Analisi dati

Metodi di analisi statistica dei dati di esperimenti di fisica delle particelle

Argomenti

- Introduzione
- Dalle misure sperimentali ai risultati finali

- Variabili aleatorie e distribuzioni statistiche
- Medie e varianze
- Distribuzioni binomiali, di Poisson e di Gauss

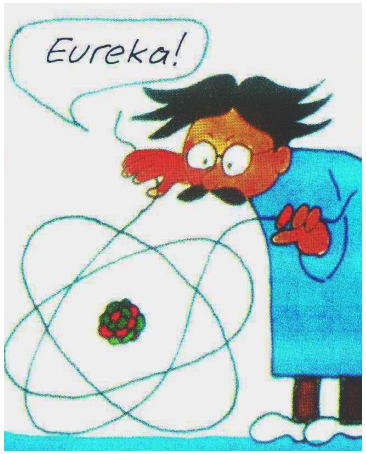
- Modelli teorici e fit sperimentali
- Misure di vita media

- Selezione dei campioni
- Livelli di confidenza e plot di esclusione
- I risultati della ricerca dell'Higgs

Introduzione

- **La fisica fondamentale e' costruita su ipotesi falsificabili**
- Le ipotesi non falsificate dagli esperimenti costituiscono la base delle teorie interpretative dei processi fisici
- Gli apparati sperimentali sono progettati per ottenere informazioni stringenti sulla natura dei processi di interesse
- La sensibilita' degli esperimenti risulta definita dalle caratteristiche dell'apparato
- L' interpretazione dei dati si basa
  sullo studio e controllo della strumentazione
  sulle conoscenze pregresse
  su ipotesi e modelli teorici

… cosa si misura nella HEP ?



Parametri fondamentali

- Branching ratio (**BR**)
- vita media (τ)
- massa (**m**)
- costanti di accoppiamento
- ……

$K^o \rightarrow \pi^+\pi^-, \pi^o\pi^o$

attraverso…

- quantita' di moto $\vec{q}$;
- energia E rilasciata nel calorimetro ;
- angoli e direzioni delle particelle prodotte ;
- intervalli temporali ;
- efficienza del rivelatore ;
- contaminazione in un campione ;
- ecc.

New physics ! ! !

Misure e variabili aleatorie

- I processi di misura sono per natura stocastici

  ERRORI STATISTICI

- Si ottengono distribuzioni di variabili relative al campione da studiare, contaminate da altri processi, "background"

  VARIABILI DISCRIMINANTI, TEST DI IPOTESI

- Inoltre strumentazione e le conoscenze "pregresse" dei fenomeni da considerare possono introdurre errori sistematici

  ERRORI SISTEMATICI

Distribuzioni di probabilita'

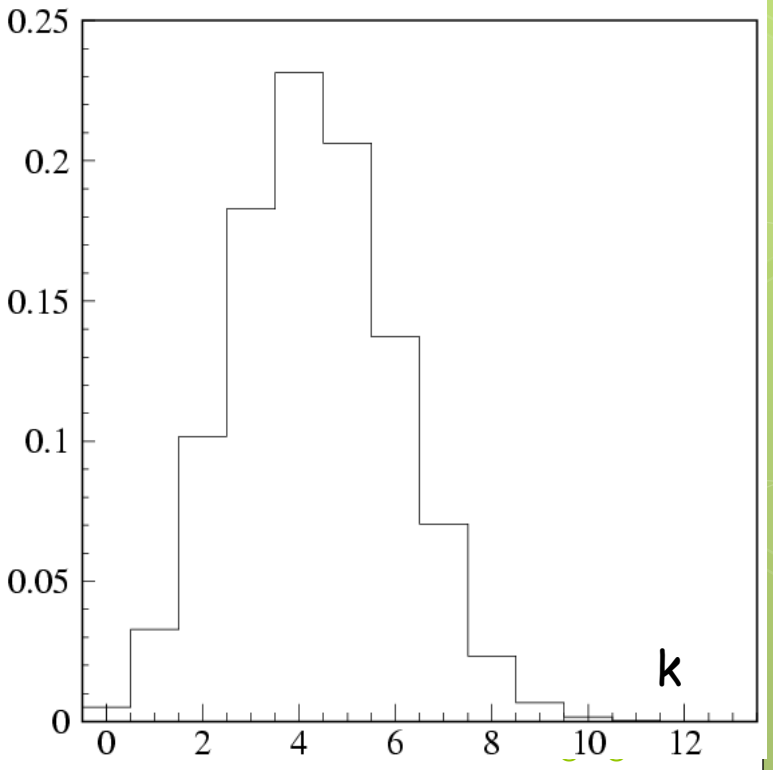
- Variabili discrete: conteggi

Se si conta classificando i conteggi come "OK" /"not-OK" questi saranno distribuiti come una binomiale

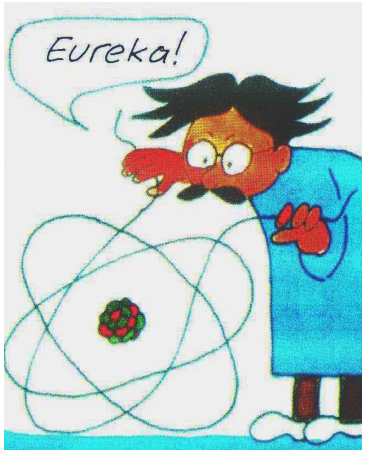
$$P(k;p,N) = \left(\begin{array}{c} N \\ k \end{array} \right) p^k (1-p)^{N-k}$$

Se i conteggi sono di un processo con frequenza μ, saranno distribuiti come una distribuzione di Poisson

$$P(k;\mu) = \frac{e^{-\mu}\mu^k}{k!}$$



$K^o \rightarrow \pi^+\pi^-, \pi^\circ\pi^\circ$



K = # $K^o \rightarrow \pi^\circ\pi^\circ$
BR (= p) = k/N

$$m = pN$$

$$\sigma = \sqrt{Np(1-p)}$$

$$\frac{\sigma}{m} = \frac{\sqrt{Np(1-p)}}{pN} \approx \frac{1}{\sqrt{N}}$$

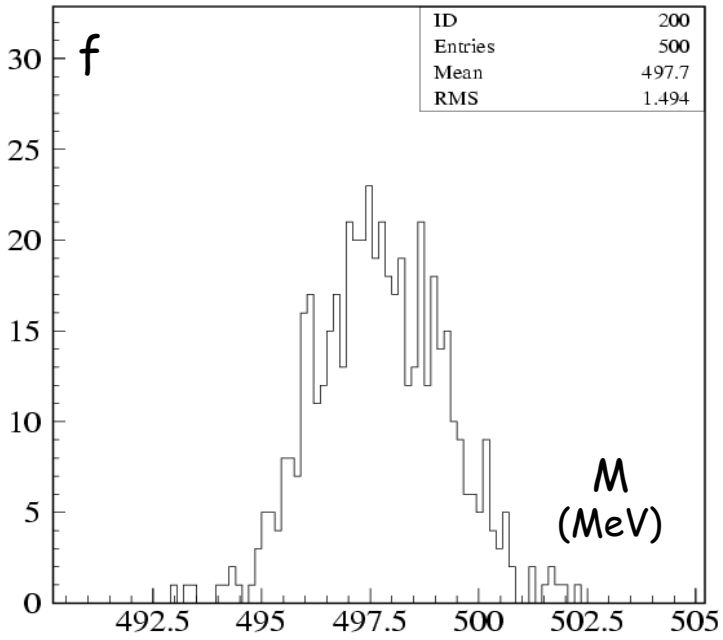
- **Precisione misura BR**

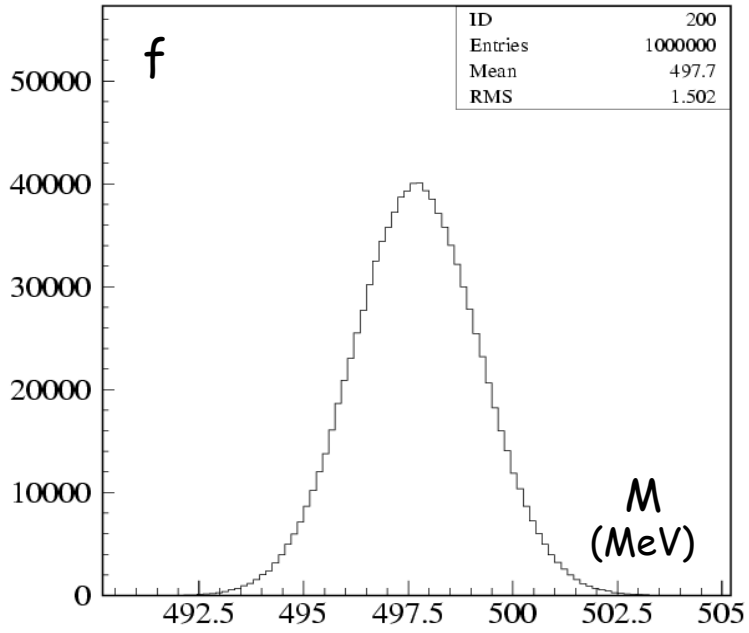
N=50 -> σ/m → 10%
$N=10^3$ -> σ/m → 2.4%
$N=10^6$ -> σ/m → 0.08%

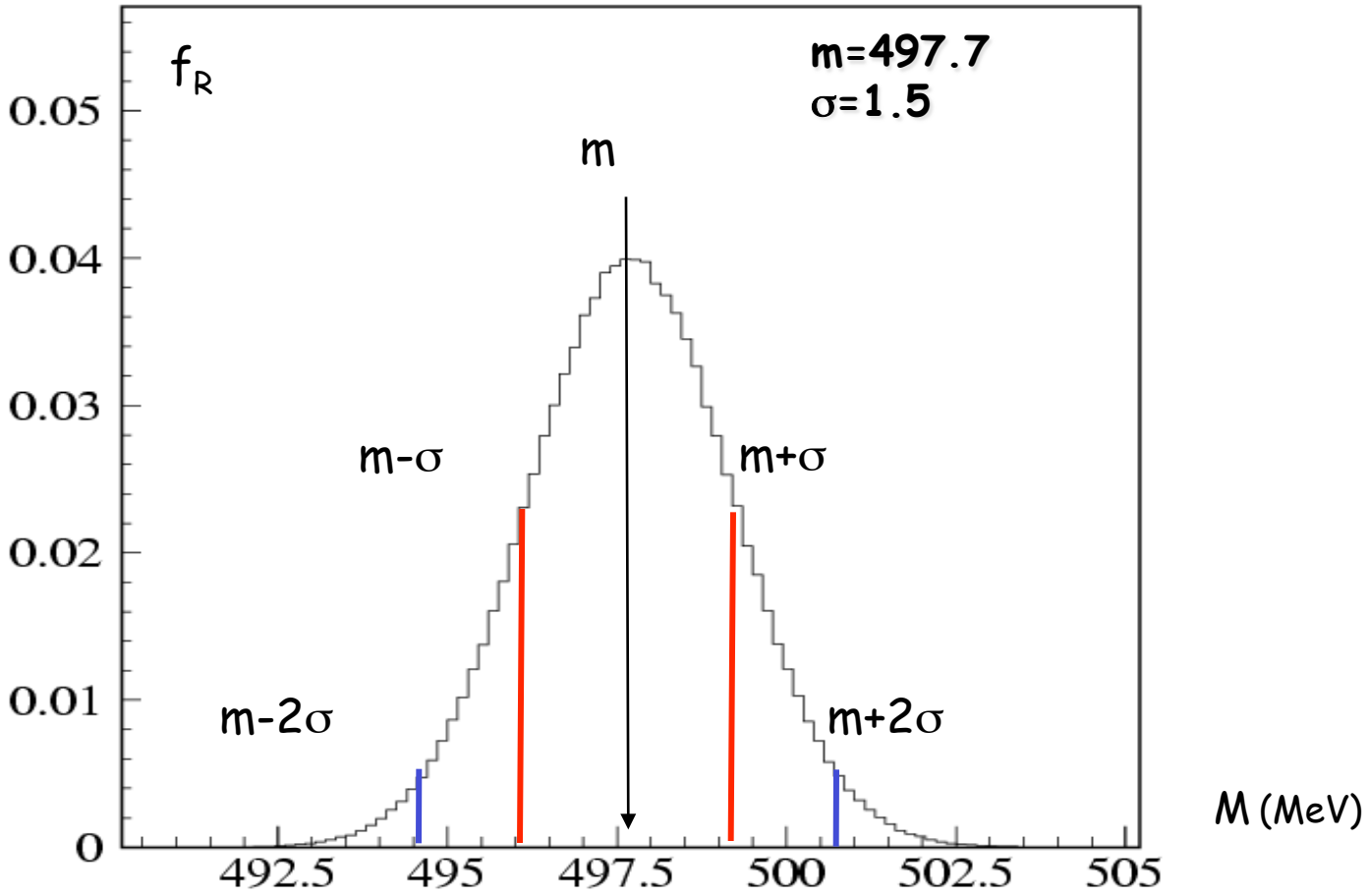
Distribuzione di Gauss

- Teorema del limite centrale

$$f(x;m,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

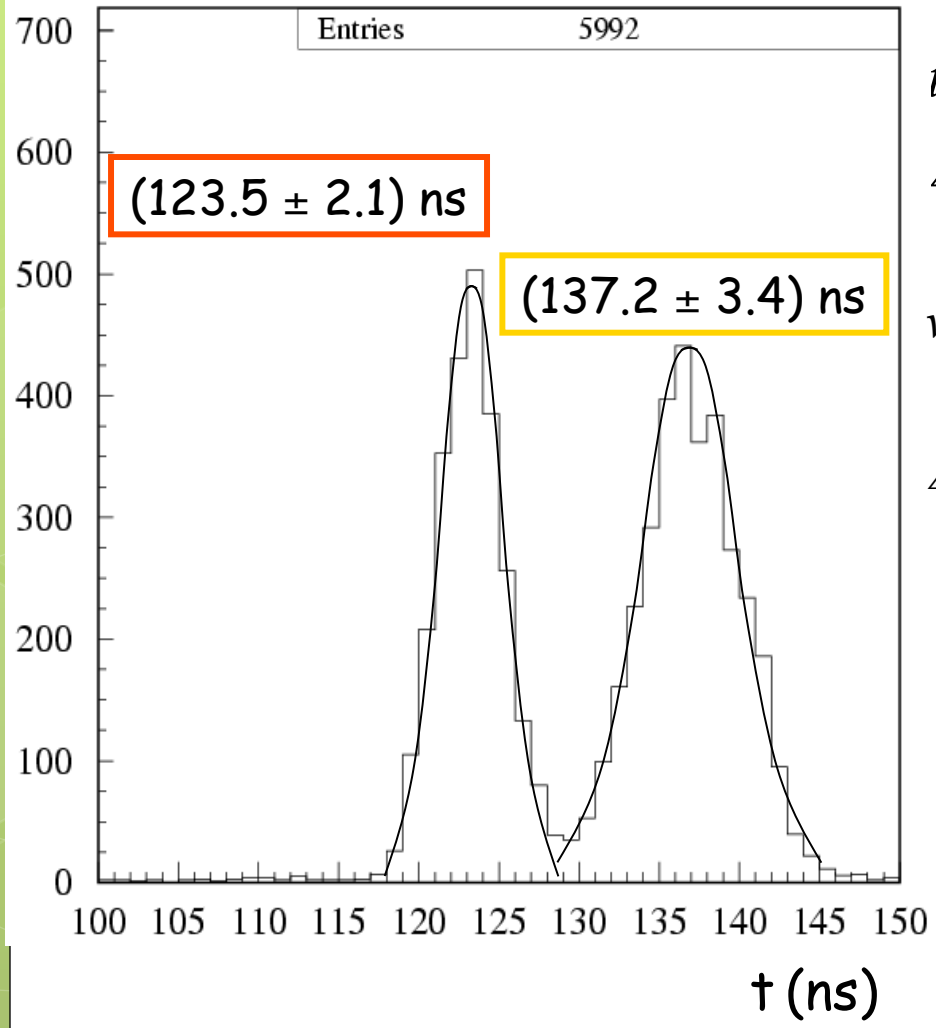






by *Marco Dreucci*

Tempi di volo



$$t_{12} = \bar{t}_2 - \bar{t}_1 = 13.7ns$$

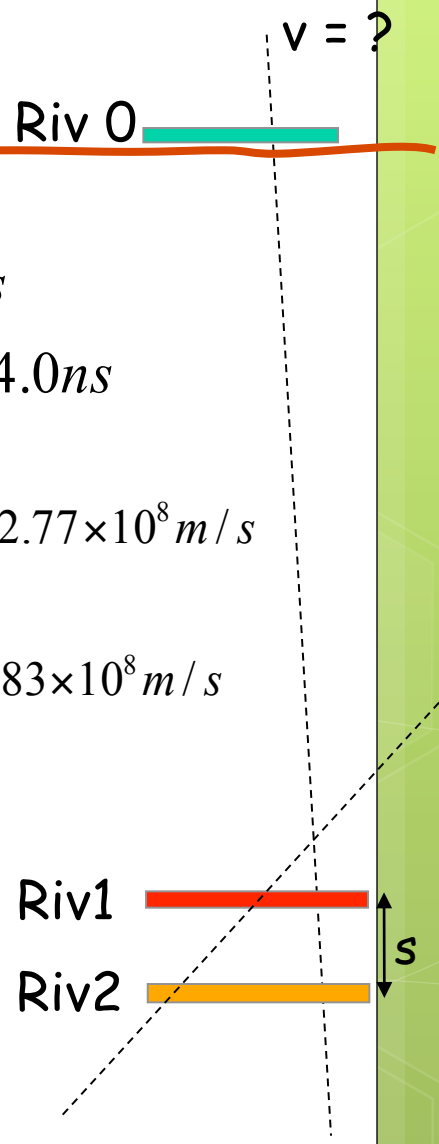
$$\Delta t_{12} = \Delta t_1 \oplus \Delta t_2 = 4.0ns$$

$$v = \frac{s}{t_{12}} = \frac{3.8m}{13.7\times10^{-9}s} = 2.77\times10^{8}\,m/s$$

$$\Delta v = v\left(\frac{\Delta s}{s} \oplus \frac{\Delta t_{12}}{t_{12}}\right) = 0.83\times10^{8}\,m/s$$

~ 2%

~ 30%



by Marco Dreucci

Binomiale

Il numero di **successi** in N prove e' una variabile aleatoria: $0 \leq k \leq N$.

risultato lancio di un dado/moneta
conteggi / efficienza
conteggi/ selezione

Distribuzione di Poisson

campionamenti idealmente infiniti

nascite al mese
decadimenti di una sostanza radioattiva in un intervallo di tempo fissato

Distribuzione di Gauss

innumerevoli fonti di errori casuali

spessori, ...
masse, impulsi, tempi, ...

Media e varianza

Valor medio

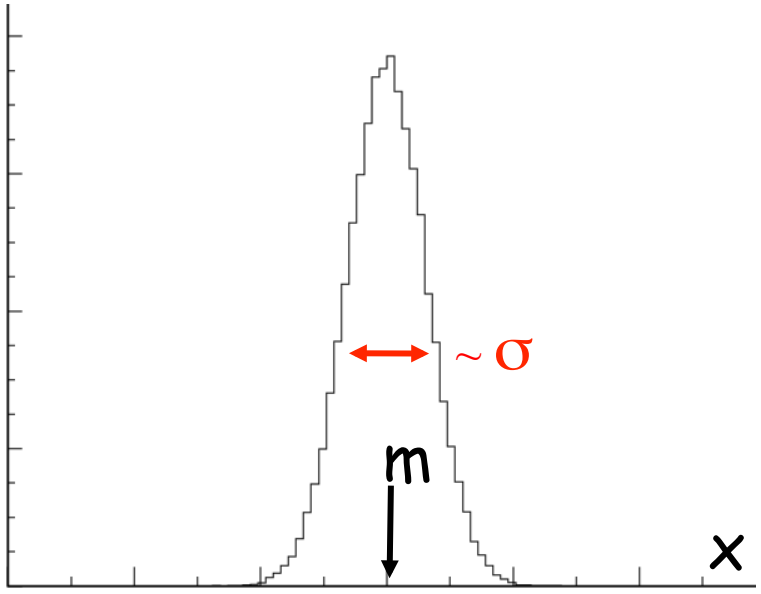
$$m = \frac{\sum x_i}{N} \qquad \bar{x} = \int x \cdot f(x)\,dx$$

Deviazione standard

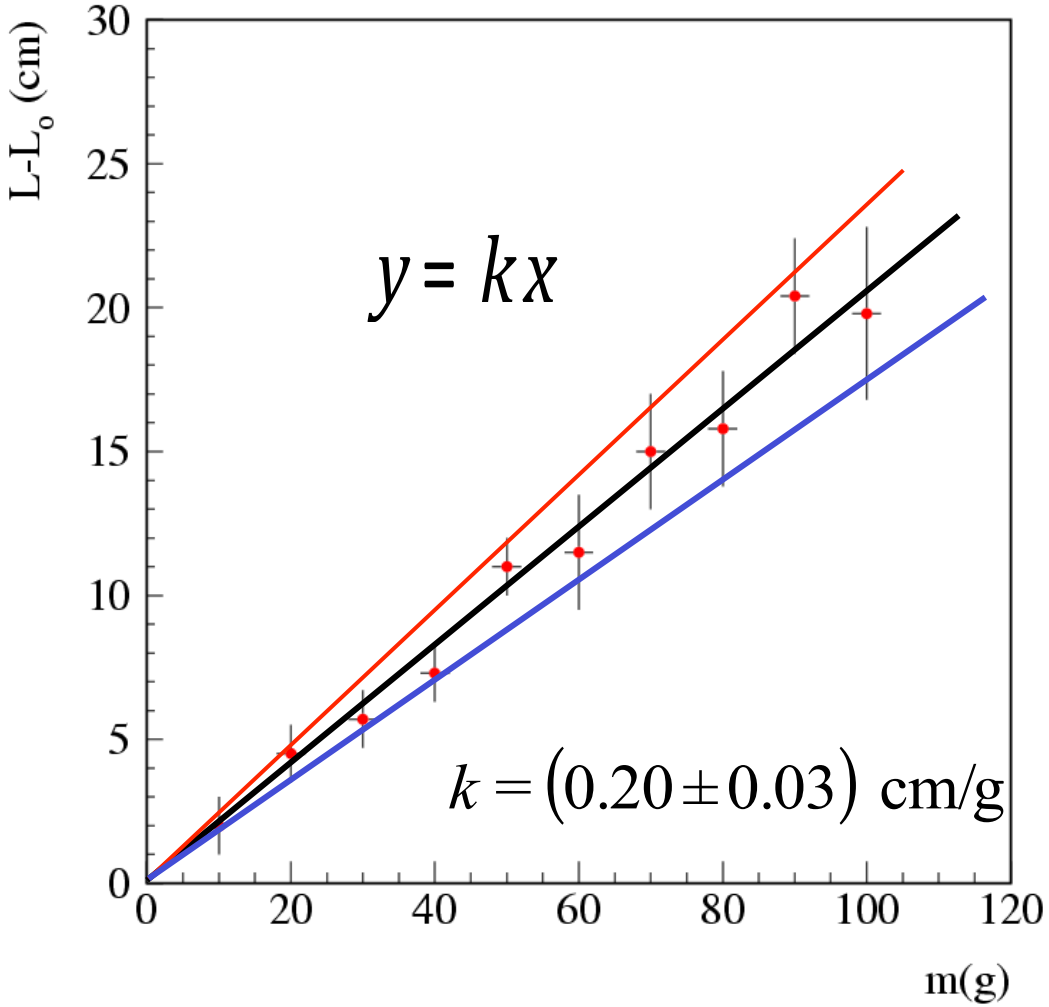
$$\sigma = \sqrt{\frac{\sum (m - x_i)^2}{N - 1}} \qquad \sigma^2 = \int (x - \bar{x})^2 \cdot f(x)\,dx$$

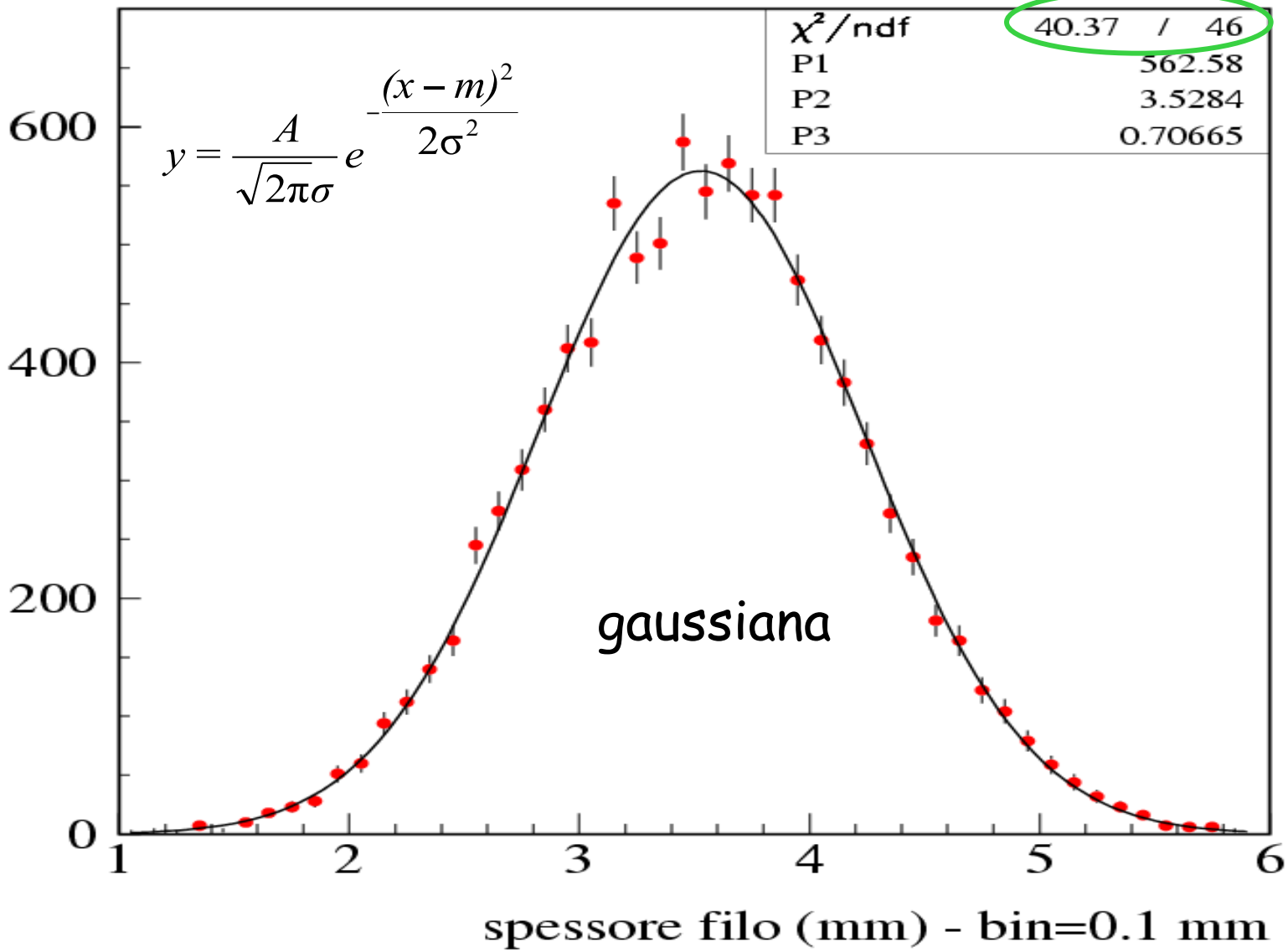
$x = m \pm \sigma$

$e_x = \sigma / m$



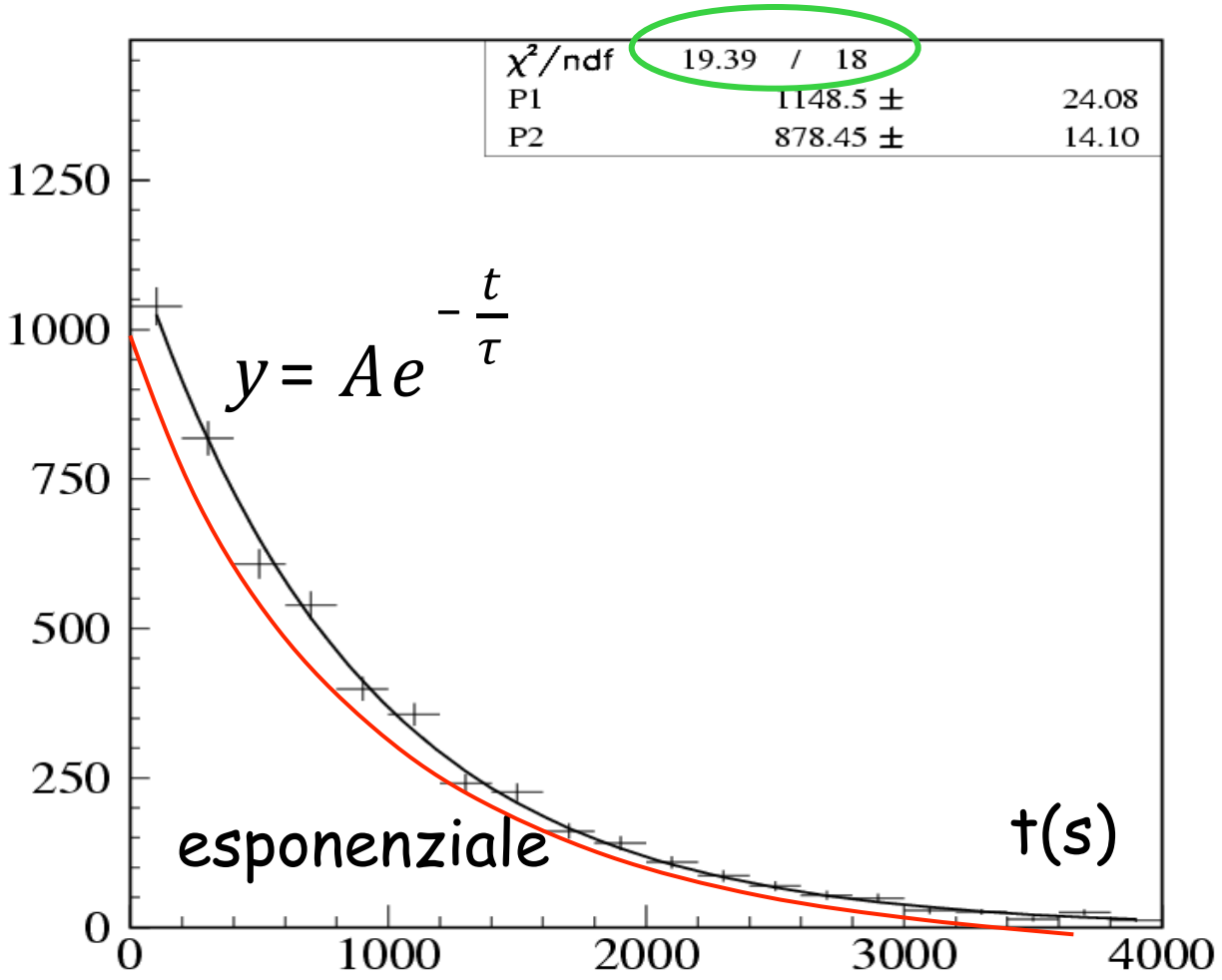
Test di ipotesi, Fit





by Marco Dreucci

Misura vita media



Funzione del χ^2

• Se il fenomeno in studio e' ben noto, la funzione $y = f(x; \alpha,\beta)$ e' gia' decisa. Lo scopo del *fit* e' allora di determinare i valori dei parametri (α e β in questo caso) che corrispondono al **miglior adattamento della curva ai dati sperimentali** .

• I valori che corrispondono al miglior adattamento sono quelli che rendono minima la quantita' seguente :

valore **sperimentale**

valore **teorico**

$$\frac{\chi^2}{N-2} = \frac{\sum_{i=1}^{N} \left(\frac{y_i - f(x_i;\alpha,\beta)}{\sigma_i} \right)^2}{N-2}$$

~ 1 → **O.K.!**

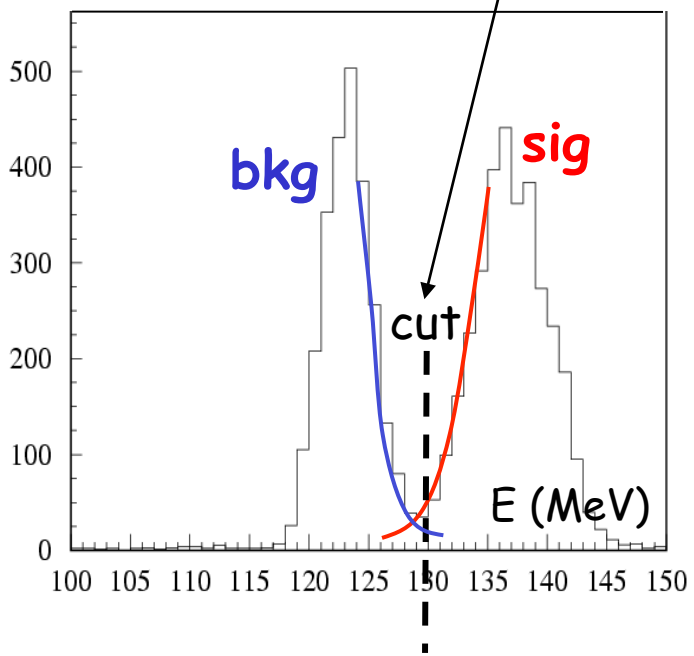
> 1 → **K.O.!**

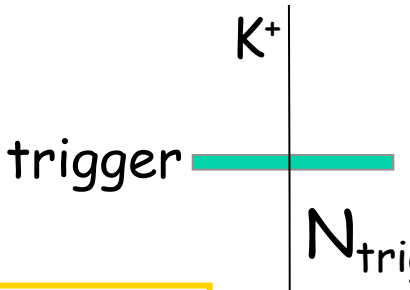
Van der Waals

Branching ratio

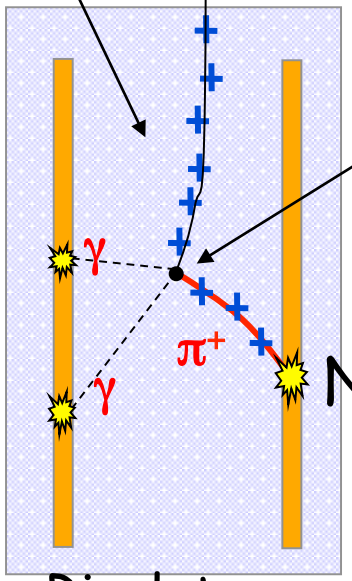
$$K^+ \rightarrow \begin{cases} \pi^+ \pi^- \pi^+ \\ \pi^+ \pi^0 \pi^0 \\ \pi^0 \pi^+ \\ \pi^0 \mu^+ \nu_\mu \\ \pi^0 e^+ \nu_e \\ \mu^+ \nu_\mu \end{cases}$$

Efficienza
$\varepsilon_{CUT} = 0.94$





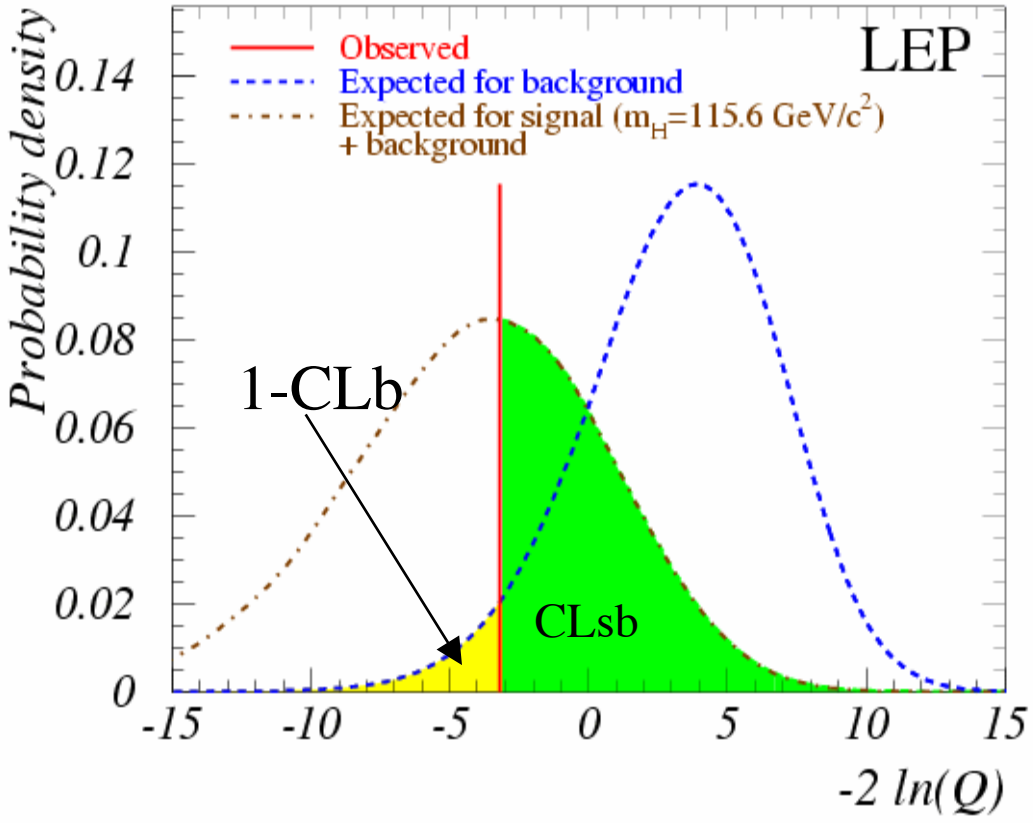
Efficienza
$\varepsilon_{SEL} = 0.52$



$$BR\left(K^+ \rightarrow \pi^0 \pi^+\right) = \frac{N_{SIG}}{N_{trig}} = \frac{\dfrac{\left(N_{OBS} - N_{BKG}\right)}{\varepsilon_{SEL}\varepsilon_{CUT}}}{N_{trig}} = \frac{\left(N_{OBS} - N_{BKG}\right)}{N_{trig} \cdot \varepsilon_{SEL} \cdot \varepsilon_{CUT}}$$

by Marco Dreucci

Variabili discriminanti, Likelihood



Confronto tra due ipotesi: H_0 e H_1

$Q = L(H_1 \mid \text{dati}) / L(H_0 \mid \text{dati})$

L'ipotesi H_0 generalmente non e' associata a parametri liberi

La distribuzione nell'ipotesi H_1 cambia quando cambiano i parametri liberi della teoria

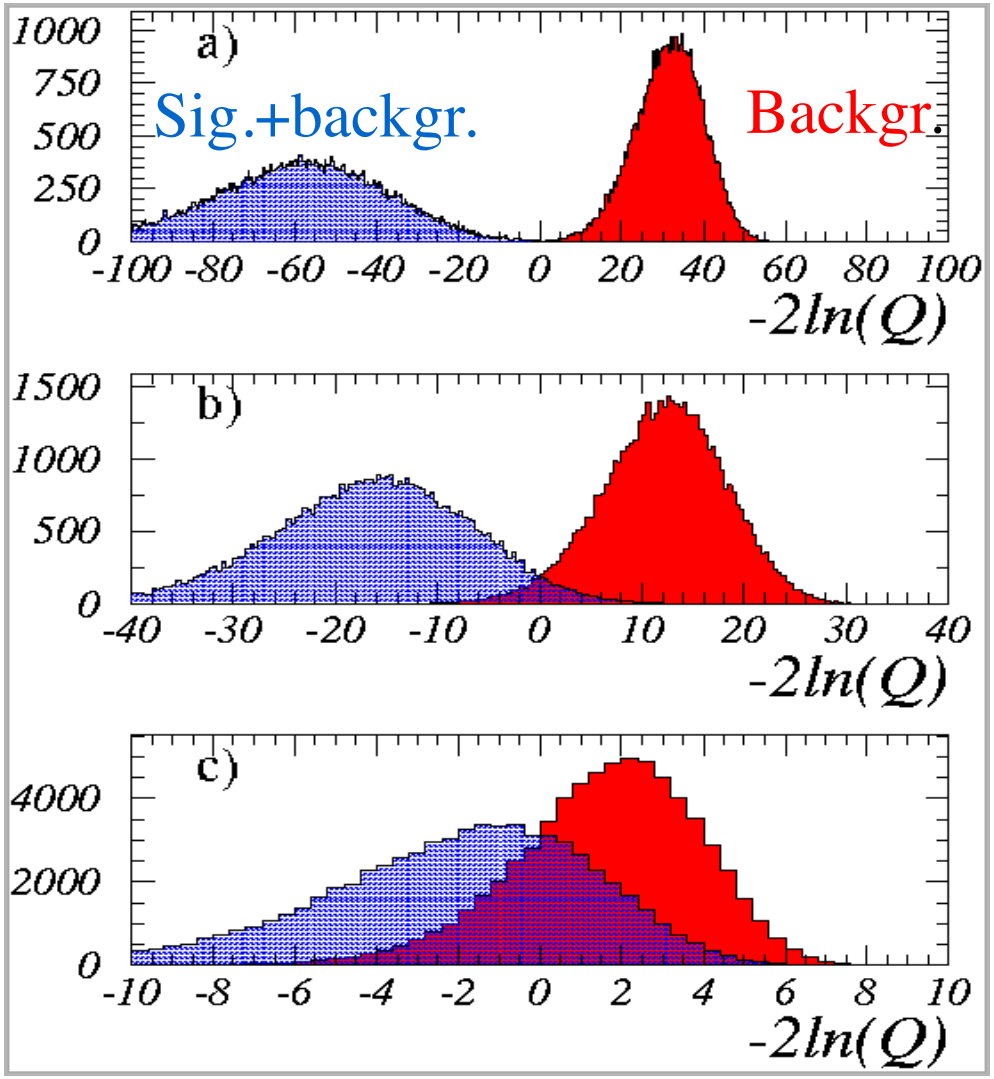
CLb: conf. nel background

1-CLb: significativita' del test

CLsb: conf nel sig.+bg.

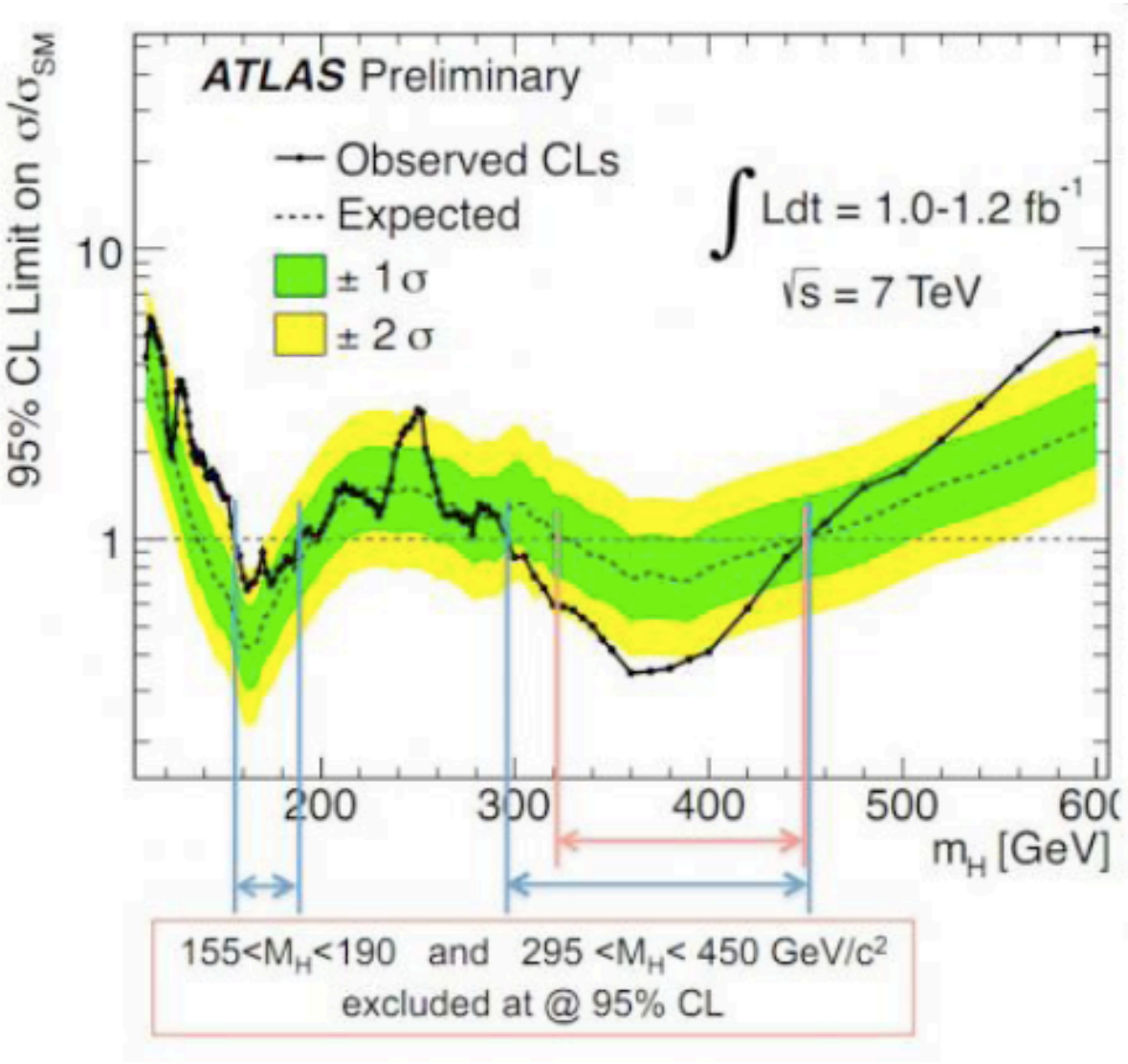
CLs=CLsb/CLb: approssimazione della conf. nel sig.

Sensibilita' sperimentale



- Ottima sensibilita'. Ottime possibilita' di scoperta o in caso di assenza del fenomeno, esclusione del segnale o limiti molto stringenti
- Buona sensibilita' . Probabile un risultato di esclusione con buoni livelli di confidenza
- **Scarsa sensibilita', probabili risultati ambigui, importante la scelta di metodi statistici raffinati**

Ricerca dell'Higgs: risultati di ATLAS



Conclusioni

- Metodi avanzati per il trattamento statistico dei dati sono necessari per l'interpretazione delle misure sperimentali
- La sensibilita' ai fenomeni di interesse stabilisce quanto critica sia la modalita' d'analisi sui risultati

- Restano comunque tre gli step necessari alla misura:
  - studio della componente stocastica delle distribuzioni
  - valutazione e trattamento del background
  - valutazione e trattamento degli errori sistematici