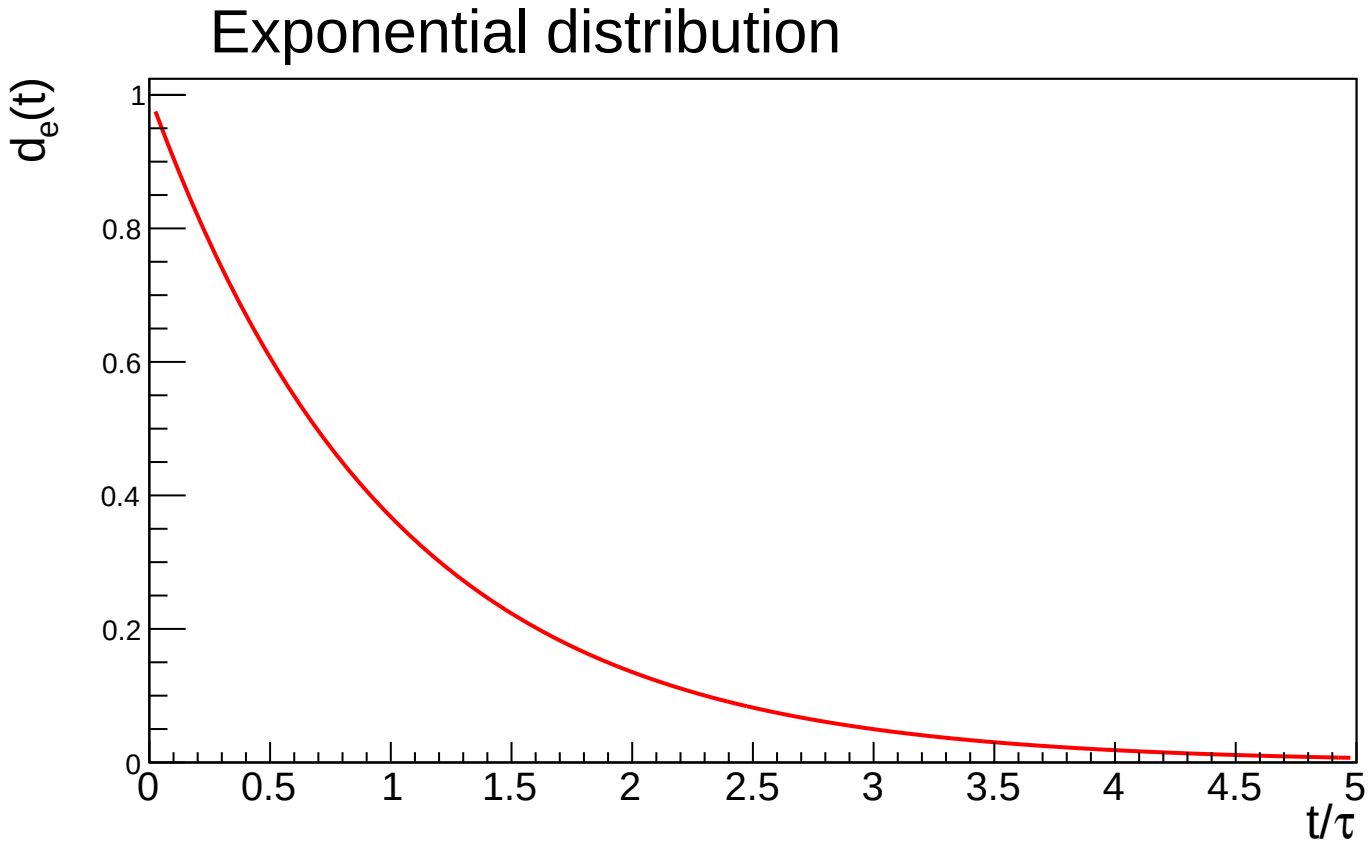
Experimental observables are Random variables

The fundamental reason is quantum mechanics

Example: decay time of an unstable particle. We can't predict when each decay will occur, but we can predict the probability $P(t) = (1 - e^{-t/\tau})$ to occur within a time t.

The probability density function (PDF) for the variable t is

$$d_e(t) = \frac{dP(t)}{dt} = \frac{1}{\tau} e^{-t/\tau}$$



Suppose to observe N decays. The distribution of the decay times can be shown with an histogram, where the number of decays k_i occurring in each time interval Δ_i is counted. The probability that a decay occurs in a given interval is

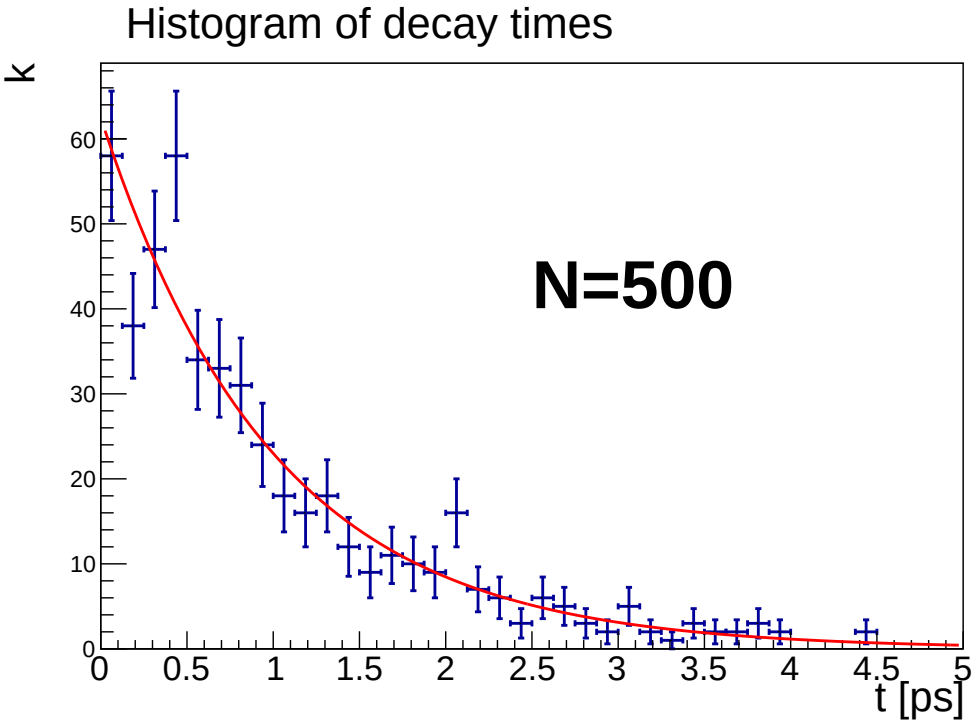
$$p_i = \int_{\Delta_i} d_e(t)dt$$

Each n_i is itself a random variable, following a binomial distribution

$$d_B(k_i) = \frac{k_i!}{N!(N-k_i)!} p_i^{k_i}(1-p_i)^{N-k_i}$$

whose expected value is obviously Np_i:

$$E(k_i) \equiv \sum_0^N k_i d_B(k; p_i, N) = Np_i \equiv \lambda_i$$



Its random fluctuation around the expected value can be quantified by the standard deviation

$$\sigma(k_i) \equiv \sqrt{E\left[(k_i - E(k_i))^2\right]} = \sqrt{Np_i(1-p_i)}$$

In the limit $p_i \ll 1$ the binomial distribution can be approximated to the Poisson distribution

$$d_B(k_i) \to d_P(k_i) = \frac{e^{-\lambda_i}\lambda_i^{k_i}}{k_i!}$$

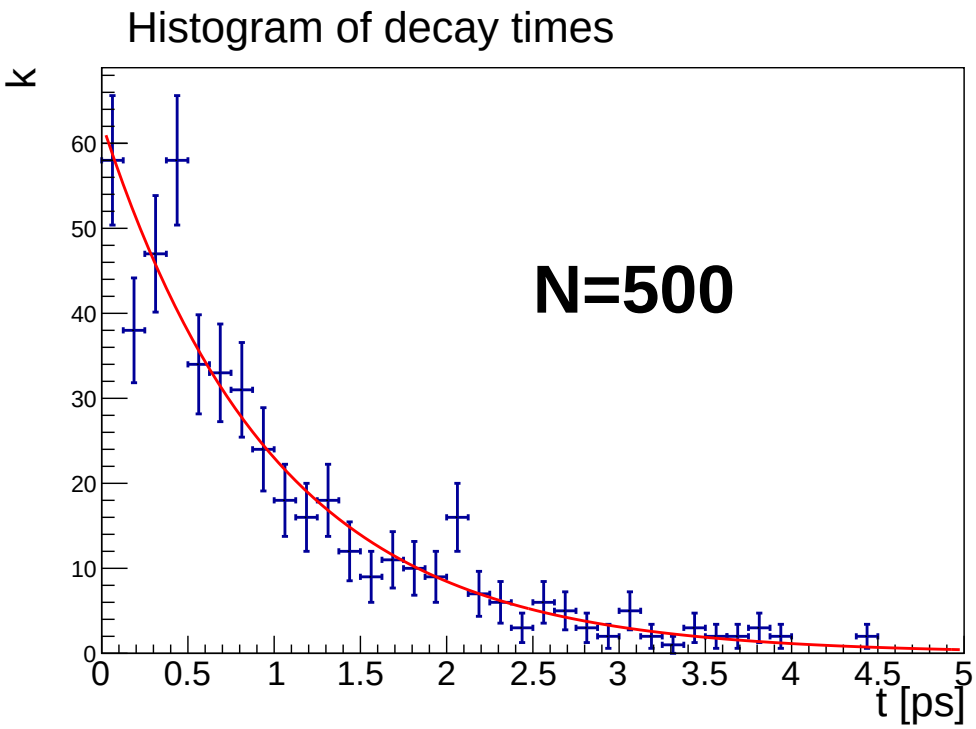
$$\sigma(k_i) \to \sqrt{\lambda_i}$$

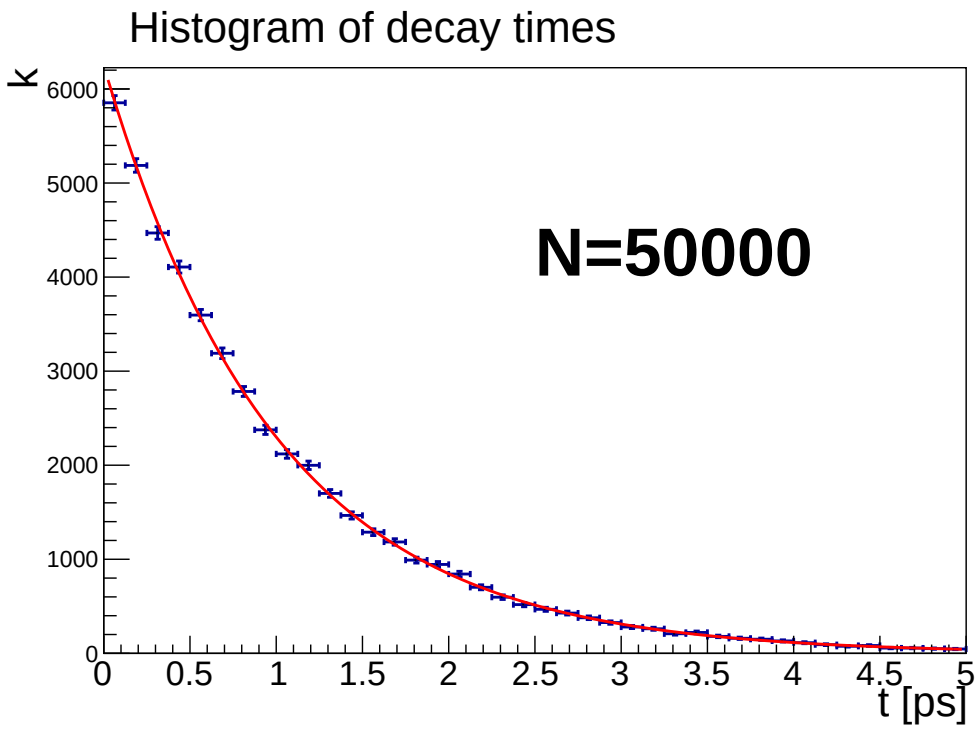
When trying to infer the model (in this case, the lifetime τ) from the observations, we are limited by the available statistics.

p_i in each interval can be estimated from the frequency $f_i = k_i/N$, whose standard deviation is

$$\sigma(f_i) = \sqrt{\frac{p_i(1-p_i)}{N}}$$

Its value will thus converge to p_i for $N \to \infty$, a concept known as the "law of large numbers"





The response of any measuring device is usually affected by many sources of random uncertainty (temperature or vibrational effects, digital sampling, ...). When several effects sum up linearly, we can expect that, at least to a good approximation, the resulting PDF is Gaussian, according to the central limit theorem (CLT) :

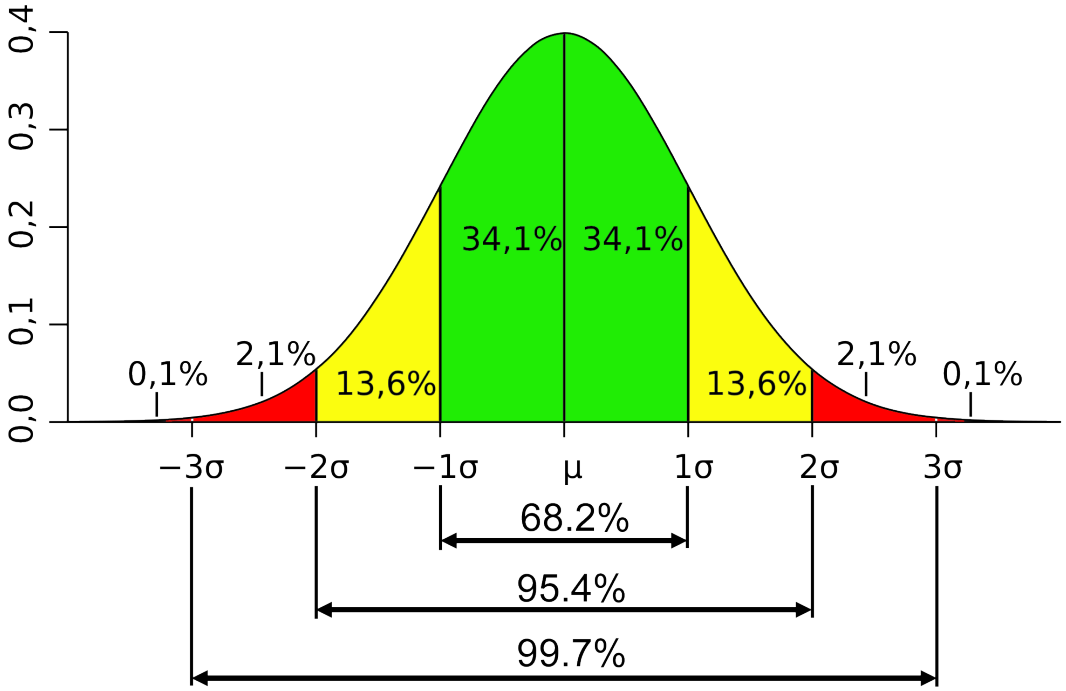
For a set of M independent random variables x_i, distributed according to (almost) any distribution, the PDF of the sum $X = \sum_i x_i$ converges to a Gaussian (or normal) distribution

$$d_G(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right]$$

where $\mu = \sum_i E(x_i)$ and $\sigma^2 = \sum_i \sigma(x_i)^2$ are the expected value and variance (squared standard deviation) of the distribution

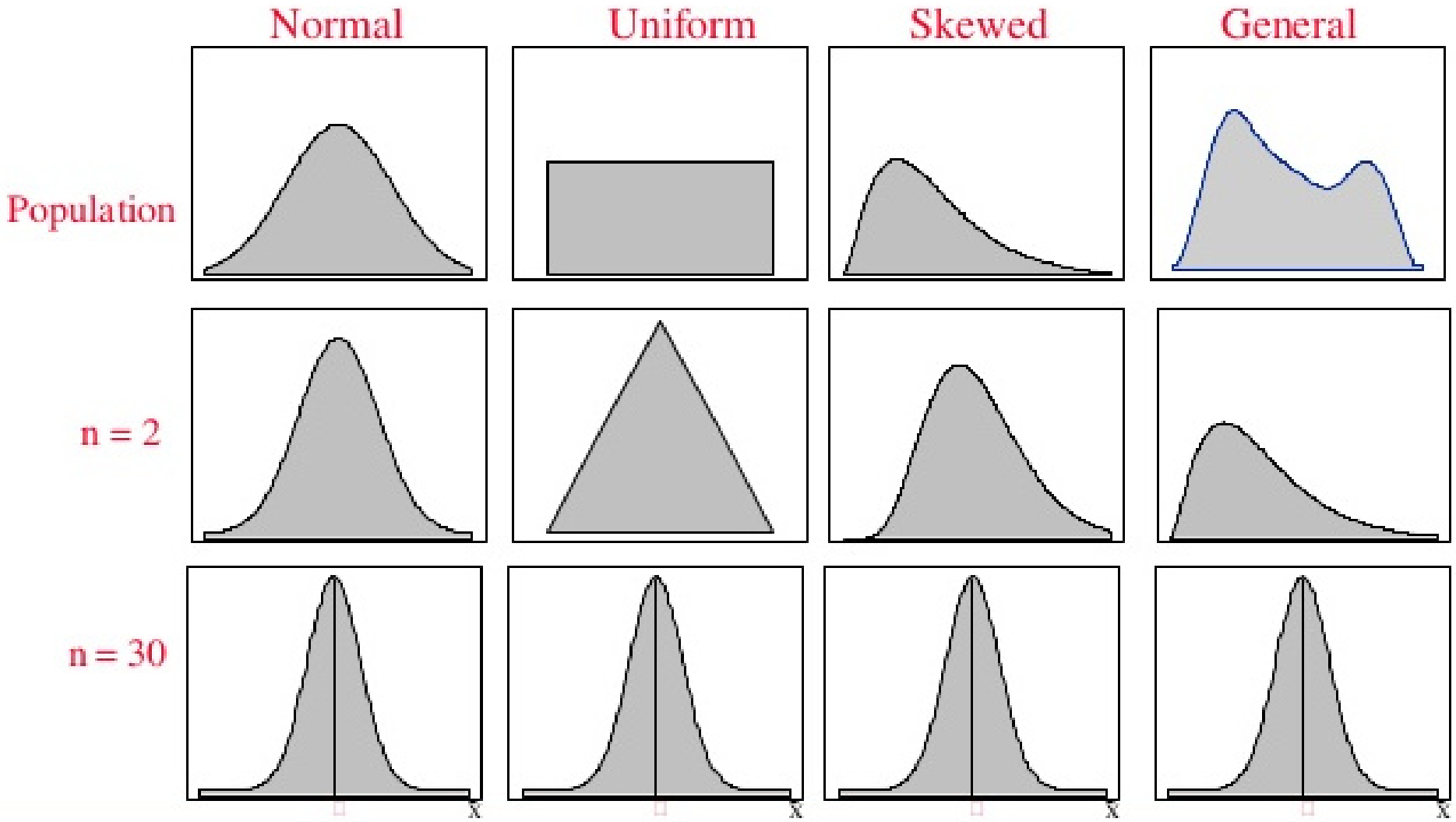
It is costumary to express central intervals corresponding to a given probabiltiy in terms of "number of σs":

$$\begin{aligned} P(|X-\mu| < n\sigma) = \quad & 68.27\% \quad \text{for } n=1 \\ & 95.45\% \quad \text{for } n=2 \\ & 99.73\% \quad \text{for } n=3 \end{aligned}$$

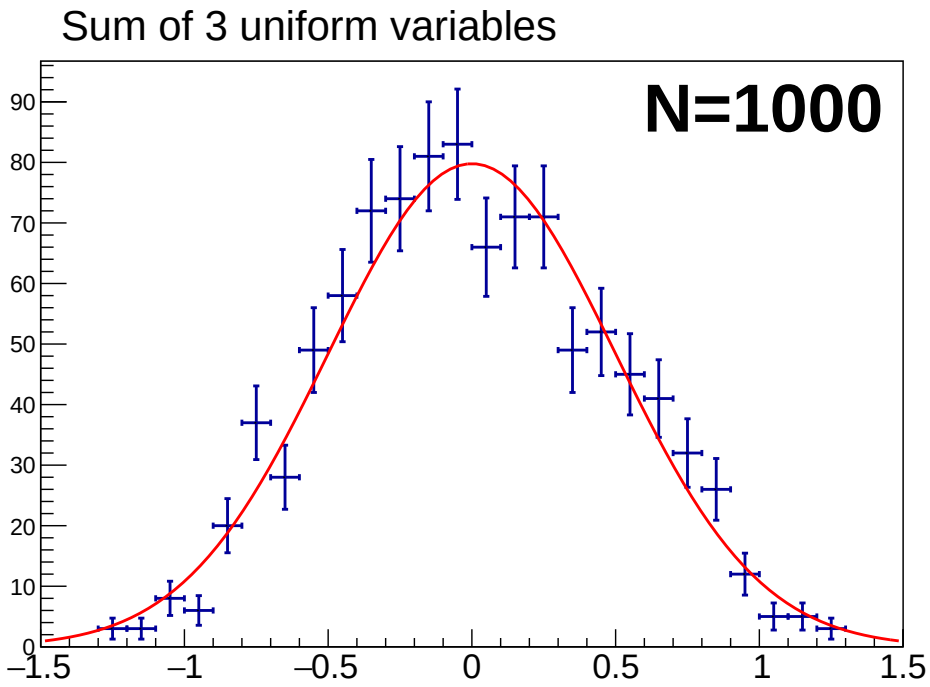


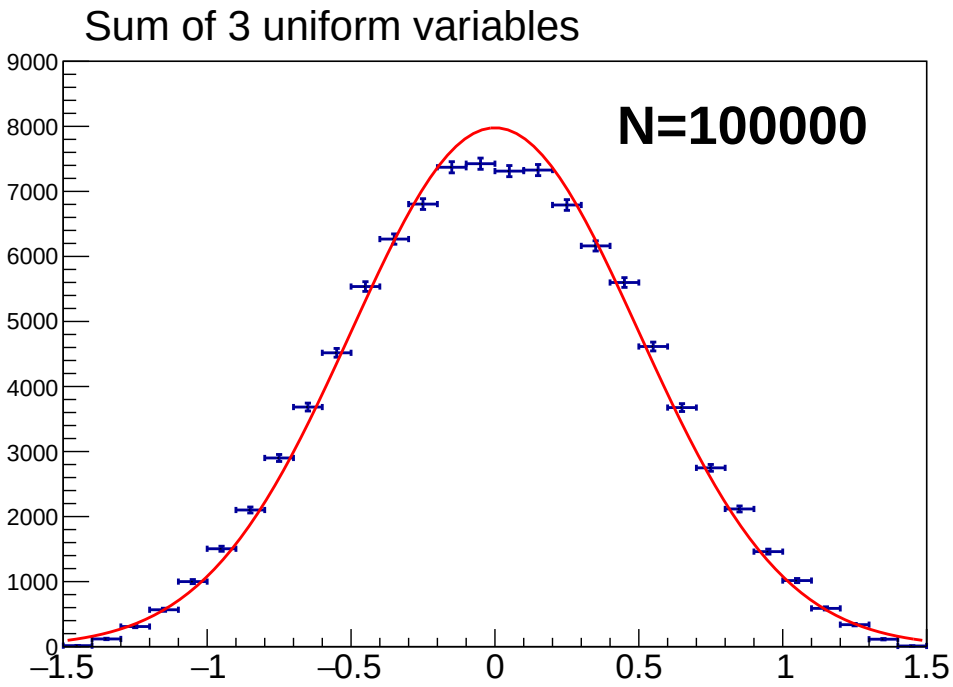
Note that the value of $P(|X-\mu| < n\sigma)$ is distribution dependent,
though a limit valid for any distribution is the Chebyshev's inequality $P(|X-\mu| > n\sigma) < 1/n^2$

The CLT limit is strictly valid only for $n \to \infty$ variables, but is still very powerful in practice.
Try yourself with a computer simulation!



Take for example M variables uniformly distributed in $[-0.5, 0.5]$.
The PDF of the sum can be obtained by a convolution of uniform distributions, resulting in polynomials of degree M-1. Already for M=3, a large statistics is needed to distinguish the observed distribution from a gaussian.



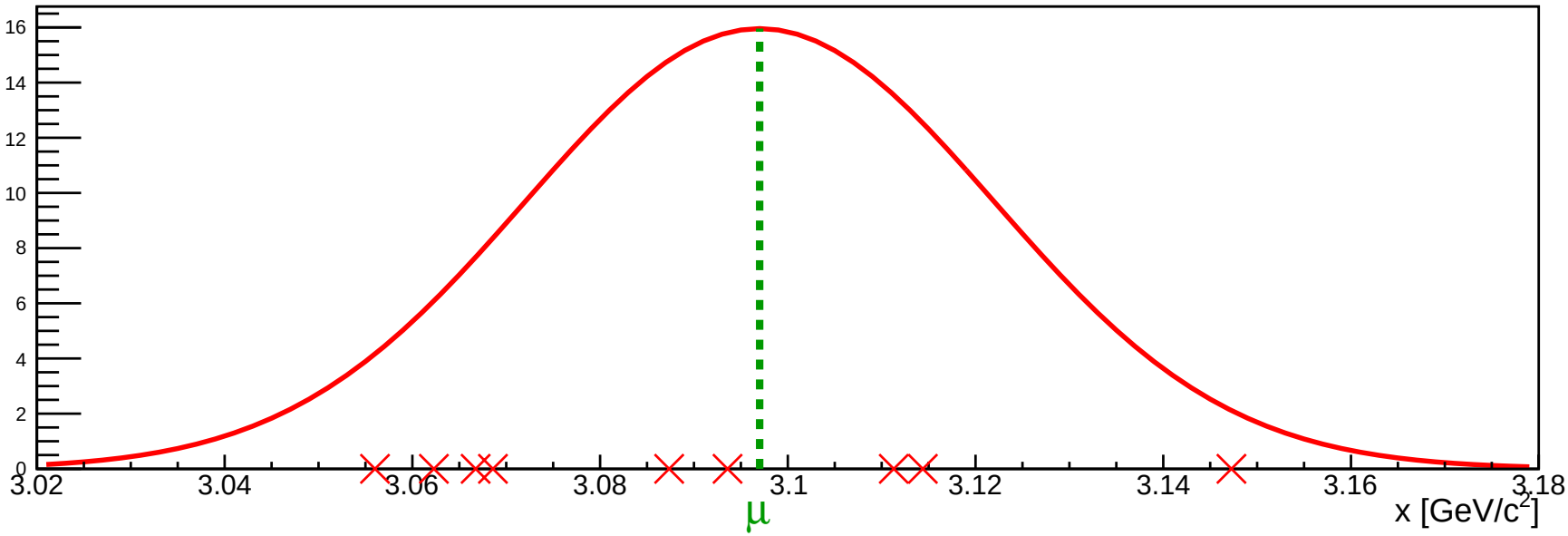


Still, with very large statistics, a non-gaussian model may be needed to describe the data

Estimators

Now let's take the experimentalist's viewpoint.
From the observations, we want to estimate the parameters of the underlying model.

Example: we have a set $[x_i] \quad i = 1, \ldots N$
of independent measurements of the mass of a particle μ, and we can assume that the uncertainty is gaussian.



What is the best estimation of μ?

A common estimator of the expected value is the average

$$\overline{x} = \frac{\sum_i x_i}{N}$$

which has the nice property of being unbiased for any PDF of x:

$$E(\overline{x}) = \frac{\sum_i E(x_i)}{N} = E(x) = \mu$$

its standard deviation is

$$\sigma(\overline{x}) = \sqrt{\frac{\sum_i \sigma^2(x_i)}{N^2}} = \frac{\sigma(x)}{\sqrt{N}}$$

If $\sigma(x)$ is not known, it can be estimated through the standard root mean square estimator:

$$\overline{\sigma^2} = \frac{\sum_i (x_i - \overline{x})^2}{N-1}$$

So we have an estimation of μ and its uncertainty.
Are we allowed to stop here?
Did we use all the information about μ contained in our N measurements?

For example, we could have used two other unbiased estimators of μ:

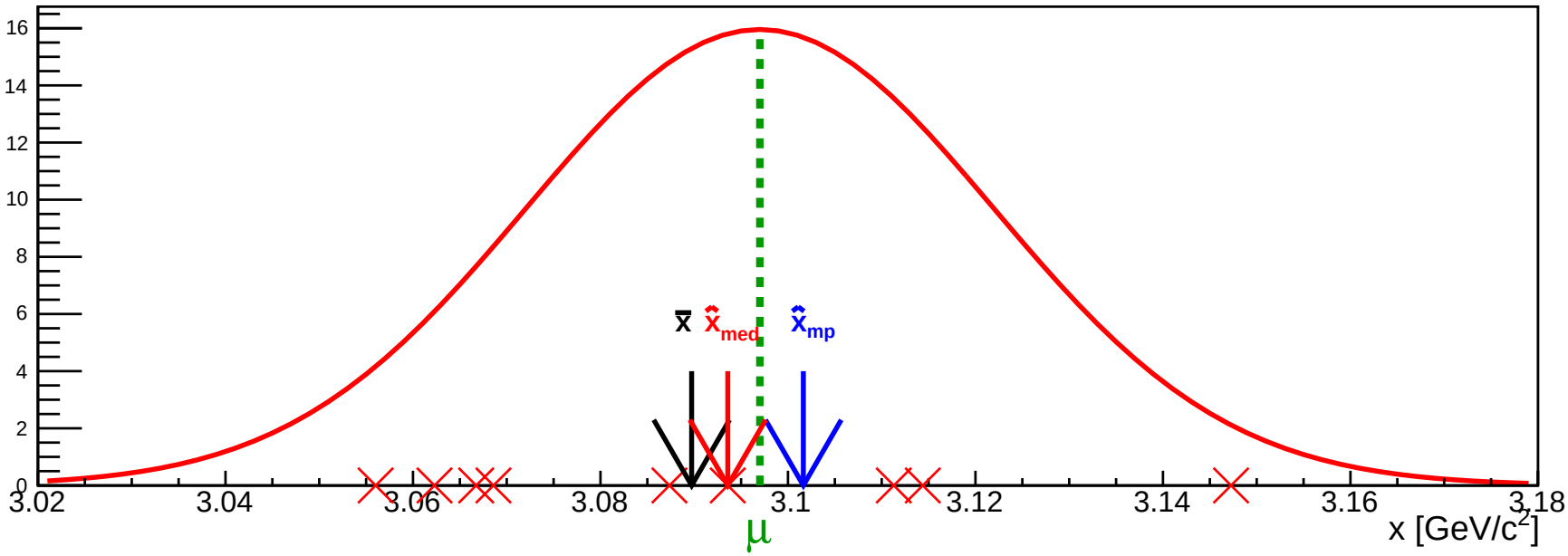
The median: taking the ordered list of the x_i

$$\hat{x}_{med} = \begin{cases} \dfrac{x_{N/2} + x_{N/2+1}}{2} & N \quad \text{even} \\ x_{(N+1)/2} & N \quad \text{odd} \end{cases}$$

The midpoint:

$$\hat{x}_{mp} = \frac{max([x_i]) + min([x_i])}{2}$$

For a symmetric distribution (as the Gaussian), $\overline{x}$, $\hat{x}_{med}$ and $\hat{x}_{mp}$ are all unbiased estimators of μ



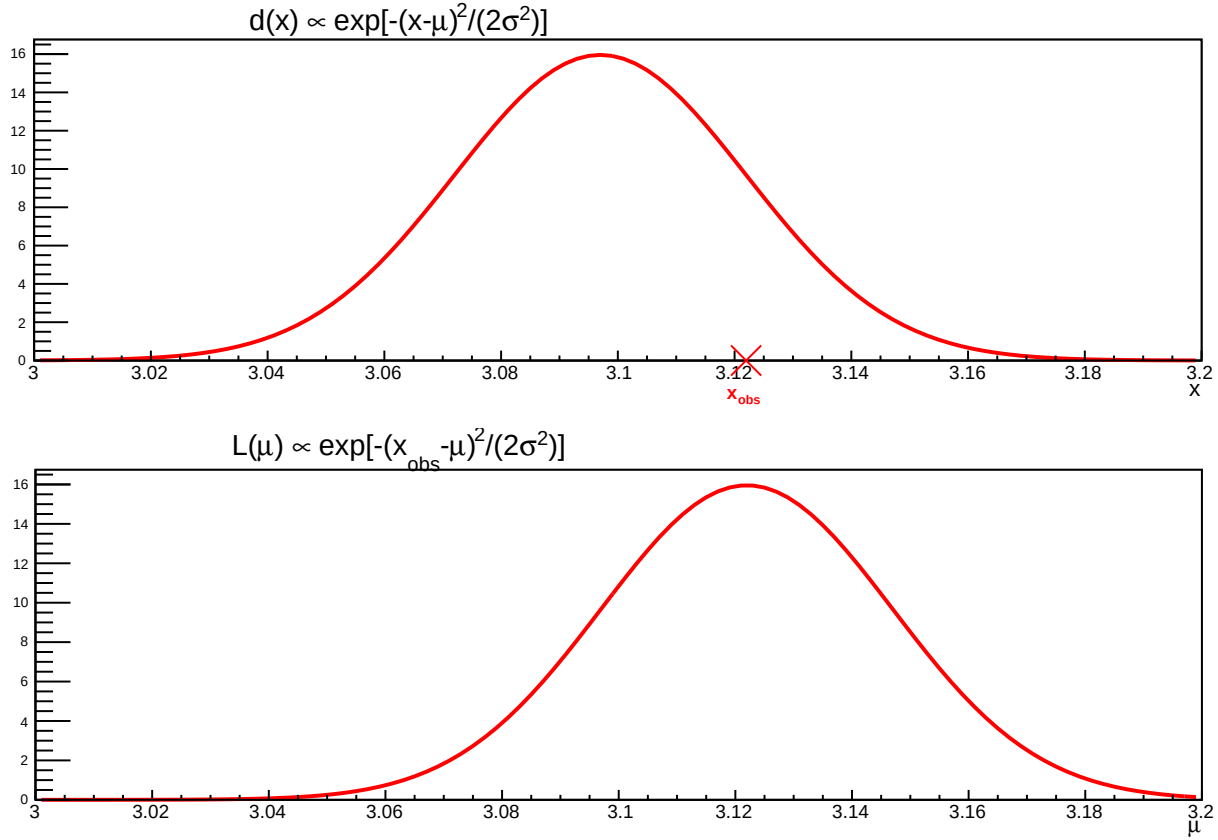
Of course, the one with lowest variance ("the most efficient") should be preferred.

Is there a general way to choose the best estimator?

The Likelihood function is a way to quantify how a given hypothesis for the value of the wanted parameter(s) is compatible with the observations. It is simply the PDF of the observed random variables, interpreted as a function of the parameter, while the random variables are fixed to the observed values.

$$\mathcal{L}(\mu) = d([x_i] \mid \mu)$$

$$\text{if measurements are independent} \implies \mathcal{L}(\mu) = \prod_i d(x_i|\mu)$$



The "principle of maximum likelihood" consists in choosing the estimator which maximes $\mathcal{L}$. For example, for a single measurement x_{obs} of a gaussian variable the ML estimate will be $\mu = x_{obs}$

In our example of N Gaussian measurements we have

$$\mathcal{L}(\mu) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

It is convenient to find the maximum of $\log(\mathcal{L})$:

$$\begin{aligned}
& \frac{\partial \log \mathcal{L}(\mu)}{\partial \mu}\Big|_{\overline{\mu}_{ML}} = 0 \\
\Longrightarrow \quad & \frac{\partial}{\partial \mu} \sum_i -\frac{(x_i-\mu)^2}{2\sigma^2} = 0 \\
\Longrightarrow \quad & \sum_i \frac{2(x_i - \overline{\mu}_{ML})}{2\sigma^2} = 0 \\
\Longrightarrow \quad & \overline{\mu}_{ML} = \frac{\sum_i x_i}{N}
\end{aligned}$$

So, the average is a ML estimator of $E(x)$ in the Gaussian case.

It can be shown that, under very general assumption, for any estimator $\hat{x}$ of a parameter μ

$$\sigma^2(\hat{x}) \geq \frac{\left(\frac{\partial E(\hat{x})}{\partial \mu}\right)^2}{I_\mu} \qquad \text{where} \qquad I_\mu = E\left[\left(\frac{\partial \log \mathcal{L}(\mu)}{\partial \mu}\right)^2\right] = -E\left(\frac{\partial^2 \log \mathcal{L}(\mu)}{\partial \mu^2}\right)$$

this is the Cramèr–Rao bound.
I_μ is called Fisher's information and quantifies the amount of information carried by the data sample on the μ parameter

For an unbiased estimator $\boxed{\sigma^2(\hat{x}) \geq 1/I_\mu}$

So the "perfect" estimator should be unbiased and having minimum variance.

Note that, for independent and identically distributed measurements $\mathcal{L}(\mu) = \prod_{i=1}^{N} d(x_i|\mu)$
$I_\mu = E\left(-\sum_{i=1}^{N} \partial^2 log(d(x_i|\mu))/\partial\mu^2\right) \propto N \implies$ the minimum σ decreases as $\sqrt{N}$

In our example

$$\frac{\partial}{\partial \mu} \log \mathcal{L}(\mu) = \sum_{i=1}^{N} \frac{-(x_i - \mu)}{\sigma^2}$$

$$\implies \quad \frac{\partial^2}{\partial \mu^2} \log \mathcal{L}(\mu) = -\frac{N}{\sigma^2}$$

$$\implies \quad I_\mu = \frac{N}{\sigma^2}$$

which is eactly the inverse variance of the $\overline{x}$ estimator!
$\Longrightarrow$the ML estimator of μ has minimum variance

The choice of $\overline{x}$ as best estimator for our case could have been simply justified by noticing that the conditional probability

$$d([x_i] \mid \overline{x})$$

is not dependent on μ. Namely, once the value of our estimator is fixed, the particular configuration of our data points doesn't carry any additional information on μ.
In this case the estimator is said to be sufficient, and the likelihood can be factorised as

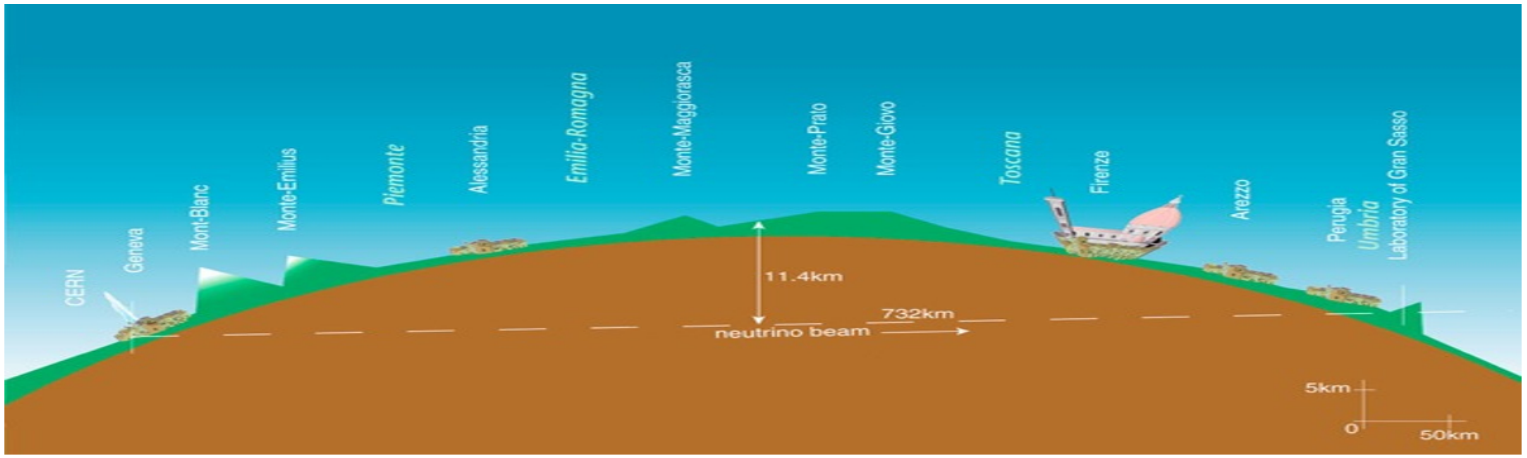
$$\mathcal{L}([x_i], \mu) = g(\hat{x}, \mu) h([x_i]) \qquad \text{with} \qquad dh/d\mu = 0$$

In our example

$$\begin{aligned} \mathcal{L}([x_i], \mu) &= \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{-N\mu^2}{2\sigma^2} + \frac{N\mu\overline{x}}{\sigma^2}\right) \cdot \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(\frac{-\sum_i x_i^2}{2\sigma^2}\right) \\ &= g(\hat{x}, \mu) \cdot h([x_i]) \end{aligned}$$

An unbiased and sufficient estimator will also be an unbiased efficient one.
Clearly, if a sufficient estimator exists, the maximum likelihood estimator is a function of it.

Let's now consider a set $[t_i]$ following a uniform distribution
Example: arrival time of the neutrinos from CERN to Gran Sasso



We need the average arrival time T, knowing that the data follow a uniform PDF with limits $[T-\delta, T+\delta]$

$$d_u(t) = \begin{cases} 1/(2\delta) & T-\delta < t < T+\delta \\ 0 & \text{otherwise} \end{cases}$$

What is the best estimator for T?

$$\mathcal{L}(T) = \prod_i d_u(t_i, T) = \begin{cases} (1/(2\delta))^N & T-\delta < \min(t_i) < \max(t_i) < T+\delta \\ 0 & \text{otherwise} \end{cases}$$

so a sufficient estimator can only depend on $\min(t_i)$ and $\max(t_i)$. The configuration of the other data points doesn't bring any information

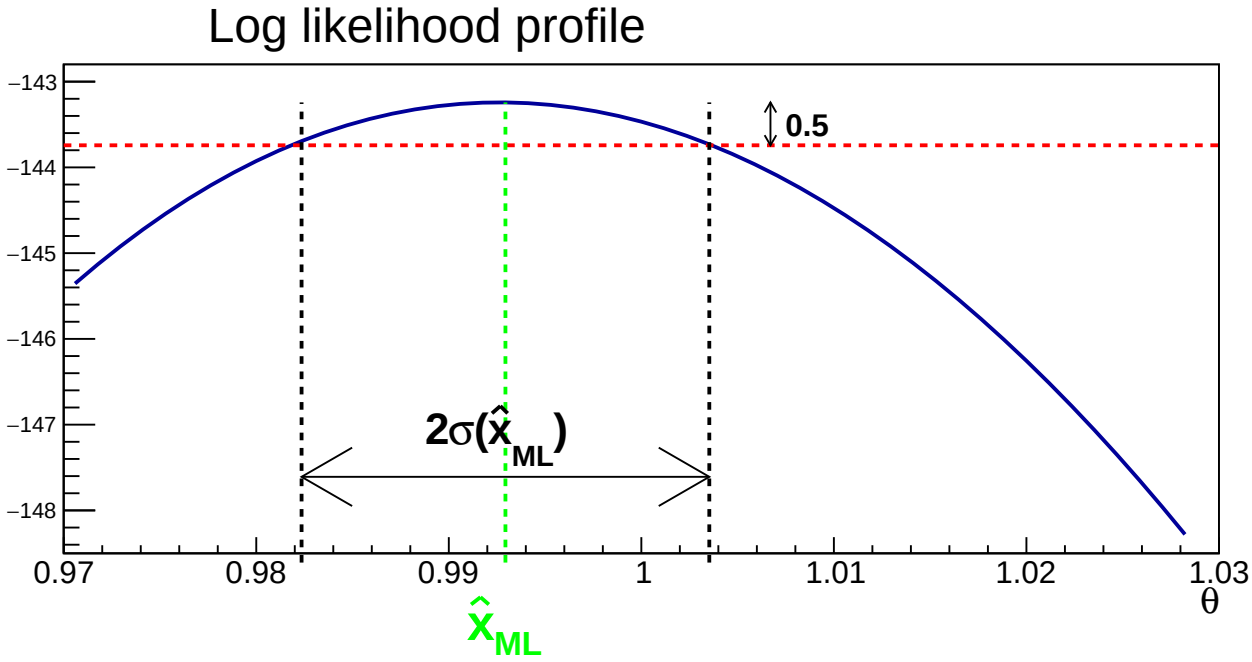
$\Longrightarrow$the unbiased estimator $\hat{x}_{mp} = (max(x_i) + min(x_i))/2$ performs better than the average $\overline{x}$

In general, ML estimators are not necessarily unbiased and efficient.
But, as a consequence of the CLT, they get both properties in the limit $N \to \infty$

$$\log \mathcal{L}(\theta) = \log \mathcal{L}(\hat{x}_{ML}) + \frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}(\hat{x}_{ML}) \cdot \frac{(\theta - \hat{x}_{ML})^2}{2} + \dots$$
$$\to \log \mathcal{L}(\hat{x}_{ML}) + E\left(\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right) \cdot \frac{(\theta - \hat{x}_{ML})^2}{2} + \mathcal{O}(1/\sqrt{N})$$
$$= \log \mathcal{L}(\hat{x}_{ML}) - \frac{I_\mu}{2}(\theta - \hat{x}_{ML})^2 + \mathcal{O}(1/\sqrt{N})$$

The profile of $\log \mathcal{L}$ tends asymptomatically to become parabolic, with a curvature equal to half the Fisher's information. The pdf of $\hat{x}_{ML}$ tends to be a Gaussian with minimum variance $1/I_\mu$

In practice, a parabolic profile is telling us if we are reasonably close to the limit, and allows to estimate the uncertainty of the estimation from the values of μ corresponding to $\max(\log \mathcal{L}) - 1/2$
(more commonly, $\min(-2 \log \mathcal{L}) + 1$)



The likelihood method can be extended to any number of parameters, by finding the maximum on the multi-dimensional space of the parameters $\underline{\theta}$.

In the limit $N \to \infty$, $\mathcal{L}(\underline{\theta})$ tends to

$$\log \mathcal{L}(\underline{\theta}) = \log \mathcal{L}(\underline{\hat{\theta}}_{ML}) - \frac{1}{2}(\underline{\theta} - \underline{\hat{\theta}}_{ML})^T F (\underline{\theta} - \underline{\hat{\theta}}_{ML})$$

where F is the Fisher information matrix

$$F_{ij} = E\left(-\frac{\partial^2 \log \mathcal{L}(\underline{\theta})}{\partial\theta_i \partial\theta_j}\right)$$

The covariance matrix of the ML estimators is obtained by inverting the matrix:

$$V_{\underline{\hat{\theta}}} \equiv E\left[(\underline{\hat{\theta}} - E(\underline{\hat{\theta}}))(\underline{\hat{\theta}} - E(\underline{\hat{\theta}}))^T\right] = \begin{pmatrix} \sigma^2(\hat{\theta}_1) & cov(\hat{\theta}_1, \hat{\theta}_2) & \dots & cov(\hat{\theta}_1, \hat{\theta}_n) \\ cov(\hat{\theta}_2, \hat{\theta}_1) & \sigma^2(\hat{\theta}_2) & & \\ \dots & & & \\ & & & \sigma^2(\hat{\theta}_n) \end{pmatrix} = F^{-1}$$

which is optimal, since the Cramèr-Rao bound becomes

$$(V_{\underline{\hat{\theta}}})_{ij} \geq (F^{-1})_{ij}$$

The χ^2 method

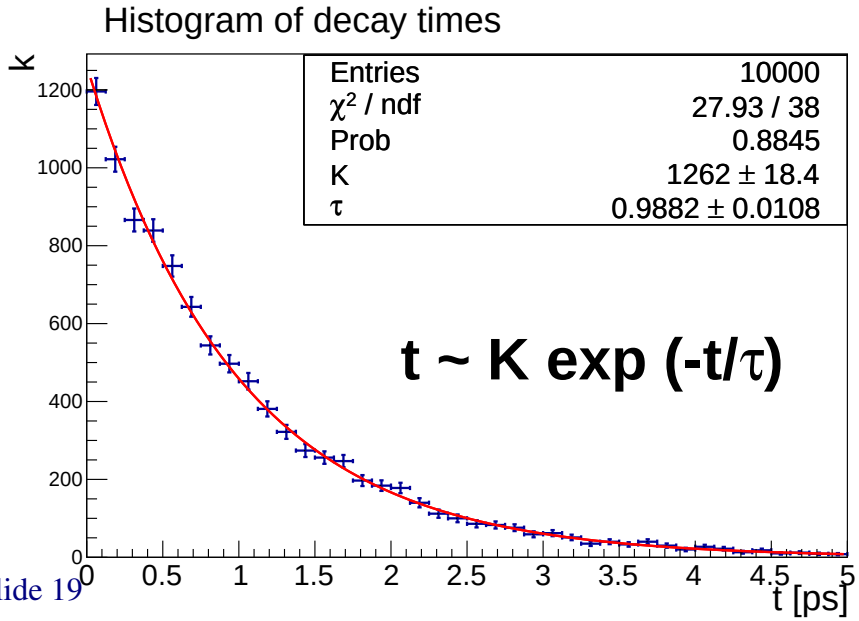
For large amounts of data, the computation of $\mathcal{L}$ can become computationally expensive, and we tend to do binned analysis, where the data space is divided in intervals and we analyze the counts k_i in each interval.
If the statistics is large enough the distribution of k_i tend to be a Gaussian with $\sigma_i = \sqrt{E(k_i)} \equiv \mu_i$, so that the likelihood function is (m is the number of intervals)

$$\mathcal{L} \sim \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\mu_i}} \exp\left(\frac{-(k_i-\mu_i)^2}{2\mu_i}\right)$$

and the parameters describing the model of μ_i can be find maximizing $\mathcal{L}$, i.e. minimizing the quantity

$$\chi^2 = \sum_{i=1}^{m} \frac{(k_i-\mu_i)^2}{\mu_i}$$

motivating the use of the χ^2, or "least square" method to "fit" the model with unknown parameters to the observed distribution



Example: lifetime measurement from our first dataset.
In this case $\mu_i \propto exp(-t/\tau)$.
The lifetime is estimated by minimizing the χ^2 with respect to τ.

The χ^2 fits also provide a "goodness-of-fit" test, since if the assumed model is correct, the distribution of the minimum χ^2 follows the well-known Pearson χ^2 distribution.

In general, for N gaussian variables $x_i \sim N(\mu_i, \sigma_i)$ the quantity

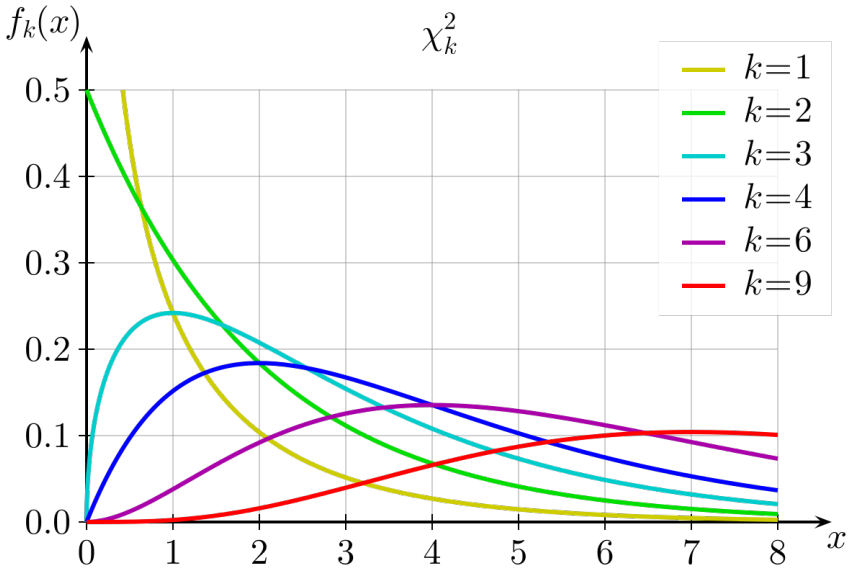
$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

follows the Pearson χ^2 distribution

$$f_{\chi^2}(\chi^2; n) = \frac{1}{2^{N/2}\Gamma(N/2)} (\chi^2)^{N/2-1} \exp(-\chi^2/2)$$

where $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$ and
N is called the "number of degrees of freedom" (ndf)

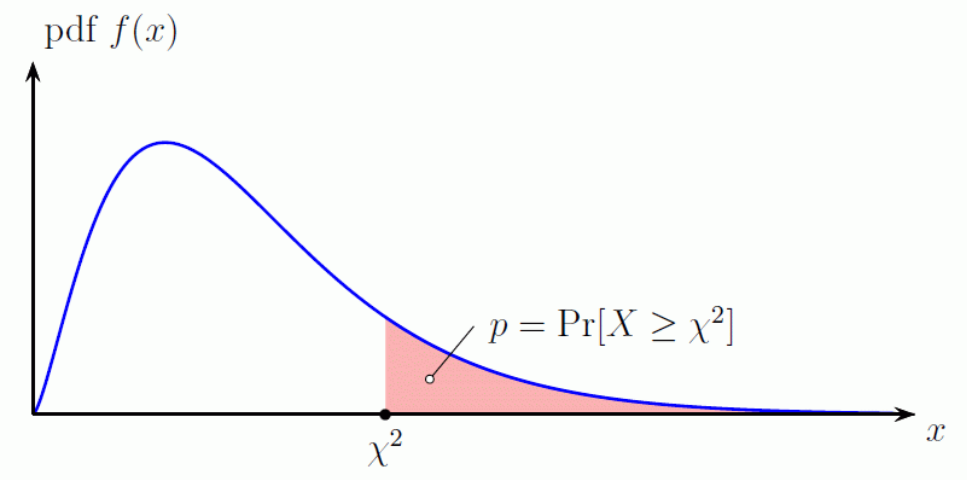
It has $E(\chi^2) = N$ and $\sigma(\chi^2) = \sqrt{2N}$

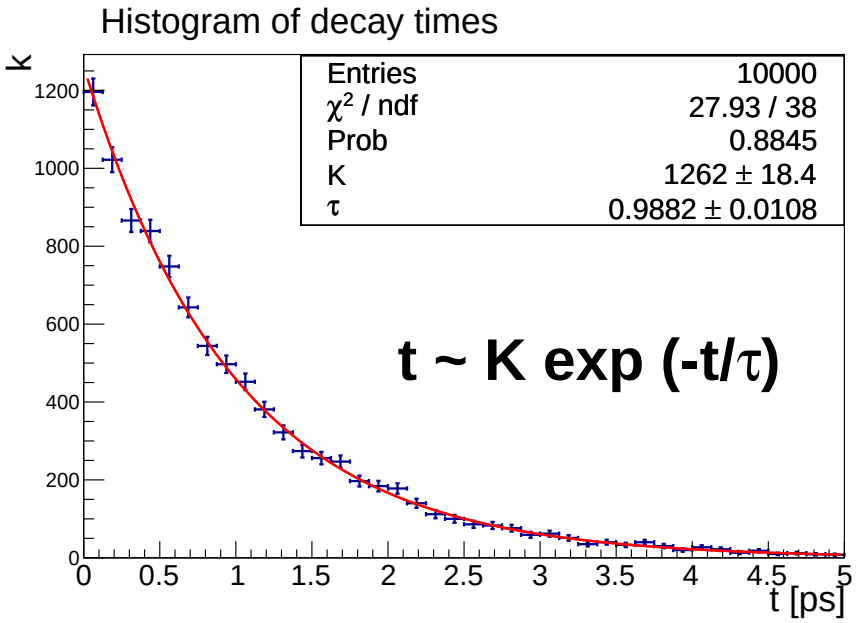


When the μ_i are fitted to your data with n_p independent parameters and normalizing the model curve to the number of observations, the $\min(\chi^2)$ still follows a Pearson distribution with $N = m - n_p - 1$

If the assumed model is not correct, the χ^2 will have larger values than expected.
So, to verify if the data are compatible with the model, one can first check if the χ^2 is close to its expected value $\chi^2/\text{ndf} \sim 1$.

More accurately, one can give a p-value, namely the probability that, if the model is correct, the χ^2 is larger than the observed value.
If the p-value is reasonably large (tipically $> 5\%$ or 10%) we can consider that the model is fitting the data.
This is the χ^2 goodness-of-fit test





In our example, ndf$= m - 1 - 1 = 38$
We find $\chi^2 = 27.93$, which is a reasonable value (p-value=88%)

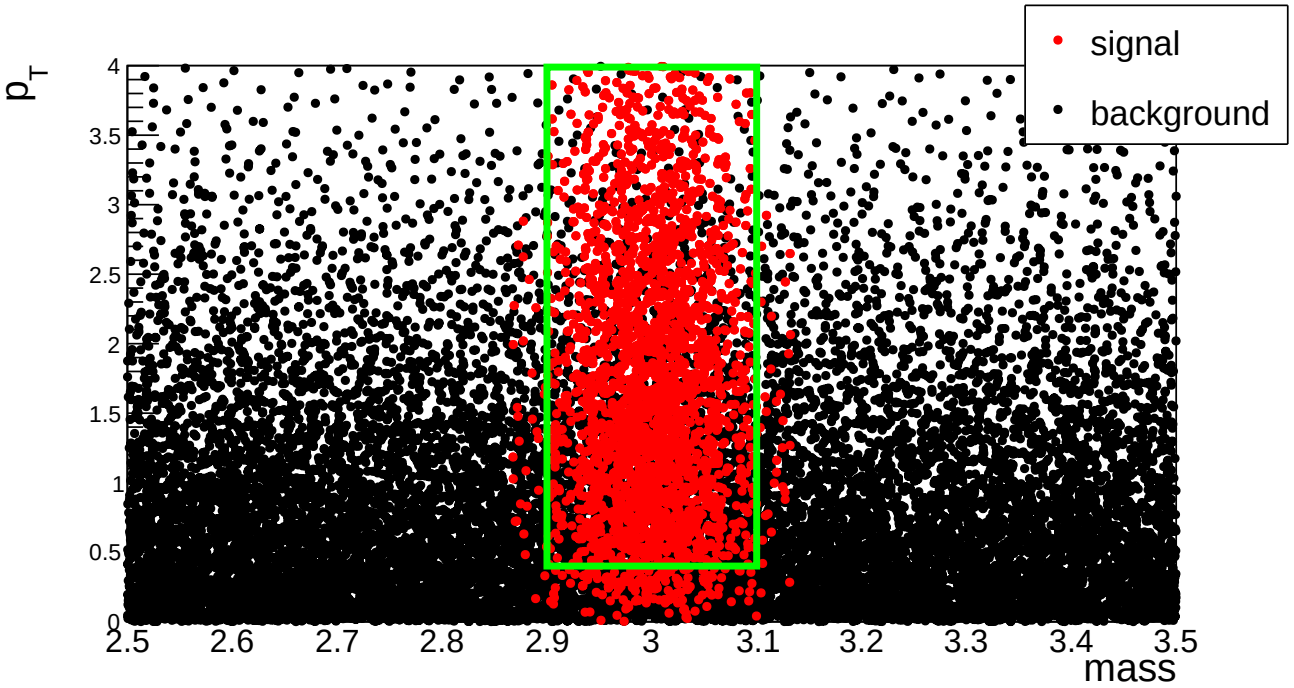
Classification

The likelihood function is also the ideal tool in classification problems, for example when we want to classify single events as signal or background from a set of observables $[x_i]$.

The "Neyman-Pearson lemma" states that the most powerful criterion to discriminate an hypothesis H_1 from an hypothesis H_0 is based on the likelihood ratio

$$\frac{\mathcal{L}([x_i] \mid H_1)}{\mathcal{L}([x_i] \mid H_0)} > k$$

where the threshold k depends on the efficiency or purity you want to have on the selection



Example: a cut on $\frac{\mathcal{L}(m, p_T \mid S)}{\mathcal{L}(m, p_T \mid B)}$ is more powerful than the rectangular selection shown on the plot

When dealing with many correlated variables, modeling the likelihood is often not possible in practice (due to the limited statistics of calibration samples), and optimizing the classification requires the use of Machine Learning tools.

Confidence intervals

The result of an estimation is usually given as $\hat{x} \pm \sigma(\hat{x})$

But what this interval actually means in terms of probability?

In general, the result should be a confidence interval with a given confidence level α, meaning that we have a probability α that the true value of our parameter is within the interval.

(i.e. if I repeat the experiment many times, a fraction α of the results will include the true value)

In the Gaussian limit, the interval $\hat{x} \pm \sigma(\hat{x})$ corresponds to $\alpha = 68.3\%$, and the interval for any value of α will be

$$\hat{x} \pm n_\sigma \sigma(\hat{x}) \quad \text{with} \quad n_\sigma = P_G^{-1}((1+\alpha)/2)$$

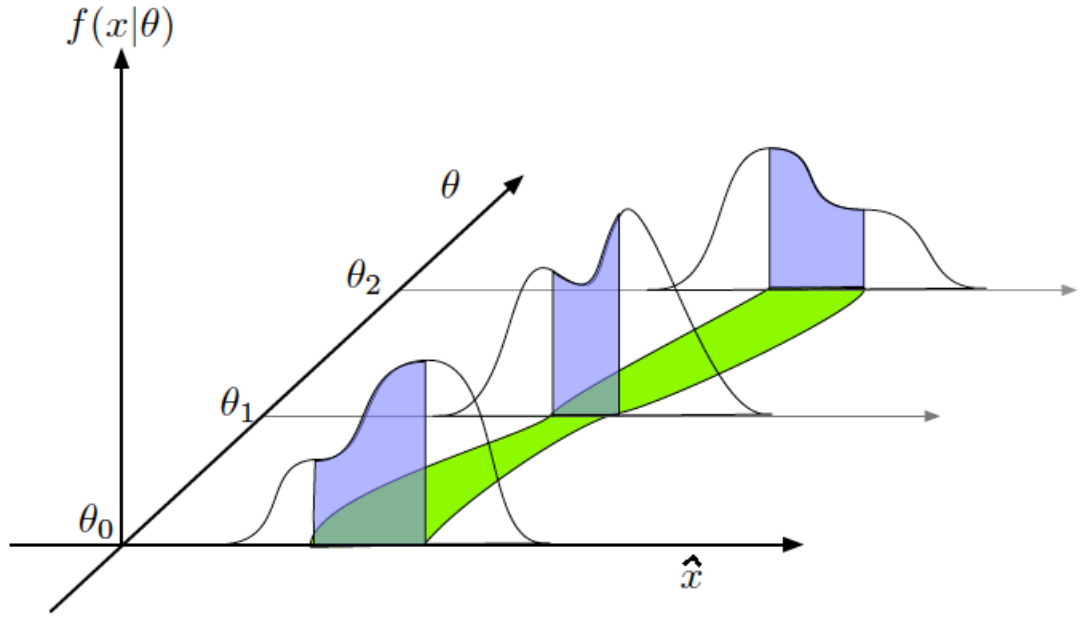
where $P_G(X) = \int_0^X \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right] dx$ is the cumulative "standard" Gaussian distribution ($\mu = 0, \sigma = 1$)

When far from the limit, we need to take into account the non-Gaussian distribution and the uncertainty on $\sigma(\hat{x})$.

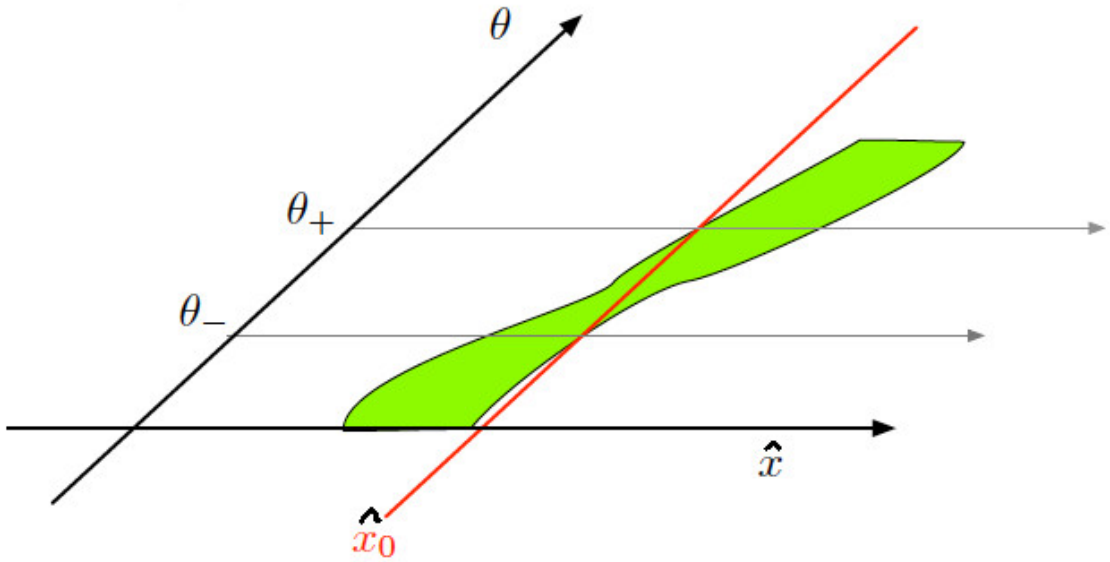
Example: consider a single observation t_{obs} of a decay time. The ML estimator of the lifetime is $\tau = t_{obs}$, but what is the 68% confidence interval?

The standard deviation of the estimator is also τ, but if we use the interval $\tau = t_{obs} \pm t_{obs}$ we are right only in 60.6% of cases
(exercise: prove it)

The general procedure for calculating accurate confidence intervals connsists in building a "confidence belt", determining an interval for the estimator values corresponding to probability α for each possible value of the parameter θ.

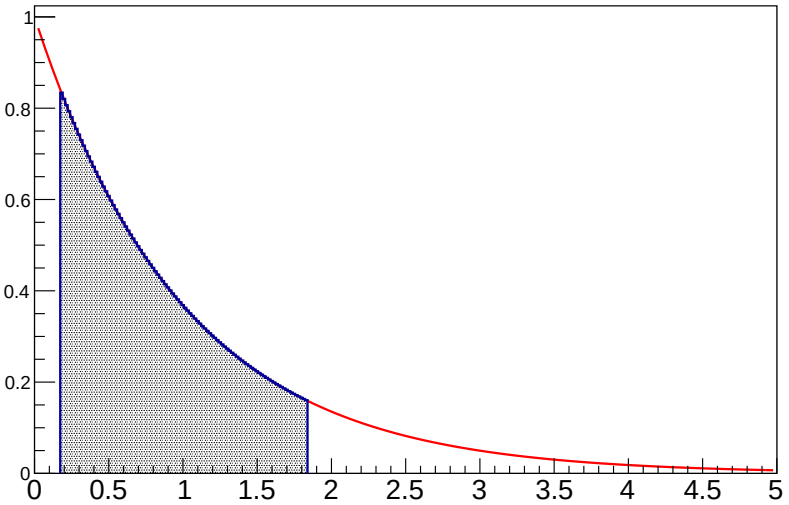


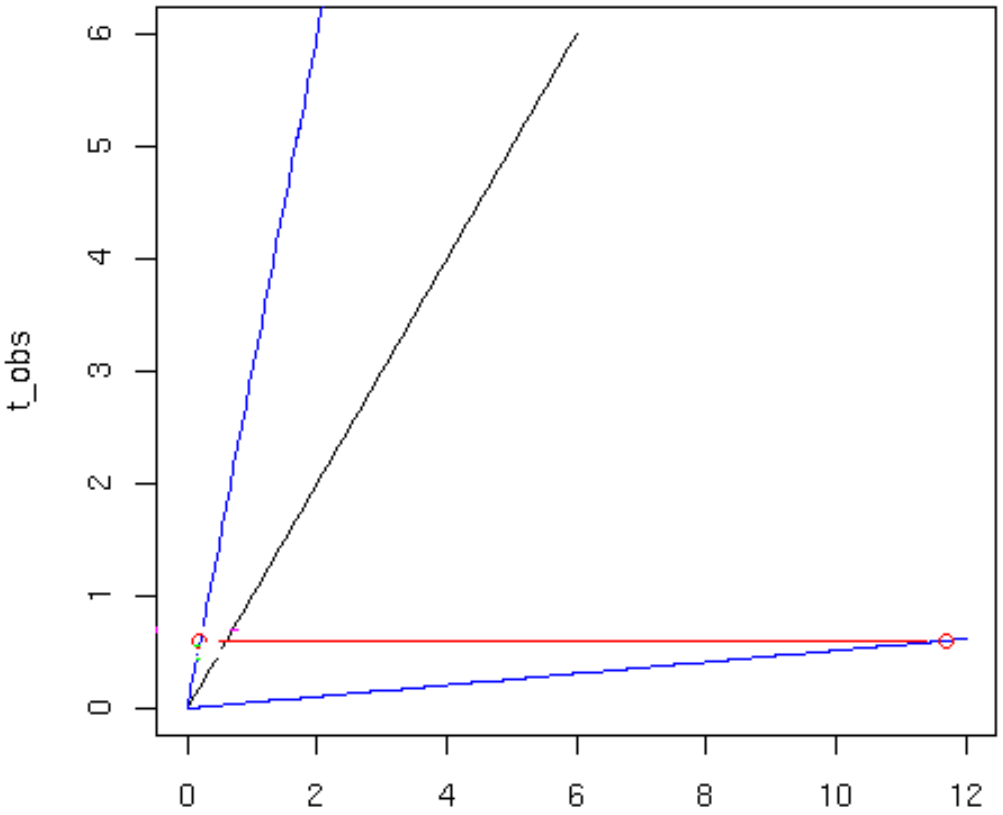
The interval in the belt corresponding to a given observation will contain the true value with probability α, independently on the value of θ



In our example of a single decay time variable, a "central" confidence belt corresponds to

$$-\tau log(1-\frac{1-\alpha}{2}) < t < -\tau log(1-\frac{1+\alpha}{2})$$





from which we get the confidence interval

$$-\frac{t_{obs}}{\log\left((1-\alpha)/2\right)} < \tau < -\frac{t_{obs}}{\log\left((1+\alpha)/2\right)}$$

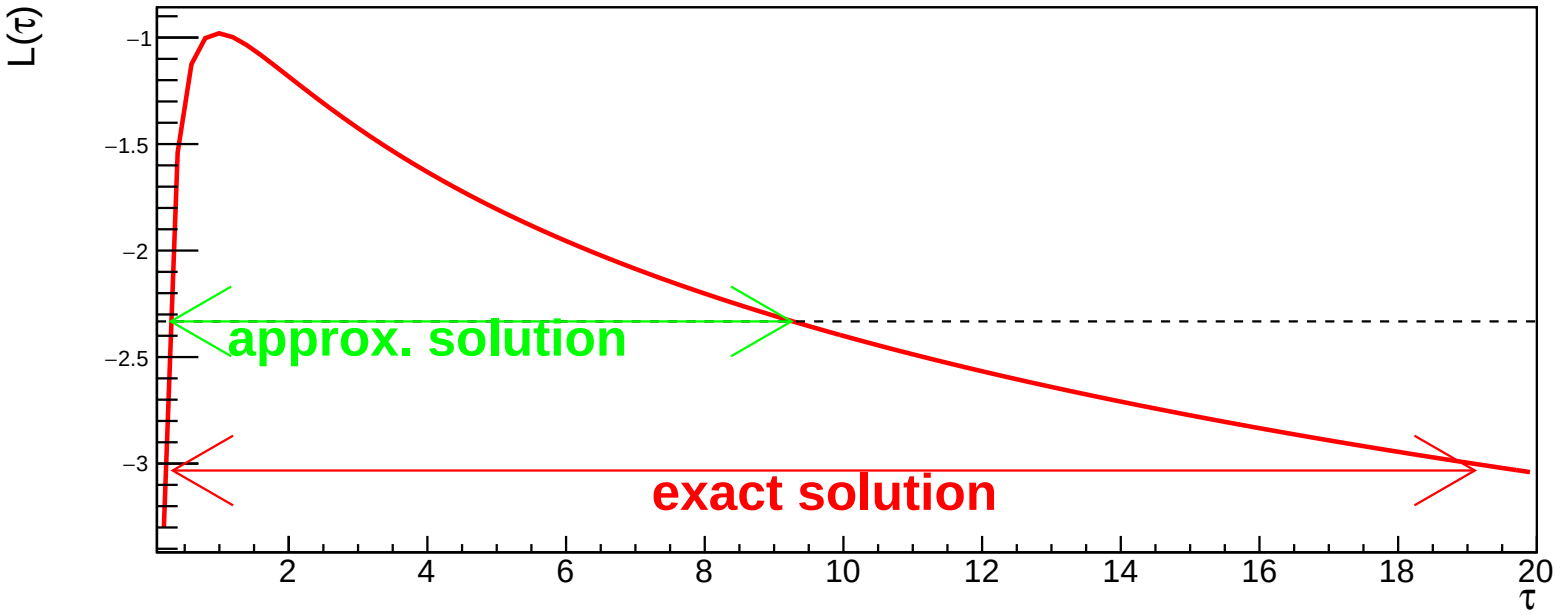
By choosing a non-central interval $t > K$ (or $t > K$) one can determine lower (or upper) limits for the parameter

The profile of the likelihood function, provides an approximate but faster way to estimate confidence intervals: even when far from the Gaussian limit, the interval corresponding to

$$\mathcal{L} > \max(\mathcal{L}) - n_\sigma(\alpha)^2/2$$

is a good proxy for the confidence interval corresponding to level α.
($n_\sigma = 1$ for $\alpha = 68.3\%, n_\sigma = 2$ for $\alpha = 95.5\%$, etc)

In our example, $\log \mathcal{L}(\tau) = -\log(\tau) - t_{obs}/\tau$
The approximate solution corresponding to $\alpha = 90\%$ ($n_\sigma = 1.64$) has an actual confidence of 87.5%.



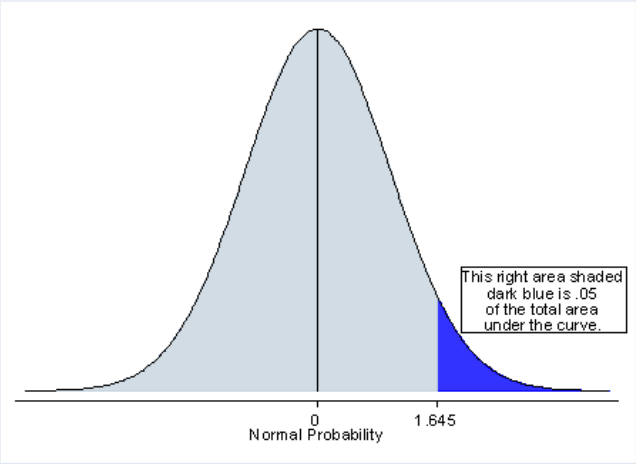
Significance

Understanding the "confidence" is particularly important for establishing a discovery.
A claim for discovery is usually based on the compatibility of the observation with the "null hypothesis" of no new signal, expressed with the p-value, namely the probability that the the data are less compatible with the null hypothesis than the observation. The confidence level for the discovery is thus $\alpha = 1 - p$
The significance is the equivalent number of Gaussian standard deviations

p-value	Z_{1t}	Z_{2t}
10%	1.28	1.64
5%	1.64	1.96
1%	2.33	2.58
0.13%	3	3.21
0.27%	2.78	3
$2.9 \cdot 10^{-7}$	5	5.13
$5.7 \cdot 10^{-7}$	4.86	5

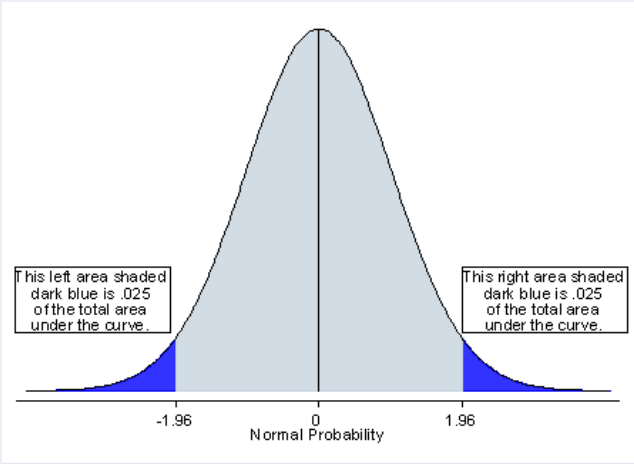
One-tailed case
(signal can appear only as an excess over background)

$$Z_{1t} = P_G^{-1}(1-p)$$



Two-tailed case
(signal can appear on both sides)

$$Z_{2t} = P_G^{-1}(1-p/2)$$



It is custumary in HEP to consider seriously ("evidence") significances above 3 σ, and to claim for discovery for $Z > 5$ ("observation")

As an example, consider a counting experiment with (known) b expected background events and n observed.

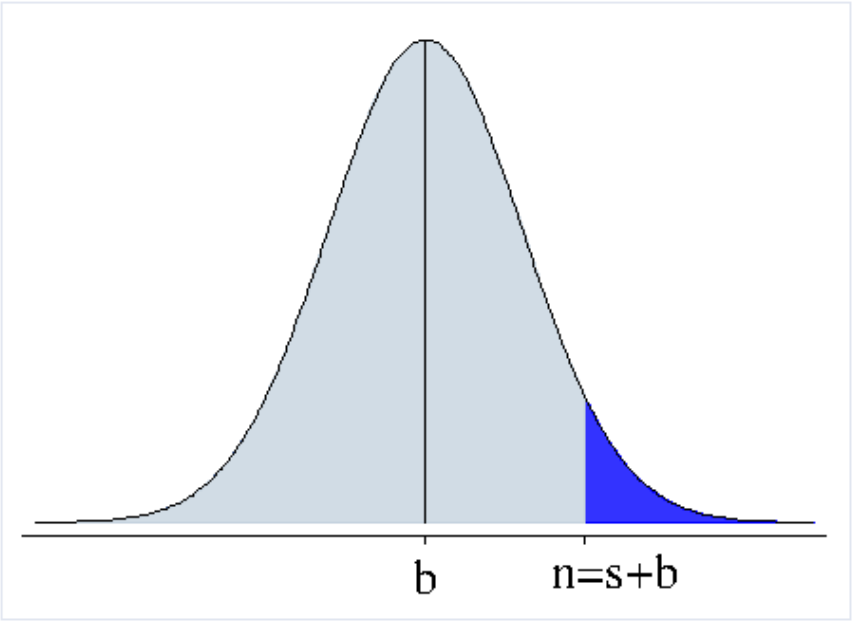
To look for a signal, we look for an excess of the number of counts with respect to b (one-tailed test):

$$s = n - b$$

A negative s can only be due to a downward background fluctuation, so one usually computes the significance only for $s > b$ (and sets $Z = 0$ otherwise).

In the limit where b is large enough so that its pdf can be approximated with a Gaussian with $\sigma = \sqrt{b}$ the significance is simply

$$Z = \frac{\max(s, 0)}{\sqrt{b}}$$



In a counting experiment with Poissonian statistics, one usually obtains the significance from the statistics

$$q_0 = \begin{cases} -2\,ln\left(\dfrac{\mathcal{L}(s=0)}{\mathcal{L}(s=\hat{s})}\right) = 2(n\,log(n/b) + b - n) & \hat{s} > 0 \\ 0 & \hat{s} < 0 \end{cases}$$

where $\hat{s} = n - b$ is the ML signal estimator and we used $\mathcal{L}(s) = \dfrac{(s+b)^n}{n!}\exp[-(s+b)]$

If statistics is reasonably large, the Wilk's theorem states that, in the null hypothesis of no signal, the likelihood ratio of nested hypothesis is approximatively distributed as a χ^2 with a number of degrees of freedom equal to the additional parameters (1 in this case: the signal count). In this limit the significance Z is simply $\sqrt{q_0}$:

$$Z \sim \begin{cases} \sqrt{2(nlog(n/b) + b - n)} & \hat{s} > 0 \\ 0 & \hat{s} < 0 \end{cases}$$

One often considers the median significance of an experiment for a given expected signal s. To a good approximation, it is equal to the significance computed for the data set where the signal count is set equal to the expected value s (the "Asimov data set"), thus:

$$med(Z \mid s) = \sqrt{q_A} = \sqrt{2((s+b)log(1+s/b) - s)}$$

These formula can be readily extended to the case of a set of counters $[n_i]$, and the $\sqrt{q_A}$ method can be used for any likelihood function

Some Bibliography

- G. Cowan, “Statistical Data Analysis”, Oxford University Press
- R. Barlow, “Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences”, Wiley
- S.Brandt, “Statistical and Computational Methods in Data Analysis”, North-Holland
  oldish, but still good for the classical theory
- I. Narsky and F. Porter, “Statistical Analysis Techniques in Particle Physics”, Wiley
  more recent and advanced, introduces Machine Learning techniques