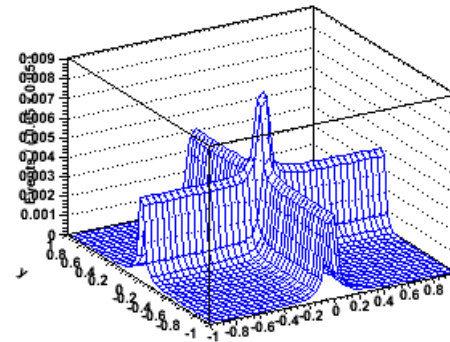


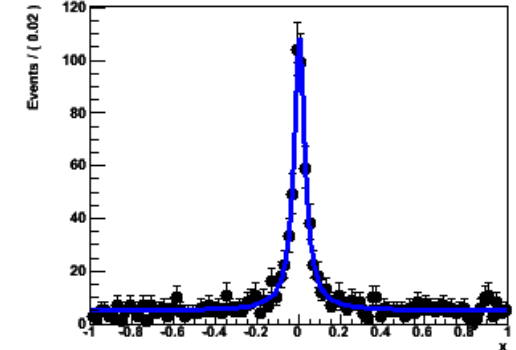
Analysis Methods

An experimentalist's view

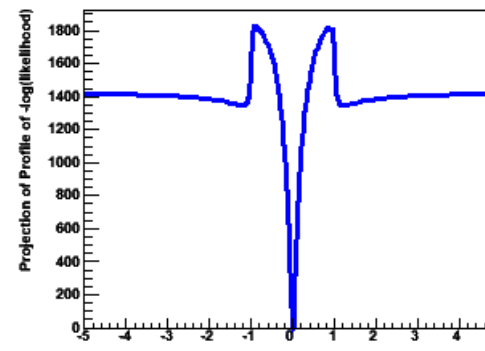
PDF model(x,y)



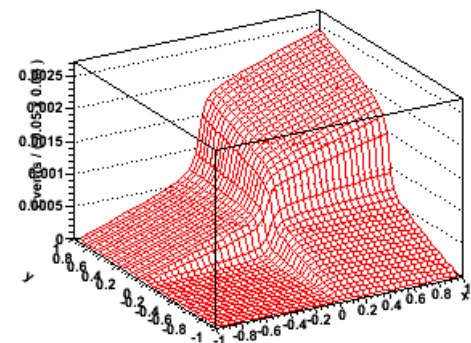
Data and p.d.f. projected on X



Profile likelihood in parameter a



CDF model(x,y)



C. Delaere
UCLouvain – CP3

Outline

- Probability and Statistics, basic concepts
- Monte Carlo techniques
- Event classification
- Parameter estimation
- Limits, confidence intervals, significance
- Closing remarks

Outline

- Probability and Statistics, basic concepts
 - Axioms, Frequentist vs Bayesian approaches
 - Mean, variance, covariance
 - (Some) Basic distributions
 - Central Limit Theorem & error propagation
- Monte Carlo techniques
- Event classification
- Parameter estimation
- Limits, confidence intervals, significance
- Closing remarks

- Particle Physics is all about matching experiment and theory
 - What Theory describes the Data ?
 - How Data can discriminate between Model X and Y ?
- Due to the intrinsic nature of the processes studied, the proper question is often:

Are theory and experiment statistically compatible ?



Probability and statistics are very hot topics, constantly improved.
In four hours, we will only scrape the surface !

- Probability: from theory to data
 - For a given model, what are the possible outcomes for experiments ? => predictions
- Statistics: from data to theory
 - This is „solving the inverse problem” : from a set of measurements infer the right model => experimental data analysis
- There are various ways to interpret probabilities.
 - Axiomatic
 - Frequentist
 - Objective probability
 - Bayesian probability

What probability is about ???

- Kolmogorov axioms

n.b. : other sets of axioms exist

- The probability of any event E in the event space F , $P(E)$ is a non-negative real number:

$$0 \leq P(E) \leq 1 \quad \forall E \in F$$

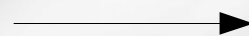
- The probability that some elementary event in the entire sample space will occur is 1.

$$P(\Omega) = 1 \text{ and } P(\emptyset) = 0.$$

- Any countable sequence of pairwise disjoint events E_i satisfies:

$$P(E_1 \cup E_2 \cup \dots) = \sum_i P(E_i).$$

While perfectly valid, that approach doesn't tell anything about what probability is.



More than a mathematical tool, we want to interpret the probabilities of physical quantities.

The Frequentist approach

- Natural definition of probability via the frequency of the corresponding event:
 - Let perform N times (identical trials) the same experiment
 - $$P(E) = \lim_{N \rightarrow \infty} \frac{k}{N}$$
- Pro: intuitive interpretation in particle physics.
- Con:
 - One cannot consider the event independently of the collective.
 - One cannot mathematically prove the convergence.
 - Not all measurement can be repeated under identical conditions.
 - ? probability that the top mass is in $[171.2, 174.0] \text{ GeV}/c^2$



The Bayesian approach

- Probability is seen as a **degree of belief**.
 - Credibility of a statement -> taking into account the past
- Bayesian approach is all about the probability of an hypothesis or theory.
- $P(E) = P(E|I)$ is the state of our knowledge and depends on the information we have.
- This is in opposition to the frequentist approach of probability $P(E)$ as a state of nature.
 - Physicists often claim „I am frequentist” -> well suited to describe quantum phenomena.
 - Bayesian approach more suited to the analysis of experimental outcome or prediction.
 - What is the „probability to reject a Higgs boson of 500GeV” ?
- Bayesian probability includes a PRIOR knowledge about the theory and tells us the influence of a new measurement.
 - Subjective probability ?

The Bayes theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

$$P(A|B) P(B) = P(A \cap B) = P(B|A) P(A).$$

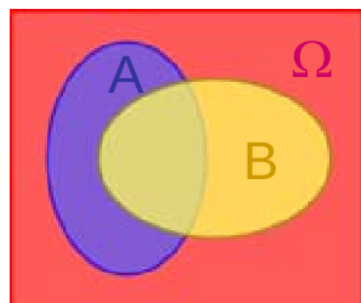
$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}.$$

This is the Bayes' Theorem:
it gives the probability for A to be true if B is true.



REV. T. BAYES

Bayes theorem



$$P(A) = \frac{\text{Area of } A}{\text{Area of } \Omega} \quad P(B) = \frac{\text{Area of } B}{\text{Area of } \Omega}$$

$$P(A|B) = \frac{\text{Area of } A \cap B}{\text{Area of } B} \quad P(B|A) = \frac{\text{Area of } A \cap B}{\text{Area of } A}$$

From a drawing
by B.Cousins and L.Lista

$$P(A|B) P(B) = \frac{\text{Area of } A \cap B}{\text{Area of } B} \times \frac{\text{Area of } B}{\text{Area of } \Omega} = \frac{\text{Area of } A \cap B}{\text{Area of } \Omega} = P(A \cap B)$$

$$P(B|A) P(A) = \frac{\text{Area of } A \cap B}{\text{Area of } A} \times \frac{\text{Area of } A}{\text{Area of } \Omega} = \frac{\text{Area of } A \cap B}{\text{Area of } \Omega} = P(A \cap B)$$

$$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$$

From a drawing by B.Cousins

Bayes - Interpretation

Likelihood

Probability to observe data
according to the theory

Prior

independent of the measurement

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory}) \times P(\text{theory})}{P(\text{data})}$$

Posteriori probability

probability of the theory to be true

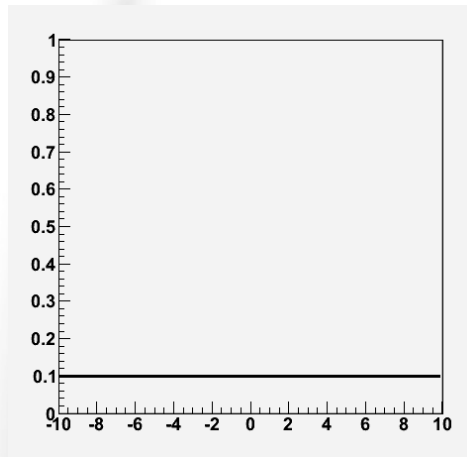
Evidence:

probability of data, assuming a model M.

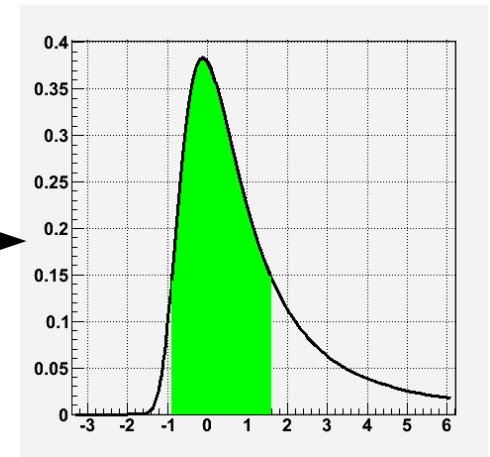
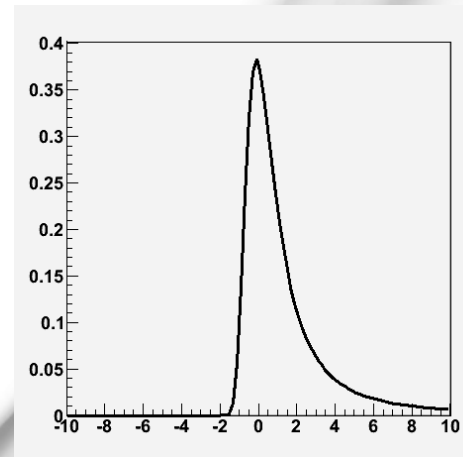
- To prove a theory, better have $P(\text{data}|\text{theory})$ large and $P(\text{data})$ small („strong evidence”)
- $P(\text{data}|\text{theory}) = 0 \rightarrow P(\text{theory}|\text{data})=0$: data allows then to reject the theory
- $P(\text{data})$ can be expressed as $\sum\{P(\text{data}|\text{theory}_i) \times P(\text{theory}_i)\} \rightarrow$ normalization factor.
- Learning interpretation: description of the evolution of $P(\text{theory})$ with new data.

Example

Measurement
from „flat prior” :

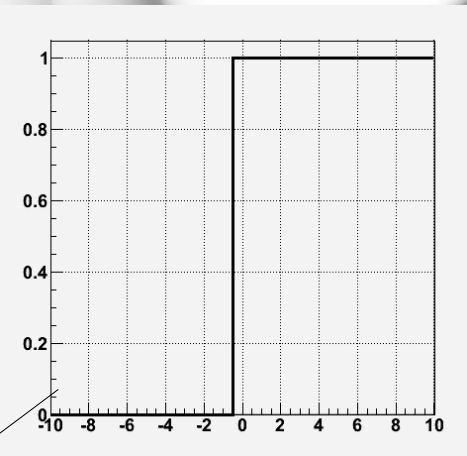


X

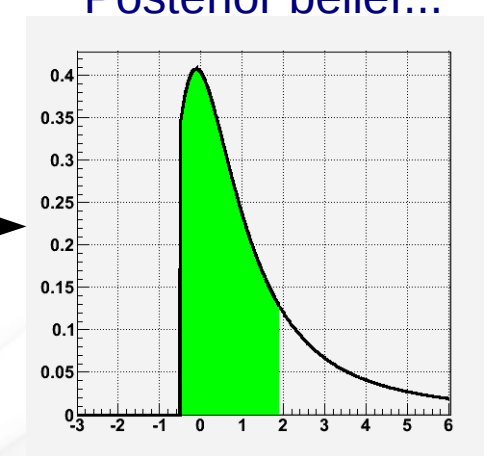
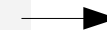
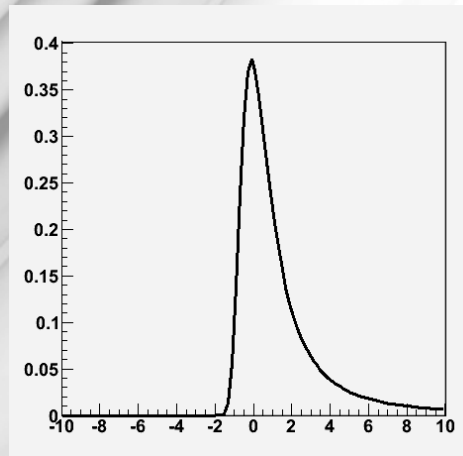


Posterior belief...

Given previous
measurement,
excluded region
put as prior :



X



Note that we could use
a more complex prior:

- Smoothed step function
- Gaussian to reflect existing measurement
- ...

Measurement: likelihood from fit.

We will come back on this later.

How to describe data ?

Quantitative measurements

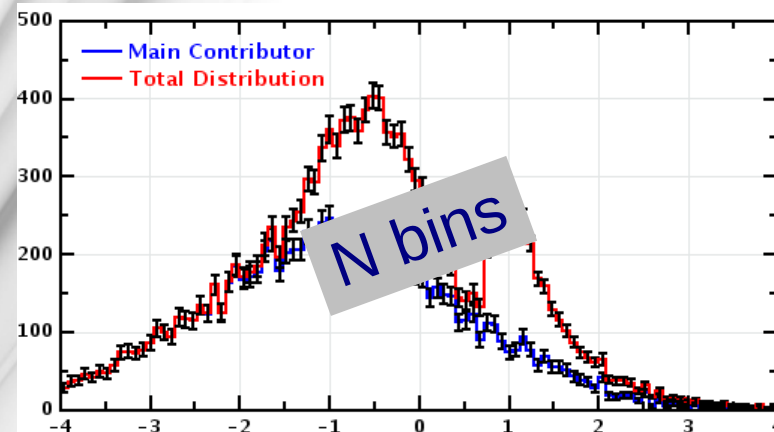
	D	E	F
1	1000.	10000.	100000.
2	2000.	20000.	200000.
3	3000.	30000.	300000.
4	4000.	40000.	400000.
5	5000.	50000.	500000.
6	6000.	60000.	600000.
7	7000.	70000.	700000.
8	8000.	80000.	800000.
9	9000.	90000.	900000.
10	10000.	100000.	1e+06.

N measurements

„n-tuples”
(unbinned data)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

Average



„histograms”
(binned data)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N n_i x_i$$

Spread

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

$$V(x) = \frac{1}{N} \sum_i (x_i^2 - 2x_i \bar{x} + \bar{x}^2)$$

$$V(x) = \frac{1}{N} \sum_i x_i^2 - \frac{1}{N} 2\bar{x} \sum_i x_i + \frac{1}{N} \bar{x}^2 N$$

$$V(x) = \bar{x^2} - \bar{x}^2$$

The **Variance** has the dimensions of *x squared*. On the contrary, the **Standard Deviation** has the *same dimension as x*.

$$\sigma = \sqrt{V(x)}$$

$$\sigma = \sqrt{\bar{x^2} - \bar{x}^2}$$

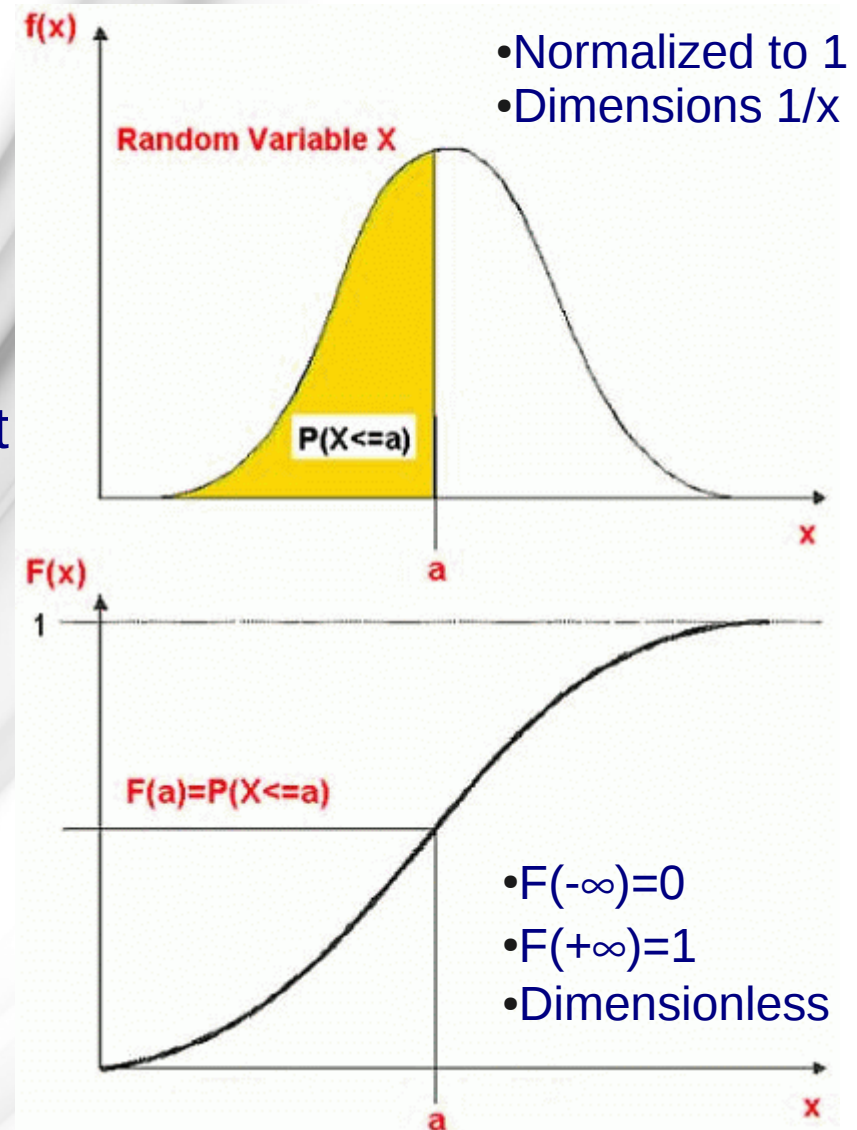
How to describe theory ?

- Definition: the probability to measure a value x in the interval $(x, x+dx)$ is given by **the probability density function $f(x)$** .
 - This is NOT a probability !
- Definition: **the cumulative distribution function $F(x)$** is the probability that we get a measurement smaller than x .

$$P(x_1 < w < x_2) = \int_{x_1}^{x_2} f(x') dx' = F(x_2) - F(x_1)$$

- The expectation value of a variable x is noted $E[x]$ or $\langle x \rangle$. For a given pdf, it is given by:

$$E[x] = \int x' f(x') dx' = \langle x \rangle$$



Estimation of mean and variance

- In general, the mean and variance of the „parent” pdf are unknown and have to be estimated.
- The law of large numbers relates the arithmetic mean of a data sample to the expectation value of the „parent” pdf:

$$\bar{x} \approx \langle x \rangle$$

- For n data points, we estimate the variance σ^2 by

- If the mean $\langle x \rangle := \mu$ is known :

$$\sigma^2 \equiv s^2 \equiv \frac{1}{N} \sum_i (x_i - \mu)^2$$

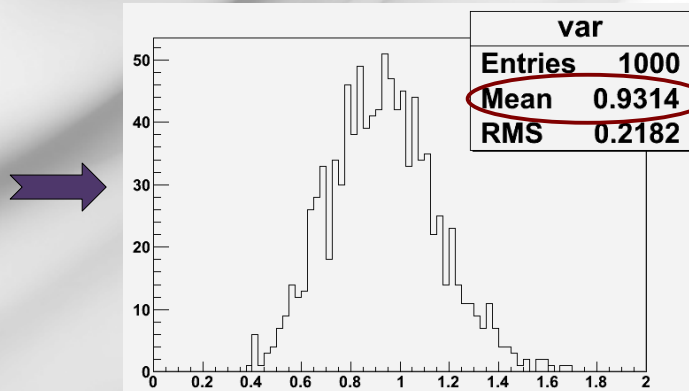
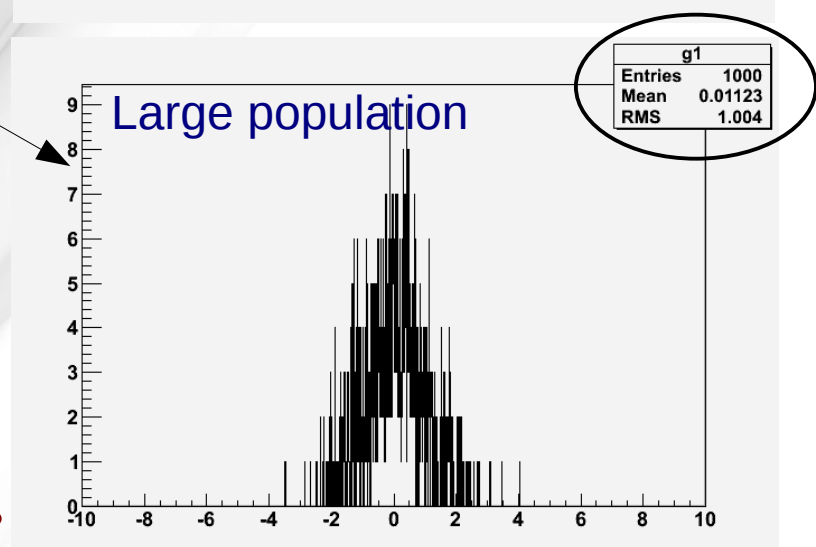
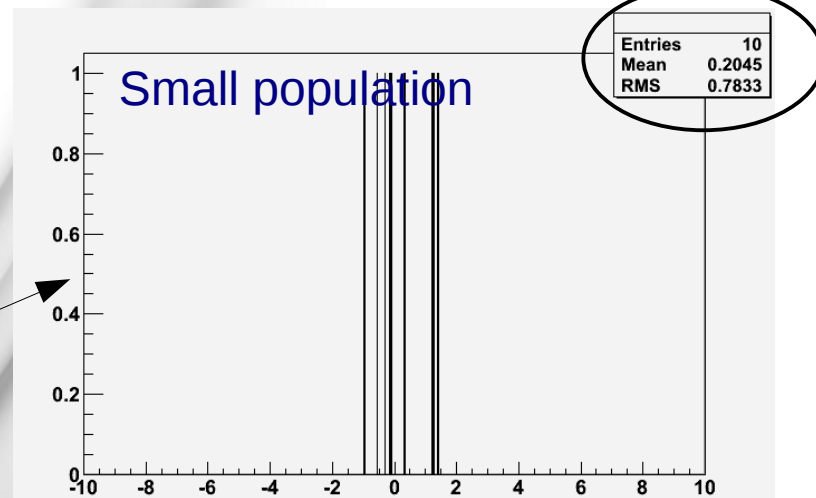
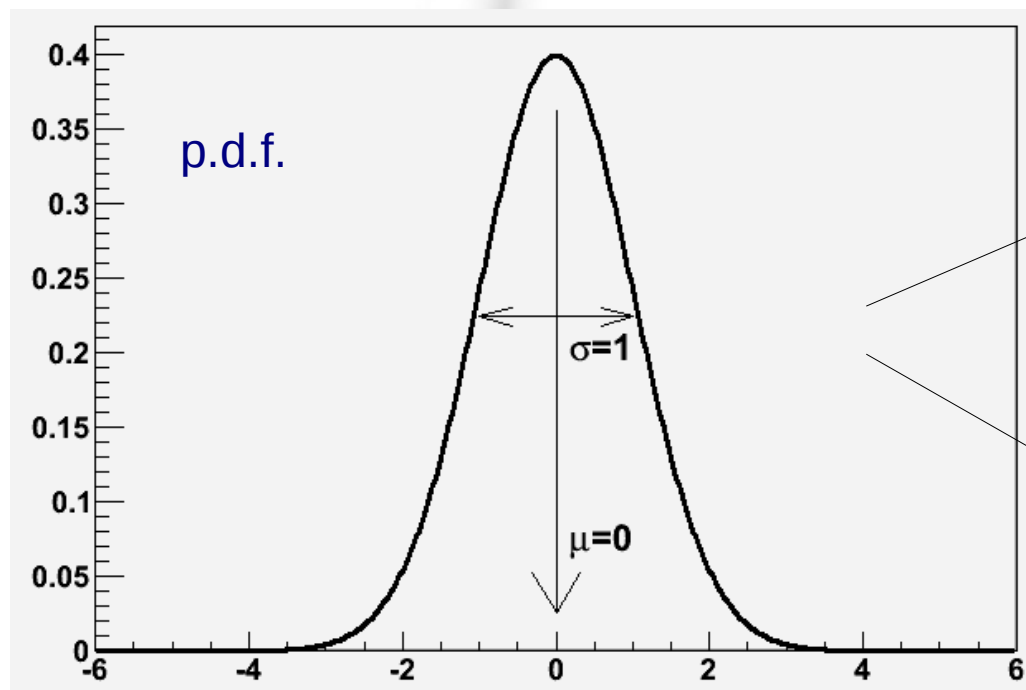
„ variance s^2 ”

- If the true mean μ is unknown :

$$s^2 \equiv \frac{1}{N-1} \sum_i (x_i - \bar{x})^2 \equiv \frac{1}{N-1} \left(\sum_i x_i^2 - \frac{1}{N} \left(\sum_i x_i \right)^2 \right)$$

„ sample variance s^2 ”

Illustration



Remember:
ROOT „RMS” is NOT
the sample variance !

Correlation, covariance

- Given two variables x, y , a dataset consists of pairs of numbers:

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

- The mean and variance for each variable are defined as usual.
- The **covariance** describes the dependence between x and y :

$$\text{cov}(x, y) = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{cov}(x, y) = (x - \bar{x})(y - \bar{y})$$

$$\text{cov}(x, y) = \bar{xy} - \bar{x} \bar{y}$$

- The dimensionless **correlation coefficient** is then defined as:

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

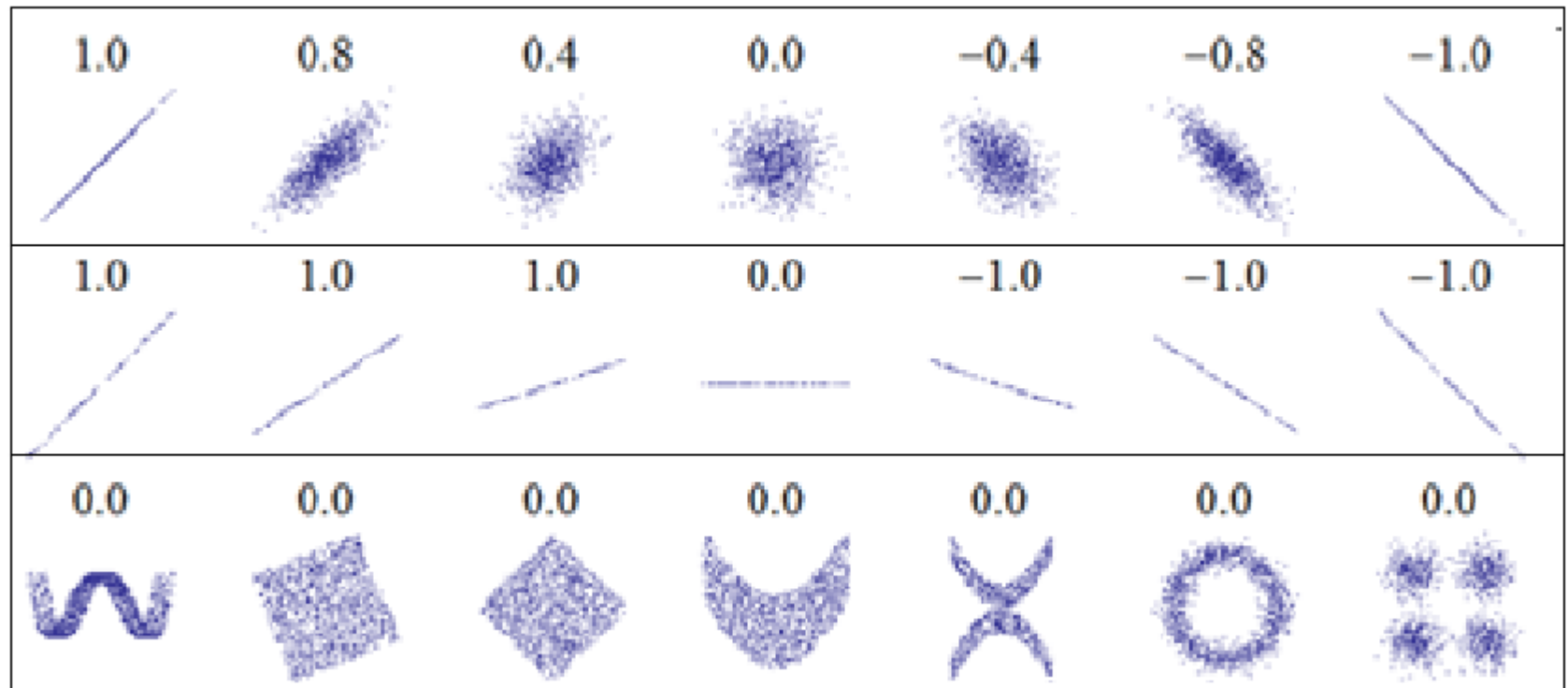
Correlation, covariance

- If the two variables are uncorrelated, $\rho=0$. The contrary is not true !

Reflects the direction of a linear relation.

Does NOT reflect the slope.

Does NOT reflect other non-linear relations



One defines the covariance matrix, often called „error matrix” :

$$\text{cov}(x, y) = \begin{pmatrix} \sigma_x^2 & V[xy] \\ V[xy] & \sigma_y^2 \end{pmatrix}$$

One can then generalize the discussion to more than two variables.
Let's denote n variables $x_{(i)}$, $i=1, \dots, n$

The covariance matrix is
 $n \times n$ symmetric:

$$V_{ij} = \text{cov}(x_{(i)}, x_{(j)}) = \langle x_{(i)} x_{(j)} \rangle - \langle x_{(i)} \rangle \langle x_{(j)} \rangle$$

Define the correlation matrix

$$\rho_{ij} = \frac{\text{cov}(x_{(i)}, x_{(j)})}{\sigma_i \sigma_j}$$

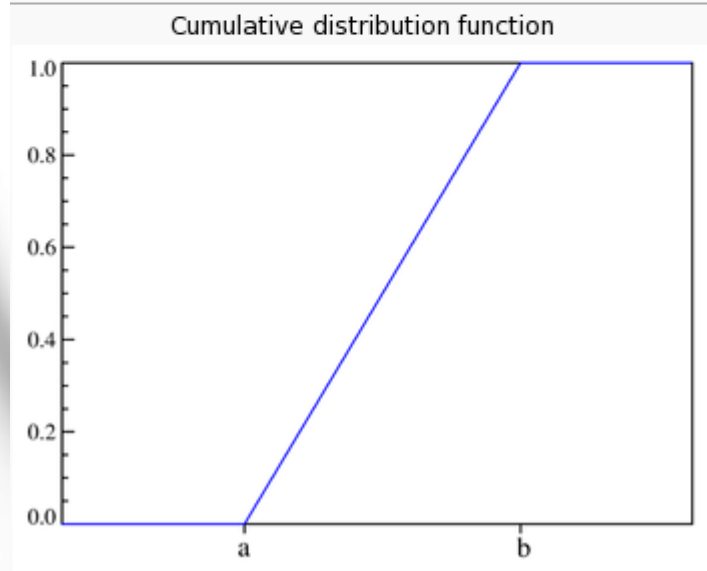
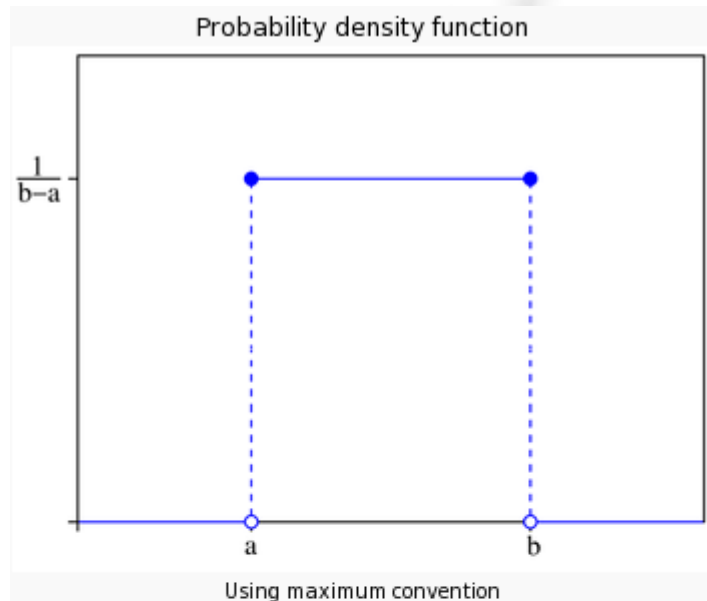
Redefine the error matrix

$$V_{ij} = \rho_{ij} \sigma_i \sigma_j$$

Some useful pdfs

- Uniform
- Binomial
- Gaussian
- Exponential
- Chi-square
- Breit-Wigner
- Landau

Uniform distribution

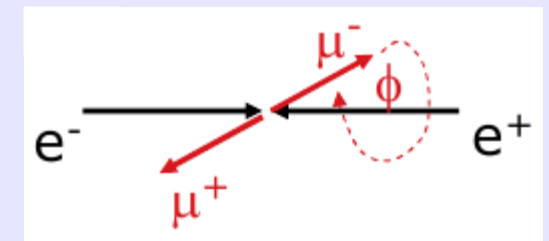
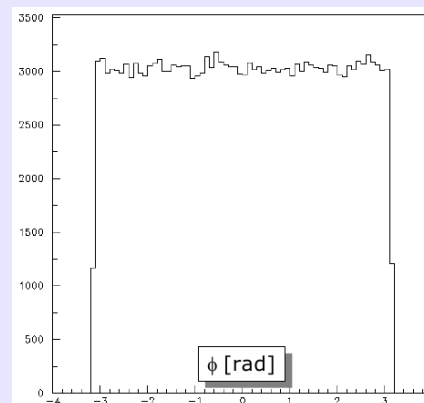


parameters:	$-\infty < a < b < \infty$
support:	$x \in [a, b]$
pdf:	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
cdf:	$\begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x \geq b \end{cases}$
mean:	$\frac{1}{2}(a + b)$
median:	$\frac{1}{2}(a + b)$
mode:	any value in $[a, b]$
variance:	$\frac{1}{12}(b - a)^2$

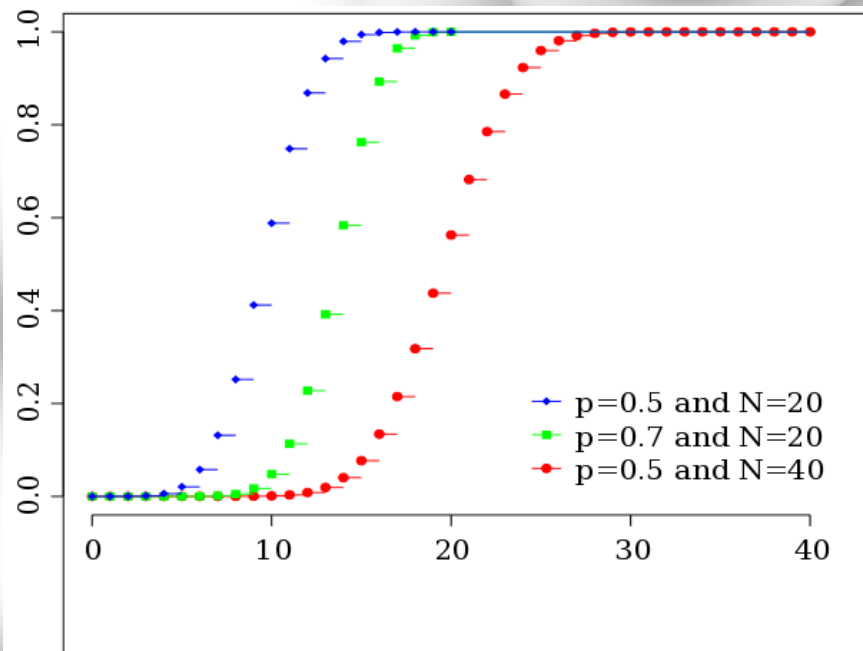
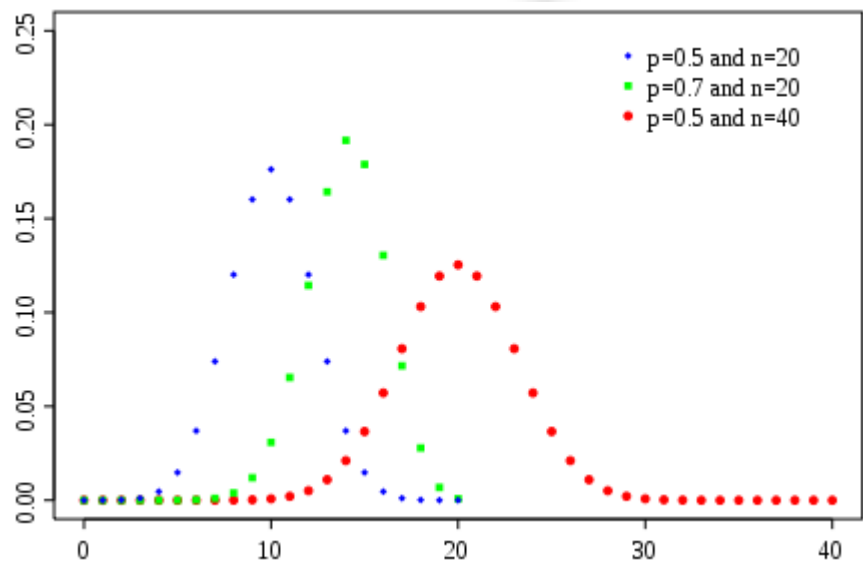


Resolution of
discrete
measurements

Example: phi distribution of muons
in Drell-Yan production.



Binomial distribution



N trials (independent processes) that can either succeed or fail.
 $B(n,p)$ represents the probability to have k successes among n.

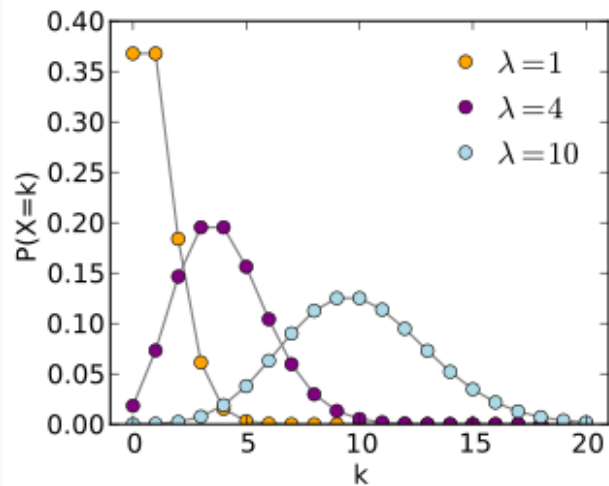
notation:	$B(n,p)$
parameters:	$n \in \mathbb{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
support:	$k \in \{0, \dots, n\}$
pmf:	$\binom{n}{k} p^k (1-p)^{n-k}$
cdf:	$I_{1-p}(n-k+1, k)$
mean:	np
median:	$\lfloor np \rfloor$ or $\lceil np \rceil$
mode:	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
variance:	$np(1-p)$

Reminder: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Example at LHC:
 8 colliding bunches in the machine
 $p(\text{interaction})=0.75$
 ? probability to have k collisions in the same „orbit” ?

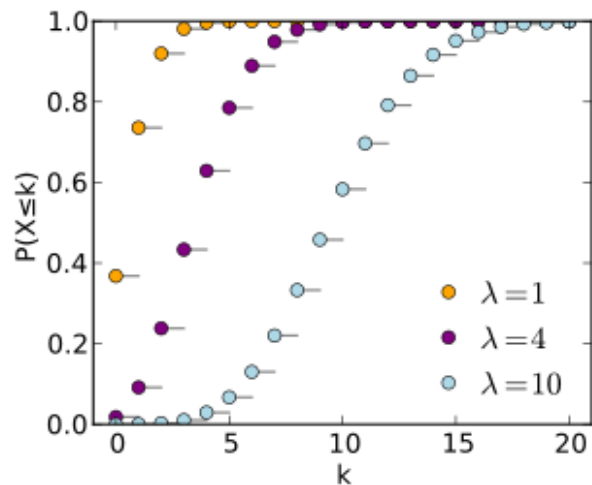
Poisson distribution

Probability mass function



The horizontal axis is the index k . The function is only defined at integer values of k . The connecting lines are only guides for the eye.

Cumulative distribution function



The horizontal axis is the index k . The CDF is discontinuous at the integers of k and flat everywhere else because a variable that is Poisson distributed only takes on integer values.

Poisson distribution describes cases of sharp events occurring in a continuum.

- The #trials is unknown
- The rate is known

It corresponds to $B(n \rightarrow \infty, p \rightarrow 0)$ with $np = \lambda$

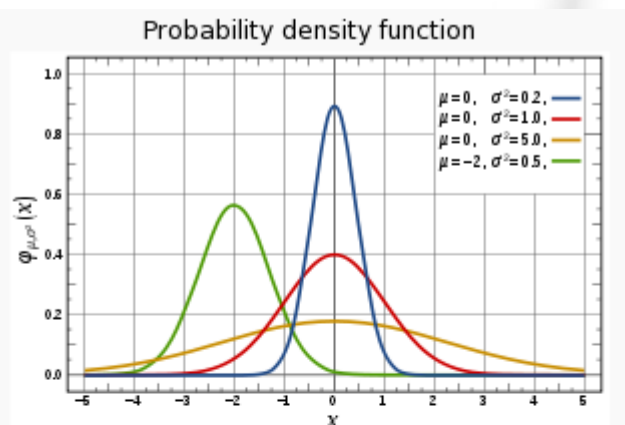
notation:	$Pois(\lambda)$
parameters:	$\lambda > 0$ (real)
support:	$k \in \{0, 1, 2, 3, \dots\}$
pmf:	$\frac{\lambda^k}{k!} \cdot e^{-\lambda}$
cdf:	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\Gamma(k+1, \lambda)}$ for $k \geq 0$ or $e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$ <p>(where $\Gamma(x, y)$ is the Incomplete gamma function and $\lfloor k \rfloor$ is the floor function)</p>
mean:	λ
median:	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
mode:	$\lfloor \lambda \rfloor$, and $\lambda - 1$ if λ is an integer
variance:	λ

$$\Delta N \sim \sqrt{N}$$

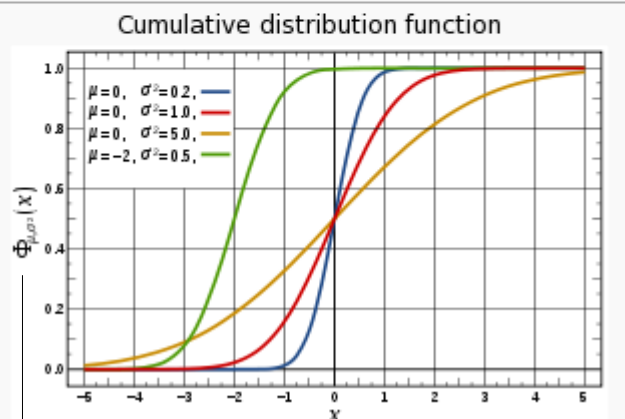
Examples:

- **Radioactive decays**
- **Observation of rare processes (Higgs decays ?)**

Gaussian distribution



The red line is the standard normal distribution



Colors match the image above

Gaussian p.d.f., or Normal p.d.f.

For $\mu=0$, $\sigma=1$, one obtains the Standard distribution.

Properties:

- Symmetric around μ
- σ characterises the width
- FWHM = $2\sigma \sqrt{2 \ln 2} = 2.355 \sigma$

The error function
being defined as:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

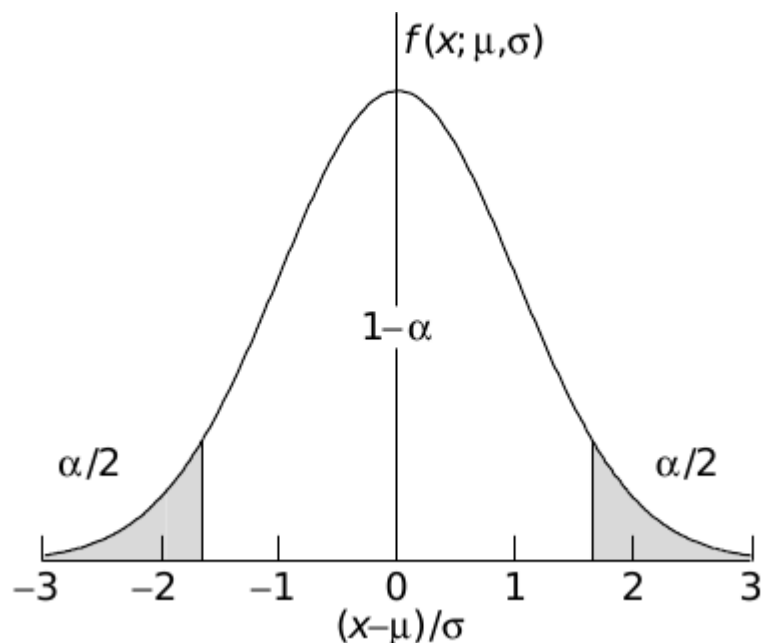
one has:

$$\Phi(x) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{x - \mu}{\sqrt{2}\sigma} \right) \right)$$

notation:	$\mathcal{N}(\mu, \sigma^2)$
parameters:	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 \geq 0$ — variance (squared scale)
support:	$x \in \mathbf{R}$ if $\sigma^2 > 0$ $x = \mu$ if $\sigma^2 = 0$
pdf:	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
cdf:	$\frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x - \mu}{\sqrt{2}\sigma} \right) \right]$
mean:	μ
median:	μ
mode:	μ
variance:	σ^2

Gaussian properties

$$1 - \alpha = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2} dx = \operatorname{erf}\left(\frac{\delta}{\sqrt{2}\sigma}\right)$$



**1/3 of the measurements
lie out of the 1σ band !**

If we do 200 measurements, the probability
to observe a 3σ effect is : $1 - 0.9987^{200} = 0.23$

When quoting the magnitude of an excess,
deviation or probability, it is usual to use
Gaussian quantiles.

Two conventions can be adopted:

- One sided
- Double sided

The difference is a factor 2

-> always precise the convention !

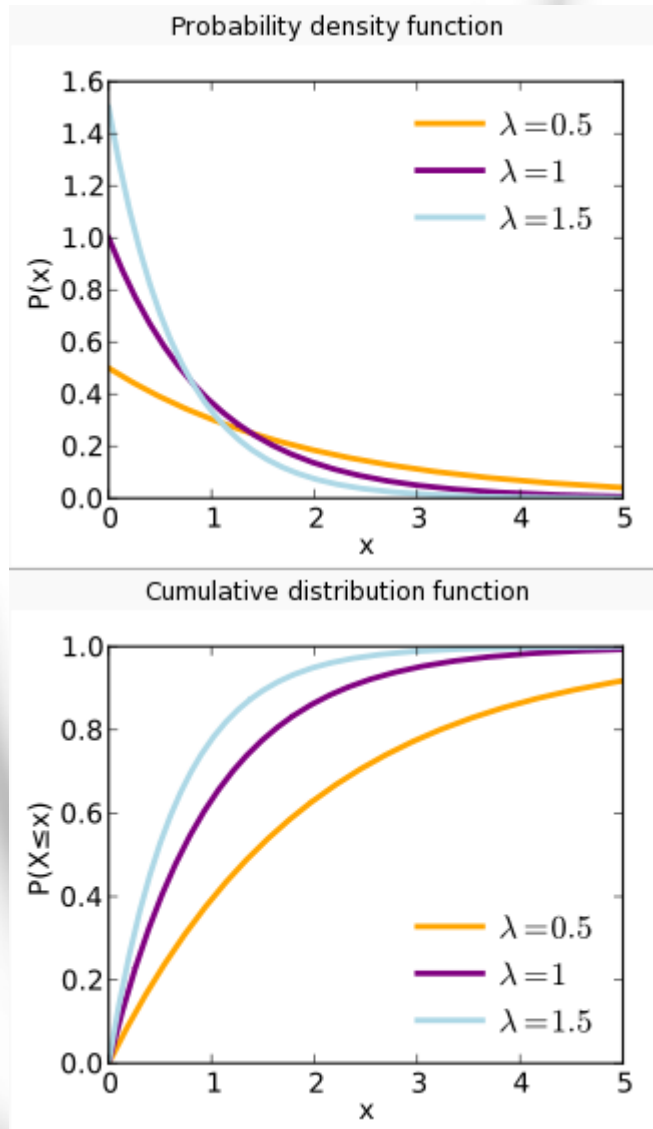
Double-sided convention:

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
4.55×10^{-2}	2σ	0.1	1.64σ
2.7×10^{-3}	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	10^{-4}	3.89σ

Beware of the Look Elsewhere Effect !

Exponential distribution

- Describes the lifetime of non-aging particles
 - No history, no aging.
 - Decay probability = constant
- Applies to particles physics (quantum physics)



parameters:	$\lambda > 0$ rate or inverse scale (real)
support:	$[0, \infty)$
pdf:	$\lambda e^{-\lambda x}$
cdf:	$1 - e^{-\lambda x}$
mean:	$\frac{1}{\lambda}$
median:	$\frac{\ln(2)}{\lambda}$
mode:	0
variance:	$\frac{1}{\lambda^2}$

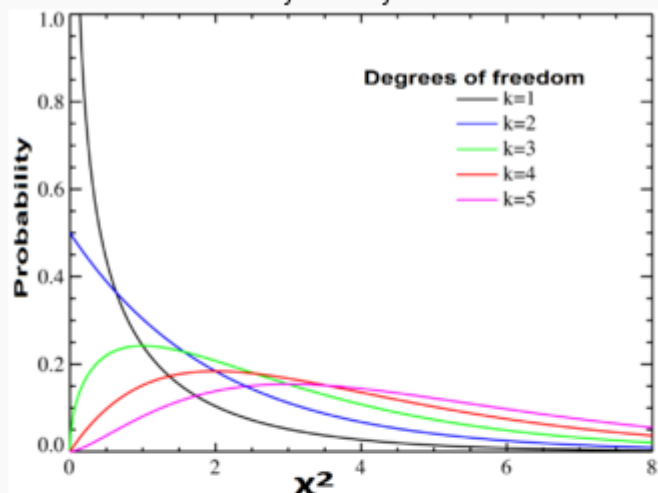
Chi-square distribution

- Arises in the context of the method of least-squares
- If $x_1 \dots x_n$ are n independent, Gaussian distributed variables, then quantity

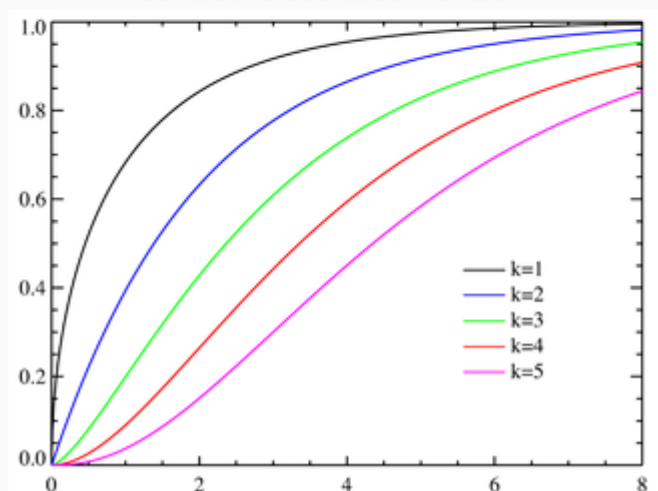
$$K = \sum_i \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

is distributed according to a χ^2 distribution.

Probability density function



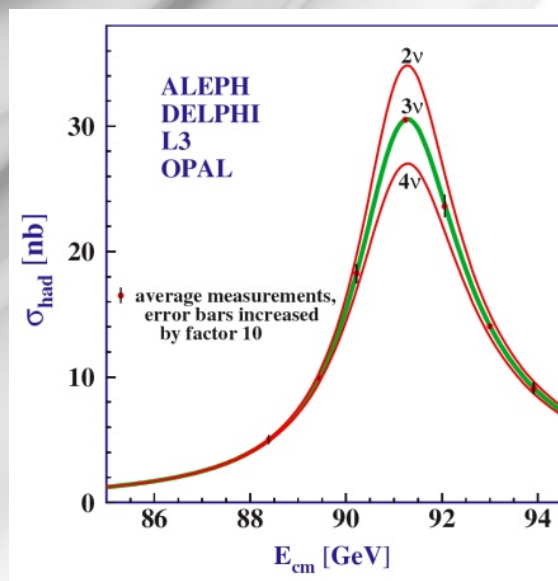
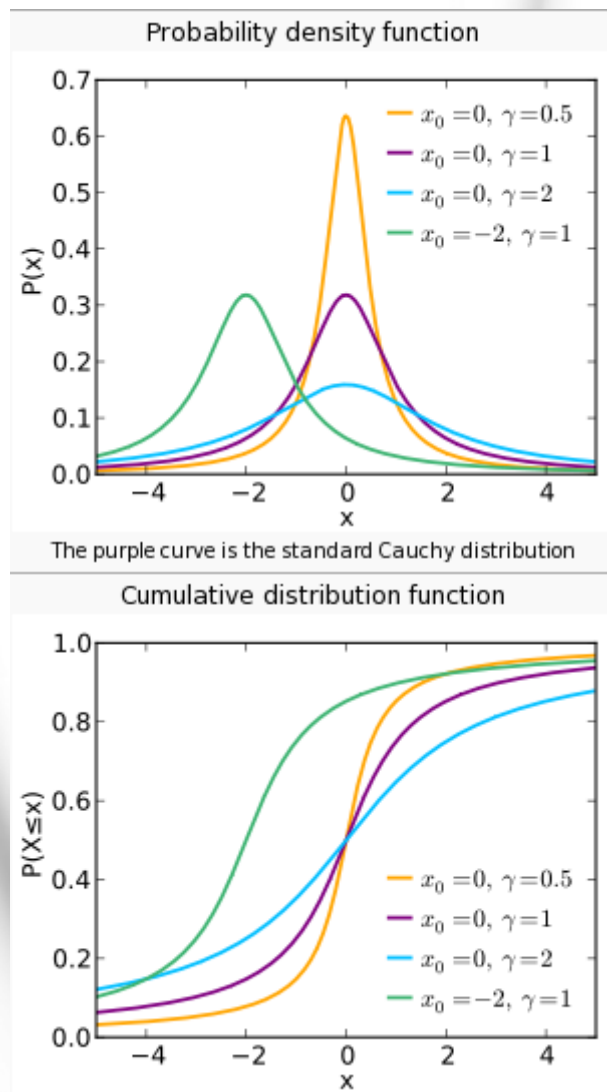
Cumulative distribution function



notation:	$\chi^2(k)$ or χ_k^2
parameters:	$k \in \mathbf{N}_1$ — degrees of freedom
support:	$x \in [0, +\infty)$
pdf:	$\frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2}$
cdf:	$\frac{1}{\Gamma(k/2)} \gamma(k/2, x/2)$
mean:	k
median:	$\approx k \left(1 - \frac{2}{9k}\right)^3$
mode:	$\max\{k - 2, 0\}$
variance:	$2k$

Breit-Wigner distribution

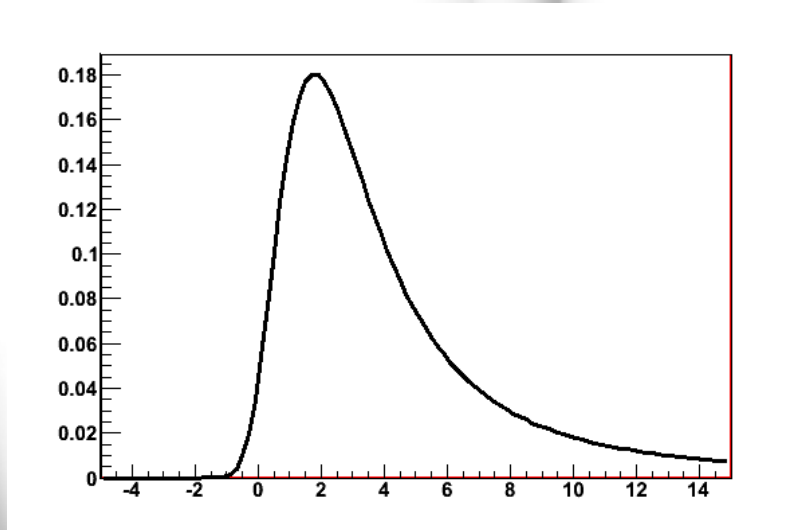
- Arises naturally from the propagator of a massive particle in QFT.
- Few unusual properties
 - μ undefined : use median/mode x_0
 - σ undefined : use $\text{HWHM}=\gamma$



parameters:	x_0 location (real) $\gamma > 0$ scale (real)
support:	$x \in (-\infty; +\infty)$
pdf:	$\frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma} \right)^2 \right]}$
cdf:	$\frac{1}{\pi} \arctan \left(\frac{x-x_0}{\gamma} \right) + \frac{1}{2}$
mean:	not defined
median:	x_0
mode:	x_0
variance:	not defined

Landau distribution

- The Landau distribution is used to describe the distribution of energy loss of a charged particle passing through a thin layer of matter.



Valid for **thin** sensors ($t \rightarrow 0$)

$$p(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{s \log s + xs} ds,$$

Thickness

Energy loss

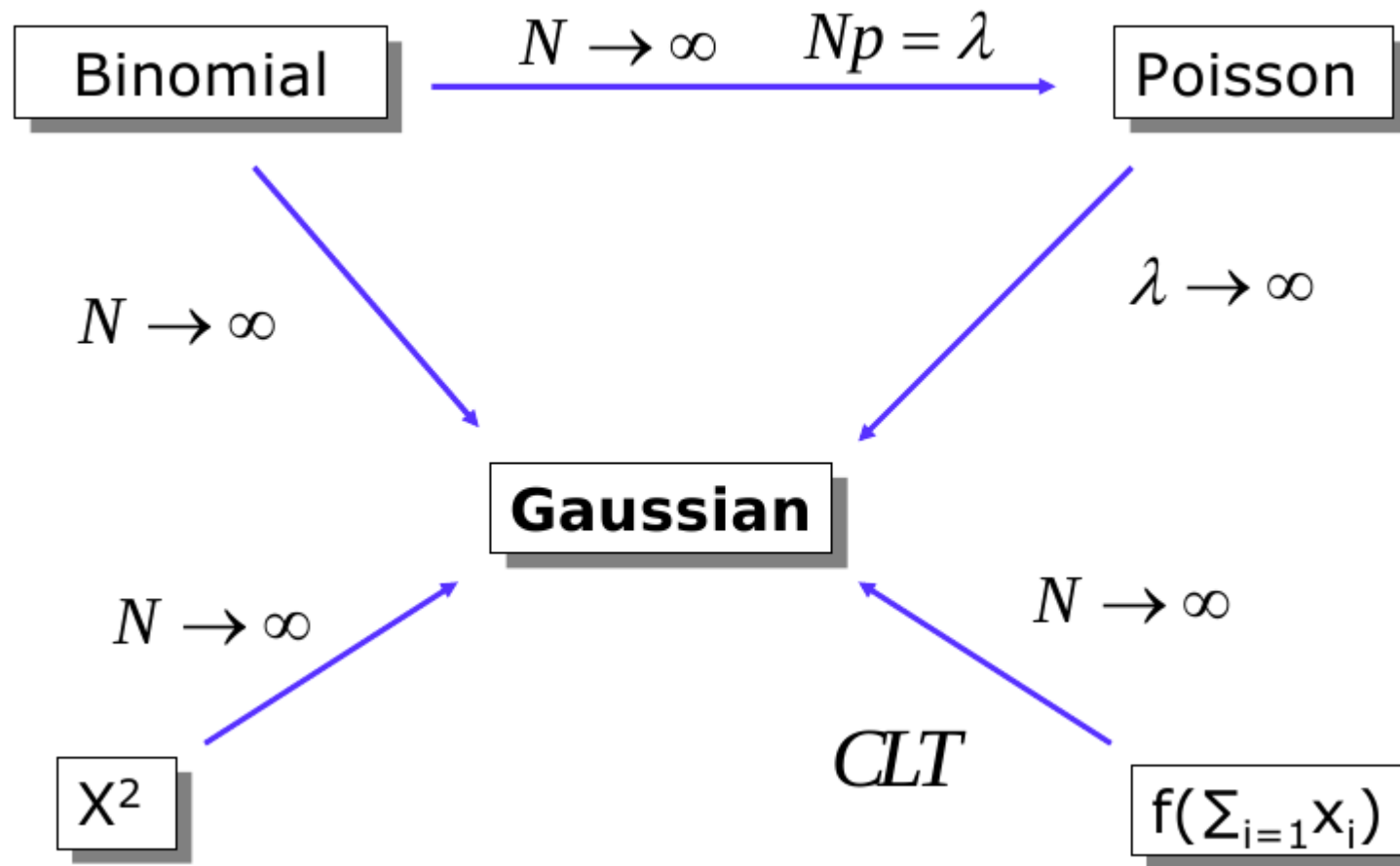
$$p(x) = \frac{1}{\pi} \int_0^\infty e^{-t \log t - xt} \sin(\pi t) dt.$$

$$\text{with } x = R (E - E_p)$$

Constant depending on
the absorber

Most probable Eloss

Central limit theorem



- Repeated measurements -> slightly different results each time
 - (changing conditions, resolution, quantum fluctuations, ...)
- **Statistical errors**
 - From frequentist definition of probability: repeated measurements give a distribution of probability for the result.
 - Quote the „spread” in addition to the central value.
- **Systematic errors (aka systematics)**
 - Uncertainty in estimating effects from systematic mistakes or from neglecting systematic mistakes.
 - Wrong method, instrument, formulae, calibration, ...

Errors

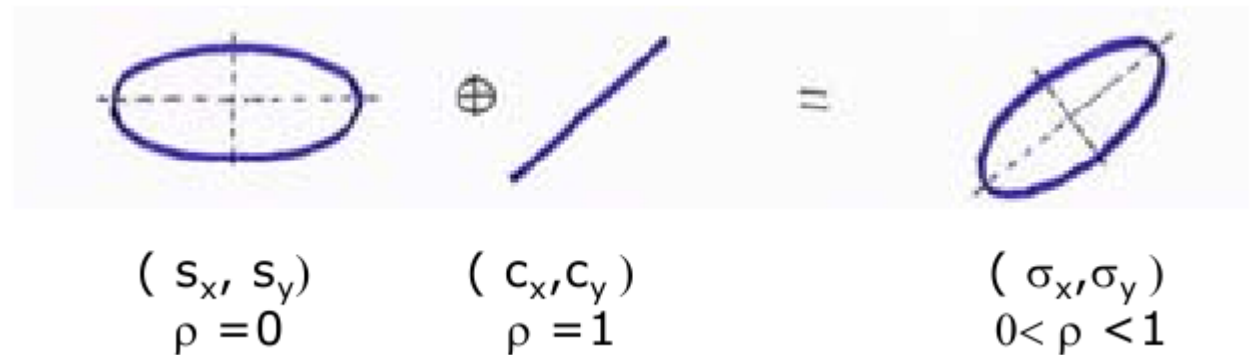
- Statistical and systematic errors will evolve differently when more data is accumulated
 - Have to be quoted separately
 - Still, can be added in quadrature, but systematics often introduce correlations among variables.

$$V_{ij}^{tot} = \begin{pmatrix} s_x^2 & 0 \\ 0 & s_y^2 \end{pmatrix} + \begin{pmatrix} c_x^2 & c_{xy} \\ c_{yx} & c_y^2 \end{pmatrix} \triangleq \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

error matrix
squared

- With the correlation coefficient :

$$\rho = \frac{c_{xy}}{\sigma_x \sigma_y}$$



Evaluating systematics

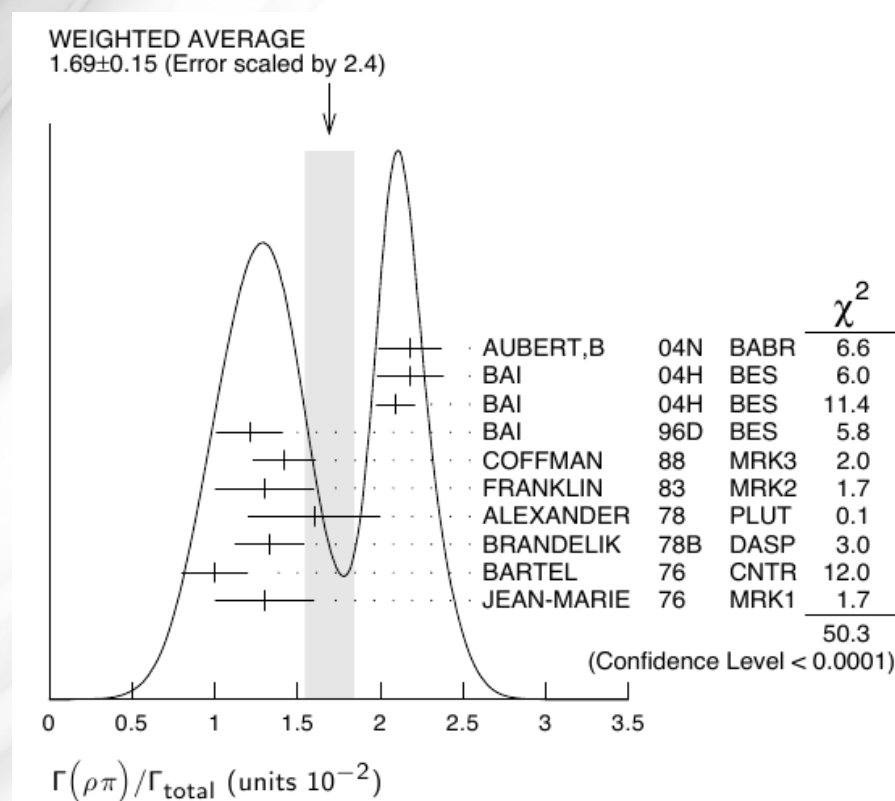
- There is nothing to be gained repeating your mistakes.
- An experiment with large systematics can look perfectly healthy, and the result is rubbish.
- Often, there are checks you can do to satisfy yourself there are no systematics
 - Be ingenuous
 - Be mildly paranoid
 - Ask a colleague talented for destructive criticism
- Evaluation of systematics may/may not be easy to evaluate.
E.g. :
 - calibration uncertainty,
 - theory error,
 - or even unknown cause !

Sanity checks

- Sanity checks are the key to evaluate systematics
- A successful check will **NOT** lead to systematics
- A sanity check is not an evaluation of systematics (should be decided beforehand)
 - A sanity check fails only in cases of mistakes
 - If the outcome is legitimately different from zero, it is a systematic uncertainty evaluation.
- If the alternate approach is better, don't use it to estimate uncertainties... just use it !
 - e.g. : If you find out there should be a 1.05 calibration factor, don't quote 5% uncertainty but use the calibration factor.
- More generally: when an effect is observed, first try to suppress/mitigate it, add a systematic only **in last resort.**

Incompatible measurements

- What to do when two measurements are incompatible ?
 - Taking the weighted mean + rms would not make sense.
 - Need special treatment.
- PDG recipe:
 - Calculate the weighted mean of measurements
 - Compute the global χ^2 w.r.t. that mean.
 - 3 cases:
 - $\chi^2/(n-1) \sim 1$: use weighted mean
 - $\chi^2/(n-1) \gg 1$: see case by case
 - $\chi^2/(n-1) > 1$: rescale errors by $\sqrt{\chi^2/(n-1)}$.



Error propagation

CLT -> errors can be treated as Gaussian in most of the cases.

Let's consider $f(x) = ax+b$

How do we compute $V(f)$ from $V(x)$?

$$V(f) = \langle f^2 \rangle - \langle f \rangle^2 = \langle (ax+b)^2 \rangle - \langle ax+b \rangle^2$$

$$V(f) = a^2 \langle x^2 \rangle + 2ab \langle x \rangle + b^2 - a^2 \langle x \rangle^2 - 2ab \langle x \rangle - b^2$$

$$V(f) = a^2 V(x) \quad \Rightarrow \sigma_f = a \sigma_x$$

1 variable

More generally, if f is locally linear, $V(f) = \left(\frac{df}{dx} \right)^2 V(x)$

Let's consider $f(x,y) = ax+by+c$

How do we compute $V(f)$?

$$V(f) = a^2 (\langle x^2 \rangle - \langle x \rangle^2) + b^2 (\langle y^2 \rangle - \langle y \rangle^2) + 2ab (\langle xy \rangle - \langle x \rangle \langle y \rangle)$$

$$V(f) = a^2 V(x) + b^2 V(y) + 2ab \text{cov}(x, y)$$

2 variables

More generally, if f is locally linear, $V(f) = \left(\frac{df}{dx} \right)^2 V(x) + \left(\frac{df}{dy} \right)^2 V(y) + 2 \left(\frac{df}{dx} \right) \left(\frac{df}{dy} \right) \text{cov}(x, y)$

$$\sigma_f^2 = \left(\frac{df}{dx} \right)^2 \sigma_x^2 + \left(\frac{df}{dy} \right)^2 \sigma_y^2 + 2 \left(\frac{df}{dx} \right) \left(\frac{df}{dy} \right) \rho \sigma_x \sigma_y$$

Addition of errors in quadrature, valid if uncorrelated ($\rho=0$).

Error propagation (2)

- For an arbitrary number of variables, the previous result can be generalised as:

$$\sigma_f^2 = \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i} \right)^2 \cdot \sigma_{x_i}^2 + \sum_{i=1}^n \sum_{j \neq i}^n \left(\frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \right) \cdot \text{cov}(x_i, x_j)$$

Normal errors (for
uncorrelated variables).

Additional terms accounting
for correlations

This can be even further expended to m functions of n variables, introducing the (symmetric) covariance matrix for functions, U.

$$U_{kl} \approx \text{cov}(f_k, f_l) = \sum_{i,j} \left(\frac{\partial f_k}{\partial x_i} \frac{\partial f_l}{\partial x_j} \right)_{x=\mu} \text{cov}(x_i, x_j)$$

By defining the nxm matrix of derivatives A,
one can write shortly $U = A V A^T$.

$$A_{ij} = \left(\frac{\partial f_i}{\partial x_j} \right)_{x=\mu}$$

- Few more useful formulas:

Let's consider $f(x) = xy$
How do we compute $V(f)$?

$$V(f) = \langle y \rangle^2 V(x) + \langle x \rangle^2 V(y) + V(x)V(y) \quad (x, y \text{ uncorrelated})$$

$$\left(\frac{\sigma_f}{f} \right)^2 \simeq \left(\frac{\sigma_x}{x} \right)^2 + \left(\frac{\sigma_y}{y} \right)^2 \quad \longrightarrow \quad \text{Add relative errors in quadrature, if } x, y \text{ are uncorrelated and relative errors are small}$$

Other useful formulas :

$$\frac{\sigma_{1/x}}{1/x} = \frac{\sigma_x}{x}$$

$$\sigma_{\ln x} = \frac{\sigma_x}{x}$$

Outline

- Probability and Statistics, basic concepts
- **Monte Carlo techniques**
 - Types of Monte Carlo generators
 - Flat Random number generators
 - The Inverse Method
 - The Rejection Method
 - General purpose MC
- Event classification
- Parameter estimation
- Limits, confidence intervals, significance
- Closing remarks

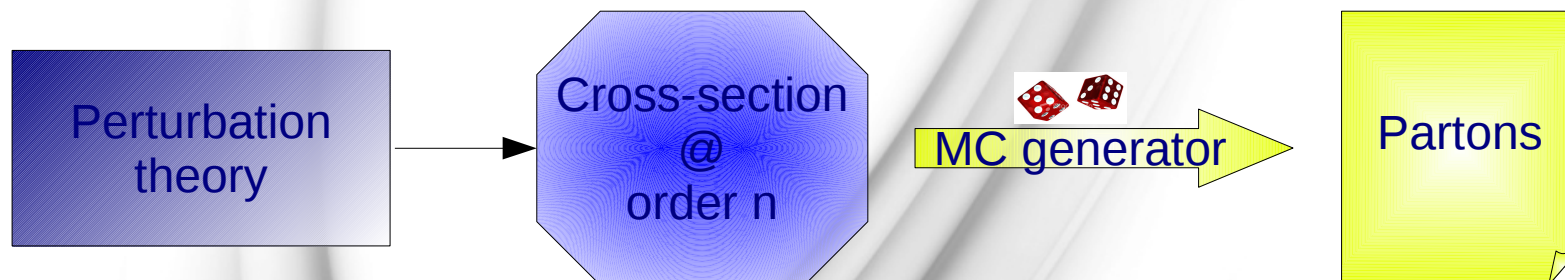
Monte Carlo Techniques

- Monte Carlo techniques play a central role in particle physics
 - Often the only practical way to evaluate difficult integrals or to sample complicated p.d.f.
 - Used to evaluate the signature of a model
 - Used to evaluate the hadronization (non-perturbative QCD)
 - Used to evaluate the detector response
- Often the key to evaluate p.d.f. of physics quantities



Fixed order Monte Carlo

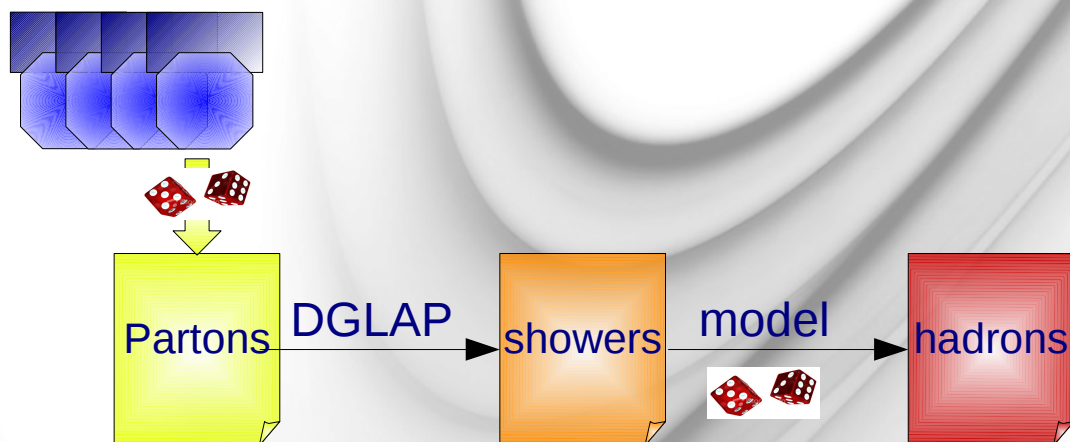
- Most straightforward approach:



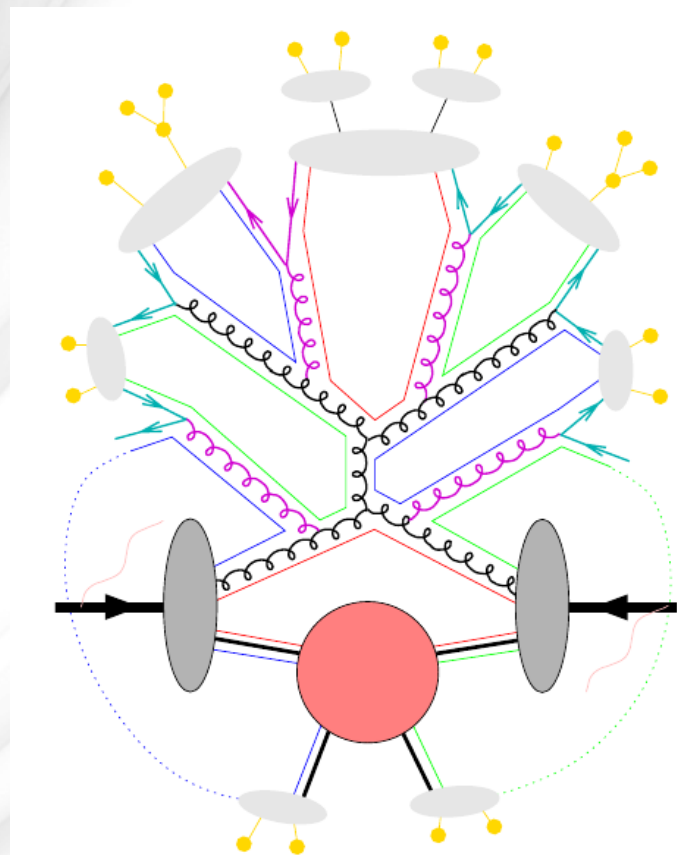
- 2 cases
 - Weighted events:
 - the weight is the matrix element squared
 - Unweighted events:
 - events are distributed according to the matrix element squared.
- Technical difficulty: avoid singularities (collinear and soft regions) at LO + and implement numerically the cancellation between N and N+1 partons contributions at $N^n\text{LO}$.

All-orders Monte Carlo

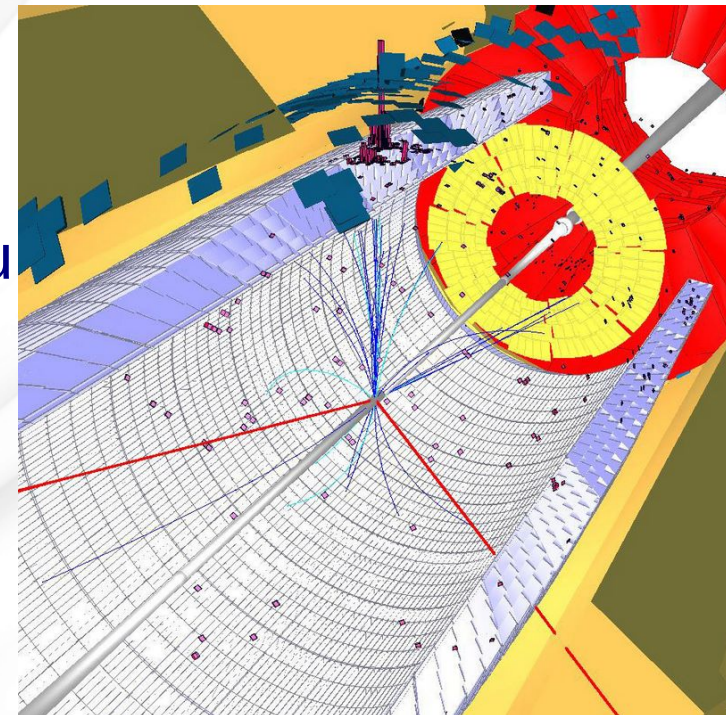
- Aim: produce not just partons but a full set of hadrons in the final state.
- Complex system to simulate utterly complex reactions.



- + underlying event
- + decay unstable particles
- PYTHIA, ARIADNE, HERWIG, ISAJET, ...



- Geant4
 - *Geant4 (for GEometry ANd Tracking) is a platform for "the simulation of the passage of particles through matter," using Monte Carlo methods.*
 - *Its areas of application include high energy, nuclear and accelerator physics, as well as studies in medical and space science."*
 - Why a Monte Carlo ?
 - Draw an energy deposit from a Landau
 - Decay particles in flight
 - Generate showers in calorimeters
 - ...



Other examples (II)

- Detector simulations

- Purpose: to go from energy deposits to „detector response”

- Acceptance
- Resolution

Why a MC ?

Use the response as p.d.f.
Add noise

...

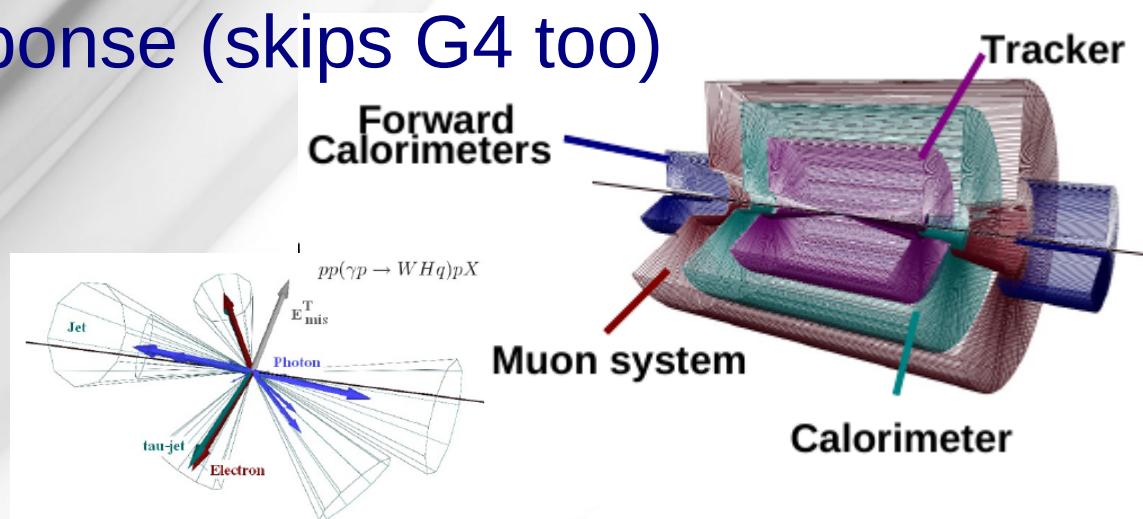
- Detector-specific simulation

- Generic detector response (skips G4 too)

- e.g. Delphes

- Toy Monte Carlo

- MC integration
- Coverage studies
- Pseudo-experiments




Delphes: <http://arxiv.org/abs/0903.2225>



Random number generators



- A random number generator that follows a $U[0,1]$ is the basic ingredient to any Monte Carlo
- Various options are available:
 - Hardware true random generators
 - Uses radioactive decays or thermal noise
 - Truly random
 - Slow and requires dedicated hardware
 - (Pseudo-)random generators 
 - Many algorithms available. Some better.
 - Quasi-random generators
 - Uses a recurrence relation to compute x_{i+1} from x_i
 - May have good coverage properties, but produces always the same sequence. If you know any x_i , you know the sequence.



This is what we want
to use in practice
(good ones)

What does NR say ?

- Ban (multiplicative) linear congruential generators
- Never use a generator with a period $T < \sim 2^{64} \sim 10^{19}$ or any generator with undisclosed T .
- Never use a generator that warns against using its low-order bits.-> sign of obsolete generator.
- Never use built-in C/C++ generators.
- Avoid generators that take $> \sim 25$ operations.
- Avoid generators designed for cryptographic uses.
- Avoid generators with $T > 10^{100}$... you don't need it.
- Generators should combine at least 2 well-understood methods.

Linear Congruential Generator

$$X_{n+1} = (aX_n + c) \bmod m$$

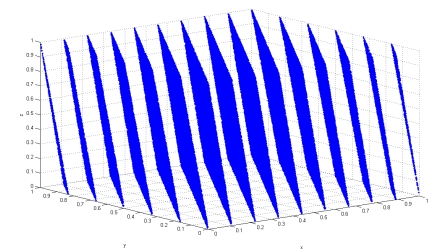
- Any seed is as good as the others
- From there, the sequence will evolve „random”.

Upper bound on the period.

Many problems/limitations of the LCG.

Most well known (**Marsaglia theorem**) : If the LCG is used to k times to obtain a point in a k-dimensional space, points will be located on $\max m^{1/k}$ (k-1)-hyperplanes, much less if the constants m and a are badly chosen.

Also, if m is a power of 2, least-significant bits are not random but have periods of maximum 2^n .



Bad example: RANDU (1960's): $a=65539$, $m=2^{31}$

Better choice (NR) : $m=2^{32}$, $a=3935559000370003845$, $c=2691343689449507681$
and keep only 32 most-significant bits out of 64.

Xorshift method

Let x be a non-zero 64-bits integer.

$$\begin{array}{ll} x \leftarrow x \wedge (x \gg a_1) & x \leftarrow x \wedge (x \gg a_1) \\ x \leftarrow x \wedge (x \ll a_2) & \text{or} \quad x \leftarrow x \wedge (x \ll a_2) \\ x \leftarrow x \wedge (x \gg a_3) & x \leftarrow x \wedge (x \gg a_3) \end{array}$$

With well chosen (a_1, a_2, a_3) , for example: $(21, 35, 4)$ or $(20, 41, 5)$.

$$\begin{pmatrix} 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & \ddots & 0 \\ \vdots & 0 & 0 & 0 & 1 & 0 & 1 \\ \vdots & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Principle:

We use bit algebra: \wedge is the **bit addition** in base 2.

Each of the 3 steps can therefore represent the **action of a matrix S_{k_i} on a vector x** .

-> one iteration : $T = S_{k_3} S_{k_2} S_{k_1}$

Max petiod: $M=(2^{64}-1)$. Will be achieved if: $T^M=1$ and $T^N \neq 1$ for each of the 7 prime factors of M : $N=3,5,17,257,641,65537, 6700417$.

This can be found by brute-force, powers of T being computed by successive squaring.

Limitations/flaws:

Only a small subset of (a_i) triplets have good randomness properties.

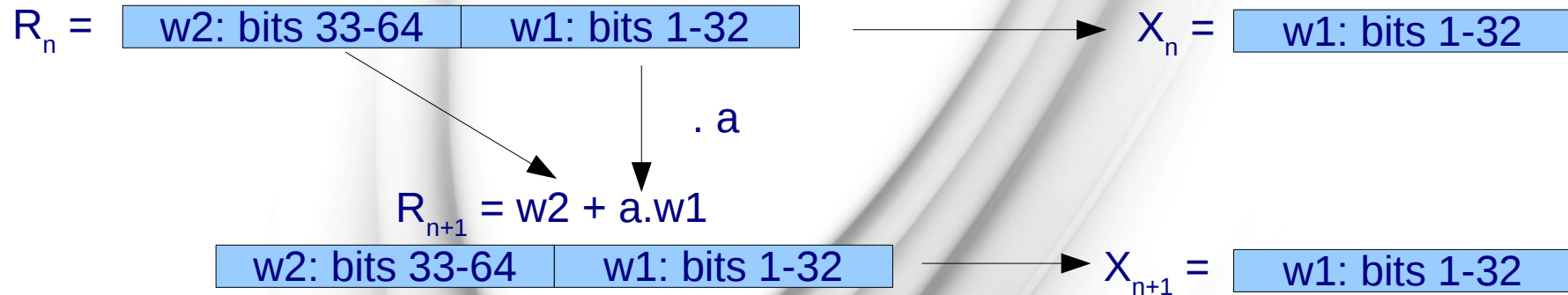
It's easy to design a test tath the Xorshift will fail:

every bit of step $i+1$ depends on max 8 bits of step i .

Very useful when combined with other methods !

Multiply with Carry

This method is easy to understand, and to implement on 64 bits architectures:



Period: $(2^{32}a-2)/2$. (a prime)

Improvements:

- r -lag MWC generators: use $w1$ from R_{n-r} . This requires $r+1$ seeds.
- The max period goes like $(a \cdot b^r - 1)$ with $b=2^{32}$ but cannot be saturated
- Complementary multiply with carry
- $R_{n+1} = (2^{32}-1) - (w2 + a \cdot w1)$: do a XOR with 0xFFFFFFFF (revert all first 32 bits)
- The max period $(a \cdot b^r - 1)$ with $b=2^{32}$ can be obtained for the right a .
- „Mother-of-all” generator: do a linear combination of >1 $w1$.

Example: $a=3636507990$, $r=1359$: $p \sim 10^{13101}$

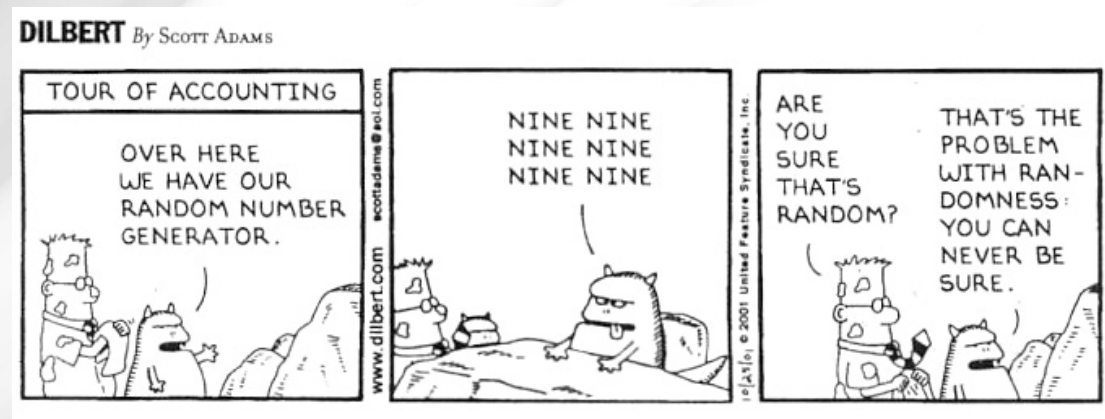
How to test the quality of a given random number generator ?

```
int getRandomNumber()
{
    return 4; // chosen by fair dice roll.
             // guaranteed to be random.
}
```

- Test on equal distribution: $\chi^2 = \sum_{i=1}^k \frac{(N_k - N/k)^2}{N/k}$ must be a χ^2
- Test on correlations: large number of hyperplanes
- Gap test: $P(\text{only last of } n \text{ in } [a,b]) = p(1-p)^{n-1}$ with $p=b-a$
- Random walk test: for $0 < a < 1$, $P(x < a)$ is binomial
- ...

-> If needed, use existing test suites (e.g. Diehard by Marsaglia)

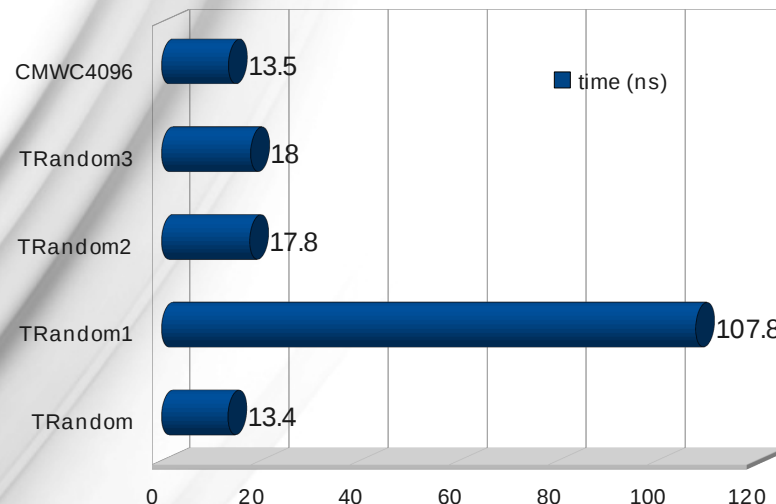
Good generators: NR's ran, CMWC, Mersene twister, ...



ROOT generators

- TRandom
 - LCG ! $T \sim 10^9$
- TRandom1
 - „RANLUX” $T \sim 10^{171}$
- TRandom2
 - „Tausworthe generator” $T \sim 10^{26}$

Intel(R) Core(TM)2 Duo CPU P8700 @ 2.53GHz:



- TRandom3
 - Mersene Twister $T \sim 10^{6000}$
 - Default in python, Ruby, Matlab, ...

**Recommended
by ROOT**

„not very elegant and is overly complex to implement”.

A simple complementary multiply-with-carry generator can have a period 10^{33000} times as long, be significantly faster, and maintain better or equal randomness.

Generating non-uniform distributions

The inverse transform method

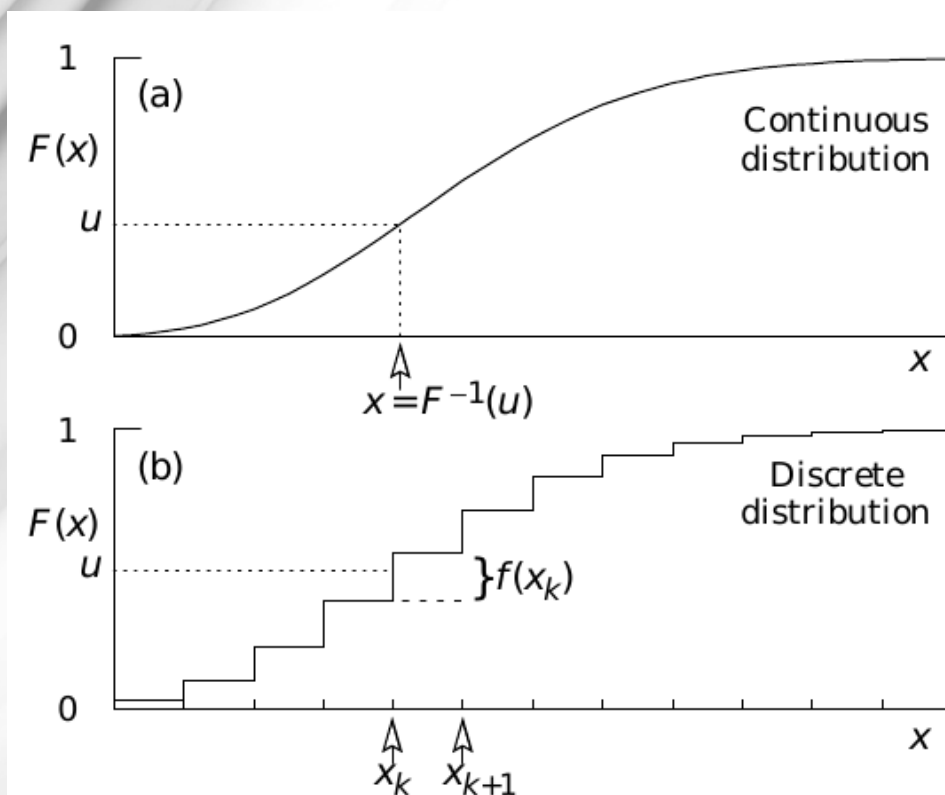
- Consider a probability density function $f(x)$ on the range $-\infty < x < \infty$, and its cumulative distribution function $F(x)$.
- If a is chosen with probability density $f(a)$, then the integrated probability up to point a , $F(a)$, is itself a uniform random variable on $[0, 1]$.

$$u = F(x)$$

$$x = F^{-1}(u)$$

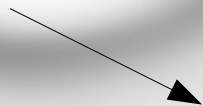
- Requires an explicit form for F^{-1} .
- May not be the fastest
- Can be applied to histograms (implemented in ROOT)

-> $\exp(x)$, $(1 - x)^n$, and $1/(1 + x^2)$



Von Neumann's method

- This is a rejection method.
- Generate a random number r_1 according to $h(x)$.
- Generate a r_2 uniformly in $[0,1]$
- If $r_2 < f(x)/Ch(x)$, keep r_1 .
Otherwise, try again.
- With f and h normalized to 1, $1/C$ is the efficiency of the method.
 - C must be close to 1.



Difficulties for narrow peaks !

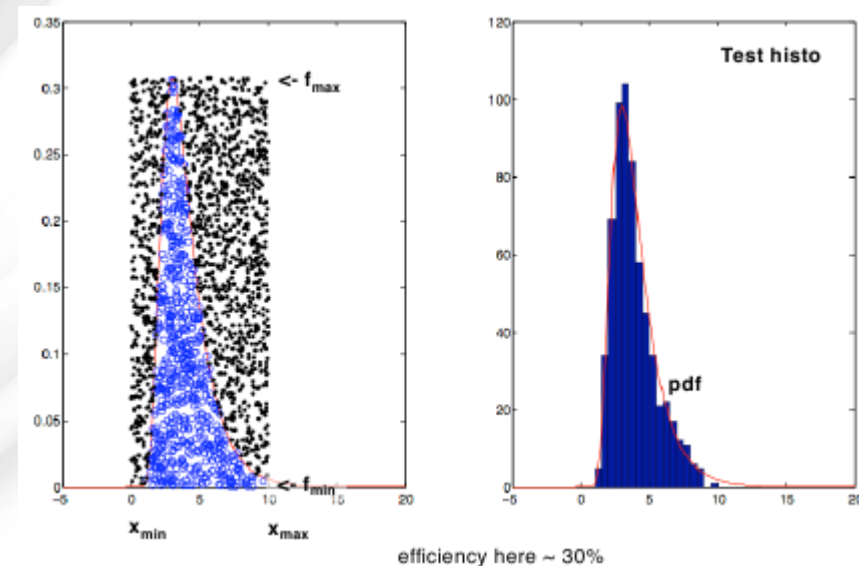
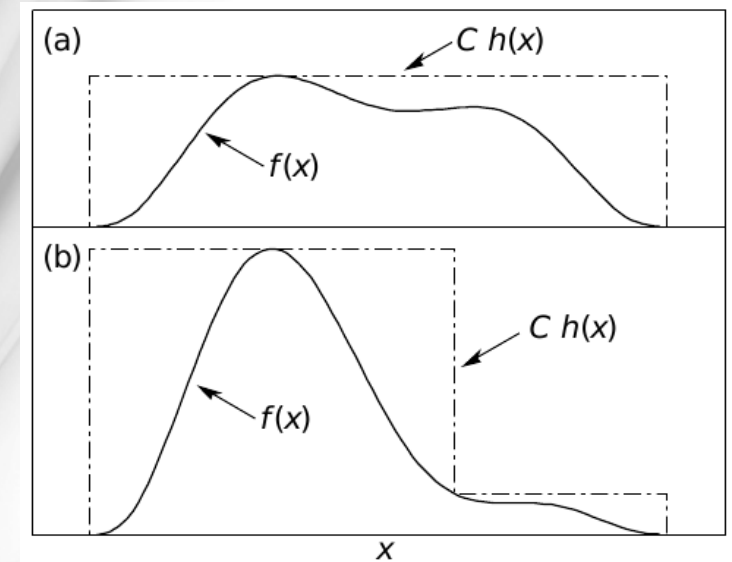
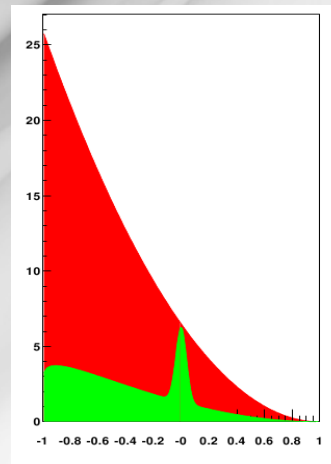


Illustration: MC integration

Simple example of MC integration

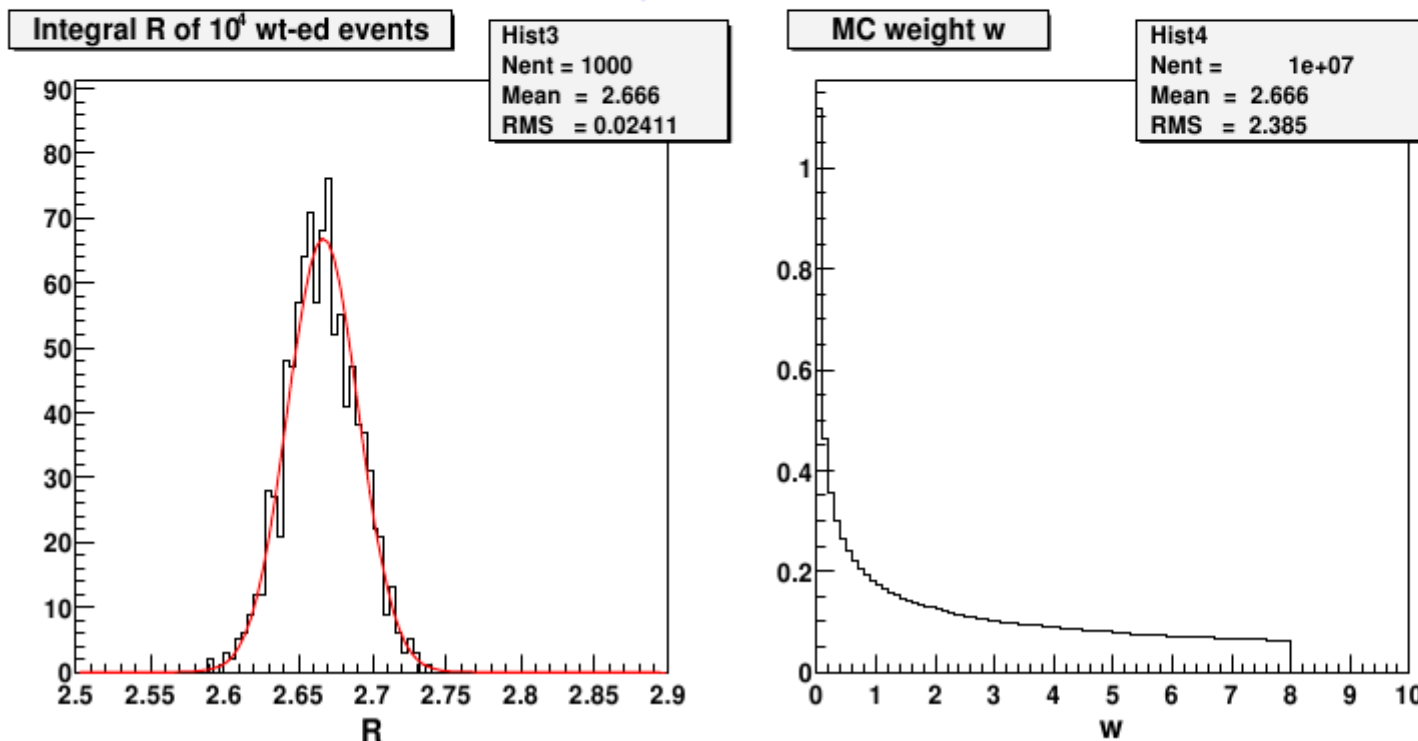
$$\rho(z) = (1+z)^2 \theta(1-z) \theta(1+z)$$

We know $R = \int \rho(z) dz = 8/3 = 2.6666\dots$ Nevertheless, we calculate it with help of the MC method.

Generate uniformly $z \in [-1, +1]$. Define MC weight $w = \rho(z)$.

$$R \simeq \langle w \rangle = \frac{1}{N} \sum_{I=1}^N w(z_I).$$

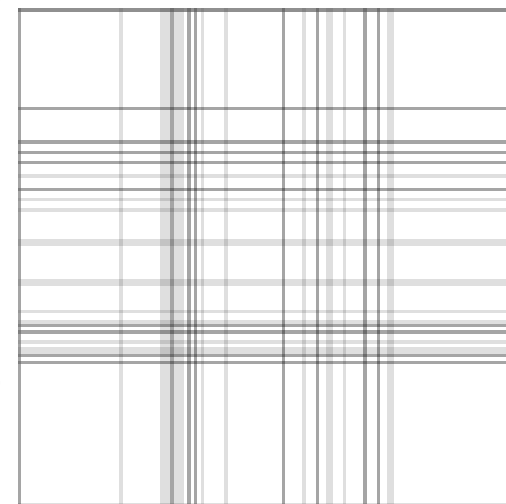
Generate $N = 10^4$ MC events. Repeat the calculation 1000 times!



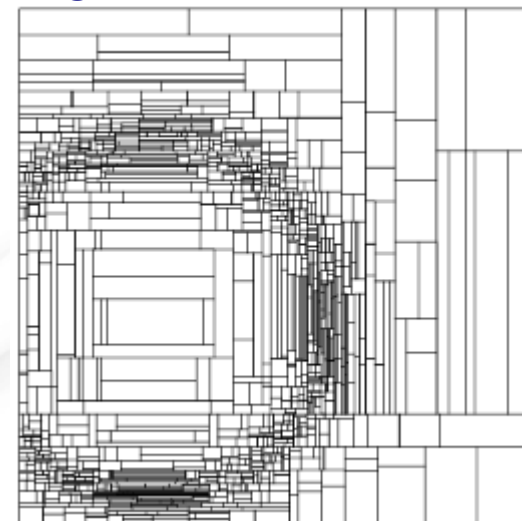
General purpose MC

- Few examples of General Purpose Monte Carlo Simulators, that is programs which work (in principle) for arbitrary integrand
 - Need of much CPU power and memory
 - only recently available/affordable.
 - Examples: VEGAS & FOAM
- VEGAS assumes the function can be factorized in terms that depends on one variable. For each variable, find the „best binning”
 - Approximation, can be pathological !
- FOAM works by dividing the integration domain in cells, where the rejection method can be efficiently used (\Leftrightarrow small variance of weights).
- In both cases, the core of the method is about finding the best „grid”.

VEGAS



FOAM



Outline

- Probability and Statistics, basic concepts
- Monte Carlo techniques
- **Event classification**
 - Introduction
 - (Optimal) cut-based selection
 - Multi-Variate Techniques (NN, BT, ...)
- Parameter estimation
- Limits, confidence intervals, significance
- Closing remarks

- **Data/Physics analysis tasks are inherently multivariate**
 - **Event selection**
 - Triggering, real time filtering, data streaming
 - **Event reconstruction**
 - Tracking/vertexing, particle Identification
 - **Signal/Background discrimination**
 - Higgs & Susy searches, ...
 - **Functional approximations**
 - Jet energy corrections, tagging efficiencies...
 - **Parameter estimations**
 - Higgs mass, top quark mass
 - **data exploration, data-mining**
 - Data-driven extraction of information
 - pattern recognition, clustering



Event classification

- **Data/Physics analysis tasks are inherently multivariate**

- **Event selection**

- Triggering

- **Event reconstruction**

- Tracking

- **Signal/Background**

- Higgs &

- **Functional**

- Jet energy

- **Parameter**

- Higgs mass

- **data exploration**

- Data-driven extraction of information
- pattern recognition, clustering



But you know, at LHC we
have very powerful magnets !
(and CMS has the biggest)

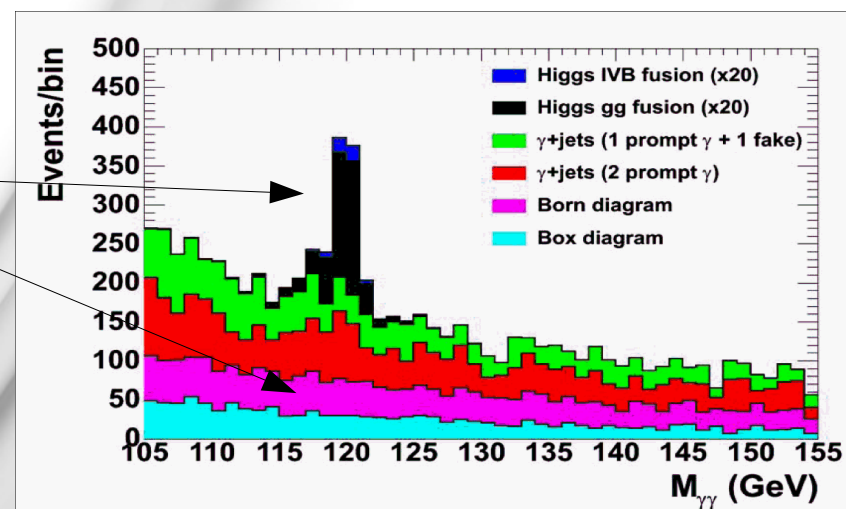




Introduction

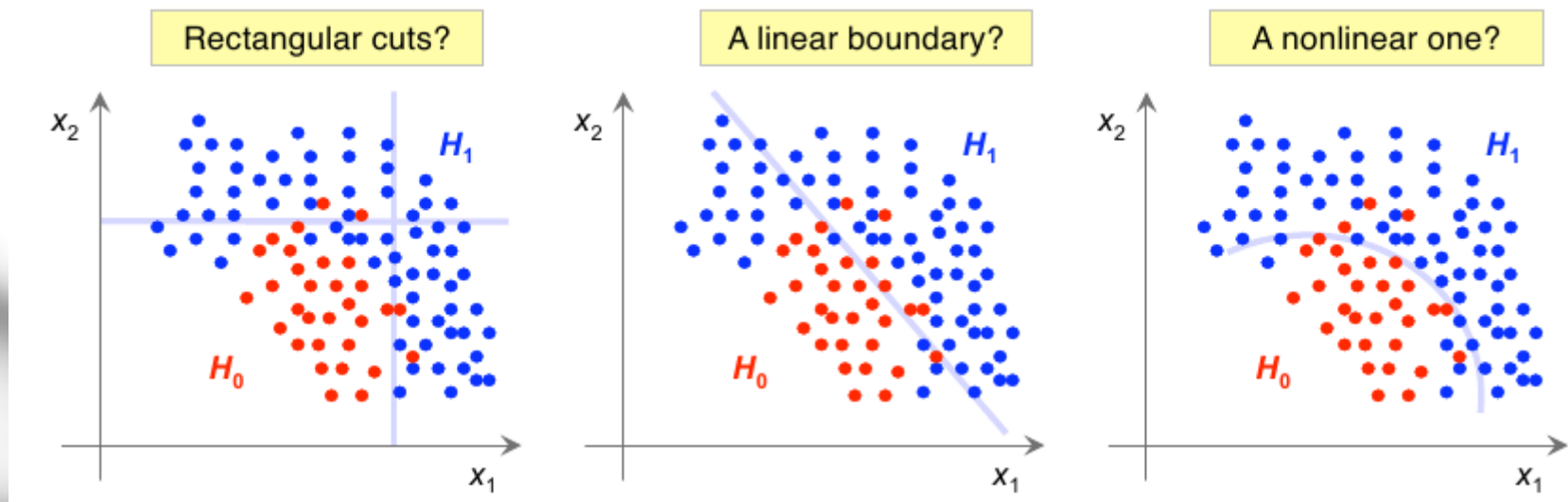


- Common task in HEP: separate „signal” from „background”.
- A typical analysis will consider many variables as levers to study data.
 - Number of jets & leptons, energy, angular distribution, invariant mass, isolation, missing (transverse) energy and momentum, ...
- Multi-variate analysis is therefore omnipresent in science
 - Event classification is performed in a N-dimensional space
 - Problem: human mind is limited to 3D (at best)
 - Various approaches:
 - Simple (consecutive) cuts
 - Compatification
 - Global approach with help of analytical or MC description.



- Using cuts sequentially
 - Generally easy
 - Little flexibility
 - Loss of information
- Compactification („MVA techniques”)
 - Combine several observables into one test statistic.
 - Computationally intense (potentially)

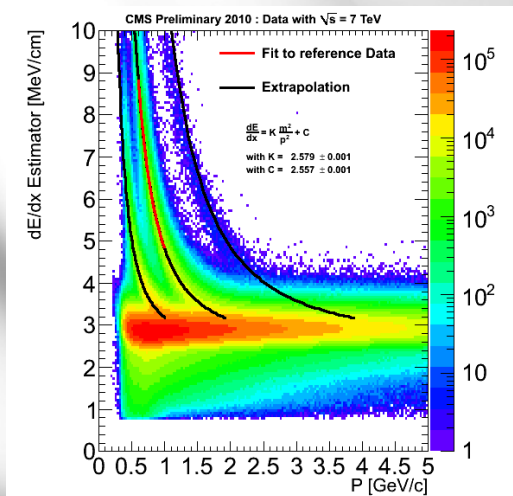
Can be improved by choosing the right combination of parameters, or the right basis.



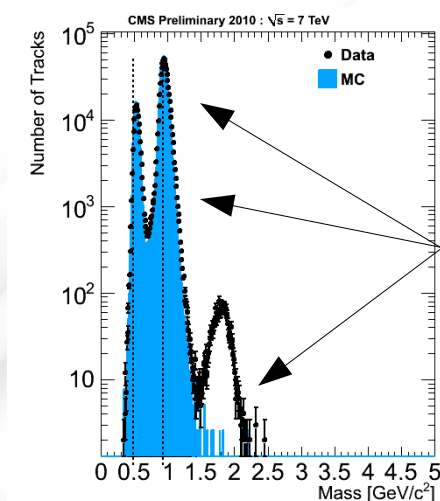
Selecting by cuts

- Applying cuts consecutively is often the simplest approach
 - But not optimal (does not take correlations into account)
 - With correlations, cut optimization is not direct.
- The situation can be greatly improved by clever choice of the observables considered
 - Never forget to be clever !
 - Observables must be motivated!

Example:

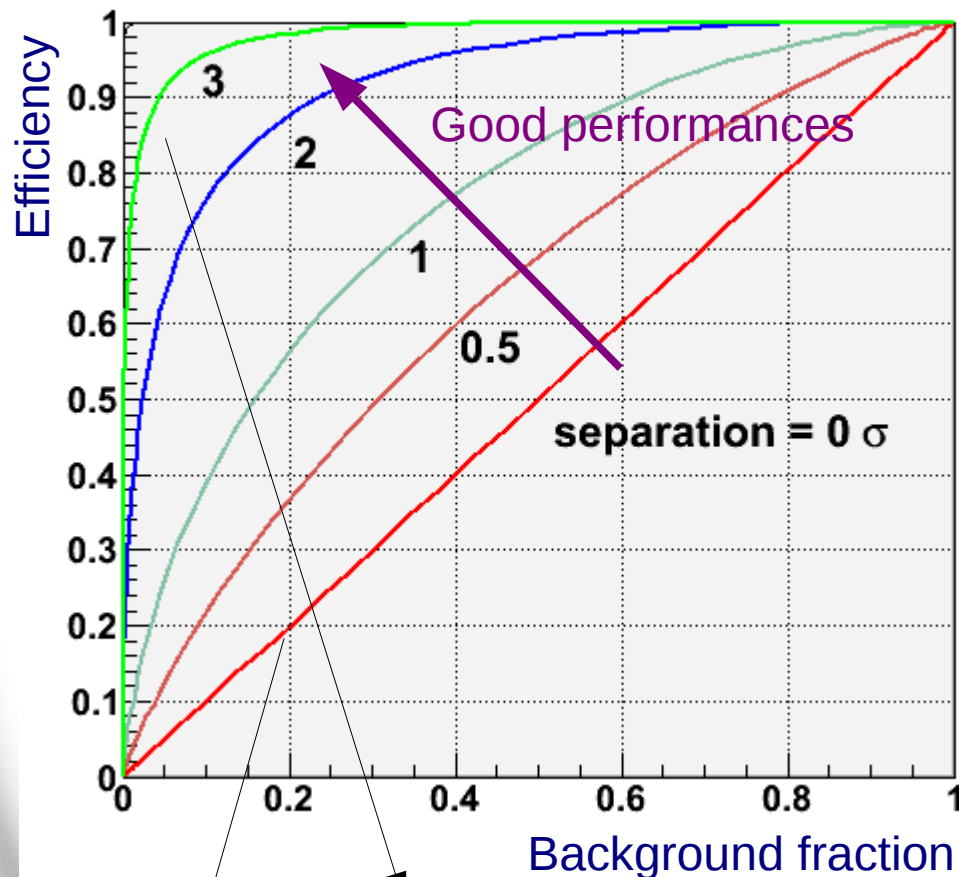


Combine momentum
and dE/dx using
Bethe-Bloch



Kaon, proton
and deuteron
signals easy to
separate !

Choice of the best cuts



Useless cut.
No discriminating power

Excellent cut

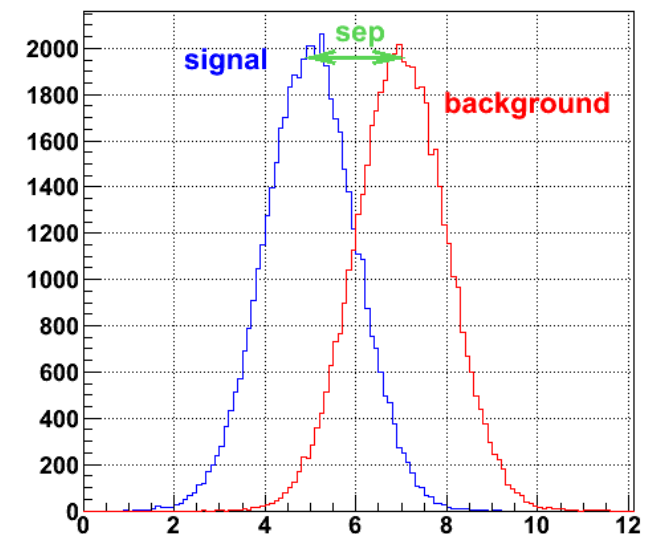
That plot doesn't tell which is the best value.
It is useful to compare the cuts on various quantities.

The value of the cut is free a priori.

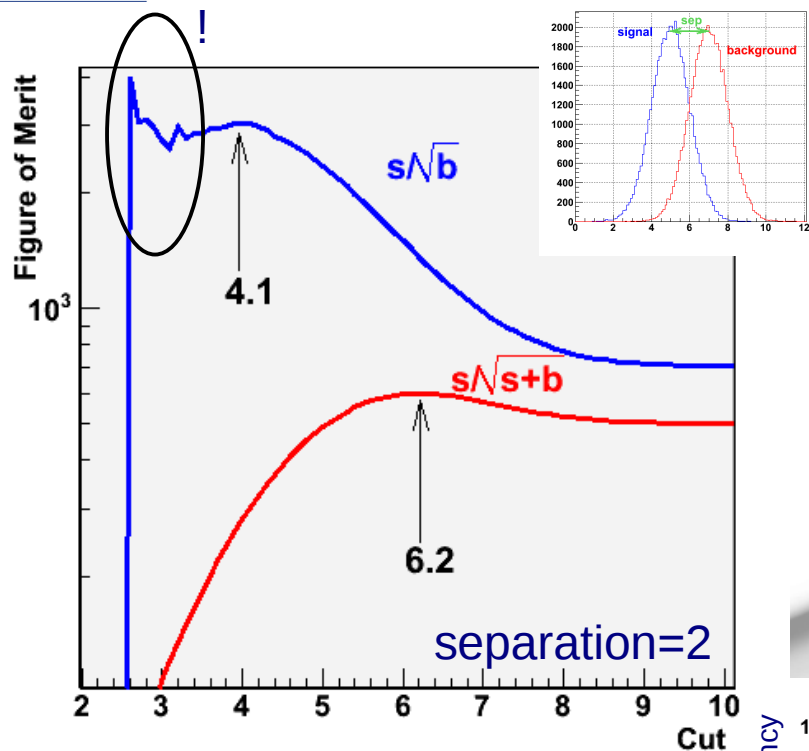
- It depends on the efficiency one wants to achieve
- It depends on the fraction of background one is ready to accept.

Useful plot:
efficiency vs background fraction

Simple
example:
upper cut



Choice of the best cuts



To decide which cut to apply, one needs a **figure of merit**.

Common choices are:

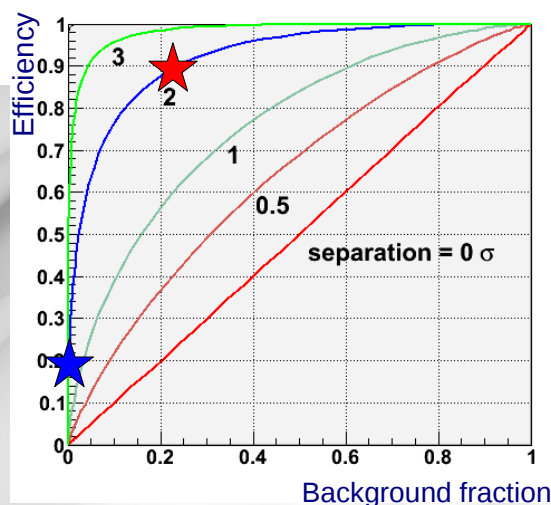
$$\frac{s}{\sqrt{b}}$$

discovery

$$\frac{s}{\sqrt{s+b}}$$

measurement

They are not (well) statistically motivated.
Could use other merit functions :



$$S_{c12} = 2(\sqrt{s+b} - \sqrt{b})$$

$$S_{cL} = \sqrt{2(s+b)\log(1+s/b) - 2s}$$

Significance
CLs

...

Note: if several cuts are applied consecutively, the working point choice should be an iterative process, to take into account correlations.

May be automatized (e.g. GARCON: Genetic Algorithm for Rectangular Cuts Optimization).

Fisher's discriminant

- Simplest: linear combination of variables
 - Cut defined by hyperplane in the space of observables
 - ? Optimal plane to separate 2 classes of „events“

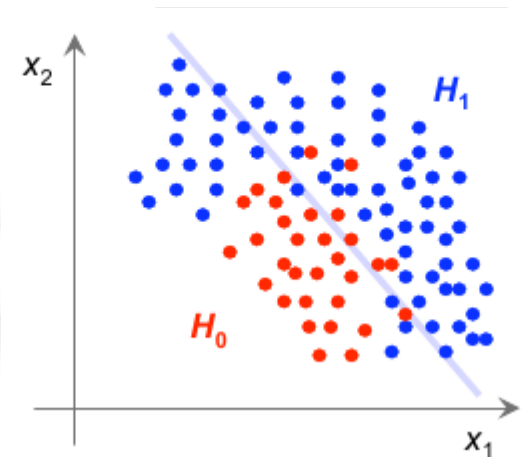
$$D(x) = a_i x^i$$

- Most common choice: Fisher's discriminant

$$F(\vec{x}) = (\vec{\mu}_s - \vec{\mu}_b)^T V^{-1} \vec{x}$$

Difference of
the means

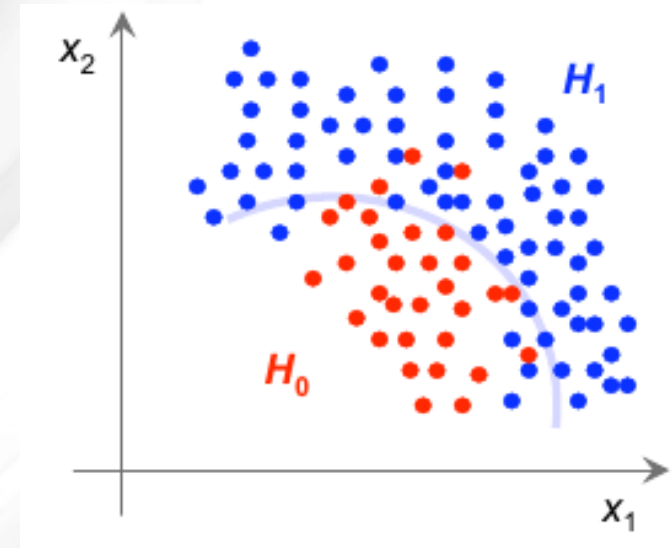
Inverse of the
variance matrix



- + Can be computed directly from the s & b distributions
- Does not consider different variances for s and b
- Linear...

Non-linear approaches

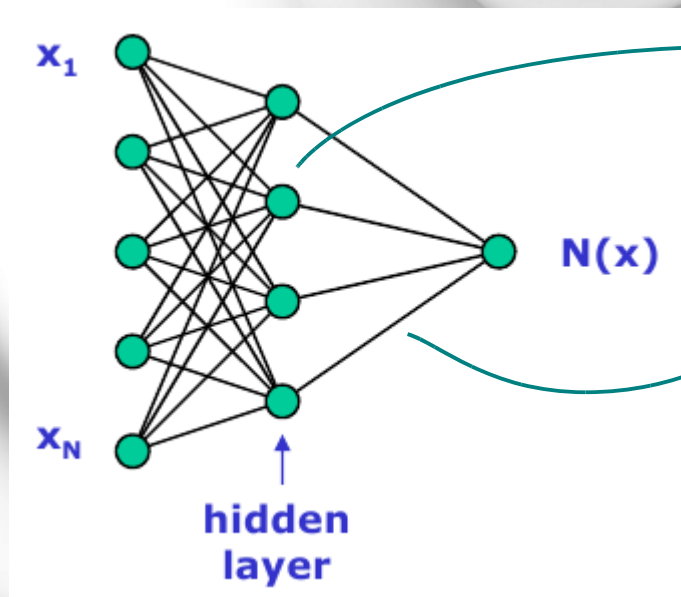
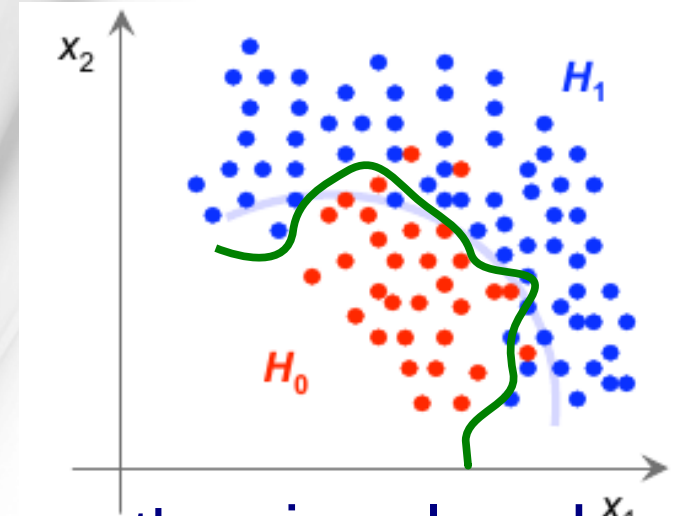
- More in general, any cut in a n-dimensional space can be expressed as a cut on test statistic.
- In that context: the test statistic is just a way to go from N observables \rightarrow 1 quantity, and then
 - Cut on it
 - Use is as a likelihood
- Non-linear cut \leftrightarrow non-linear test statistic.



This lecture

- Neural networks
- Decision trees
- Support vector machines
- Likelihood ratio
- ...

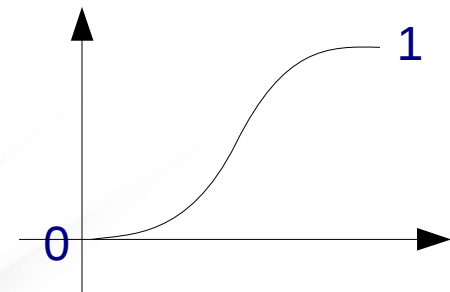
- Build a quantity N such that
 - $N=1$ in the signal region
 - $N=0$ in the background region
 - Goes smoothly from one to the other.
- The isocurve $N \sim 0.5$ is the boundary between the signal and background regions



Multi-layer perceptron

Neuron: transfer function ~

- uses the sum of inputs
- One „weight: constant term



Synapse: more or less strong link between neurons

- One multiplicative „weight” from input to output

$$N(x) = s\left(a_0 + \sum_i a_i x_i\right)$$

Common choice:
„sigmoid function”

$$s(t) = \frac{1}{1 + e^{-t}}$$



Training



- **Training/learning :**
procedure by which we get the weights „right”.

Define the error function:

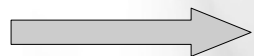
$$\epsilon^2 = \int (N(\vec{x}) - T_B(\vec{x}))^2 P_B(\vec{x}) d\vec{x} + \int (N(\vec{x}) - T_S(\vec{x}))^2 P_S(\vec{x}) d\vec{x}$$

NN value

Target

Background pdf

Same for the signal...



Evaluated on the error function on the sample (approximation/limitation)

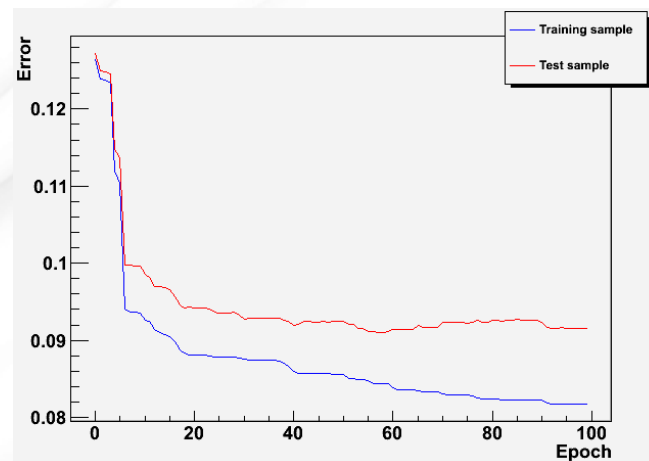
$$\epsilon'^2 = \frac{\sum_{i=1}^N (N(\vec{x}_i) - T_i(\vec{x}_i))^2}{N}$$

Difficult bit: minimization !

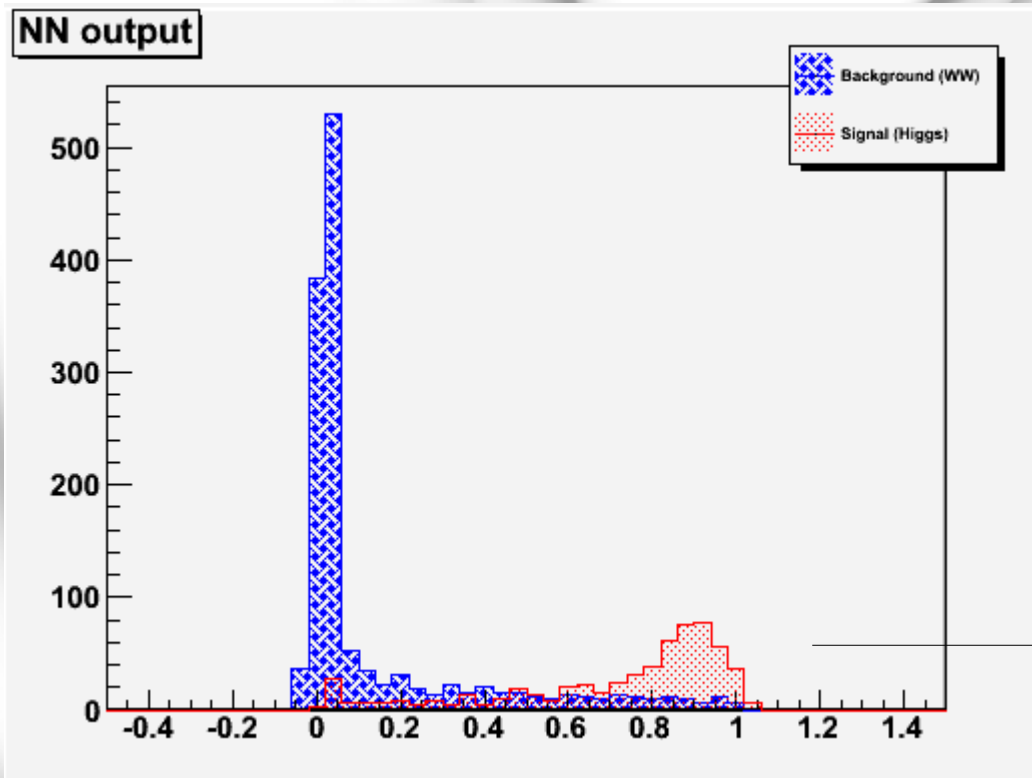
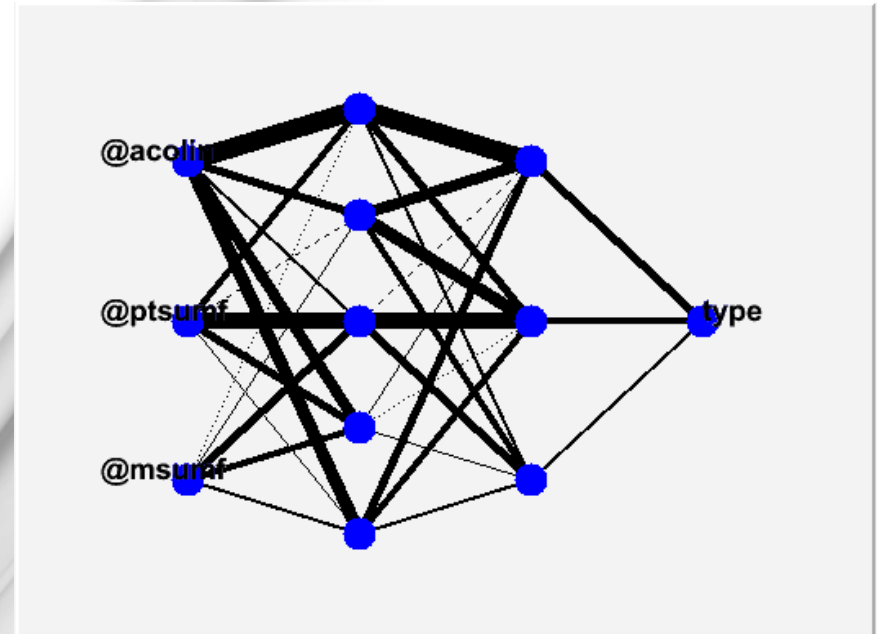
Most common technique: error back propagation

→ Follow the gradient of $d\epsilon/dw_i$ and iterate

Learning plot: error vs time during training.



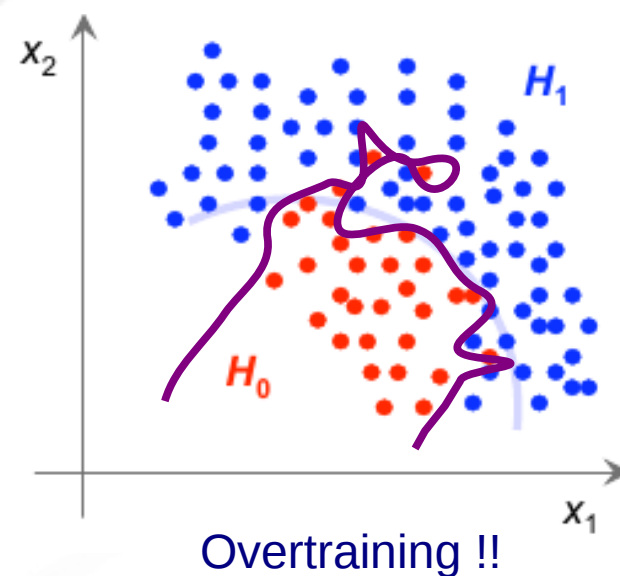
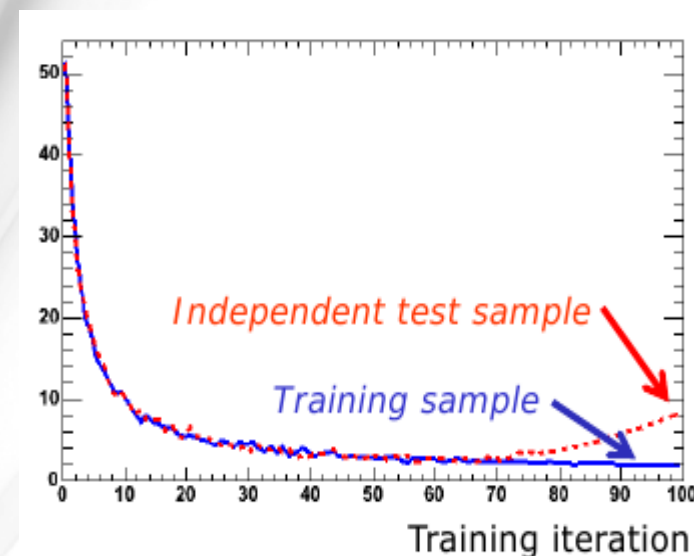
The structure of the network can be visualized.
Here, 3 input variables, 5+3 hidden neurons,
one output (0 or 1).
Weights are visualized by the thickness of the
links... at the end it often provides little information.



We have a single (compactified)
variable on which we can optimize a
cut as discussed before.

Words of caution

- Approximation of knowledge of true signal and background distributions with sample of signal and background events
 - Finite statistics limit precision (in itself usually not a problem)
- Risks of overtraining
 - Always control with an independent sample
 - Never train on data (or use a control sample)
- The result cannot be more accurate than the MC knowledge. Features not reproduced by MC will not be taken into account.



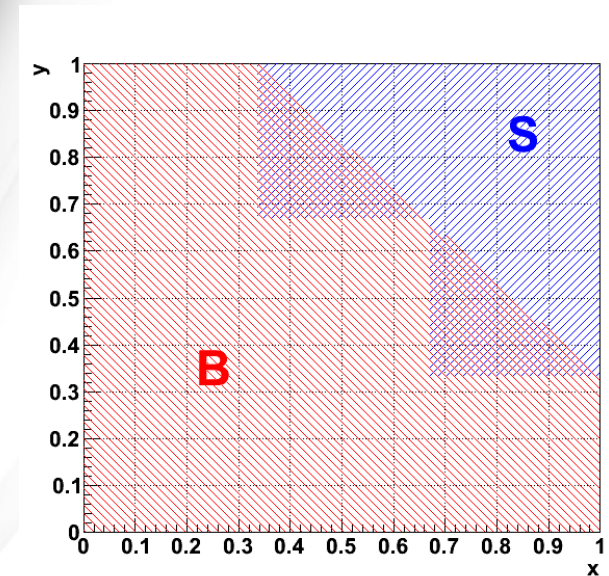
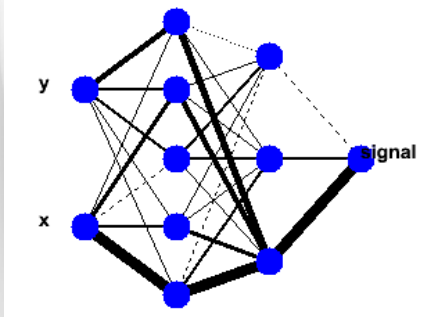
Simple Example

- Consider this example with two clear s and b regions, with some defined overlap.

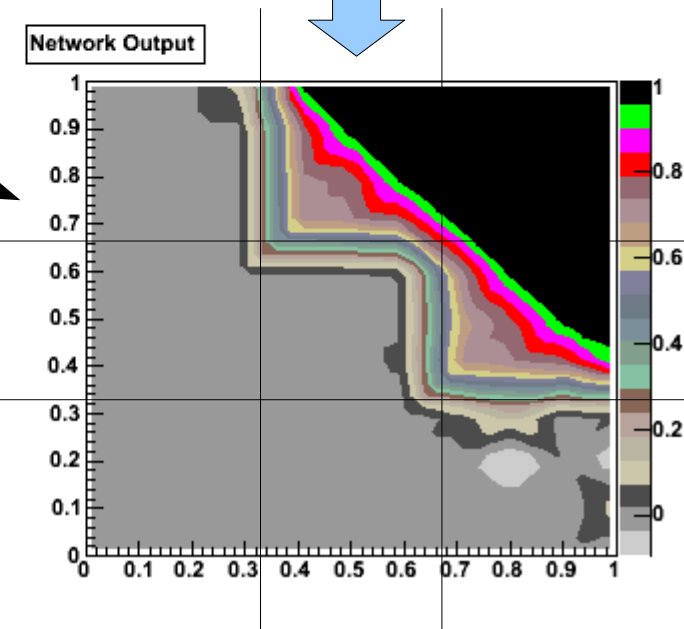
- In principle, 85% of bkg and 66% of signal can be unambiguously isolated.

- One builds and train a network with a simple structure

Structure: x,y:5:3:signal



- The NN function reproduces the input structure
- A cut on the NN output can produce a very pure (cut at ~0.9) or very efficient (cut at ~0.4) set of events.
- Performances will degrade if
 - Network structure is too simple to accomodate the shape of signal and background regions
 - The dataset is too small to give information about the shape of the two regions
 - Overtraining, ...



Simple Example

- Consider this example with two clear s and b regions, with some defined overlap.

- In principle, 85% unambiguous

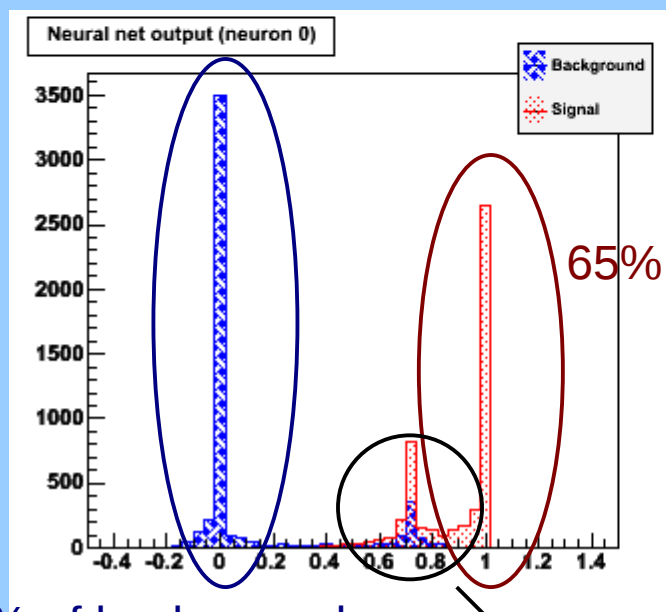
- One builds and t

- The NN functi

- A cut on the N (cut at ~ 0.9) o

- Performances

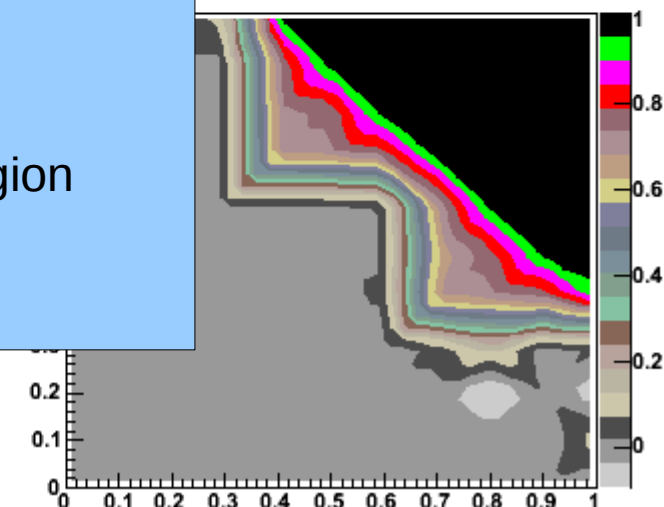
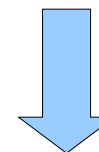
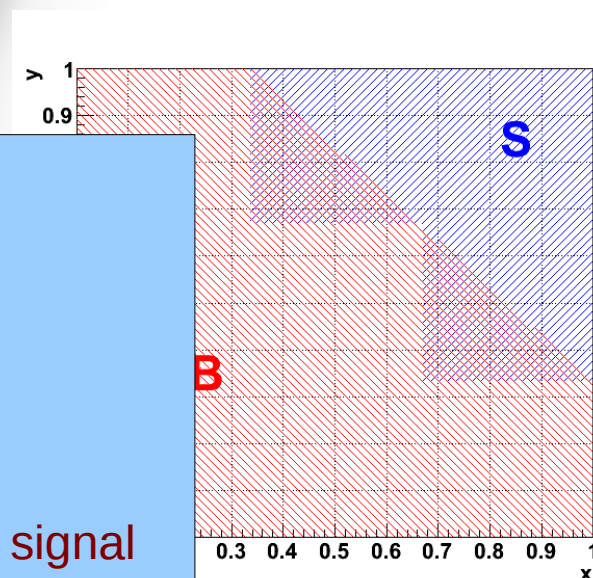
- Network str
- signal and background regions
- The dataset is too small to give information about the shape of the two regions
 - Overtraining, ...



82% of background

65% of signal

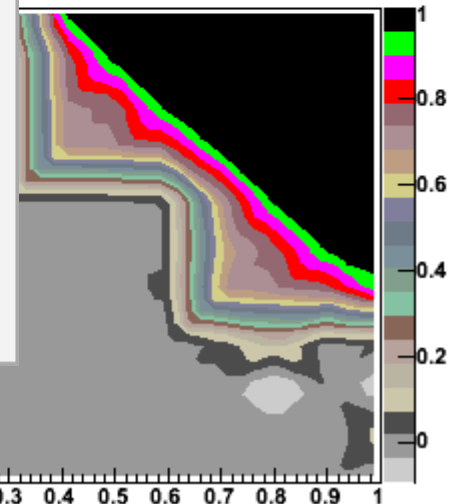
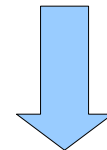
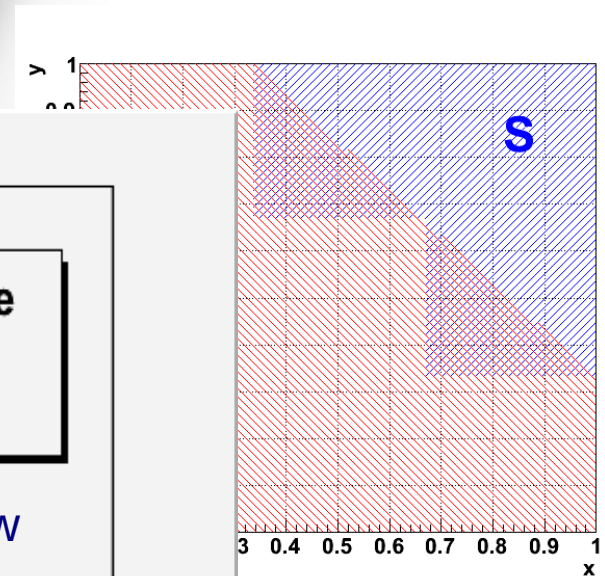
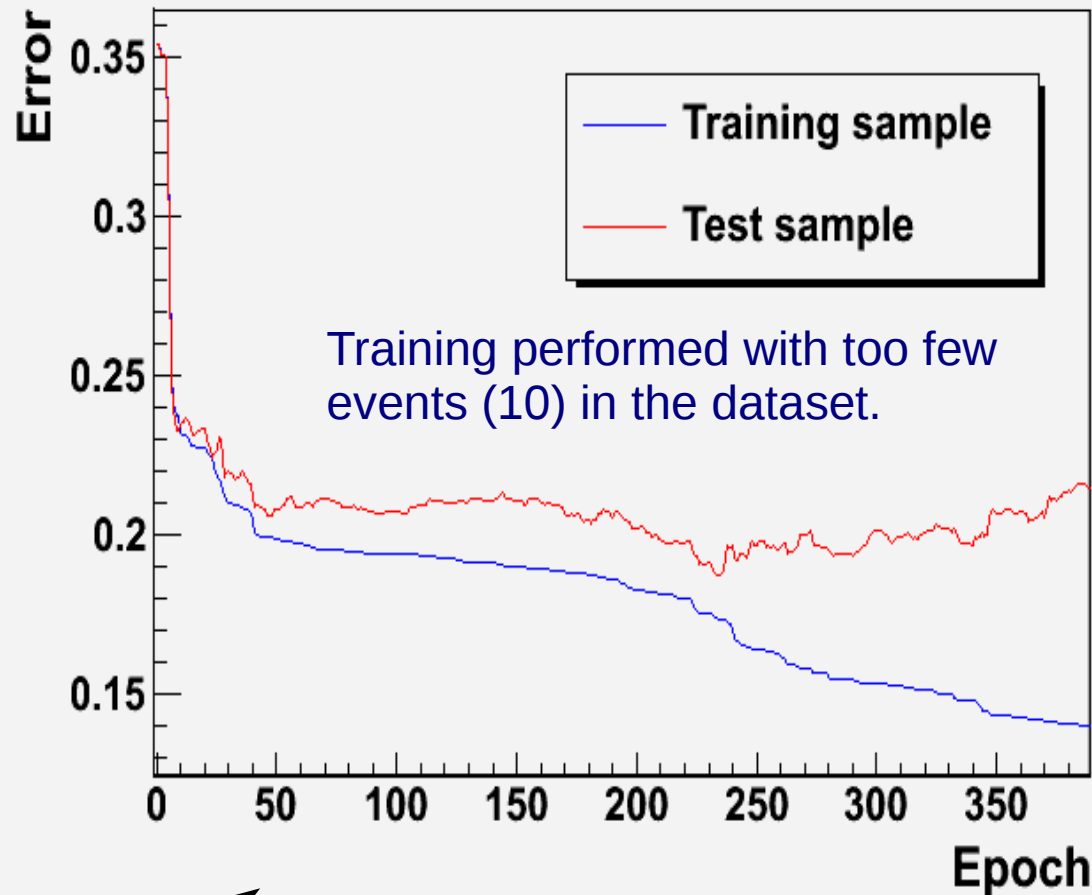
Mixed region



Simple Example

- Consider this example with two clear s and b regions, with some defined overlap.

- In principle, unambiguous
- One builds and

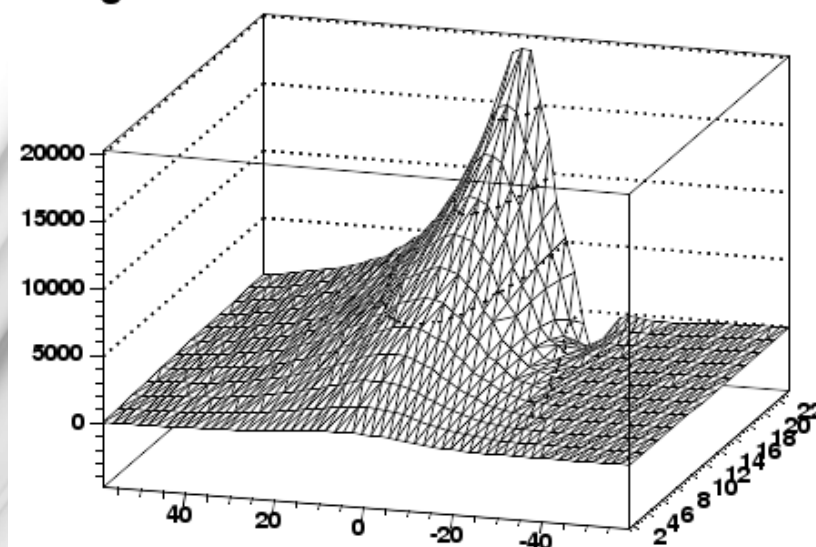


- The NN function
- A cut on the (cut at ~ 0.9)
- Performance
 - Network s
 - signal and background regions
 - The dataset is too small to give information about the shape of the two regions
 - Overtraining, ...

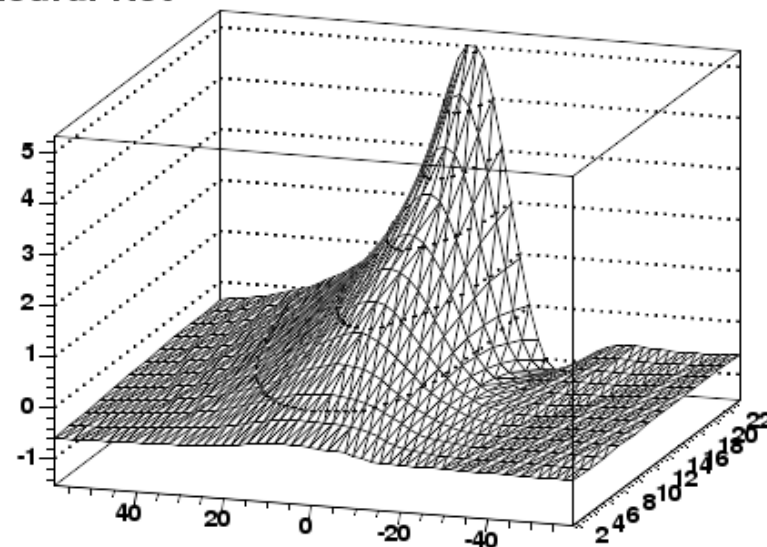
Aside note: using NN to fit functions

- For some application with a cylindrical symmetry, a magnetic field simulation gives as output the radial component of the B field on a grid.
- One want to fit those distributions with a function in order to plug them into a Geant simulation code.
 - One could try polynomial fits, but it seems difficult to reach the desired precision over the full range.
 - One could also use a spline interpolation between known points. In all cases, the resulting field would not be C-infinite.
- NN takes time to be trained (once) but then provides a C-infinite function well suited for many applications. There is no need for an a priori knowledge of the form of the function.

Original



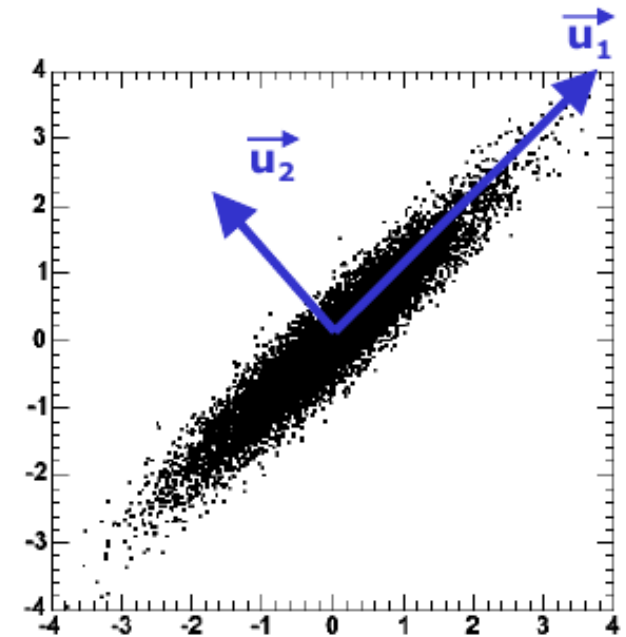
Neural Net



- **Clever choice of inputs**
 - Use well-understood observables
 - Don't put useless inputs
 - Prefer inputs in $[0,1]$, or normalize them
 - Avoid highly correlated inputs (-> decorrelation)
 - Avoid strongly peaked distribution (-> gaussianization)
- **Clever choice of the network structure**
 - No rule
 - People sometimes try first $N_{\text{inputs}}/2 + N_{\text{inputs}}/4$ hidden neurons, without motivation
 - Network structure should match the complexity of the phase space, not the number of inputs.
 - Generally, try to start simple, and extend if needed
 - This will also make the training faster

Decorrelation

- Removal of linear correlations by rotating input variables
 - Cholesky decomposition: determine *square-root* C' of covariance matrix C , i.e., $C = C'C'$
 - Transform orig (x) into decorrelated variable space (x') by: $x' = C'^{-1}x$
- Principal component analysis
 - 1) Compute variance matrix $\mathbf{Cov}(\mathbf{X})$
 - 2) Compute eigenvalues λ_i and eigenvectors \mathbf{v}_i
 - 3) Construct rotation matrix $\mathbf{T} = \mathbf{Col}(\mathbf{v}_i)^T$
 - 4) Finally calculate $\mathbf{u}_i = \mathbf{T}\mathbf{x}_i$



W. Verkerke

Gaussianization

- Decorrelation can be improved by applying a transformation to each observable that results in a Gaussian distribution
 - Can Gaussianize either signal or background sample (not both...)
- Two-step transformation
 - First apply rarity transform \rightarrow Creates uniform distribution

$$x_k^{\text{flat}}(i_{\text{event}}) = \int_{-\infty}^{x_k(i_{\text{event}})} p_k(x'_k) dx'_k, \quad \forall k \in \{\text{variables}\}$$

Measured value
PDF of variable k

Rarity transform of variable k

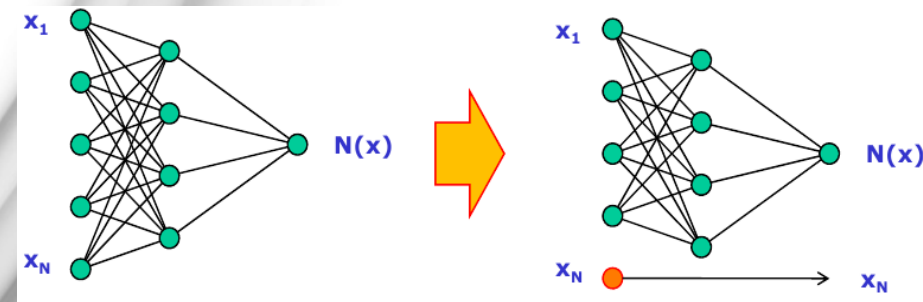
- Second: make Gaussian via inverse error function: $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

$$x_k^{\text{Gauss}}(i_{\text{event}}) = \sqrt{2} \cdot \text{erf}^{-1}(2x_k^{\text{flat}}(i_{\text{event}}) - 1), \quad \forall k \in \{\text{variables}\}$$

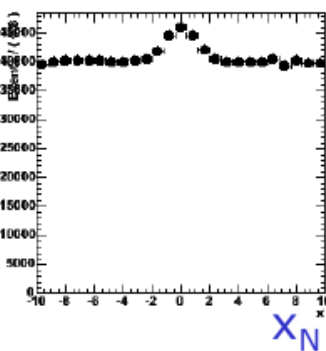
Study of network performances

The Neural Network should not be used blindly. You have to assess its performances and stability.

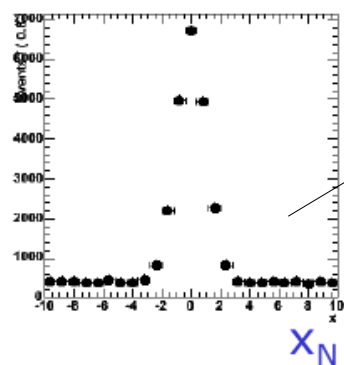
- Remove one variable
 - Maximally uncorrelated to the rest
 - Maybe the most discriminating variable
- After the NN selection, in principle, the signal appears clearly, **and the remaining background can be measured.**



No cut on $N(x)$

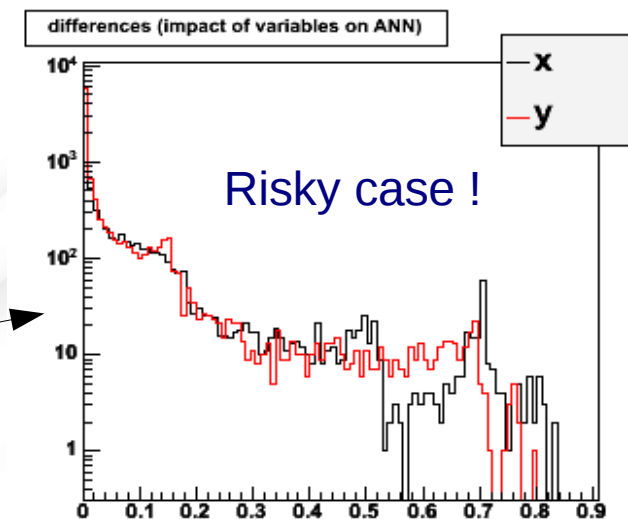


After cut $N(x) > \alpha$



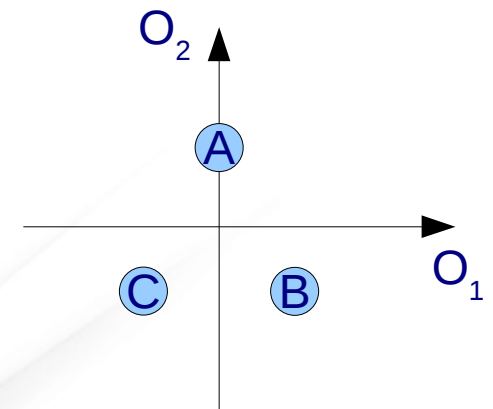
„side-band” background measurement.

- Study the stability: dN/dx_i
 - Sensible inputs
 - Risks of large systematics: control



Additional comment

- We discussed here a simple use case of NN with only one output neuron... one can go beyond that
 - OCR (1 neuron per character, take the best and estimate „risk of mistake” or „second choice”)
 - Distinguish between N categories using N-1 output neurons (and take the closest type)
 - ...

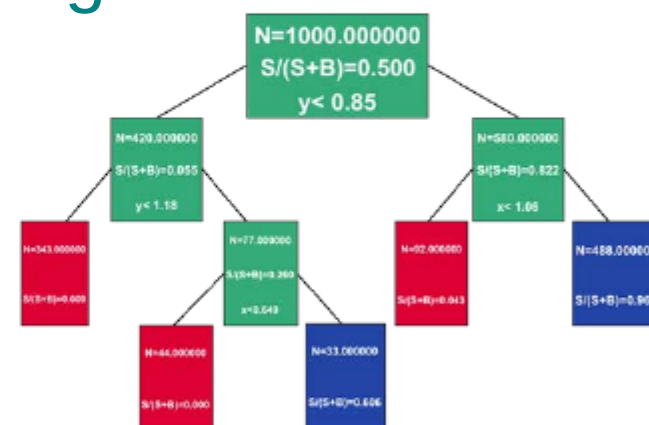




Decision trees

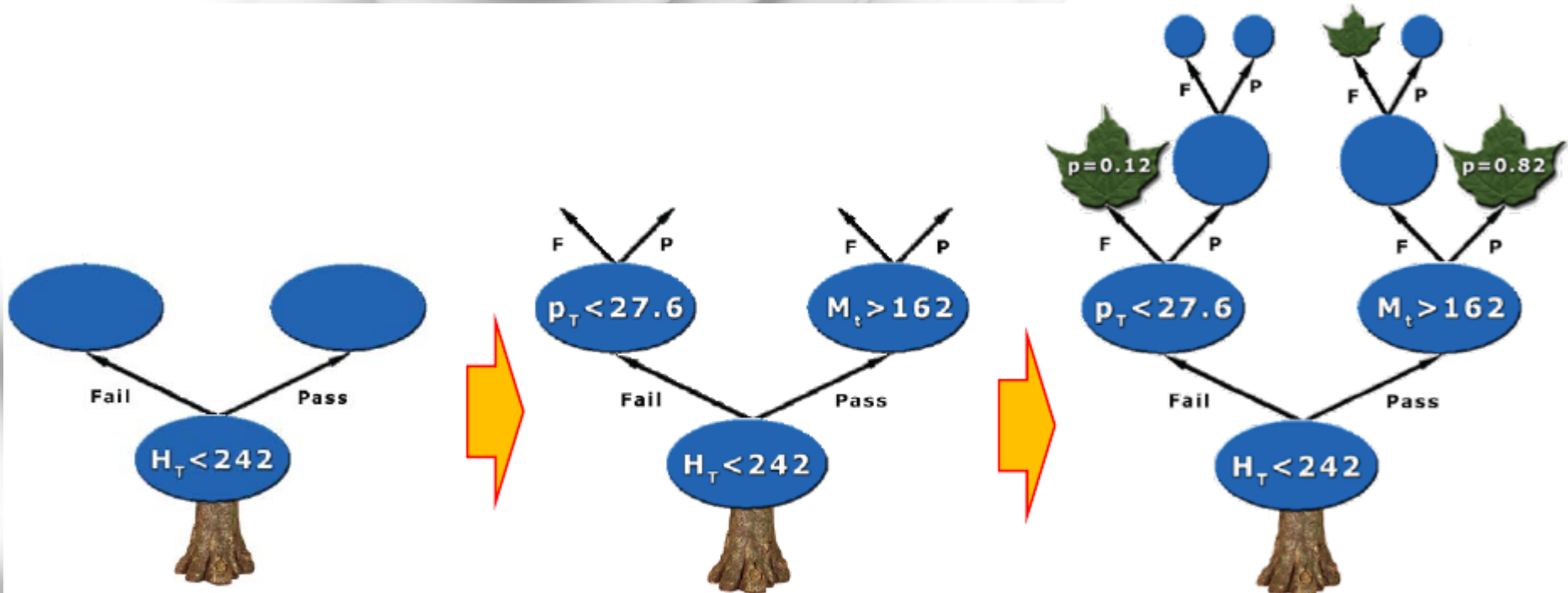


- **Decision trees** is another approach that became popular in 2005 with MiniBooNe & Tevatron Run II.
- Basic idea: **sequential rectangular cuts**
 - At each step: split data in 2 using the „best” single cut
 - Requires a metric to decide ($s/\sqrt{s+b}$, s/\sqrt{b} , „Gini”, ...)
 - Choice made independantly for each outcome of the previous step.
 - Repeat splitting until some stopping criteria is fulfilled.
 - Purity is high enough
 - N different cuts applied
 - ...
- Theoretically well motivated



- Breiman, et al (1984), Classification and regression trees, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, ISBN 978-0412048418

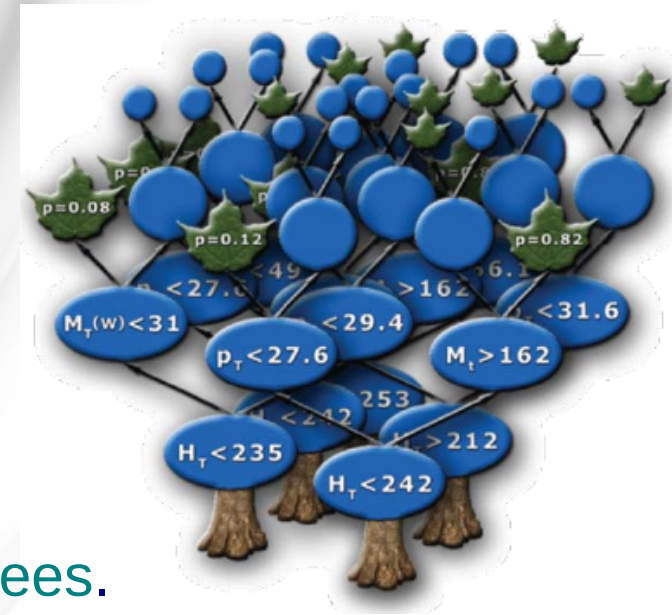
- This is an extension of the simple cut-based analysis
 - Do not (automatically) reject an event that fails only one of the criterias.
 - Either classify in two categories, or assign a probability to be A or B using $s/(s+b)$ in the corresponding leaf.



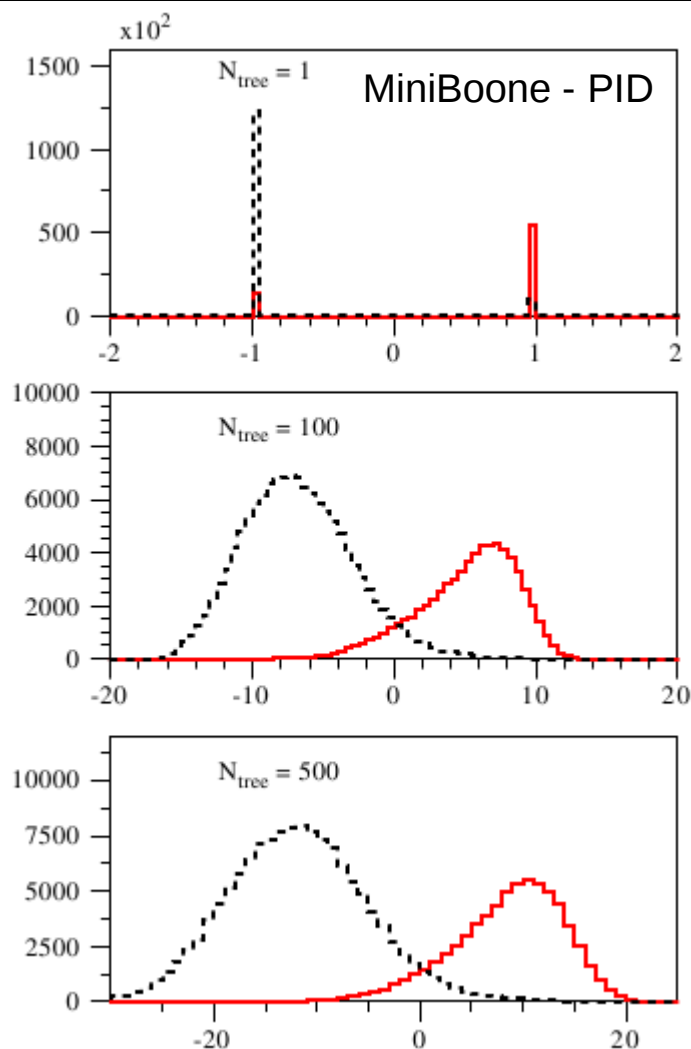
Pros and cons

- **Decision tree advantages**
 - Simple to understand and interpret.
 - Requires little data preparation (no normalization, gaussianization, ...)
 - Able to handle equally real, integer and boolean inputs
 - Perform well with large data in a short time.
- **Limitations**
 - Practical decision-tree learning algorithms cannot guarantee to return the globally optimal decision tree.
 - Genetic algorithms could be a good solution
 - Unstable w.r.t training sample.
 - Overtraining. Mechanisms such as pruning are necessary.

- **Principle:**
 - Build a first decision tree
 - Look at **misclassified events** and increase their weight
 - Build a **new decision tree** and iterate
 - As output, take the (weighted) **mean** of all N trees.
- **Boosting is a generic method that can be applied to any classifier.**
 - First idea in 1990 by Schapire (majority vote among 3 decision trees)
 - Variation in 1995 by Freund using >3 trees
 - Both joined their effort and developed adaboost in 1996.
- **Advantages:**
 - Increased discrimination power
 - Increased stability (w.r.t. training sample)



Boosted trees



arXiv:physics/0508045v1

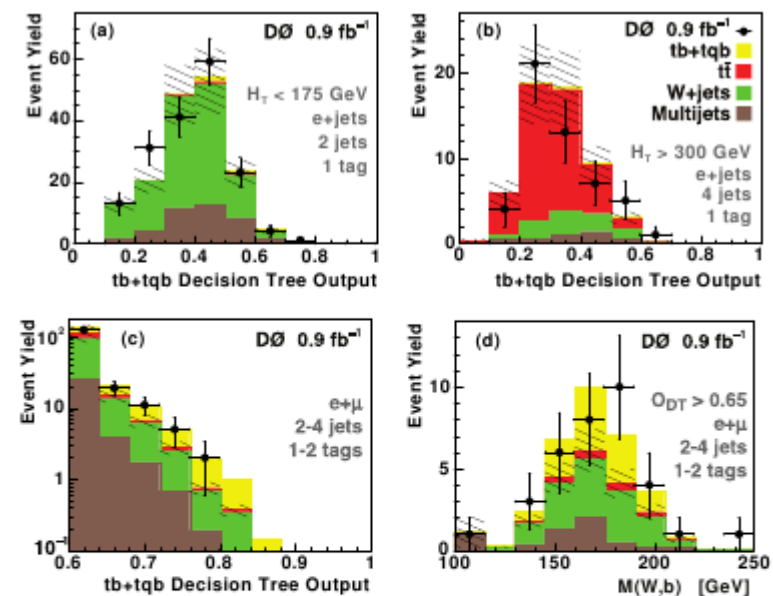


FIG. 2: Boosted decision tree output distributions for (a) a W +jets-dominated control sample, (b) a $t\bar{t}$ -dominated control sample, and (c) the high-discriminant region of the sum of all 12 $tb+tbq$ DTs. For (a) and (b), $H_T = E_T^\ell + \cancel{E}_T + \sum E_T^{\text{all jets}}$. Plot (d) shows the invariant mass of the reconstructed W boson and highest- p_T b -tagged jet for events with $O_{DT} > 0.65$. The hatched bands show the ± 1 standard deviation uncertainty on the background. The expected signal is shown using the measured cross section.

arXiv:hep-ex/0612052v2

Typical policy for Boosted trees

- Split criteria:

Define $P = \frac{\sum_s W_s}{\sum_s W_s + \sum_b W_b}$ as the purity

$$Gini = \left(\sum_{i=1}^n W_i \right) P(1 - P)$$

$Criterion = Gini_{father} - Gini_{son_1} - Gini_{son_2}$ is maximized at each step

- Boosting method:

- Adaboost (adaptative boost)

For each tree: $err = \frac{\sum_{events \text{ badly classified}} w_i}{\sum_{events} w_i}$ $\alpha = \beta \ln((1 - err)/err)$

$w_i \rightarrow w_i e^{\alpha}$ for badly assigned events, and renormalize the weights.

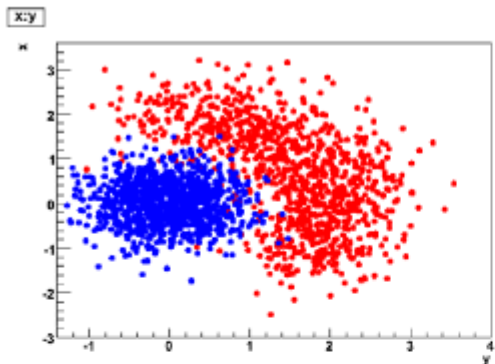
Event score: α -weighted average over the trees

- ϵ -boost (shrinkage)

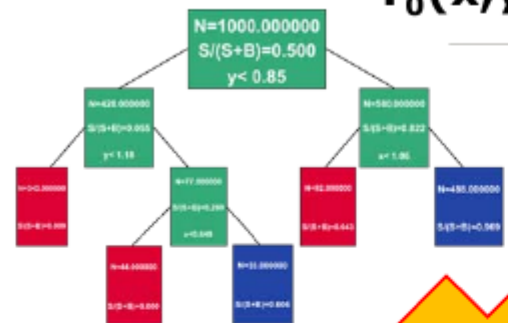
$w_i \rightarrow w_i e^{2\epsilon}$ for badly assigned events, and renormalize the weights.

Event score: renormalized but unweighted sum over the trees

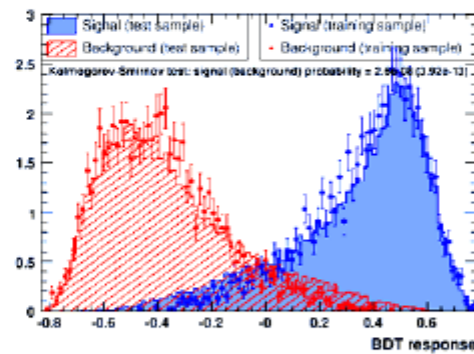
Example



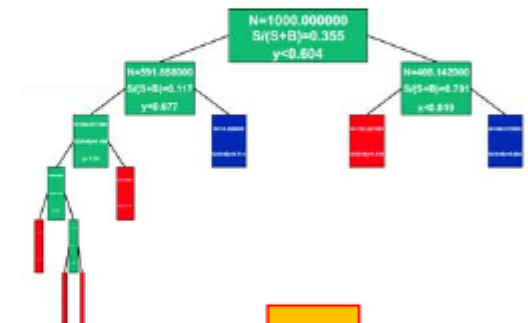
$T_0(x, y)$



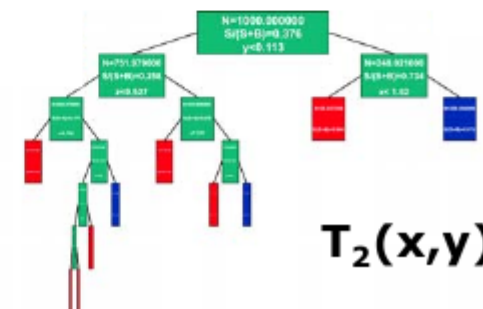
$$B(x, y) = \sum_{i=0}^4 \alpha_i T_i(x, y)$$



$T_1(x, y)$



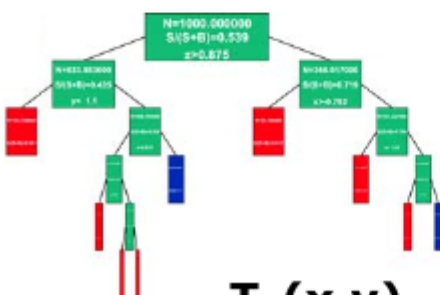
$T_2(x, y)$



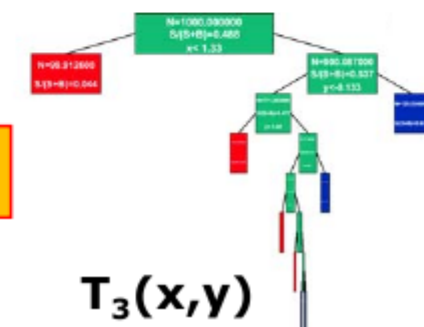
Wouter Verkerke, NIKHEF



$T_4(x, y)$



$T_3(x, y)$

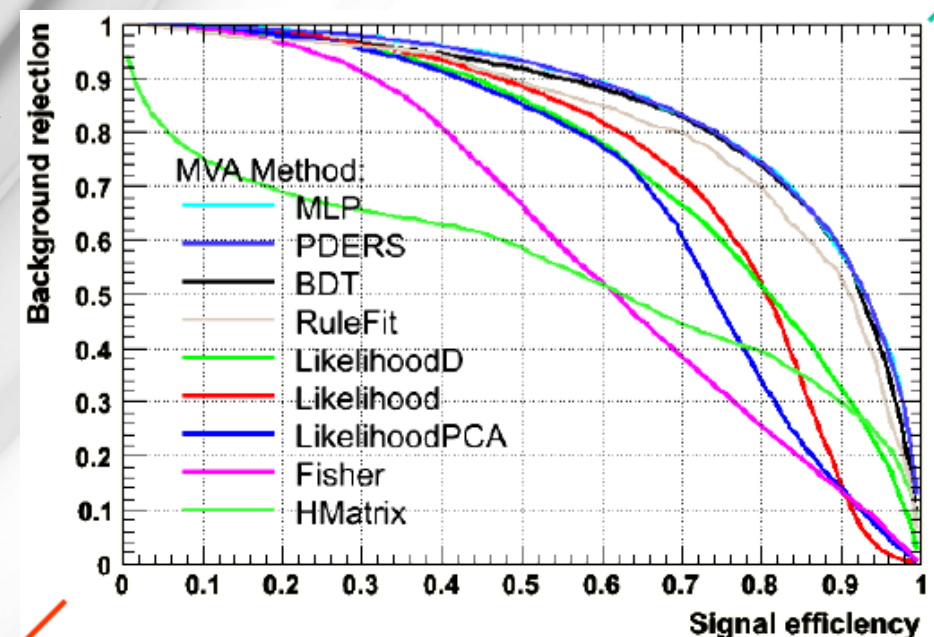
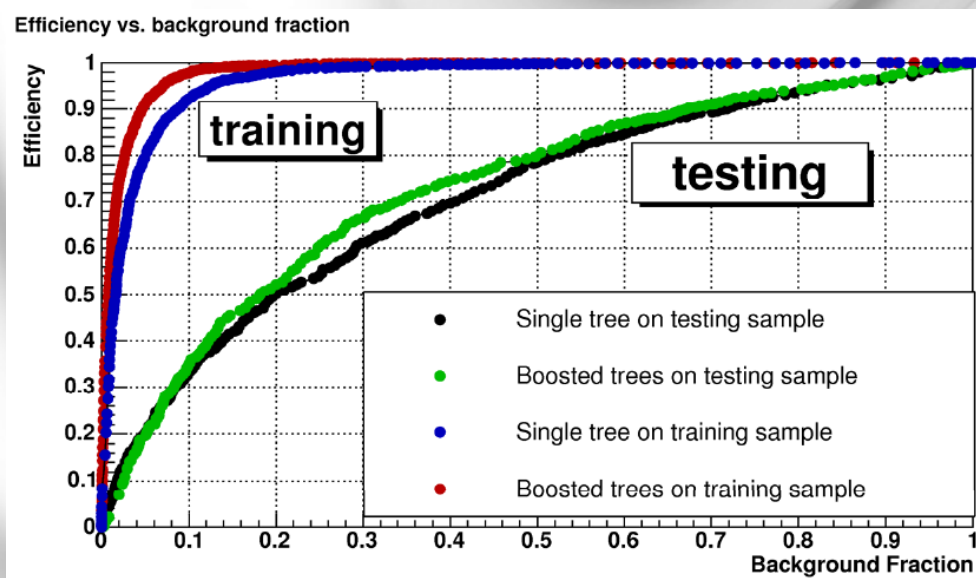


Comparing – Figure of merit

- The compactified output of a MVA is a single number on which we cut. We can use the same approach as for the evaluation of simple cuts.
- Efficiency vs rejection plot
- Figure of merit

Comparison of methods ←

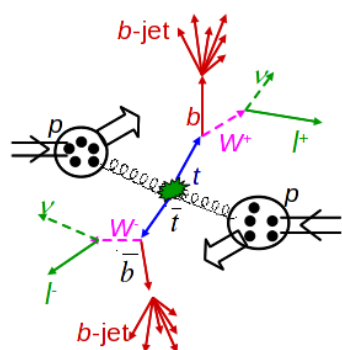
→ Comparison of training/test samples



Outline

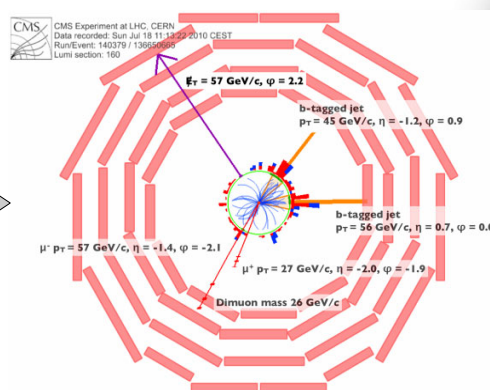
- Probability and Statistics, basic concepts
- Monte Carlo techniques
- Event classification
- **Parameter estimation**
 - χ^2 and ML estimators
 - Understanding MINUIT -> Do a proper fit with ROOT
 - Fit validation
- Limits, confidence intervals, significance
- Closing remarks

Parameter estimation



Theory

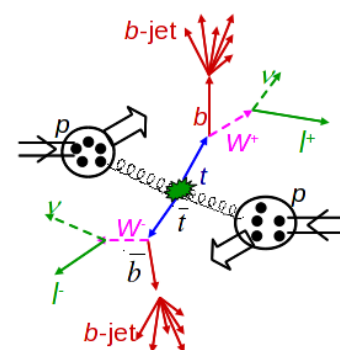
Probability



Experiment

Statistics

- Parameter estimation
- Fit



Theory

- Very common task: determine the underlying distribution for a measurement.
 - Determine the parameters of a pdf. -> parameter estimation
 - Common technique: fit
- χ^2 fit is well-known. Why does it work ?

- Definition: estimation is a procedure that leads to a result with a known imprecision
 - an ESTIMATION is NOT an approximation
- Parameter estimation is a test statistic and hence a random variable.
- Choice of an estimator requires judgement for particular application; there is no such thing as an “ideal estimator”
 - Think of the estimator of the mean (see later):
 - $1/N * \Sigma(x_i)$ for normally distributed measurements
 - $1/2 (\text{Max} + \text{Min})$ for uniformly distributed measurements
 - Truncated mean for energy loss measurements

- A perfect estimator is

- **Consistent** $\lim_{N \rightarrow \infty} \hat{a} = a$

- It approaches asymptotically the true value for large number of measurements.
- Convergence in the sense of probability: $\forall \epsilon, \lim_{N \rightarrow \infty} P(|\hat{a} - a| > \epsilon) \rightarrow 0$

- **Unbiased** $\langle \hat{a} \rangle = a$

- The expectation value of the estimator is the true value.
- An estimator that doesn't fulfill that criteria is said biased (or asymptotically unbiased).

- **Efficient**

- The variance is as small as possible.
- Meets the „minimal variance bound” (see later)

- A perfect estimator is
 - **Robust**
 - The estimator is insensitive against wrong data or wrong assumptions.
 - **Sufficient**
 - $dp(x|\hat{a})/da = 0$
 - Intuitively: it contains all information in the data concerning the parameter of interest.
- There is no perfect estimator...

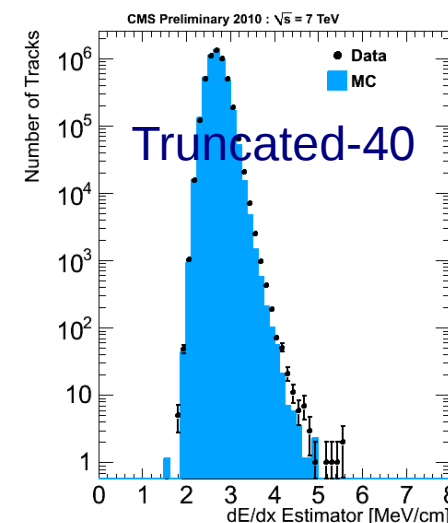
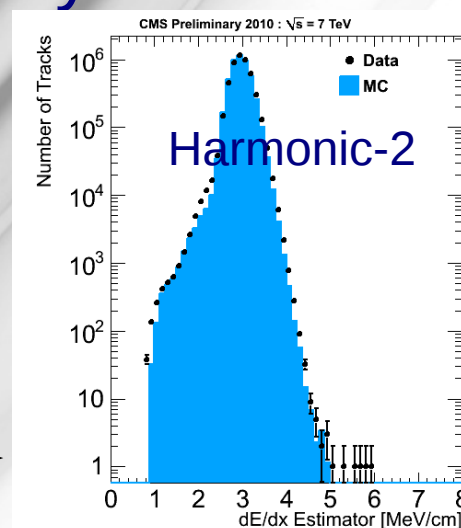
Example: mean and variance

- The arithmetic mean of a sample
 - is an **unbiased** estimator of the mean. $\langle \hat{\mu} \rangle = \mu$
 - Has **variance** given by $V(\hat{\mu}) = \frac{1}{N} \sigma^2$
- Other estimators are better in some cases:
 - Central value for uniformly distributed measurements are **more efficient**.
 - Truncated mean are **more robust** for dE/dx measurements

Check @ home!



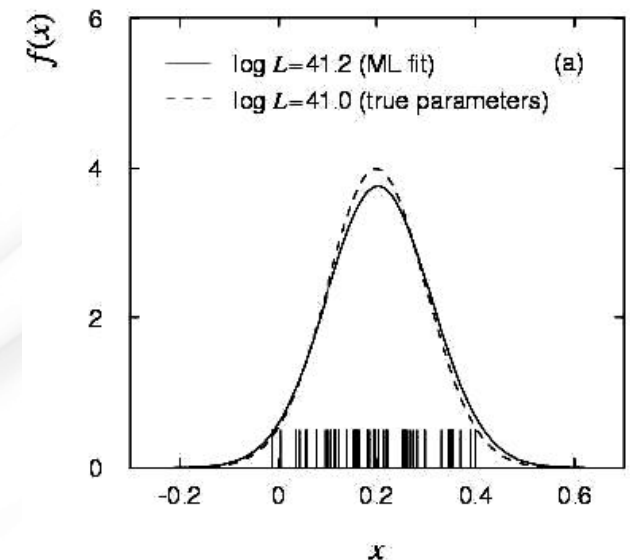
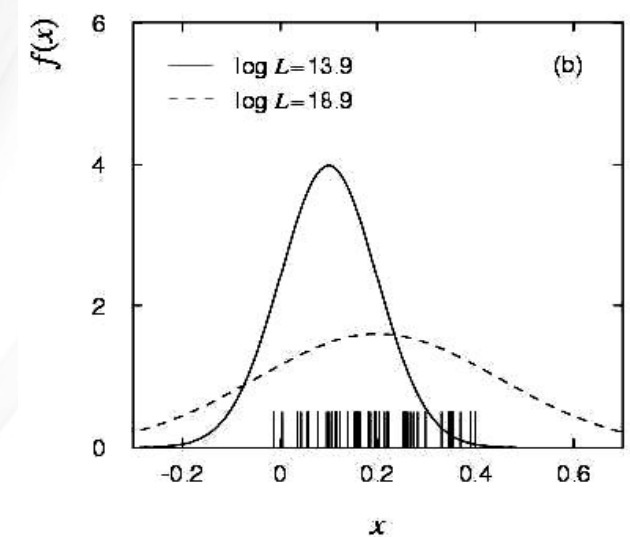
Standard CMS
dE/dx estimators:



- Beware variance vs sample variance: both are consistent but only sample variance is unbiased.

- The likelihood is the value of the pdf evaluated at the measured value.
 - For a dataset made of multiple points, the likelihood is the product of the individual likelihoods (joint pdf).
 - It is not a probability (that's why it is called likelihood).
 $\text{Probability} = L * \prod dx_i$
 - It is a test statistics that depends on the measurements.
 - Measures the probability to obtain exactly these data points x_i for a given parameter λ (assuming a known pdf).

- One defines the maximum likelihood (ML) estimate to be parameter value for which the likelihood is maximum
 - Might not be unique.
 - Gives the value of the parameter for which the data is the most likely (not the opposite)
 - Bayesian interpretation is different.
 - No goodness-of-fit.
 - The absolute value of L doesn't tell anything. If no value of p describes the data, the best value is still defined.





Some more definitions

- For practical reasons, one often considers the (negative) natural logarithm of the likelihood function.
- Definitions:

$L = \prod_i f(x^i, \lambda)$ is the likelihood function

↙ ↘ The probability: $dP = \prod_i f(x^i, \lambda) dx = L dx^N$

$l = \ln L = \sum_i \ln f(x^i, \lambda)$ is the log-likelihood

└─► $l' = \sum_i \frac{d}{d\lambda} \ln f(x^i, \lambda) = \sum_i \frac{f'}{f} = \sum_i \phi(x^i, \lambda)$

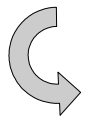
With ϕ the logarithmic derivative of f wrt λ

- Maximizing L is equivalent to minimizing $-\ln L = -l$
 - Sum over the measurements of the $\ln f(x^i|\lambda)$
- Sometimes, $-2 \ln L$ is considered, so that 1 standard deviation corresponds to an increase by 1 of that quantity (see in few slides).

Information inequality

- **Information inequality:** connection between bias and variance.
 - It's easy to achieve $\sigma^2(S) = 0$: take a constant value for S .
- Let's consider an estimator S from N measurements x^N .

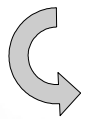
derive
w.r.t λ



$$E(S) = B(\lambda) + \lambda = \int S \prod_i f(x^i, \lambda) dx^i$$

$$E'(S) = B'(\lambda) + 1 = \int S \left(\sum_i \frac{f'(x^i, \lambda)}{f(x^i, \lambda)} \right) \prod_i f(x^i, \lambda) dx^i = E(Sl')$$

Cauchy-
Schwarz



$$1 + B'(\lambda) = E(Sl') = E(Sl') - E(S)E(l') = E[(S - E(S))l']$$

$$(1 + B'(\lambda))^2 \leq E[(S - E(S))^2] E(l'^2)$$

$$\sigma^2(S) \geq \frac{(1 + B'(\lambda))^2}{I(\lambda)}$$

Where we defined the **Fisher Information** $I(\lambda)$, having noticed that it is indep. of the dataset:

$$I(\lambda) = E(l'^2) = -E(l'') = E\left(\left(\sum \phi(x^i, \lambda)\right)^2\right) = E\left(\sum \phi(x^i, \lambda)^2\right) + E\left(\sum_{i \neq j} \phi(x^i, \lambda) \phi(x^j, \lambda)\right)$$

$$= N E\left(\left(\frac{f'(x, \lambda)}{f(x, \lambda)}\right)^2\right) \quad \underbrace{\hspace{10em}}_{=0 \text{ (} E(\phi)=0 \text{)}}$$



Minimum variance estimators

- The information inequality is also called the Cramer-Rao bound.

- It is saturated when the likelihood has the form:

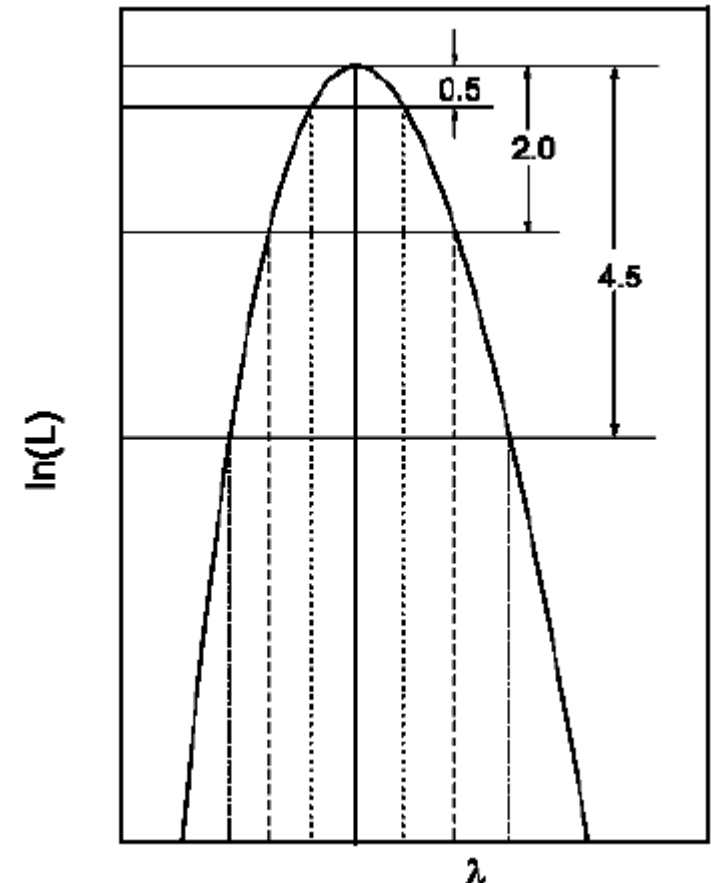
$$l' = A(\lambda)(S - E(S)) \quad \text{i.e.} \quad L = d \exp\{B(\lambda)S + C(\lambda)\}$$

→ „Minimum variance” estimators,
„Efficient” estimators

- In which case, $\sigma^2(S) = \frac{1}{|A(\lambda)|}$
- ML estimators
 - Are as efficient as it can be
 - If there is an efficient estimator, it will be found in most cases.
 - The variance of the estimator is then minimal.
 - Are consistent.
 - Are often asymptotically unbiased.

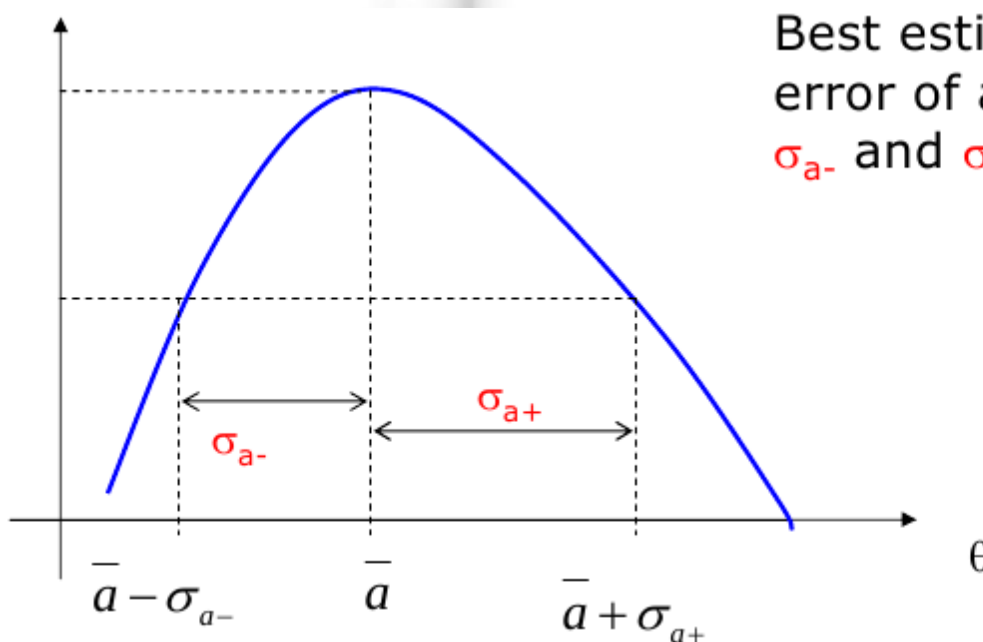
- ML is well suited to error estimate
 - As a first approximation, the information inequality gives a lower bound.
 - Alternatively, note that the likelihood function is Gaussian near the maximum
 - Taylor expansion of $-\ln L$ with 1st derivative equal to 0
 - The variance is the inverse of the second derivative estimated at the maximum.

$$l(\hat{a} \pm n.\sigma) = l(\hat{a}) - \frac{n^2}{2}$$



Note on errors

- Graphical method is thus very useful in practice



Asymmetric errors obtained by taking the points at $\ln L = \max - \frac{1}{2}$

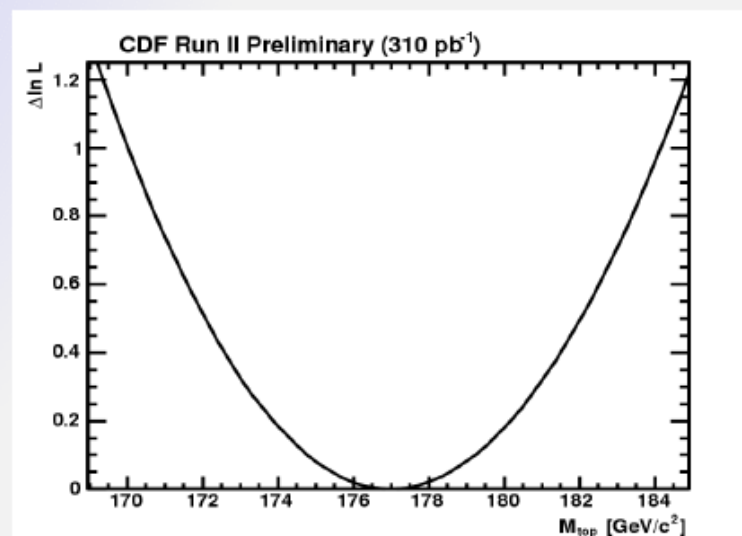
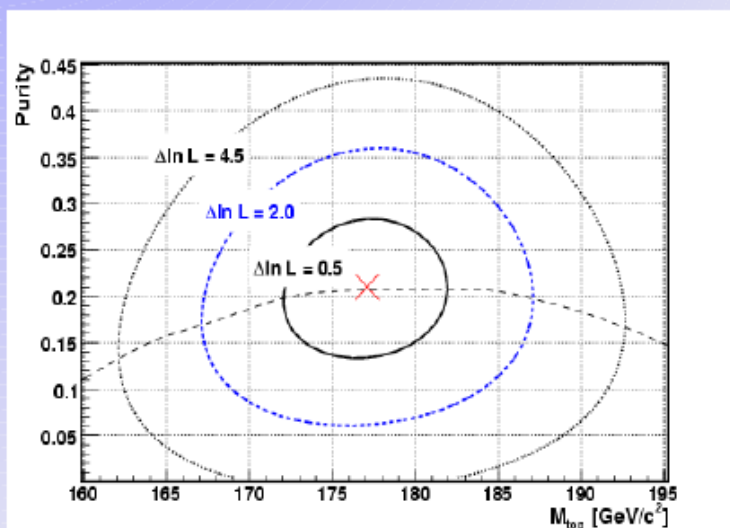
Can be easily extended to multi-dimensional cases

→ Error ellipses or L iso-contours.

- However, be careful in interpreting such intervals as 68% confidence intervals as coverage is not guaranteed. (see later discussion on intervals)

Results

- ★ From the observed 290 data events we fit a signal fraction $P_s = 0.21 \pm 0.07$
- ★ It corresponds to 61 ± 20 signal events
- ★ Compatible with the all hadronic cross section measurement and the purity fitted using Monte Carlo



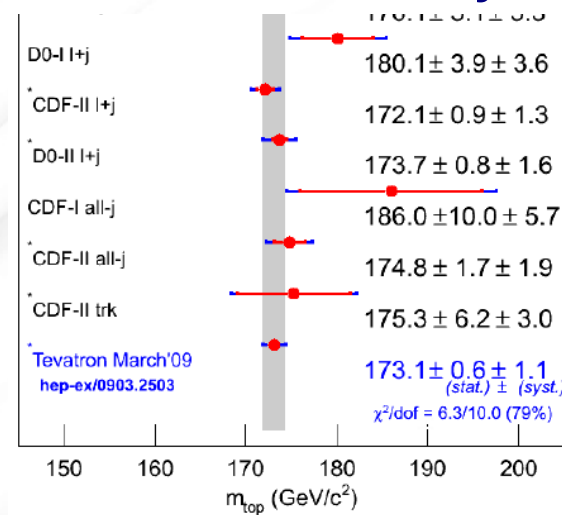
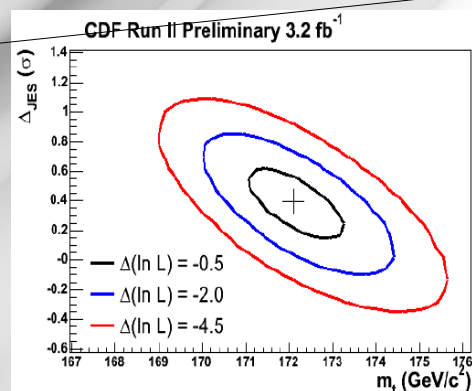
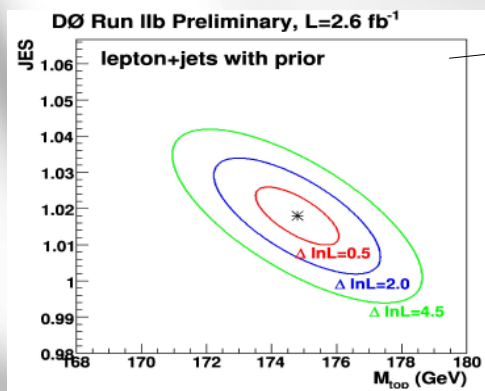
- ★ The measured mass value amount to

$$M_t = 177.1 \pm 4.9 (\text{stat}) \pm 4.7 (\text{syst}) \text{ GeV}/c^2$$

- ★ Break down of systematics will follow

Combining measurements

- The use of the likelihood function makes the combination of independent measurements easy:
 - 2 or more experiments do independent measurements
 - Can be different quantities (different likelihood functions) as long as they depend on the same parameter(s)
 - Combine them by multiplying the likelihood functions
 - Alternatively sum the log-likelihoods.
 - If measurements are compatible, the error will be obviously reduced.



Extended ML method

- Maximum Likelihood method does **not** consider the **normalization** of the function.
 - The likelihood is made of normalized pdfs.
- For cases where the normalization matters, i.e. **when the mean total number of events itself is a parameter**, the Extended Likelihood function is needed:

$$L(\nu, a) = \frac{\nu^n}{n!} e^{-\nu} \prod_i f(x_i, a) = \frac{e^{-\nu}}{n!} \prod_i \nu f(x_i, a)$$

$$\ln L = -\nu(a) + \sum_i \ln(\nu(a) f(x_i, a)) \quad \nu \text{ might depend on } a !$$

- Data might also contain **contributions from different sources** (signal, background, ...), in which case the total number of events is described as a sum of individual contributions:

$$\ln L(\mu) = \sum_{j=1}^m \mu_j + \sum_{i=1}^n \ln \left(\sum_{j=1}^m \mu_j f_j(x_i) \right) \quad \longrightarrow \text{Can fit the yield of a signal !}$$

Binned ML method

- Evaluating the likelihood on a large statistics can be expensive -> **Binned likelihood fit.**
 - Fill an histogram
 - Consider each bin as an independent Poisson-distributed measurement.

$$L = \prod_i \frac{\nu_i^{d_i}}{d_i!} e^{-\nu_i}$$

with d_i the data content of bin i and $\nu_i(\theta)$ the expected bin content.

$$-2 \ln L = 2 \sum_i \nu_i - d_i + d_i \ln \frac{d_i}{\nu_i}$$

Where we used $\ln(d!) \approx d \ln d - d$

- **Benefits/properties:**
 - Goodness-of-fit test is possible
 - Empty bins are properly handled
 - Fit integral is fixed: $\sum_i \nu_i = \sum_i d_i$

From the ML to the χ^2 estimator

- In the limit of high statistics in each bin, we can do the same using a Gaussian pdf:

$$L = \prod_i \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(d_i - \mu_i)^2}{2\sigma_i^2}}, \text{ with } \mu_i = F(x_i, \theta)$$

$$-2\ln L = \underbrace{\sum_i \ln(2\pi\sigma_i)}_{\text{Cte w.r.t. } \theta} + \sum_i \frac{(d_i - \mu_i)^2}{2\sigma_i^2}$$

- Minimizing the log-likelihood is then equivalent to minimizing the χ^2 !
- It gives an input for the χ^2 validity: $N > 5 \sim 10$ in each bin.
- Practically, one will also compute the standard deviation as:

$$\sigma_i = \frac{1}{\sqrt{\sum_j w_j}} = \frac{1}{\sqrt{N}} \quad \text{for unweighted entries.}$$



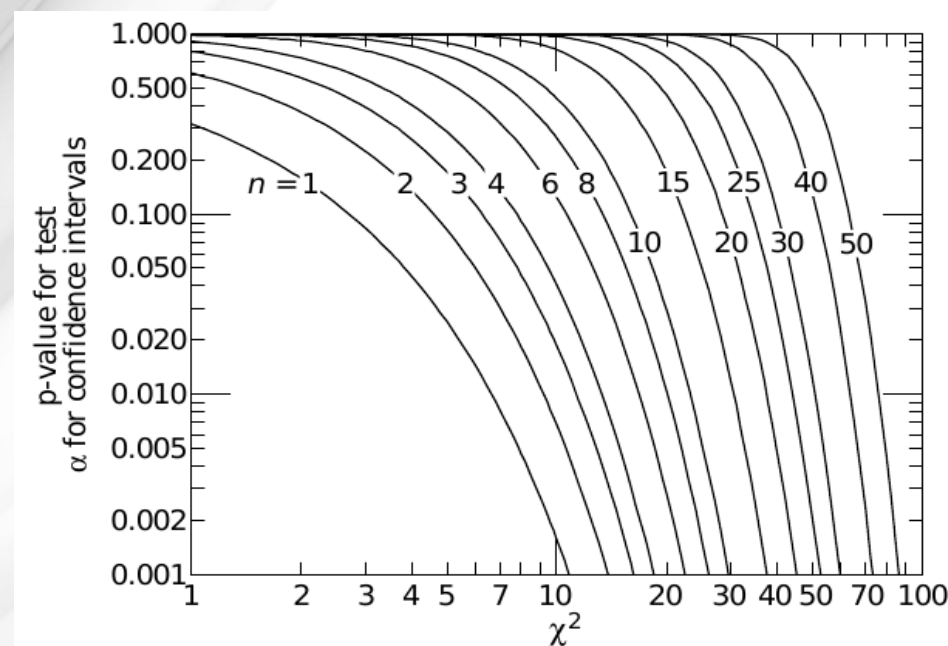
Goodness of fit



- The minimum value of the χ^2 is a test statistics that can be used as a measure of the discrepancy between the data and the model used for the fit.
- More precisely, one uses the Pearson's χ^2 statistic:

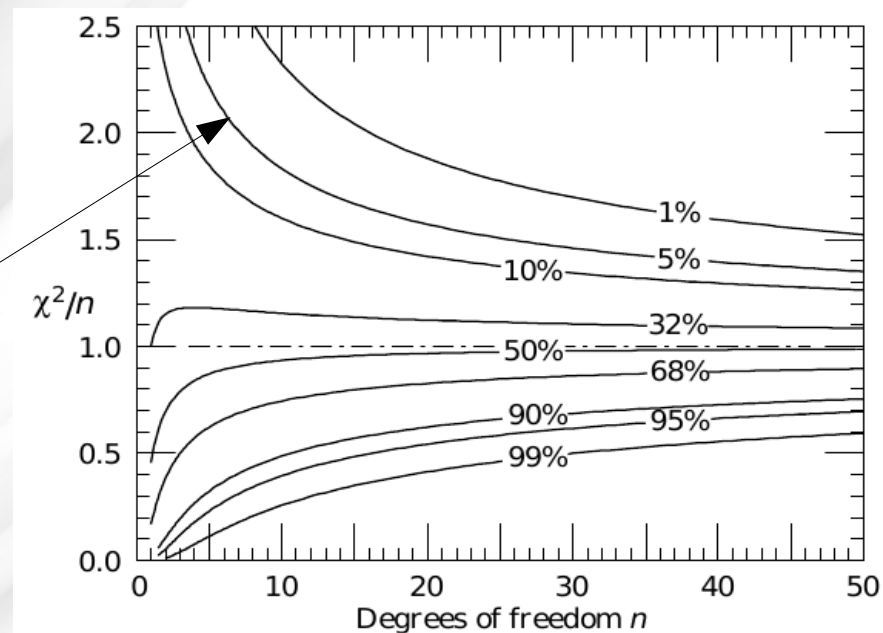
$$\chi^2 = \sum_i \frac{(n_i - \nu_i)^2}{\nu_i}$$

- If $n_i > 5 \sim 10$, that statistics follows the χ^2 distribution
 - Interpretation: use the p-values from the χ^2 (TMath::Prob).
- If the condition do not hold, the pdf has to be determined (e.g. by toy-mc) before computing the p-value.

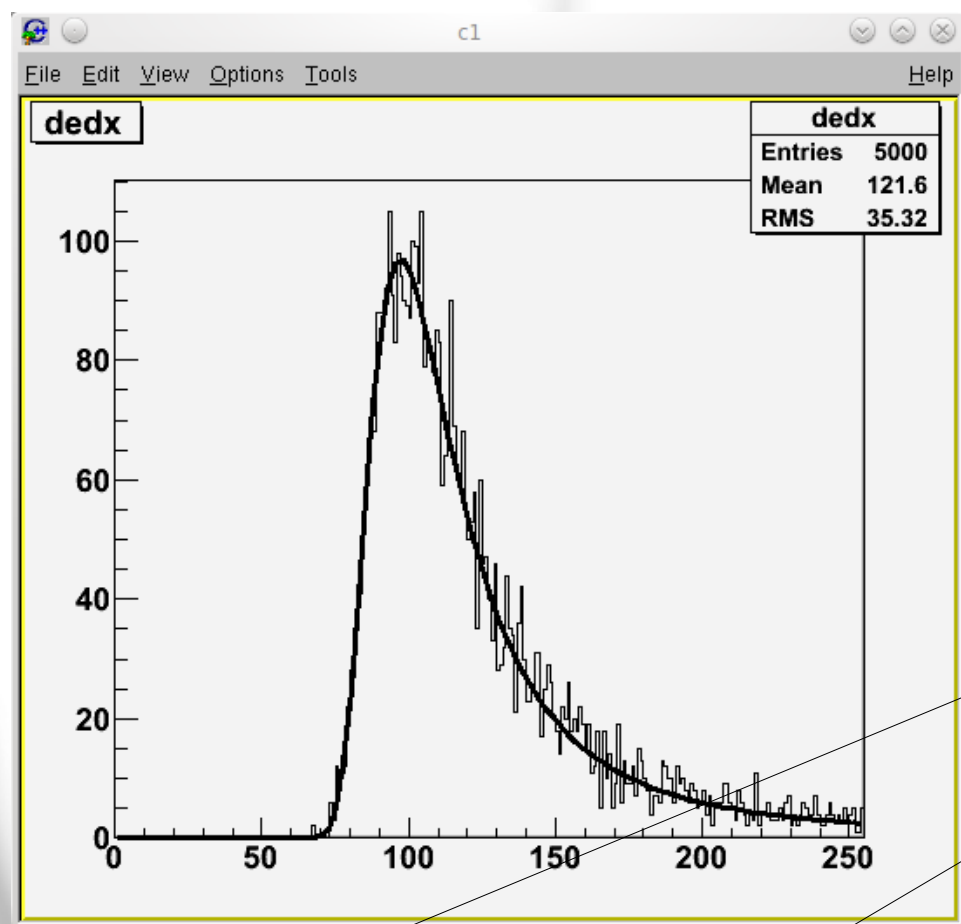


Goodness of fit (2)

- Too large χ^2 -> might be
 - Wrong model
 - Wrong estimate of the uncertainties (+ systematics)
 - In this case, the parameters estimate might be fine.
- Too small χ^2 -> overestimated uncertainties ?
- It's a common mistake to quote only χ^2/ndof as a measure of the fit quality.
 - Quantiles of the normalized χ^2 still depend on ndof !



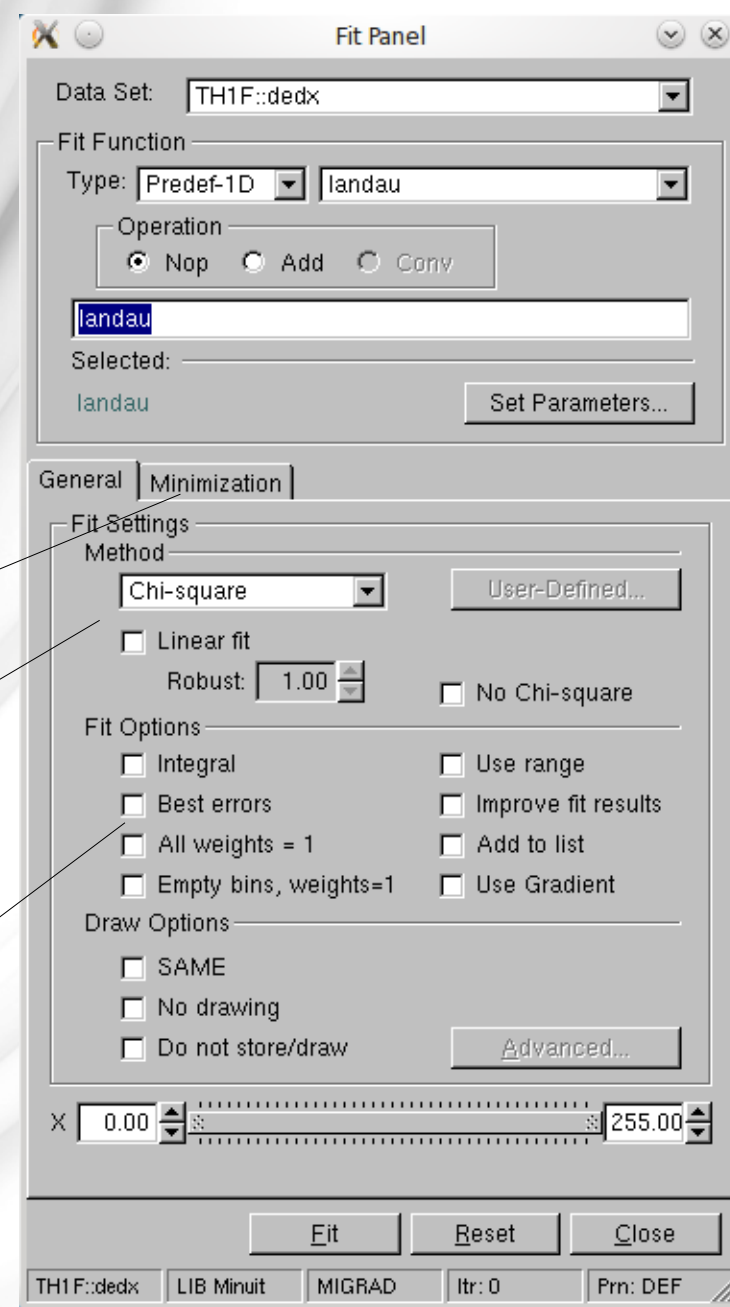
Fitting in ROOT



Verbosity switch
hidden there.

Optionnally do a more robust
computation of the errors.

Use χ^2 or binned ML fit



Fit Panel

Data Set: TH1F::dedx

Fit Function

Type: Predef-1D landau

Operation

☒ Nop ☐ Add ☐ Conv

landau

Selected:

landau

Set Parameters...

General Minimization

Fit Settings

Method

Chi-square

User-Defined...

☐ Linear fit

Robust: 1.00

☐ No Chi-square

Fit Options

☐ Integral ☐ Use range

☐ Best errors ☐ Improve fit results

☐ All weights = 1 ☐ Add to list

☐ Empty bins, weights=1 ☐ Use Gradient

Draw Options

☐ SAME

☐ No drawing

☐ Do not store/draw

Advanced...

X 0.00 255.00

Fit Reset Close

TH1F::dedx LIB Minuit MIGRAD ltr: 0 Prn: DEF

MIGRAD function: find minimum

Minimized function.
Likelihood or χ^2

Fit status: converged, failed or failed.
Error matrix: accurate or approximate
Estimated distance to minimum: small...

```
FCN=- 14027 FROM MIGRAD STATUS=CONVERGED 49 CALLS 50 TOTAL
EDM=1.07153e-09 STRATEGY= 1 ERROR MATRIX ACCURATE

EXT PARAMETER
NO. NAME VALUE ERROR STEP FIRST
1 Constant 5.31237e+02 1.08972e+01 6.14438e-01 -1.55459e-07
2 MPV 1.00106e+02 3.13941e-01 1.93106e-02 -5.71430e-05
3 Sigma 1.01814e+01 1.66848e-01 1.35878e-04 -6.79975e-03
ERR DEF= 0.5

EXTERNAL ERROR MATRIX. NDIM= 25 NPAR= 3 ERR DEF=0.5
1.187e+02 -1.617e+00 -1.317e+00
-1.617e+00 9.856e-02 3.452e-02
-1.317e+00 3.452e-02 2.784e-02
PARAMETER CORRELATION COEFFICIENTS
NO. GLOBAL 1 2 3
1 0.72436 1.000 -0.473 -0.724
2 0.65905 -0.473 1.000 0.659
3 0.80860 -0.724 0.659 1.000
```

Parameter values, plus their error.
Also shown is the error definition.
Here: 0.5 for a likelihood fit.

MIGRAD function: find minimum

```
FCN=- 14027 FROM MIGRAD    STATUS=CONVERGED    49 CALLS    50 TOTAL
                        EDM=1.07153e-09    STRATEGY= 1    ERROR MATRIX ACCURATE

EXT  PARAMETER
NO.   NAME      VALUE      ERROR      STEP      FIRST
      NAME      VALUE      ERROR      SIZE      DERIVATIVE
  1  Constant   5.31237e+02  1.08972e+01  6.14438e-01 -1.55459e-07
  2  MPV        1.00106e+02  3.13941e-01  1.93106e-02 -5.71430e-05
  3  Sigma      1.01814e+01  1.66848e-01  1.35878e-04 -6.79975e-03

                        ERR DEF= 0.5
```

```
EXTERNAL ERROR MATRIX.    NDIM= 25    NPAR= 3    ERR DEF=0.5
  1.187e+02 -1.617e+00 -1.317e+00
-1.617e+00  9.856e-02  3.452e-02
-1.317e+00  3.452e-02  2.784e-02
PARAMETER CORRELATION COEFFICIENTS
  NO.  GLOBAL      1      2      3
    1  0.72436    1.000 -0.473 -0.724
    2  0.65905   -0.473  1.000  0.659
    3  0.80860   -0.724  0.659  1.000
```

Approximate error matrix.
Approximate covariance matrix.

HESSE function

Error computed from the second derivative of $-\ln L$ or χ^2

→ Quadratic approximation...

```

FCN=-14027 FROM HESSE      STATUS=OK      16 CALLS      60 TOTAL
                        EDM=5.16509e-12    STRATEGY= 1    ERROR MATRIX ACCURATE
EXT PARAMETER
NO.   NAME      VALUE      ERROR      INTERNAL      INTERNAL
      STEP SIZE  VALUE
  1 Constant    5.31237e+02  1.08945e+01  2.45775e-02  5.31237e+02
  2 MPV         1.00106e+02  3.13847e-01  9.82439e-04  1.00106e+02
  3 Sigma       1.01814e+01  1.66786e-01  5.43504e-06  -1.22952e+00
ERR DEF= 0.5
EXTERNAL ERROR MATRIX.  NDIM=  3  NPAR=  3  ERR DEF=0.5
  1.187e+02 -1.615e+00 -1.316e+00
 -1.615e+00  9.850e-02  3.448e-02
 -1.316e+00  3.448e-02  2.782e-02
PARAMETER CORRELATION COEFFICIENTS
NO.  GLOBAL      1      2      3
  1  0.72419    1.000 -0.472 -0.724
  2  0.65880   -0.472  1.000  0.659
  3  0.80844   -0.724  0.659  1.000
    
```

Covariance matrix computed from

$$V_{ij} = \left(\frac{d^2(-\ln L)}{dp_i dp_j} \right)^{-1}$$

Symmetric errors from the second derivative of $-\ln(L)$ or χ^2

MINOS function

- MINOS errors are calculated by 'hill climbing algorithm'.
 - In one dimension find points where $\Delta L = +0.5$.
 - In >1 dimension find contour with $\Delta L = +0.5$. Errors are defined by bounding box of contour.
 - In $>>1$ dimension very time consuming, but more in general more robust.
- Optional – activated by option “E”

```

FCN=-14027 FROM MINOS      STATUS=SUCCESSFUL      40 CALLS      160 TOTAL
                        EDM=5.16509e-12      STRATEGY= 1      ERROR MATRIX ACCURATE

EXT  PARAMETER
NO.   NAME      VALUE      PARABOLIC      MINOS ERRORS
      NAME      VALUE      ERROR      NEGATIVE      POSITIVE
  1  Constant    5.31237e+02    1.08945e+01    -1.07972e+01    1.09924e+01
  2  MPV         1.00106e+02    3.13847e-01    -3.12329e-01    3.15371e-01
  3  Sigma       1.01814e+01    1.66786e-01    -1.65031e-01    1.68564e-01

ERR DEF= 0.5

PARAMETER  CORRELATION COEFFICIENTS
NO.  GLOBAL      1      2      3
  1  0.72419    1.000  -0.472  -0.724
  2  0.65880   -0.472  1.000   0.659
  3  0.80844   -0.724  0.659   1.000
  
```

Parabolic errors
repeated from HESSE

Asymmetric errors
computed by hill climbing the $-\ln(L)$

Fit validation

- In general, you will have to prove/show that:
 - The fit is not biased
 - The error is properly evaluated.
- This is especially true for low statistics...
- Natural tool: toy MC.
 - Perform many ($O(100-1000)$) pseudo experiments
 - Fit and compare to the simulated value
 - Measure the difference between fit and true \rightarrow bias
 - Measure the spread \rightarrow error
 - Usual plot used to convey that information: Pull
 - $(\text{fit}-\text{true})/\text{error}$... should be a normal distribution.

What should I use ?

- **χ^2 fit, fastest & easiest**

- Gives absolute goodness-of-fit indication
- Makes (incorrect) Gaussian error assumption on low statistics bins
- Misses information with feature size < bin size



- **Maximum Likelihood estimators, most robust**

- Valid at low statistics
- No information lost due to binning
- Gives best error of all methods (especially at low statistics)
- No intrinsic goodness-of-fit measure
- Can be computationally expensive for large N



- **Binned Maximum Likelihood, in between**

- Much faster than full Maximum Likelihood
- Correct Poisson treatment of low statistics bins
- Misses information with feature size < bin size



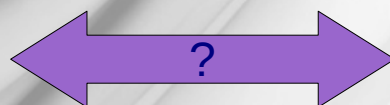
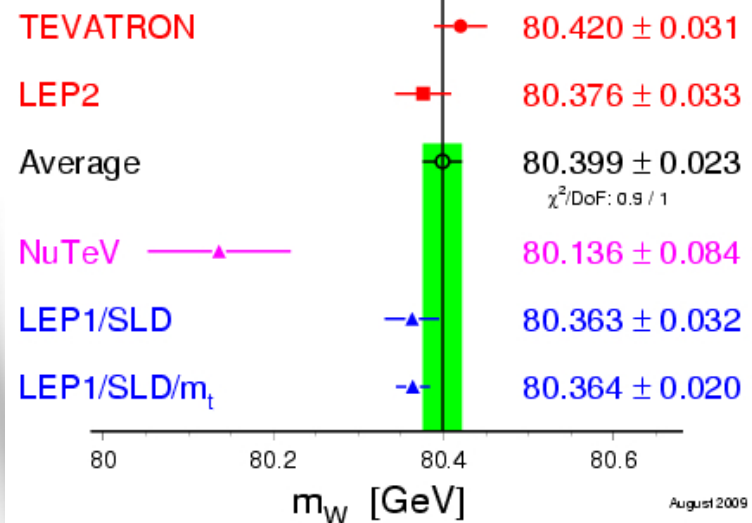
Outline

- Probability and Statistics, basic concepts
- Monte Carlo techniques
- Event classification
- Parameter estimation
- Limits, confidence intervals, significance
 - Confidence. Definition and simple example.
 - Frequentist approach(es)
 - Bayesian intervals
 - Likelihood intervals
 - The „CLs” method
- Closing remarks

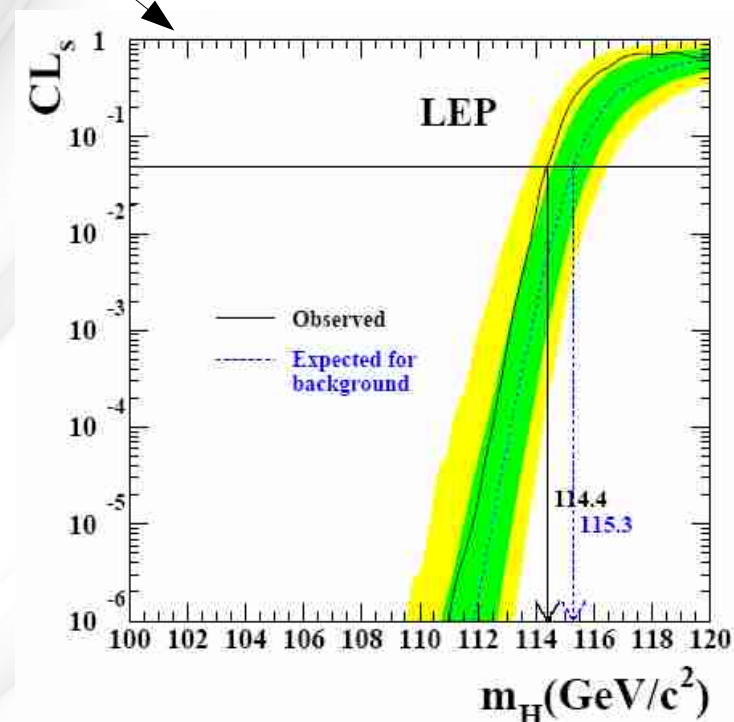
The problem...

For a given „measurement”,
determine a confidence interval or a limit

W-Boson Mass [GeV]



When to go for one
or the other ?



How do we DEFINE the interval/limit ?

How to INTERPRET the result ?

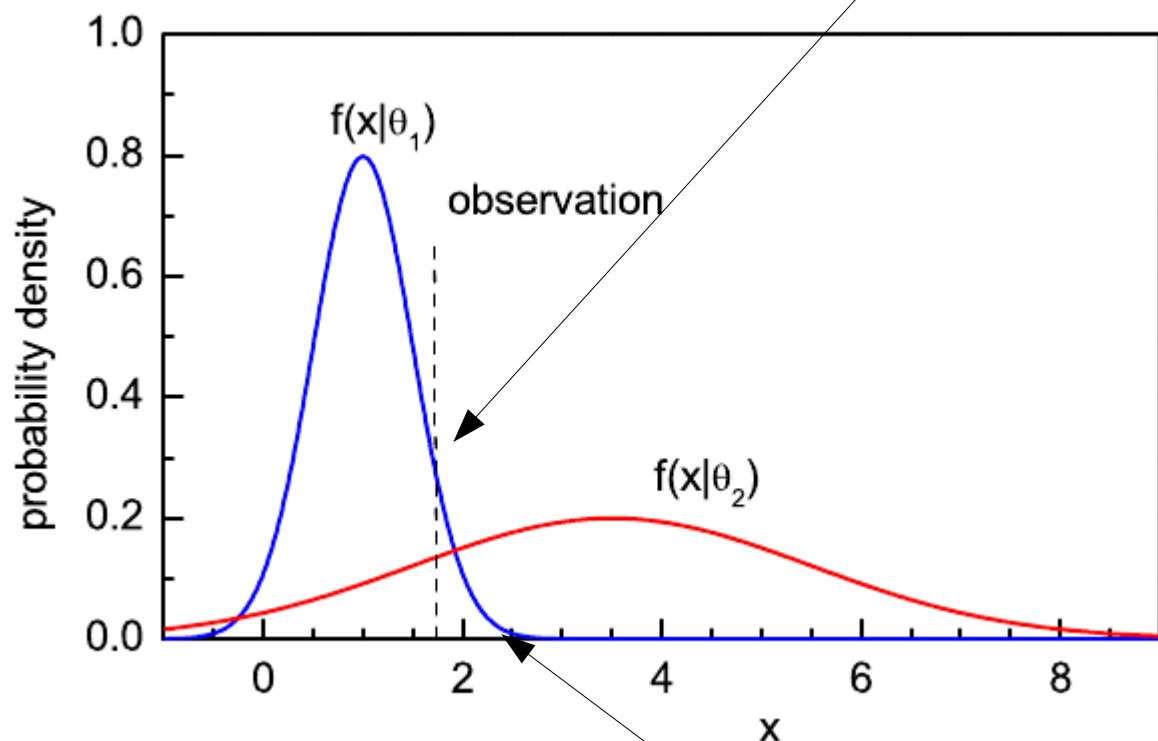
What are the limitations of the various approaches ?

► Very lively field of research !

The difficulty...

As a simple example we imagine an observation x of a variate (random variable) X and a probability distribution function (pdf) $f(X|\theta)$ depending on an unknown parameter θ which we estimate from x . How should we select the range of parameters which we consider compatible with data ?

The likelihood of θ_1 is higher than the one of $\theta_2 \rightarrow$ favor θ_1 .



This will naturally lead to very different confidence intervals, depending on the chosen approach.

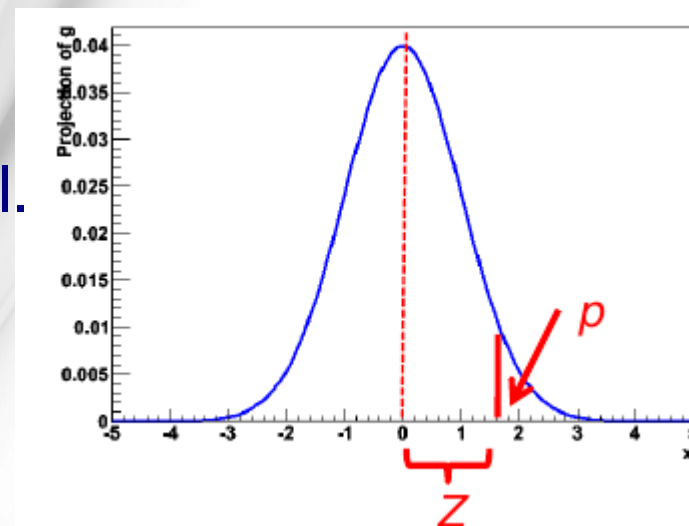
The measurement x is more than 2σ apart in the θ_1 hypothesis.
 \rightarrow favor θ_2 .

Calculating significance

- We are interested by the „probability to be wrong”
 - p-value is the probability to sit in the tail.
 - Expressed as quantiles of a normal distribution (sometimes known a Z-value).

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z) \quad \text{TMath::Prob}$$

$$Z = \Phi^{-1}(1 - p) \quad \text{TMath::NormQuantile}$$



This defines single-sided significance.

Note: for a measurement where deviations are possible on each side, consider double-sided significance (factor 2).

Discovery ?

5σ !

HEP tradition... very subjective.

$$P = 2.87 \cdot 10^{-7}$$

Might be too much for known processes.

- CMS „saw” the jj with few std deviations.
- Might be too low for totally unexpected effects
- SETI signal ? Telepathy ?

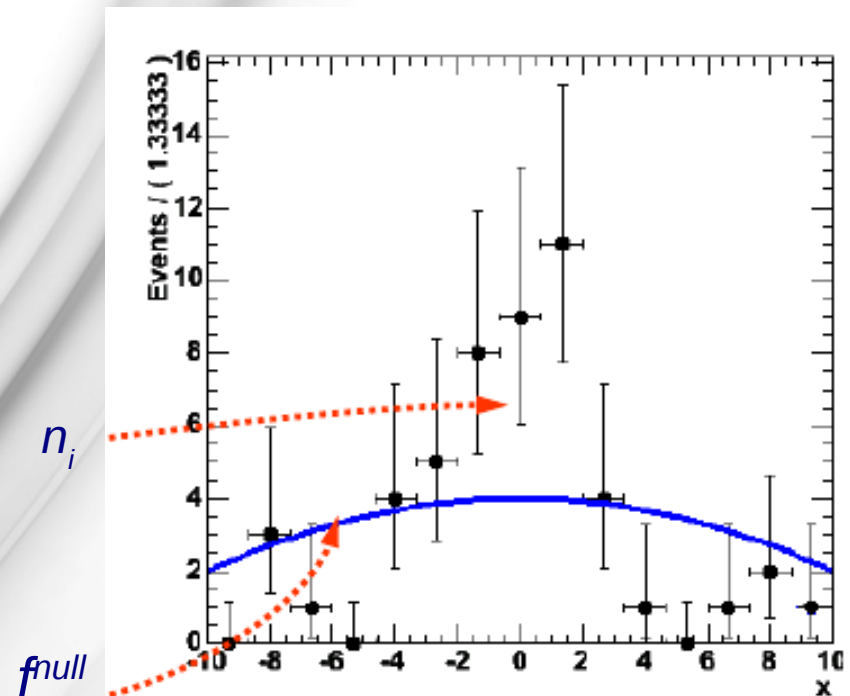
Simple example: Pearson's test

- Pearson's χ^2 test
 - Calculate χ^2 of data w.r.t. null hypothesis ($s=0$)

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - f_i^{null})^2}{f_i^{null}}$$

- The P-value is given by

$$P(\chi^2; N) = \int_{\chi^2}^{\infty} p(\chi^2; N) d\chi^2$$

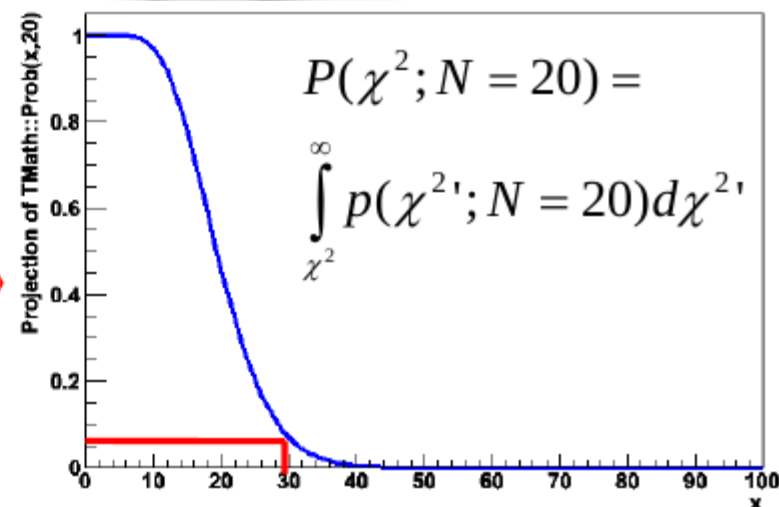
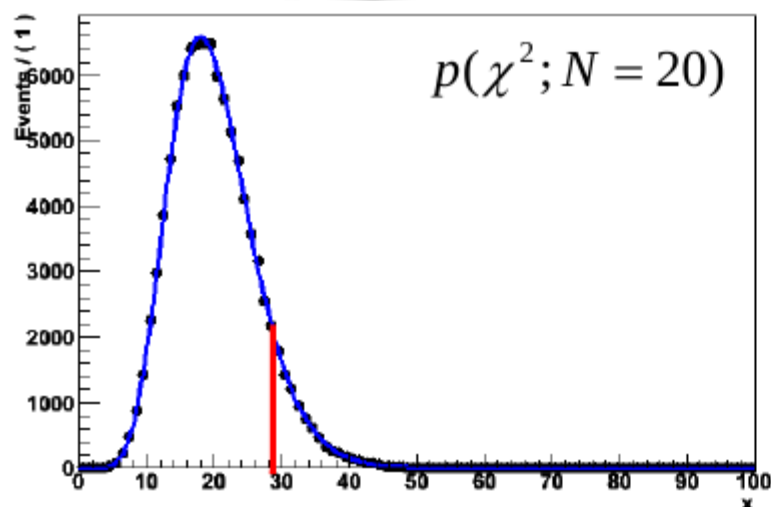


Computed with TMath::Prob()

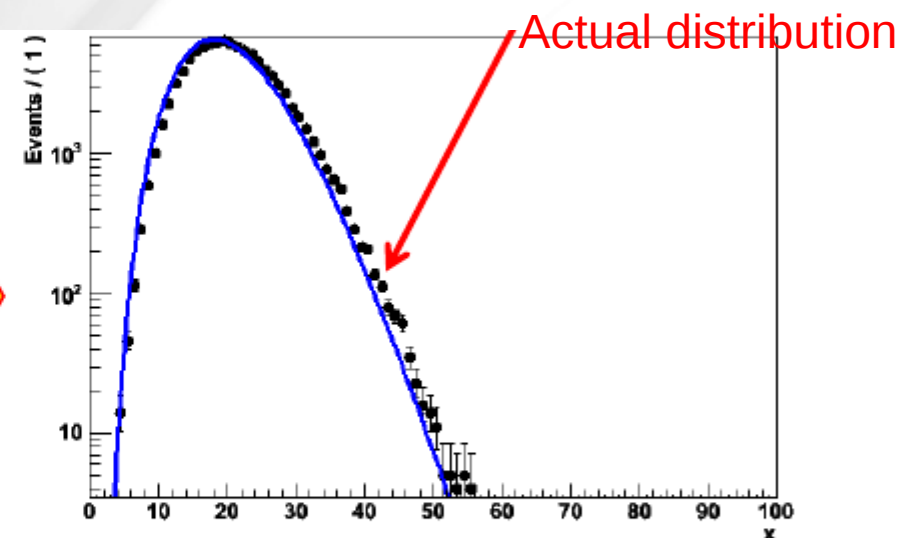
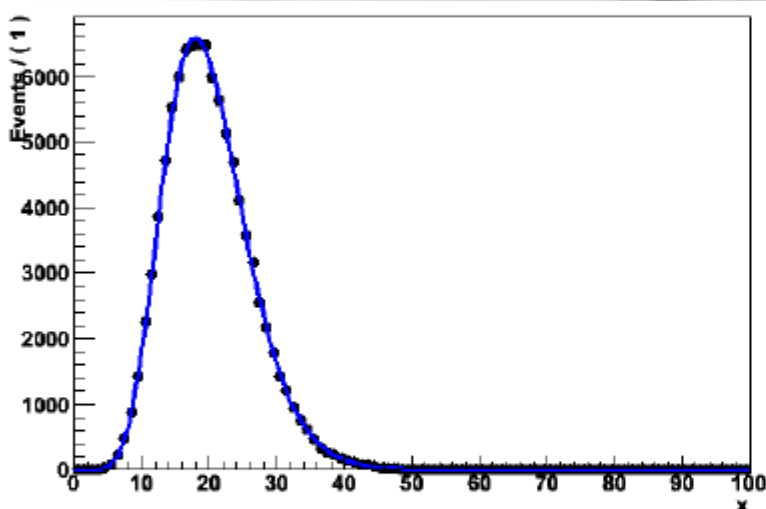
- That $P(\chi^2)$ probability interpretation is only valid for normal sampling. This is not valid if the CLT does not hold.
 - If the n_i are not Gaussian distributed, p will not follow a χ^2 pdf.

P-value

The p-value calculation made before assumes a χ^2 pdf.



If it's not the case, you will get a wrong value ! → MC calibration possible.



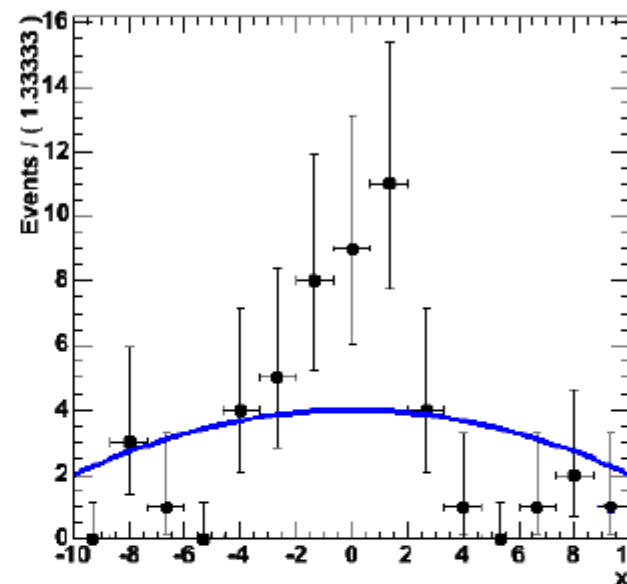


Calibration



Will usually get worse in the tails !!!

- Result of calibration applied to example
 - In p-values $0.073 \rightarrow 0.11$
 - In Z-values $1.45\sigma \rightarrow 1.22\sigma$
- Why was this particular example so bad?
 - χ^2 test assumes Gaussian errors.
 - Poisson errors on bins with $N < 10$ deviate quite strongly from Gaussian assumption
- How much MC do you need to calibrate?
 - You need roughly $100/p$ -value experiments to get reasonable accuracy in calibration at that p-value
 - In this example $p=0.1 \rightarrow$ About 1000 toy experiments (generated data + fit data) are sufficient, quite doable
- How about $Z=5$?
 - $P=2 \cdot 10^{-7} \rightarrow$ You need $O(500 \text{ million})$ experiments! At 1 second per expt and 100 dedicated CPUs this still takes 2 months!

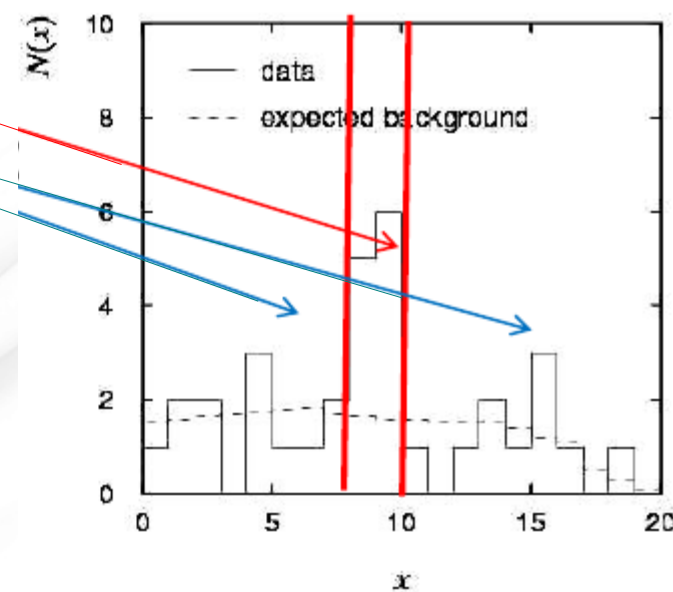




Significance at high Z



- Calibrating any method at high Z can be **very expensive at high Z**
 - Excludes practical application of the method.
- In the recent years, lots of significance calculations on MC (Higgs, susy, LEDS, ...)
 - Calculation depends on the problem. Cannot be done once for all.
 - Other approaches often used, e.g. **Sideband subtraction**.
- Example:
 - **Signal region**: $N_{\text{obs}} = s+b = 11$
 - **Sideband region**: $N'_{\text{obs}} = t.b = 25$ ($t=18/2$)
 - Less sensitivity, but can be applied to high significance.
 - There, b is a *nuisance parameter*
 - Random variable correlated with the signal, but of no peculiar interest.



Profile likelihood (1)

Profile likelihood ratio is one possible way to cope with nuisance parameters. It is constructed from the likelihood function as:

$$\lambda(s) = \frac{L(s, \widehat{b}(s))}{L(\hat{s}, \hat{b})}$$

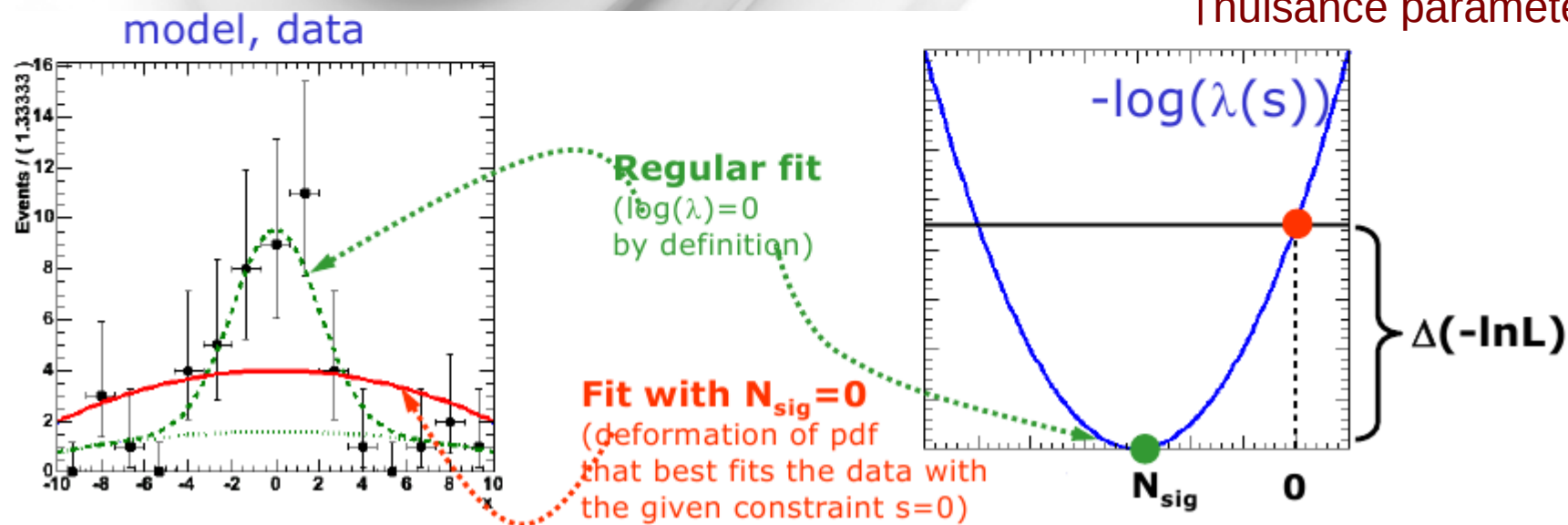
Minimum of L for a given value of s .

Global minimum of L

➔ Easy to figure out in the case of a fit with signal and background shapes

$$L(s, b, \vec{p}, \vec{q}) = \prod_{i=1}^N [s \cdot F_s(\vec{x}; \vec{p}) + b \cdot F_B(\vec{x}; \vec{q})]$$

Parameters of the shape functions are additional nuisance parameters !



Profile Likelihood (2)

- Calculate significance assuming normal sampling distributions

$$\Delta \log L = \frac{1}{2} Z^2$$

- Profile Likelihood works very well for example with shapes, but is not proven to be well calibrated at $Z=5$
- Can also apply technique to counting exp with sideband

$$L = \text{Poisson}(x_{\text{obs}} | s + b) \cdot \text{Poisson}(x'_{\text{obs}} | \tau \cdot b)$$

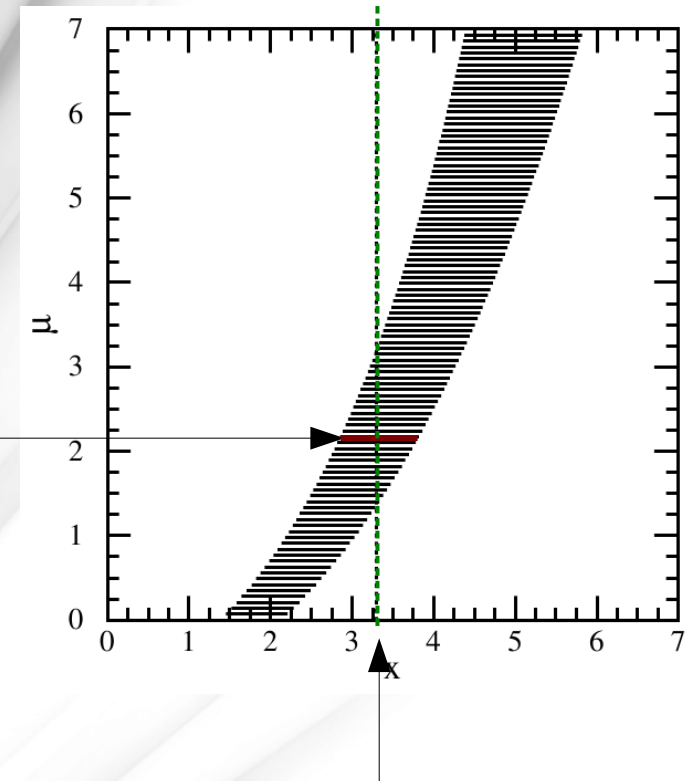
- Parameter of interest = s
- Nuisance parameter = b (τ is assumed to be known exactly)
- Advantage: standard form that can be tested / calibrated independent of experimental details up to high Z
- Example result
 - **$N_{\text{obs}}=178$ ($s+b$), $N_{\text{obs}}'=100$ (b) $\rightarrow Z= 5.0$**

This is one example of a whole class of methods... let's be more general...

Coverage: In statistics, the coverage probability of a confidence interval is the proportion of the time that the interval contains the true value of interest. If a large number n of experiments perform measurements of a parameter with confidence level α , in the limit $n \rightarrow \infty$, the fraction α of the limits has to contain the true value of the parameter inside the confidence limits.

For each possible value of the parameter θ , we fix a probability interval $[X_1(\theta), X_2(\theta)]$ such that:

$$P(X_1 \leq X \leq X_2 | \theta) = \int_{X_1}^{X_2} f(X | \theta) dX = \alpha$$



For an observation x , one then finds $\theta_{\text{low}}, \theta_{\text{high}}$ such that $x_1(\theta_{\text{low}}) = x_2(\theta_{\text{high}}) = x$

This is called „Neyman's construction”.

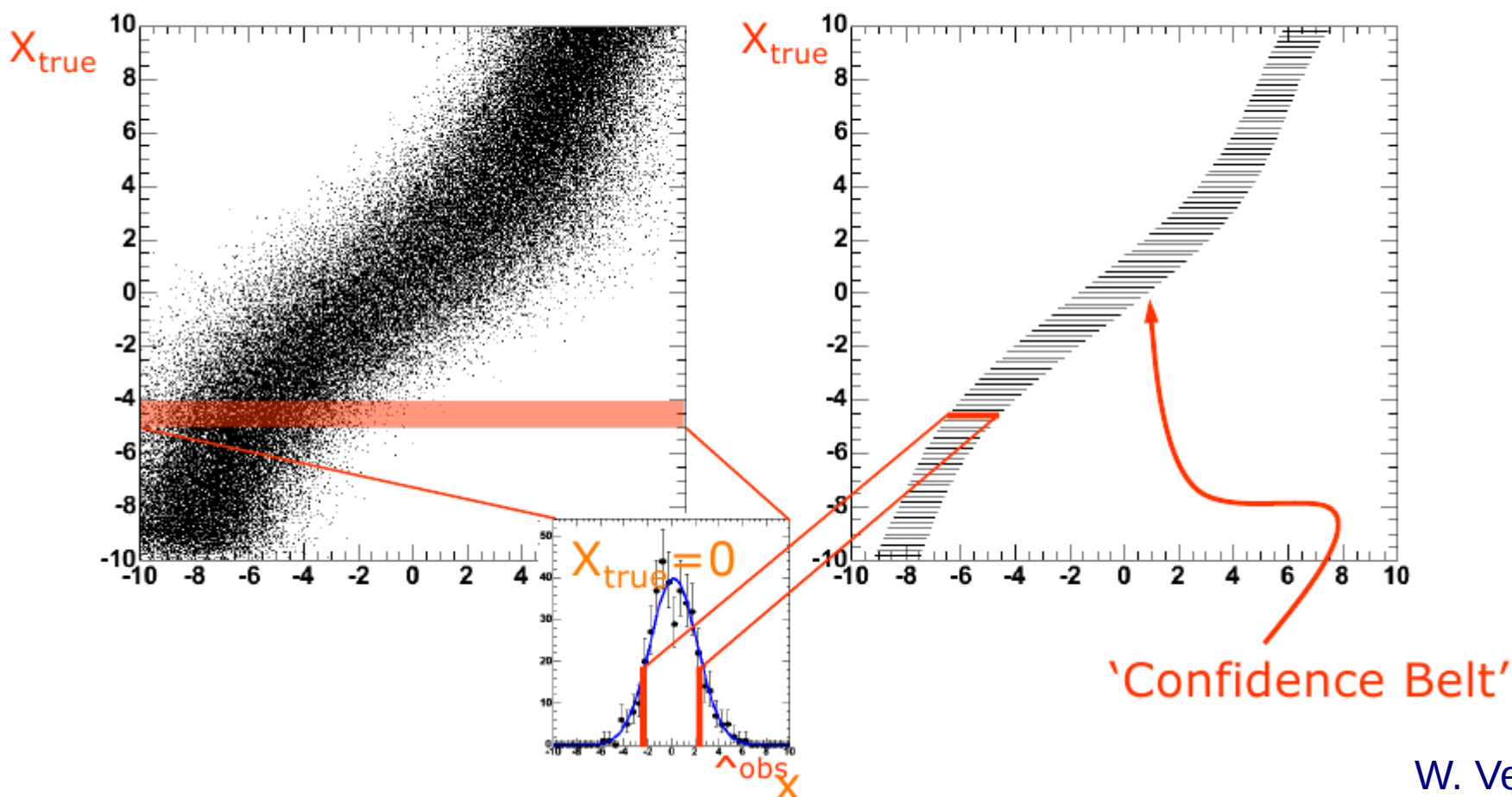
Whatever the values of the parameters realized in nature are, measurements will produce with probability α a confidence contour which contains these parameters.

Toy MC construction

- The confidence belt can be constructed from toy Monte Carlo events, for cases where no analytical calculation is possible.

Each point measurement x_{obs} from
a MC dataset generated with X_{true}

Intervals that contains 68%
of values of x_{obs} for each X_{true}



W. Verkerke

For each possible value of the parameter θ , we fix a probability interval $[X_1(\theta), X_2(\theta)]$ such that:

$$P(X_1 \leq X \leq X_2 | \theta) = \int_{X_1}^{X_2} f(X | \theta) dX = \alpha$$

For an observation x , one then finds $\theta_{low}, \theta_{high}$ such that $x_1(\theta_{low}) = x_2(\theta_{high}) = x$

There are many ways to define the interval.
Some popular choices are :

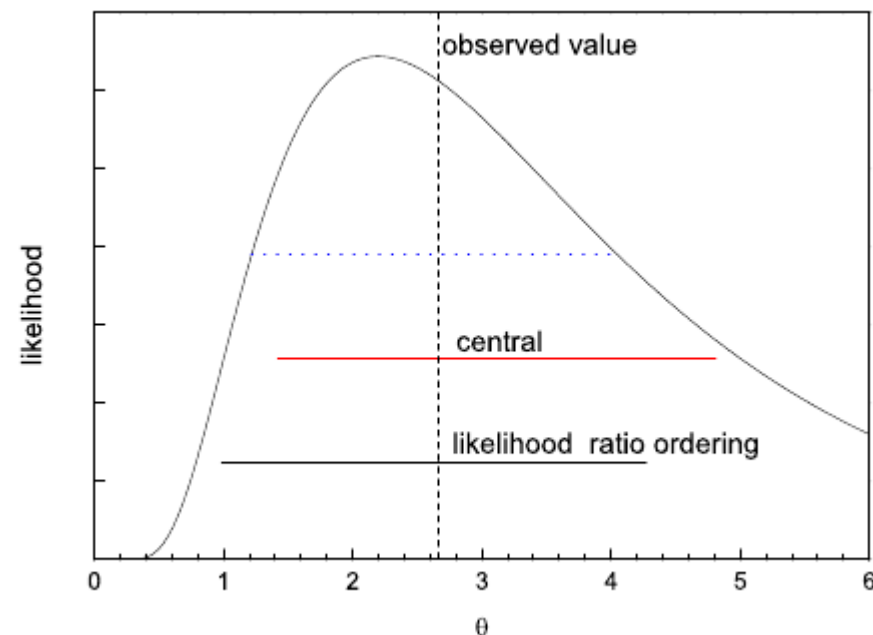
central interval	$P(X \leq X_1 \theta) = P(X \geq X_2 \theta) = (1 - \alpha)/2$
equal probability densities	$f(X_1 \theta) = f(X_2 \theta)$
minimum size	$\theta_{high} - \theta_{low}$ is minimum
symmetric	$\theta_{high} - \hat{\theta} = \hat{\theta} - \theta_{low}$
likelihood ratio ordering	$f(X_1 \theta) / f(X_1 \theta_{best}) = f(X_2 \theta) / f(X_2 \theta_{best})$
one-sided	$\theta_{low} = -\infty$ or $\theta_{high} = \infty$

Which one is the best ?

No single response. It might depend on the application.

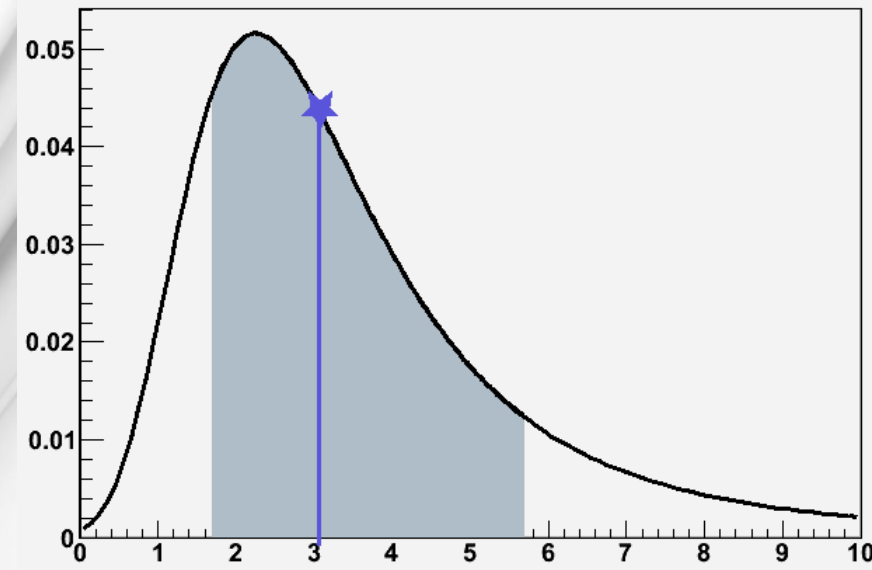
That's why one should always precise how the interval has been constructed.

Let's look at the options in more details...



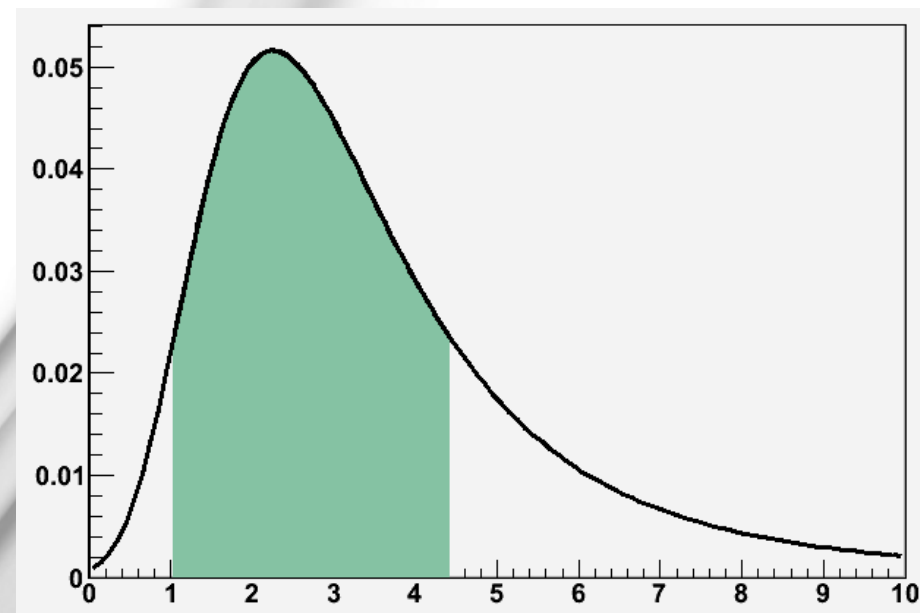
- This is the **standard choice**
 - For long, advised by the PDG
- Invariant against parameter changes
- Restricted to the case of one variate and one parameter
- The obvious choice for the **parameter estimate** is the **median of the likelihood distribution**
 - This is the limit $\alpha \rightarrow 0$ of the interval, which often doesn't coincide with the maximum likelihood value.
 - Note that it often differs from the maximum likelihood estimate (mode of the distribution)

$$P(X \leq X_1 | \theta) = P(X \geq X_2 | \theta) = (1 - \alpha)/2$$



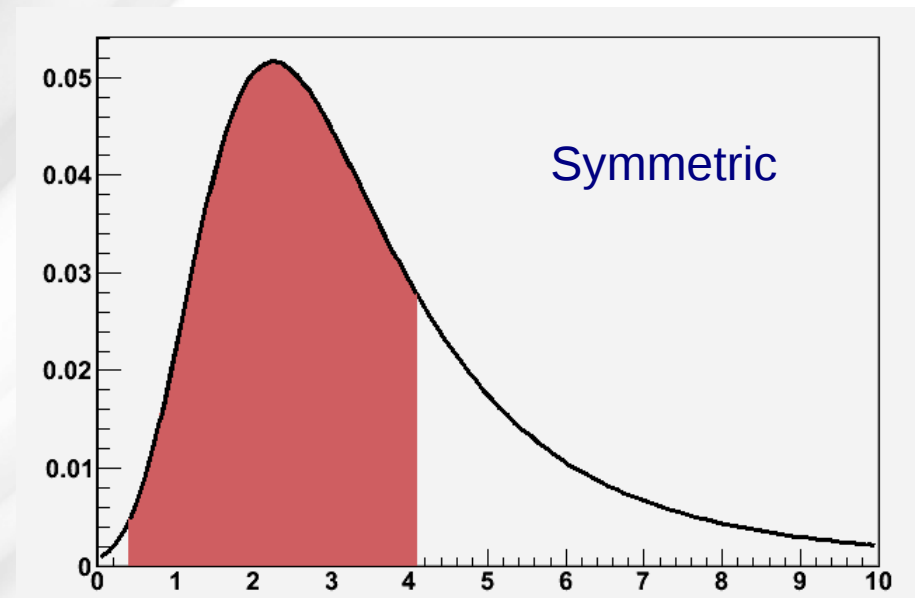
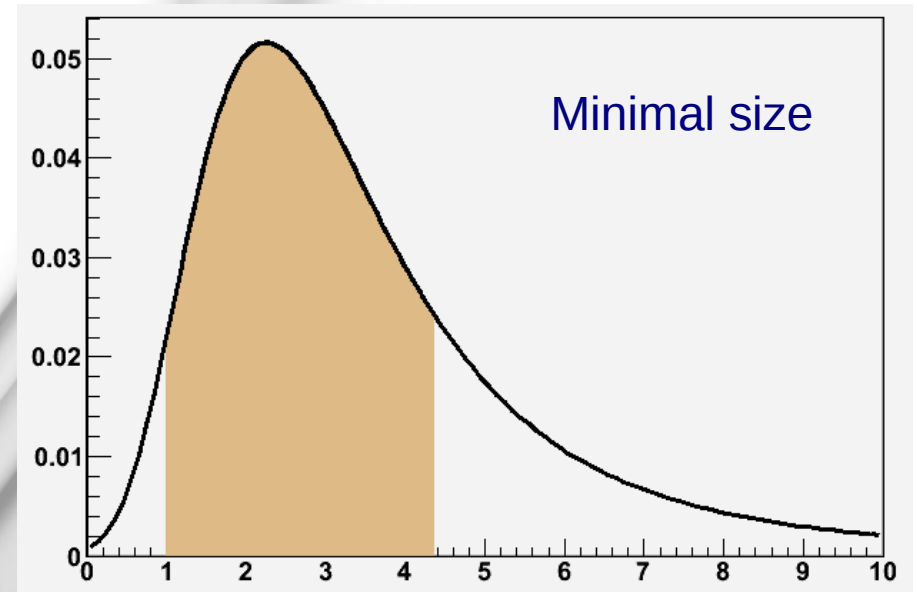
Equal probability density intervals

Equal probability density intervals are obtained by taking points of same pdf on each side of the measurement, such that the integral in between is α .



- Often shorter and less biased.
- Is applicable to multidimensional cases.
- Coincide to the central intervals for symmetric distributions.
- NOT invariant under variate transformations

- Minimal size intervals
 - Attractive at first.
 - Very difficult to compute, if possible at all.
 - Depends on the parameter choice.
- Symmetric intervals
 - Easy to handle (symmetric errors)
 - Difficult to compute.
 - Depends on the parameter choice.





Likelihood ratio ordering



Additional motivation: try to minimize the probability to contain wrong parameter values.

Class of interval:
Most Selective unbiased (MSU)

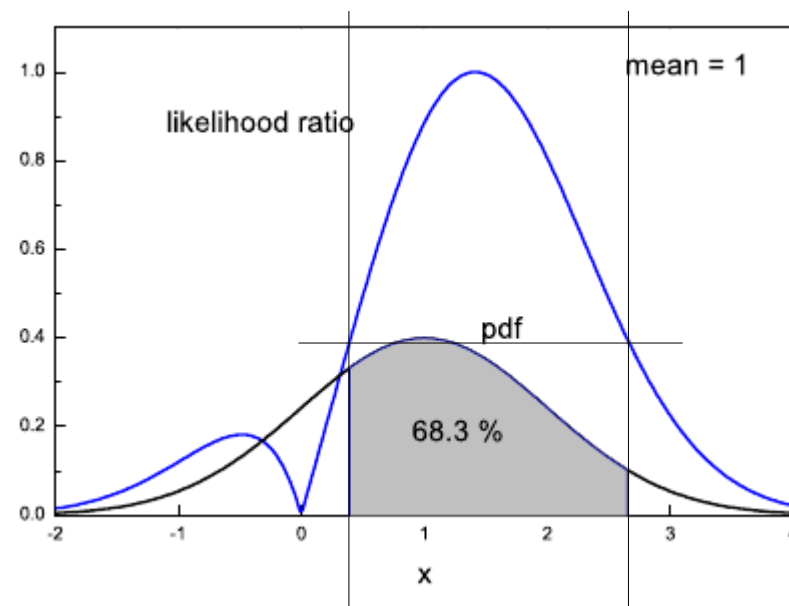
Define:

$$R(X|\theta) = \frac{f(X|\theta)}{f(X|\theta_{best})}$$

θ_{best} : Maximum likelihood estimate for a fictitious observation X .

Exemple:

$$f(X) = \frac{1}{\sqrt{2\pi c\theta}} \exp \left[-\frac{(X - \theta)^2}{2c\theta} \right]$$



The interval is then defined by $R(X_1|\theta) = R(X_2|\theta)$. It requires a significant programming effort and large CPU, and in some pathological cases can lead to non-continuous intervals.

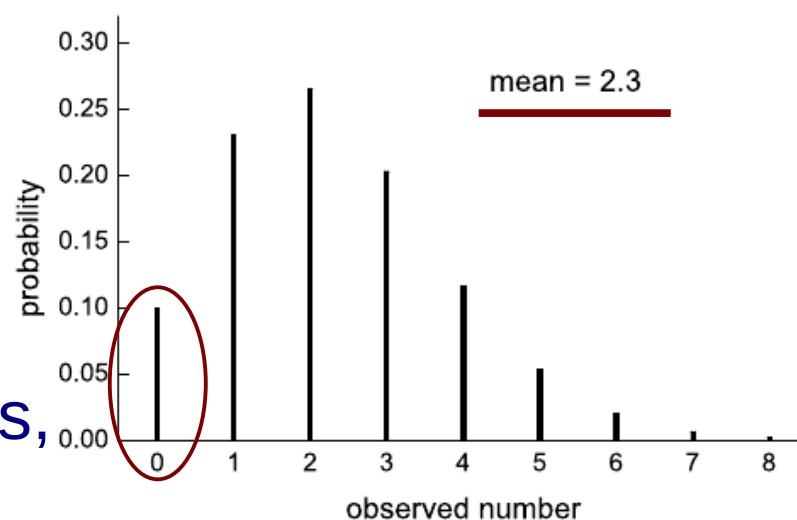
- Usually shorter than the central intervals
- Usually close the likelihood intervals



Upper and lower limits



- The classical approach to upper and lower limits is to quote the value such that
$$P(X > X_1 | \theta) = 1 - \alpha \text{ or } P(X < X_2 | \theta) = 1 - \alpha$$
- The physicist has to *decide*, preferably *before looking at the data*, whether to produce a confidence interval or a limit.
- Most common case: Poisson distributed signals in the presence of background.
 - Discrete pdf \rightarrow overcoverage cannot be avoided
 - Common case for searches:
no observation \rightarrow 90%CL limit on the mean in this case is 2.3 events, in the absence of background.



Upper limit for the Poisson case

Assuming the background expectation b is precisely known the probability to find k events (background plus signal) is

$$W(k) = \sum_{i=0}^k P(i|\mu)Q(k-i|b)$$

If the background follows a Poisson distribution too,

$$W(k) = \sum_{i=0}^k P(i|\mu)P(k-i|b) = P(k|\mu+b)$$

Then the probability to find less than or equal to n events is

$$1 - \alpha = \sum_{k=0}^n P(k|\mu+b)$$

Coverage is NOT
what you could think !

Solving the last equation for μ , we get the upper limit with confidence α .

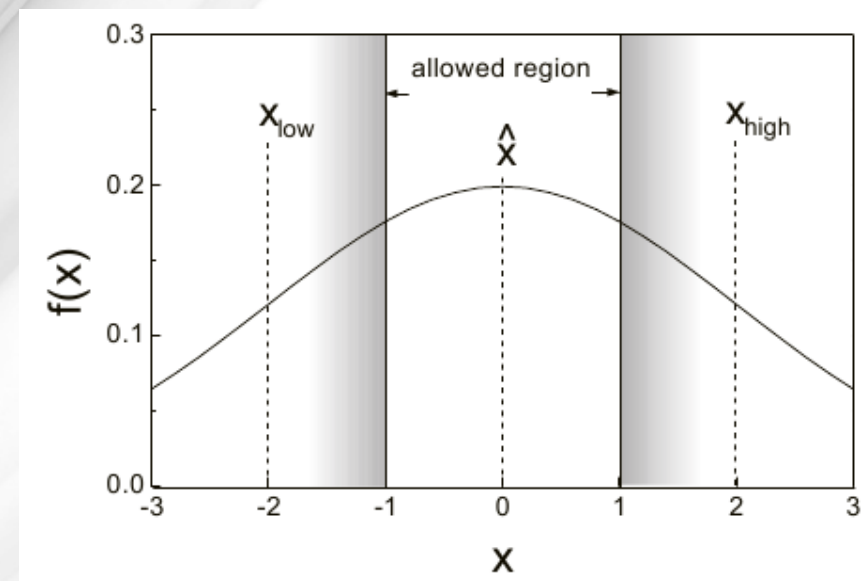
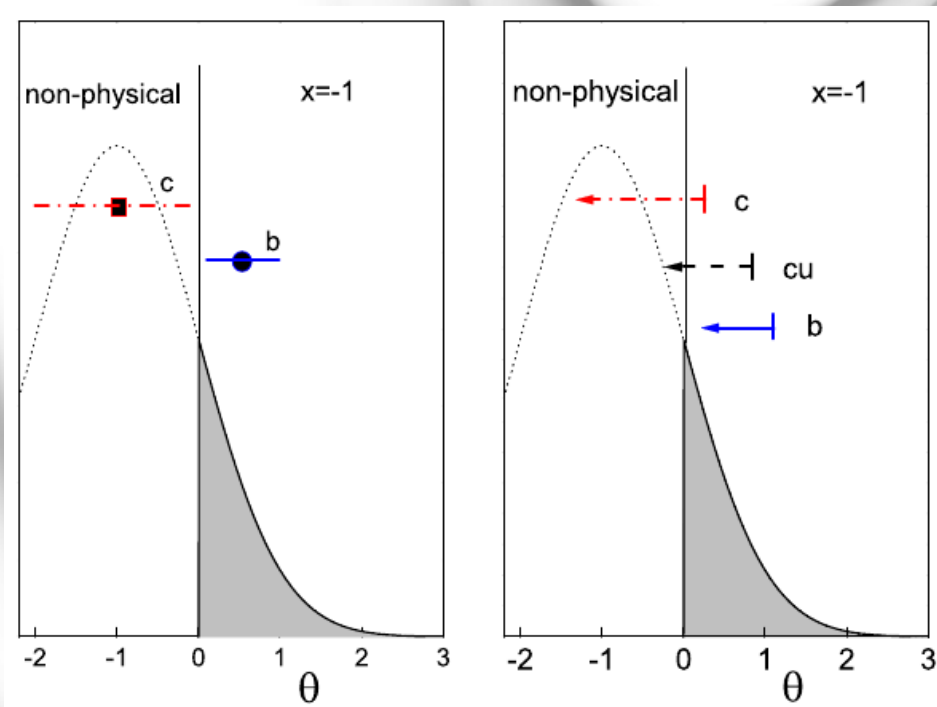
Problem: $b \nearrow$ implies that the **limit** \searrow , this contradicts the intuition. Still, this is correct from the frequentist point of view. Also, background fluctuations might lead to negative or zero-length limits.

Possible solution: renormalization of the background
(„Zech's modified frequentist approach”)

$$\begin{aligned} \longrightarrow Q'(k|b) &= \frac{Q(k|b)}{\sum_{i=0}^n Q(i|b)} \text{ for } k \leq n \\ \longrightarrow 1 - \alpha &= \frac{\sum_{k=0}^n P(k|\mu+b)}{\sum_{k=0}^n P(k|b)} \end{aligned}$$

Classical approach: problem 1

- External constraints
 - Classical interval & limits do not behave properly in the presence of constraints on the parameter space.
- Example 1: measurement in a non-physical region
- Example 2: low resolution measurement in a constrained narrow region.





Classical approach: problem 2

- When to go for a limit? When to go for a confidence interval ?
 - No problem when decided a priori
 - In practice, driven by data (is there a visible signal?)
 - In that case the coverage is not granted !

- Example: Flip-flopping:

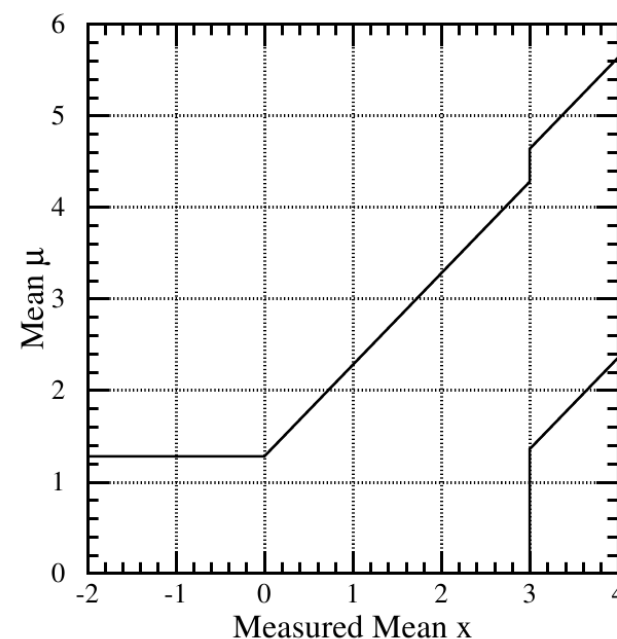
- Consider a physical positive quantity (mass?) measured with a Gaussian resolution.

$$P(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp(-(x - \mu)^2/2).$$

- Policy: If the result x is less than 3, state a limit. Otherwise, state a central confidence interval. If one measure negative values, we will pretend 0 when quoting the confidence interval.



For $\mu=2.0$, the acceptance interval contains only 85% of $P(x|\mu)$

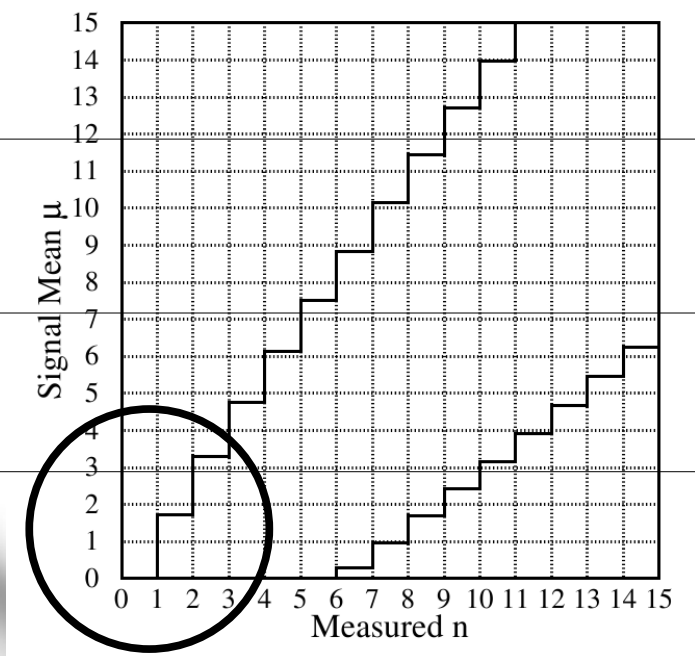


The Unified approach

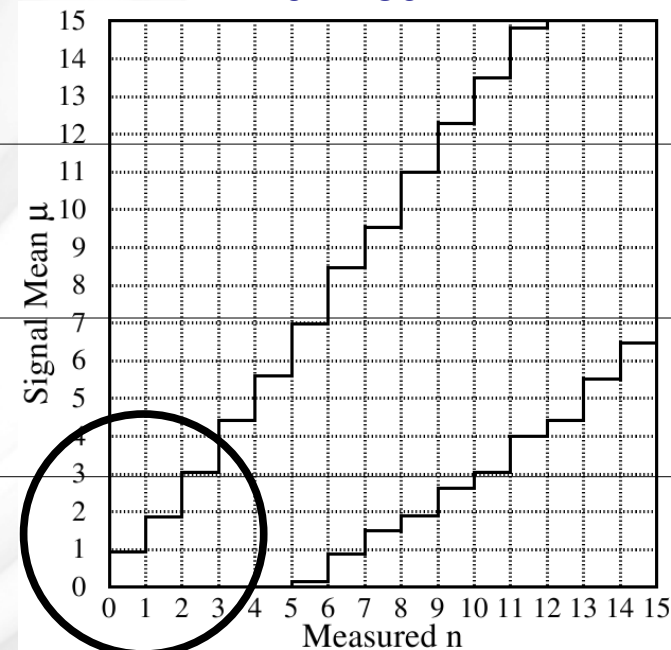
(Feldman & Cousins)

- Build a classical interval using the likelihood ratio ordering, with the additional constrain that θ_{best} sits in the physically allowed domain.
- Whenever one of the bounds is unphysical, go for a upper or lower limit.

classical



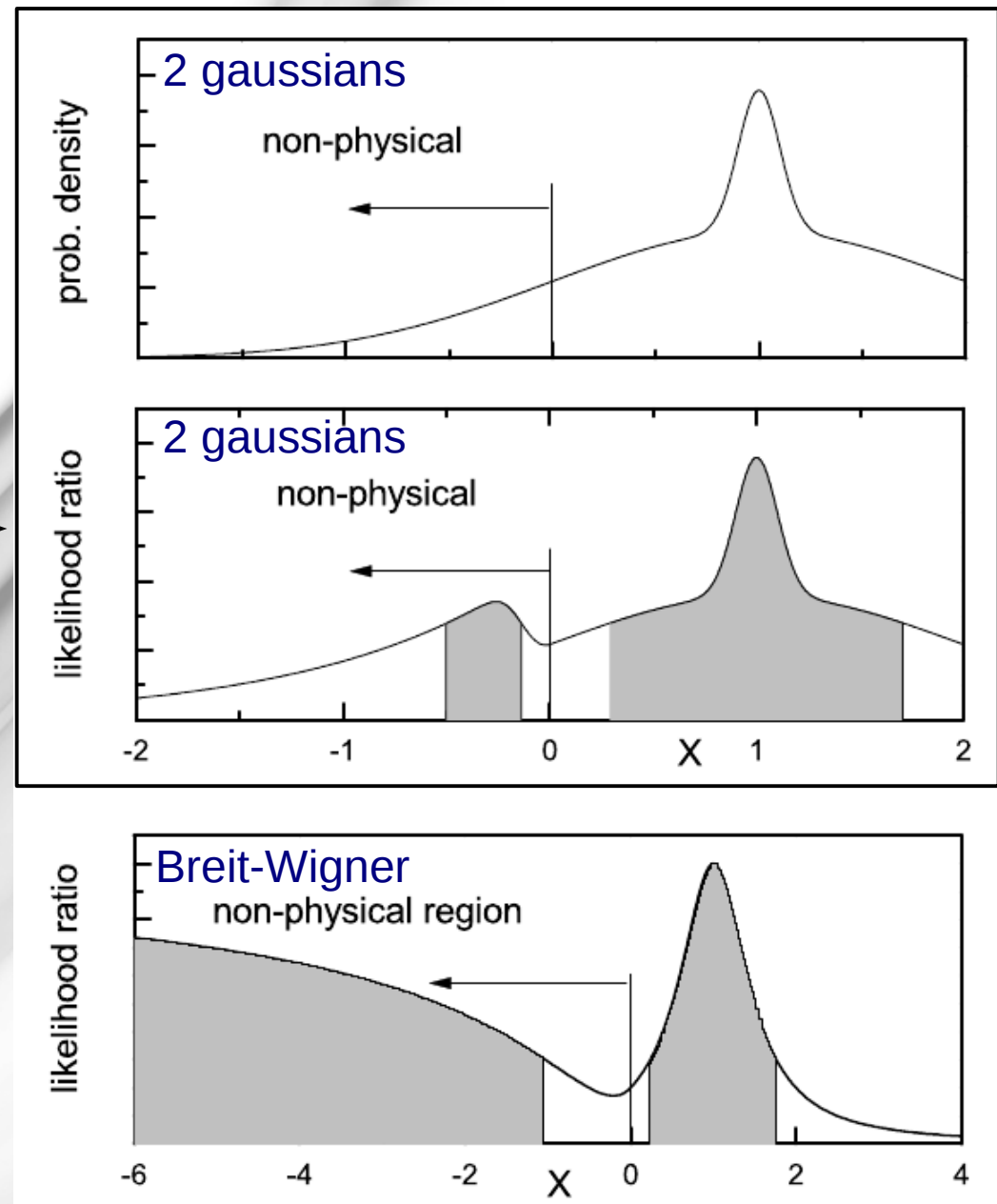
unified



- Removes the undercoverage problem due to the interval \leftrightarrow limit choice
- Reduces the unpleasant behavior for downwards fluctuations of background

Problems of the unified approach

- Two-sided constraints
 - Can produce at the same time upper & lower limits. Leads to complete coverage.
- External constraint & distribution with tails
- Does not change the situation for upper Poisson limits in the presence of background.
 - Correct from a frequentist point of view.
 - Wrong from a bayesian point of view.



Bayesians treat **parameters as random variables**. The combined probability density $f(X, \theta)$ of the variate X and the parameter θ can be conditioned on the outcome of one of the two variates using **Bayes theorem**:

$$f_{\theta}(\theta|X) = \frac{f_x(X|\theta)\pi_{\theta}(\theta)}{\pi_x(X)} \longrightarrow f_{\theta}(\theta|x) \propto L(x, \theta)\pi_{\theta}(\theta)$$

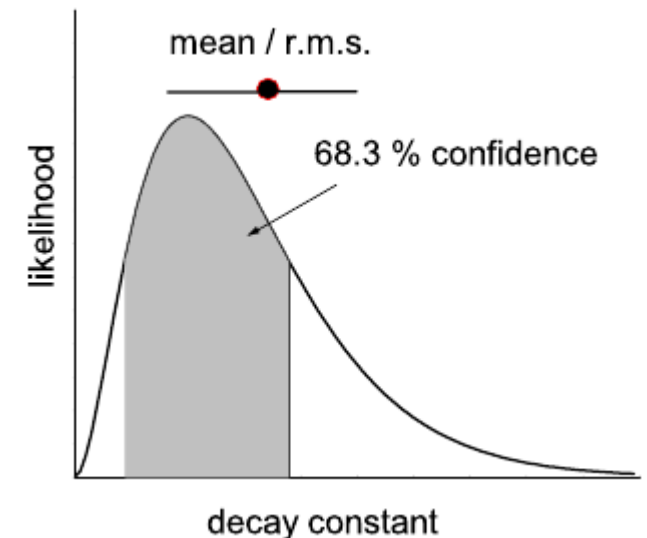
The prior density π_{θ} has to guarantee that the normalization integral is finite. Otherwise it's free.

- > Prior: uniform ? -> dependency on parameter choice
- > Fisher information ? (depends on the measurement resolution, ...)

Definition of the Bayesian limit: ~same freedom as in the frequentist case.

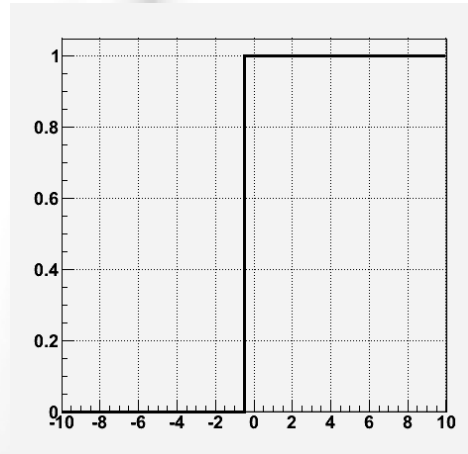
- quote the mean and the variance of the parameter
- compute intervals of a given probability (central, symmetric, ...)

—> **lack of standardization**

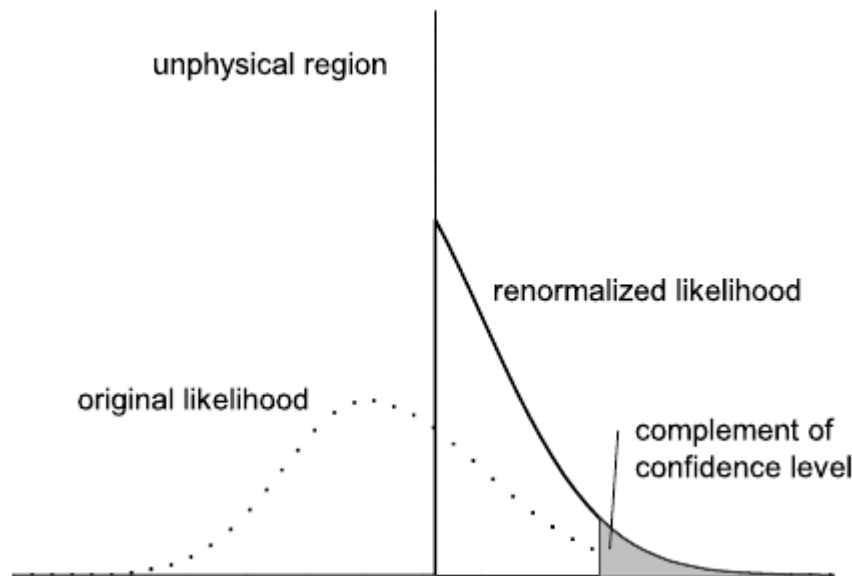
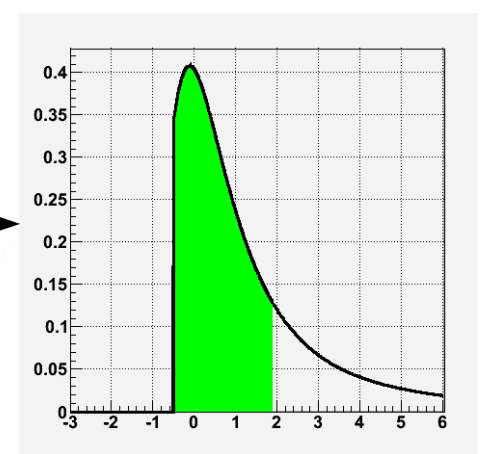
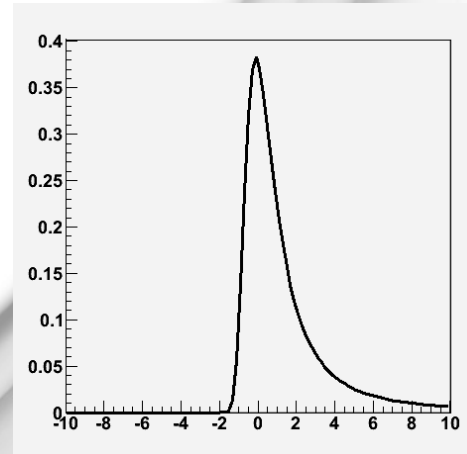


- Simple prescription:

Given previous measurement, excluded region put as prior :



X



- Easy to compute
 - Requires the likelihood & ROI.
- Gives robust confidence intervals/limits.

Likelihood intervals

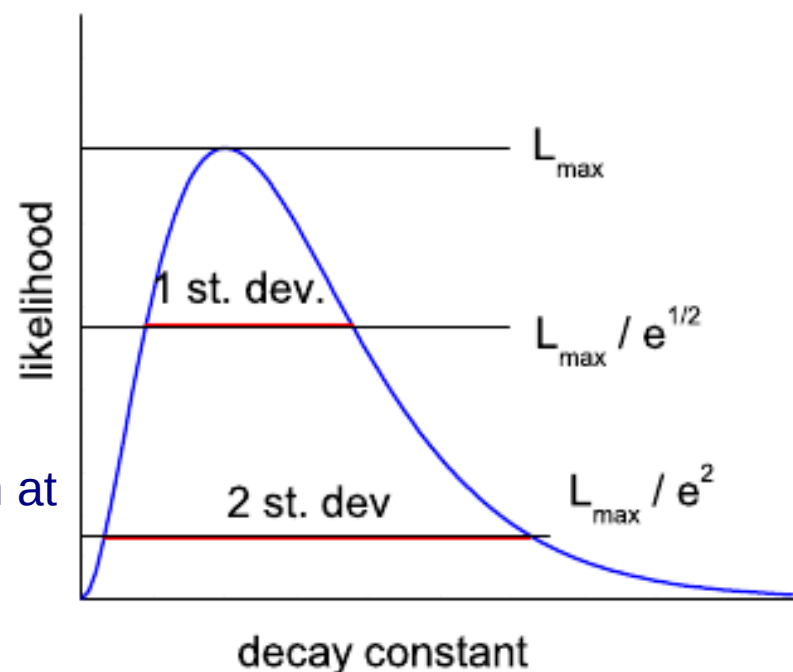
- Motivation 1: „Likelihood principle”: says that all the information is contained in the likelihood function.
- ~ Bayes with flat prior

Prescription:

$$L_{\max} / L(\theta_{\text{low}}) = L_{\max} / L(\theta_{\text{high}}) = e^{\Delta}$$
$$\ln L(\theta_{\text{low}}) = \ln L_{\max} - \Delta = \ln L(\theta_{\text{high}})$$

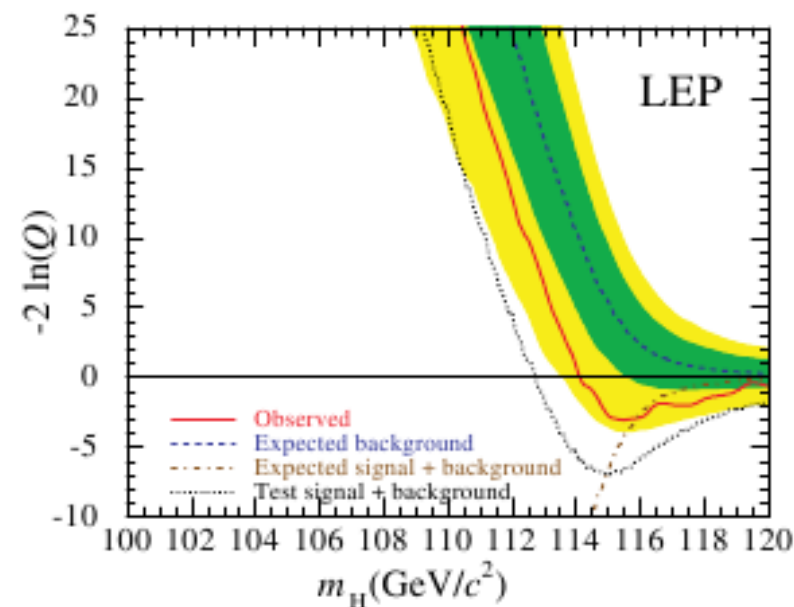
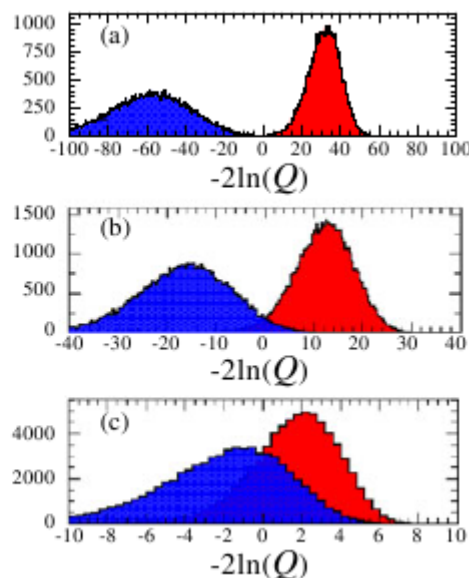
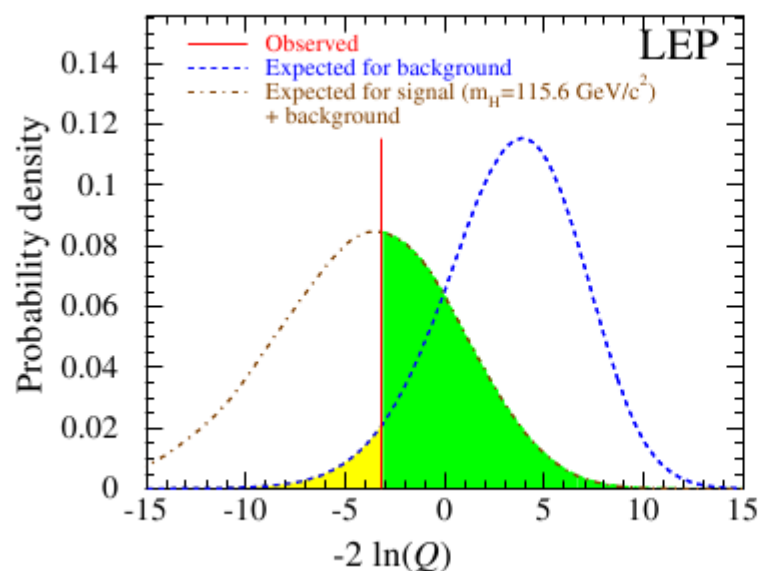
Difficulties:

- **Digital measurements** have constant likelihood functions and cannot be handled.
- The error limits for **functions with long tails** (like the Breit-Wigner pdf) are misleading.
- When the likelihood function has its mathematical maximum outside the **physical region** (L_{\max} is then at the edge of the physical region), the resulting one-sided likelihood ratio interval for $\Delta = 0.5$ may be unreasonably short.

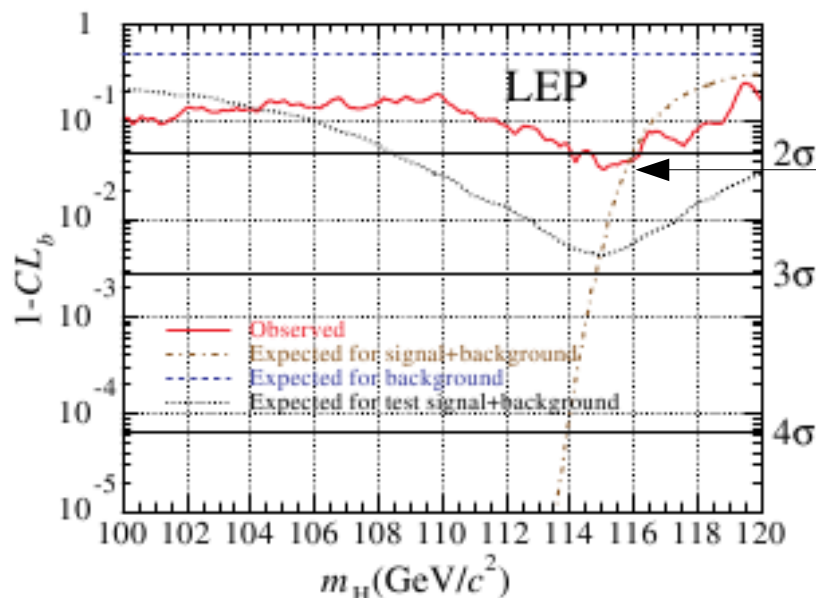


The CLs method for limits

- Equivalent method of constructing confidence intervals: consider a test of the hypothesis that the parameter's true value is θ . One then excludes all values of θ where the hypothesis would be rejected at a significance level $< \alpha$.
- One such method is the CLs method, used by the LEP Higgs group to assess limits on the Higgs boson mass.
- One defines the test statistics as $-2\ln Q$ with $Q = L(s+b)/L(b)$

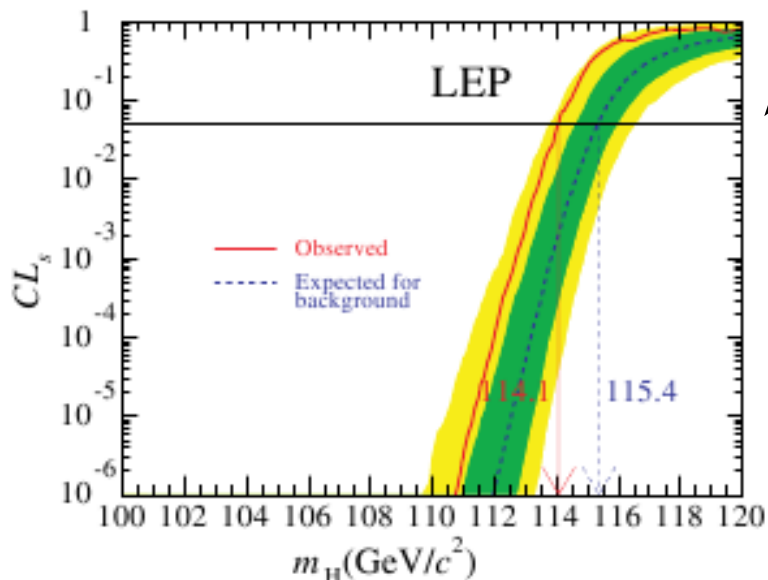


The CLs method for limits



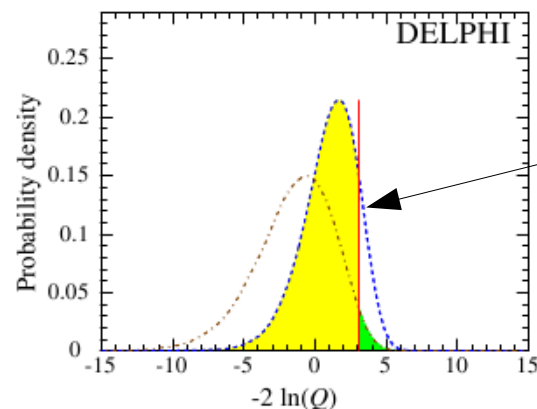
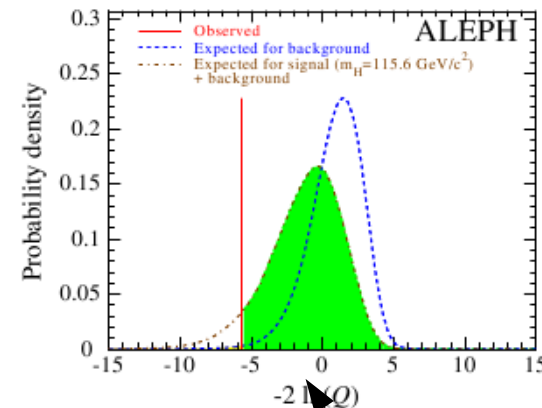
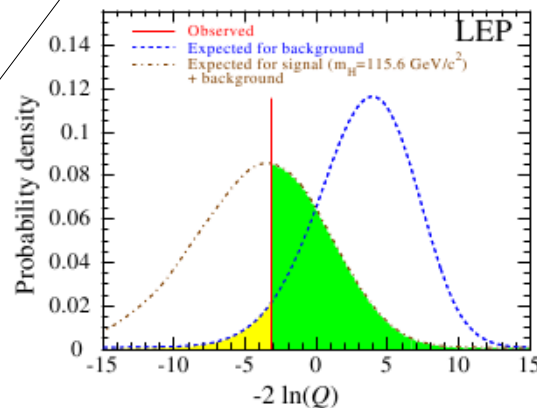
CL_b tells you about the excess w.r.t. Background-only hypothesis. Here a 2σ effect.

To obtain the limit, CL_s=CL_s+b/CL_b is used as an extension of Zech's modified frequentist approach.



$P(m_H < 114.1) < 5\%$.

No information about $P(m_H > 114.1)$!



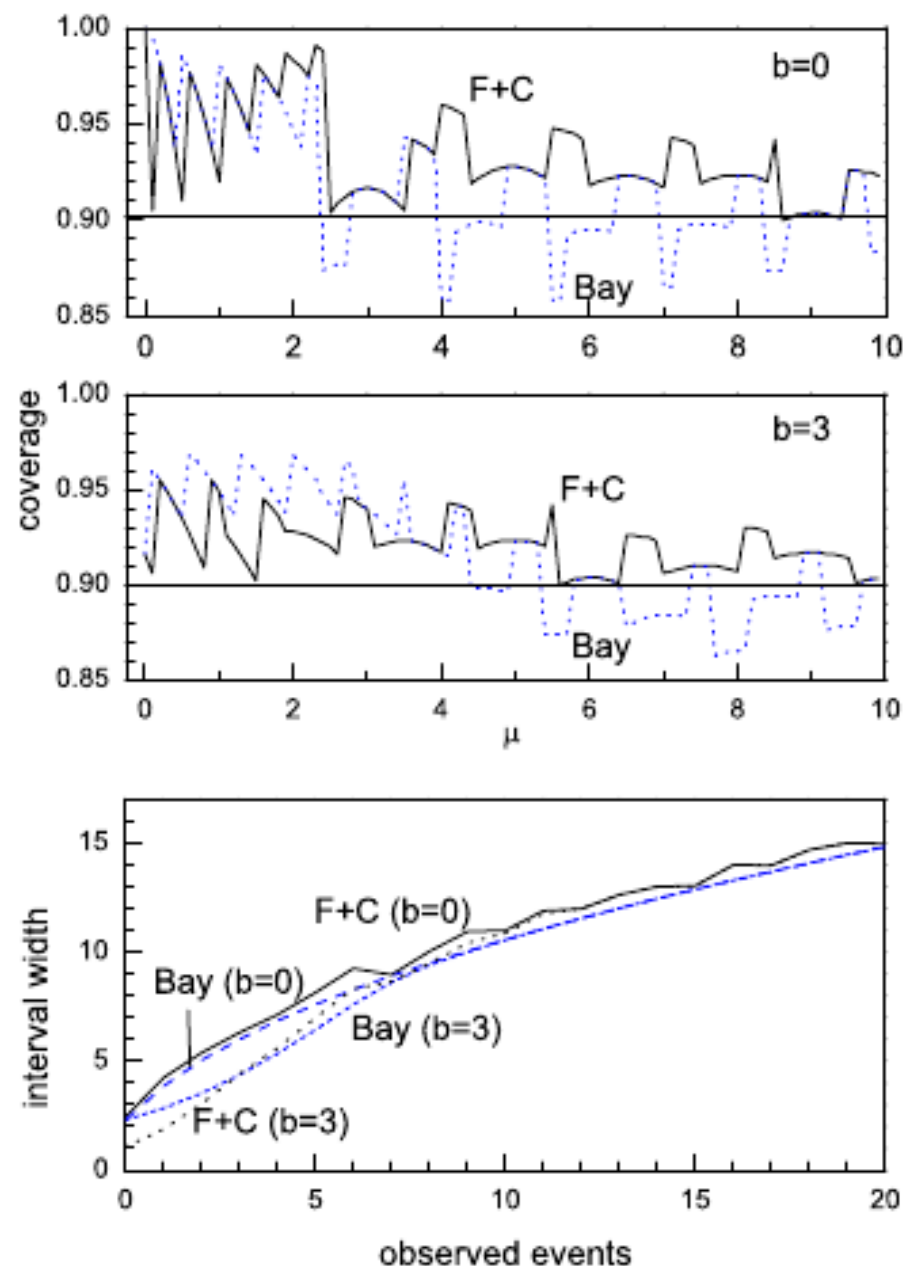
Favors
no signal

Favors
signal

The use of
likelihood makes
the combination
easy.

Comparing methods: coverage

- Coverage is a pure frequentist concept.
- Still usefull to compare methods.
- Here we see the comparison of coverage for a counting experiment (Poisson pdf) with and without background, for Feldmann-Cousins and for the Bayesian approach.



Classical, Bayesian, Likelihood...

- Frequentist confidence intervals
 - Provide ‘summary of information content’ of measurement
 - Problems and misunderstandings for low stat / near unphysical regions
- Bayesian intervals
 - Support physical interpretation of result.
 - Priors, interpretation are mixed blessing
- Likelihood intervals
 - Good coverage properties for simple cases
 - Really considered as a 3rd way in recent days. Popular in HEP statistics.

Back to the beginning...

- Which definition to prefer ?
- No definitive answer...
 - **Classical approach** definitively has deficiencies and should be avoided in potentially dangerous cases (low statistics, measurement near allowed range, etc.).
 - **Unified (Feldman-Cousins) approach** solves most of the problems of the classical approach. This is why it is often recommended (and used by most collaborations).
 - Still it has deficiencies that must not be forgotten
 - **Bayesian approach** is elegant and simple to interpret, but suffers from the freedom in the definition of the prior. Flat prior is often not justified and the method might induce undercoverage.
 - 1. Check with your collaboration.
 - 2. Be convinced yourself, and be able to explain your approach
 - 3. It's good practice to verify the sensitivity on the method.
 - One may choose to communicate an interval with good frequentist properties, and then to draw conclusions using Bayesian statistics. The two part should then be clearly separated.

Outline

- Probability and Statistics, basic concepts
- Monte Carlo techniques
- Event classification
- Parameter estimation
- Limits, confidence intervals, significance
- **Closing remarks**

Closing remarks

- As announced, we only scratched the surface.
- Despite what we could think, Probability and statistics is a lively field.
- Many improvements in the last years
 - Feldman-Cousins
 - Increased consideration for Bayesian techniques
 - Better random number generators
 - Boosted decision trees
 - ...
- Good tools are now available for the physicist (TMVA, RooStat, RooFit, ...)
- Things change quickly: stay tuned, follow the progress on preprint servers and journals, get in touch with the experts in your experiment.
- Before all, convince yourself (and others) that you are using the best approach!



References



- **Data Analysis**, Wouter Verkerke, lecture at the Joint Belgian Dutch German Graduate School 2008.
- **Statistical Methods and Analysis Techniques in Experimental Physics**. Frühlingsemester 2010. Common Lectures (ETHZ + UZH) by C.Grab, C.Regenfus, Institute for Particle Physics, ETH Zurich and Physik-Institut, University Zurich.
- Statistical data analysis, G. **Cowan**, Oxford University Press; ISBN: 0198501552
- Statistics for nuclear and particle physics, L.**Lyons**, Cambridge University Press.
- Statistics: A guide to the use of statistical methods in the Physical Sciences, R.J.**Barlow**; Wiley Verlag
- Datenanalyse, S.**Brandt**, BI Wissenschaftsverlag.
- **Unified approach to the classical statistical analysis of small signals**, Gary J. Feldman, Robert D. Cousins, Phys. Rev. D 57, 3873–3889 (1998)
- **Why isn't every physicist a Bayesian?** R. D. Cousins, Am. J. Phys. 63 (5), May 1995.
- **Frequentist and Bayesian Confidence Intervals**, G. Zech Eur.Phys.J.direct C4 (2002) 12 (arXiv:hep-ex/0106023v2)
- **Numerical Recipes: The Art of Scientific Computing, Third Edition** (2007), 1256 pp. Cambridge University Press ISBN-10: 0521880688
- **G4 - a simulation toolkit** S. Agostinelli et al, NIM A 506(3) 250-303
- **Delphes, a framework for fast simulation of a generic collider experiment** S. Ovin, X. Rouby, V. Lemaître arXiv:0903.2225v3 [hep-ph]
- **The Review of Particle Physics** K. Nakamura et al. (Particle Data Group), J. Phys. G 37, 075021 (2010)