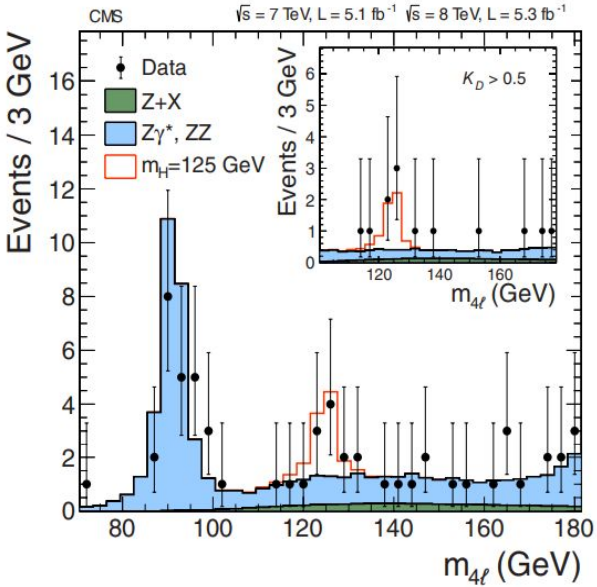
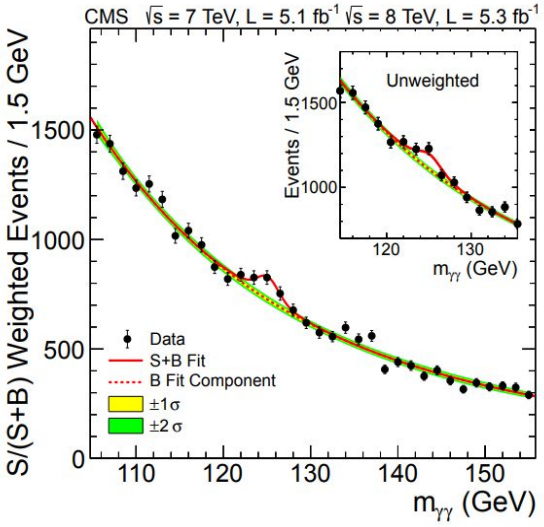
Introdução a estatística para análise de dados em HEP

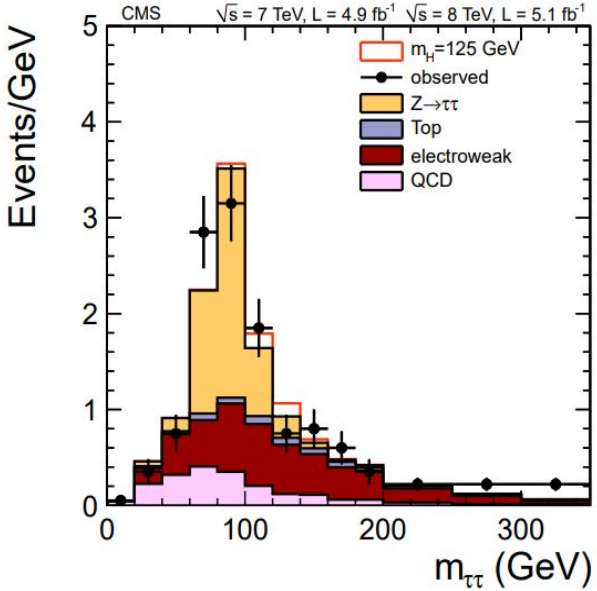
Índice

- Motivação
- Métodos estatísticos básicos
- Probabilidade e distribuições
- Ajuste de função
- Teste do χ^2

Motivação


CMS
√s = 7 TeV, L = 5.1 fb-1 √s = 8 TeV, L = 5.3 fb-1
Data
Z+X
Zγ*, ZZ
mH=125 GeV
Events / 3 GeV
KD > 0.5
m4ℓ (GeV)


CMS √s = 7 TeV, L = 5.1 fb-1 √s = 8 TeV, L = 5.3 fb-1
Unweighted
Events / 1.5 GeV
Data
S+B Fit
B Fit Component
±1σ
±2σ
S/(S+B) Weighted Events / 1.5 GeV
mγγ (GeV)


CMS
√s = 7 TeV, L = 4.9 fb-1 √s = 8 TeV, L = 5.1 fb-1
mH=125 GeV
observed
Z→ττ
Top
electroweak
QCD
Events/GeV
mττ (GeV)

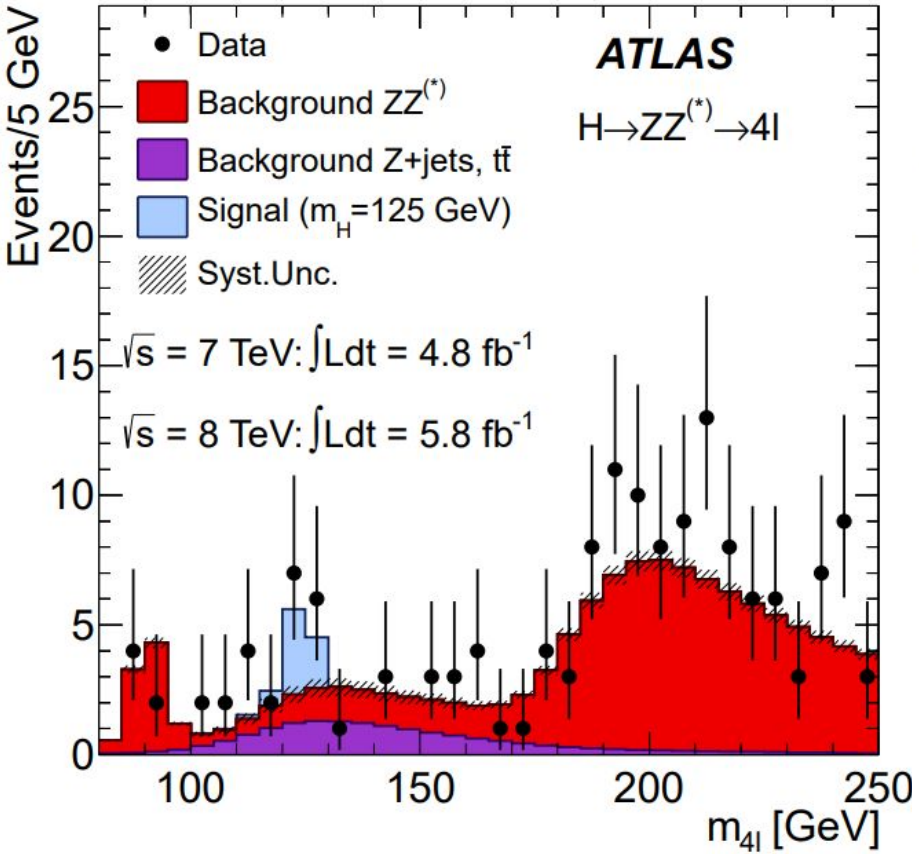
Motivação

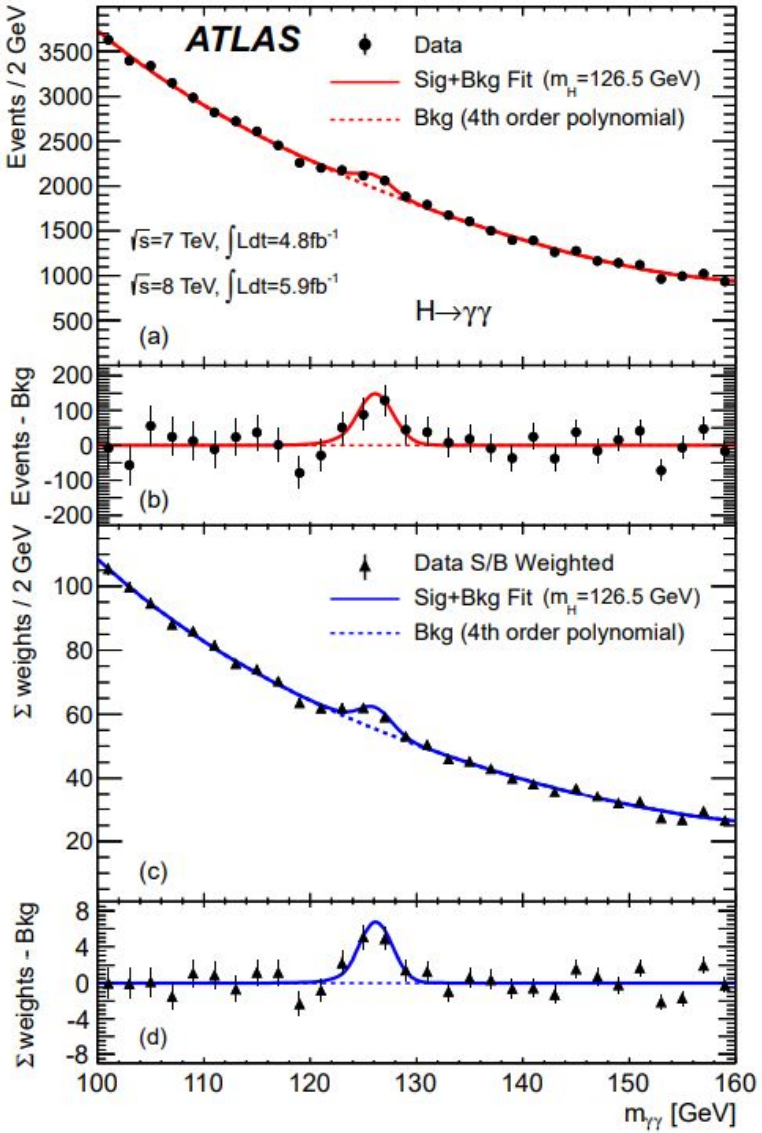
8 Conclusions

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV in the CMS experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 fb^{-1} at 7 TeV and 5.3 fb^{-1} at 8 TeV. The search is performed in five decay modes: $\gamma\gamma$, ZZ, W^+W^-, $\tau^+\tau^-$, and $\text{b}\bar{\text{b}}$. An excess of events is observed above the expected background, with a local significance of 5.0 σ, at a mass near 125 GeV, signalling the production of a new particle. The expected local significance for a standard model Higgs boson of that mass is 5.8 σ. The global p-value in the search range of 115–130 (110–145) GeV corresponds to 4.6 σ (4.5 σ). The excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ, and a fit to these signals gives a mass of $125.3 \pm 0.4\,(\text{stat.}) \pm 0.5\,(\text{syst.})$ GeV. The decay to two photons indicates that the new particle is a boson with spin different from one. The results presented here are consistent, within uncertainties, with expectations for a standard model Higgs boson. The collection of further data will enable a more rigorous test of this conclusion and an investigation of whether the properties of the new particle imply physics beyond the standard model.

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

Motivação


ATLAS
Data
Background ZZ(*)
Background Z+jets, tt̄
Signal (m_H=125 GeV)
Syst.Unc.
H→ZZ(*)→4l
√s = 7 TeV: ∫Ldt = 4.8 fb-1
√s = 8 TeV: ∫Ldt = 5.8 fb-1
Events/5 GeV
m4l [GeV]


ATLAS
Data
Sig+Bkg Fit (m_H=126.5 GeV)
Bkg (4th order polynomial)
√s=7 TeV, ∫Ldt=4.8fb-1
√s=8 TeV, ∫Ldt=5.9fb-1
H→γγ
(a)
Events / 2 GeV
Events - Bkg
(b)
Data S/B Weighted
Sig+Bkg Fit (m_H=126.5 GeV)
Bkg (4th order polynomial)
(c)
Σ weights / 2 GeV
Σ weights - Bkg
(d)
mγγ [GeV]

Motivação

10. Conclusion

Searches for the Standard Model Higgs boson have been performed in the $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ channels with the ATLAS experiment at the LHC using 5.8–5.9 fb^{-1} of pp collision data recorded during April to June 2012 at a centre-of-mass energy of 8 TeV. These results are combined with earlier results [17], which are based on an integrated luminosity of 4.6–4.8 fb^{-1} recorded in 2011 at a centre-of-mass energy of 7 TeV, except for the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels, which have been updated with the improved analyses presented here.

The Standard Model Higgs boson is excluded at 95% CL in the mass range 111–559 GeV, except for the narrow region 122–131 GeV. In this region, an excess of events with significance 5.9 σ, corresponding to $p_0 = 1.7 \times 10^{-9}$, is observed. The excess is driven by the two channels with the highest mass resolution, $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$, and the equally sensitive but low-resolution $H \to WW^{(*)} \to \ell\nu\ell\nu$ channel. Taking into account the entire mass range of the search, 110–600 GeV, the global significance of the excess is 5.1 σ, which corresponds to $p_0 = 1.7 \times 10^{-7}$.

These results provide conclusive evidence for the discovery of a new particle with mass $126.0 \pm 0.4\,(\text{stat}) \pm 0.4\,(\text{sys})$ GeV. The signal strength parameter μ has the value 1.4 ± 0.3 at the fitted mass, which is consistent with the SM Higgs boson hypothesis $\mu = 1$. The decays to pairs of vector bosons whose net electric charge is zero identify the new particle as a neutral boson. The observation in the diphoton channel disfavours the spin-1 hypothesis [140, 141]. Although these results are compatible with the hypothesis that the new particle is the Standard Model Higgs boson, more data are needed to assess its nature in detail.

Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

Motivação

Um pleno conhecimento de estatística é essencial para a extração de resultados em uma análise em Física de Altas Energias.

- Seção de choque.
- Extração de parâmetros (massa invariante, acoplamentos, etc…).
- Estabelecer limites em busca por nova física.

Outras aulas abordarão tópicos relacionados a estatística, como a aula de métodos a Monte Carlo e a aula sobre o RooFit

Métodos estatísticos básicos

- Relembrando FG
- Erros do tipo A e B
  - Estatístico e sistemático
- Propagação de erros
- Intervalo de Confiança
- Método dos Mínimos Quadrados

Métodos estatísticos básicos: **Erros do tipo A e B**

- Erros do tipo A (estatístico):
  - **Natureza**: São erros aleatórios, ou seja, flutuações estatísticas que ocorrem em medidas repetidas de uma mesma grandeza.
  - **Origem**: São variações que ocorrem devido à aleatoriedade e incertezas na medição. Esses erros resultam da variabilidade natural dos processos de medição e podem ser estimados por métodos estatísticos.
  - **Estimação**: São estimados por métodos estatísticos, como o cálculo do desvio padrão da média de um conjunto de medidas.
  - **Exemplo**: Contagem de Eventos, imagine que você está contando o número de eventos em um experimento de física de partículas. A contagem pode ter variações de uma execução do experimento para outra devido a flutuações aleatórias. Esses erros podem ser descritos estatisticamente e são frequentemente modelados usando distribuições de Poisson ou Gaussianas, dependendo da situação.

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{N}} \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2}$$ (Gaussiana)

Métodos estatísticos básicos: **Erros do tipo A e B**

- Erros do tipo B (sistemático):
  - **Natureza**: São erros sistemáticos, ou seja, erros que tendem a ocorrer sempre no mesmo sentido e com a mesma magnitude em todas as medidas.
  - **Origem**: Devem-se a causas conhecidas, como imperfeições nos instrumentos de medida, calibrações incorretas ou efeitos físicos não considerados no modelo teórico.
  - **Estimação**: São estimados por outros métodos, como a análise da especificação técnica do instrumento, a comparação com padrões de referência ou a avaliação de efeitos teóricos.
  - **Exemplo**: Calibração do Detector, em experimentos de física de altas energias, os detectores são calibrados para medir a energia ou a posição das partículas. Se o detector estiver descalibrado, todas as medições de energia terão um desvio sistemático em relação ao valor verdadeiro. Esse tipo de erro é sistemático porque afeta todas as medições da mesma maneira e precisa ser corrigido por meio de calibração e ajustes de precisão.

Métodos estatísticos básicos: **Propagação de erros**

- É um método utilizado para determinar a incerteza (ou erro) em uma quantidade calculada a partir de outras quantidades que possuem suas próprias incertezas.
- Dada uma grandeza f:

$$f = f(x_1, x_2, x_3, ..., x_n)$$

Sua incerteza será dada por:

$$\sigma_f = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \sigma_{x_2}^2 + \ldots + \left(\frac{\partial f}{\partial x_n}\right)^2 \sigma_{x_n}^2}$$

Métodos estatísticos básicos: **Intervalo de Confiança**

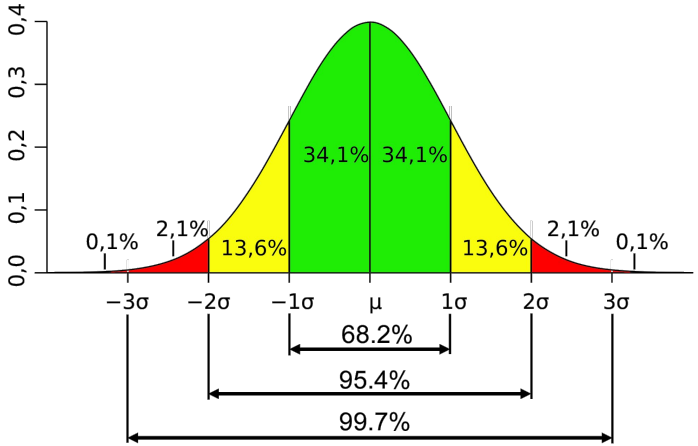
Lei dos Erros

As estimativas de erros por métodos estatísticos, a partir de uma amostra de medidas diretas de uma grandeza, baseia-se na hipótese ou constatação que, quando o número de medidas cresce progressivamente, a distribuição de freqüências dessas medidas tende à chamada distribuição normal ou distribuição de Gauss, ou ainda distribuição gaussiana

A distribuição de Gauss é dada por:

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma_x^2}}$$

onde $\mu = \overline{x}$ e σ_x^2 é a variância.

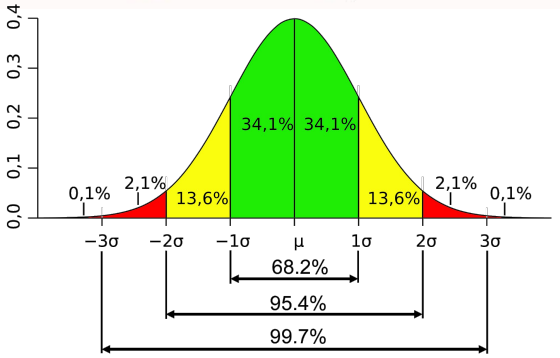


Métodos estatísticos básicos: **Intervalo de Confiança**

Do ponto de vista estatístico, o erro associado a uma estimativa define um intervalo de variação ou de incerteza ao qual se pode atribuir um nível probabilístico de confiança.

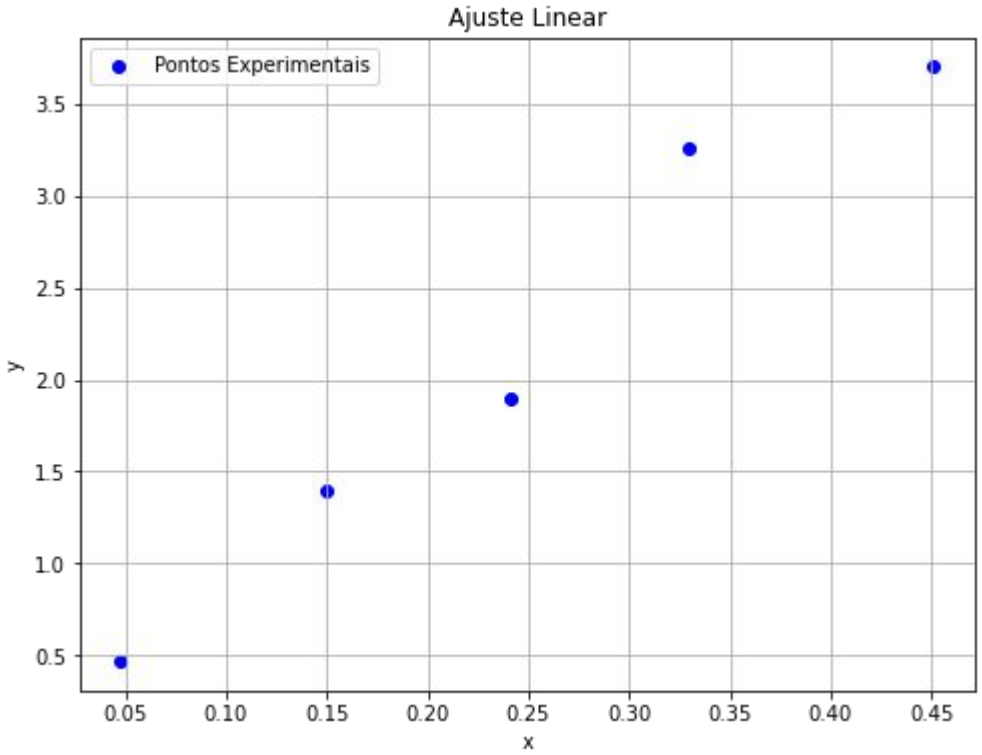
Considerar que as medidas de uma grandeza x se distribuem normalmente e caracterizar por $\sigma_{\bar{x}}$ o erro na estimativa $\bar{x}$, para o valor esperado da grandeza, significa que o *nível de confiança*[26] de que o intervalo $(\bar{x}-\sigma_{\bar{x}}\,,\ \bar{x}+\sigma_{\bar{x}})$ contenha o valor esperado para a grandeza é de 68,3%, ou que esse *intervalo de confiança* é de 68,3%.

Intervalo de Confiança	Nível de Confiança (CL)
$(\bar{x}-0{,}67\,\sigma_{\bar{x}}\,,\ \bar{x}+0{,}67\,\sigma_{\bar{x}})$	50,0 %
$(\bar{x}-1{,}00\,\sigma_{\bar{x}}\,,\ \bar{x}+1{,}00\,\sigma_{\bar{x}})$	68,3 %
$(\bar{x}-1{,}65\,\sigma_{\bar{x}}\,,\ \bar{x}+1{,}65\,\sigma_{\bar{x}})$	90,0 %
$(\bar{x}-1{,}96\,\sigma_{\bar{x}}\,,\ \bar{x}+1{,}96\,\sigma_{\bar{x}})$	95,0 %
$(\bar{x}-2{,}00\,\sigma_{\bar{x}}\,,\ \bar{x}+2{,}00\sigma_{\bar{x}})$	95,5 %
$(\bar{x}-3{,}00\,\sigma_{\bar{x}}\,,\ \bar{x}+3{,}00\,\sigma_{\bar{x}})$	99,7 %



Métodos estatísticos básicos: **Método dos Mínimos Quadrados**

- Qual é a melhor função linear que se ajusta a um conjunto de dados experimentais?

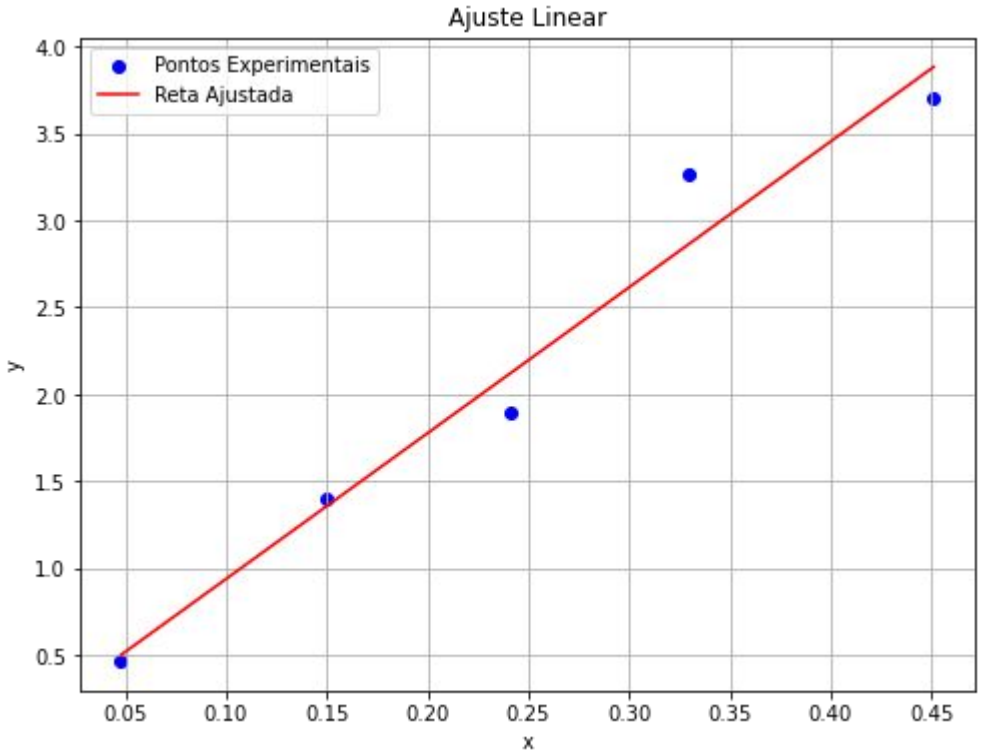


Métodos estatísticos básicos: **Método dos Mínimos Quadrados**

- Qual é a melhor função linear que se ajusta a um conjunto de dados experimentais?

- **A função que melhor se ajusta é aquela que minimiza a soma dos quadrados dos desvios**

$$y'(x) = m \cdot x + b$$



Métodos estatísticos básicos: **Método dos Mínimos Quadrados**

- Qual é a melhor função linear que se ajusta a um conjunto de dados experimentais?

- **A função que melhor se ajusta é aquela que minimiza a soma dos quadrados dos desvios**

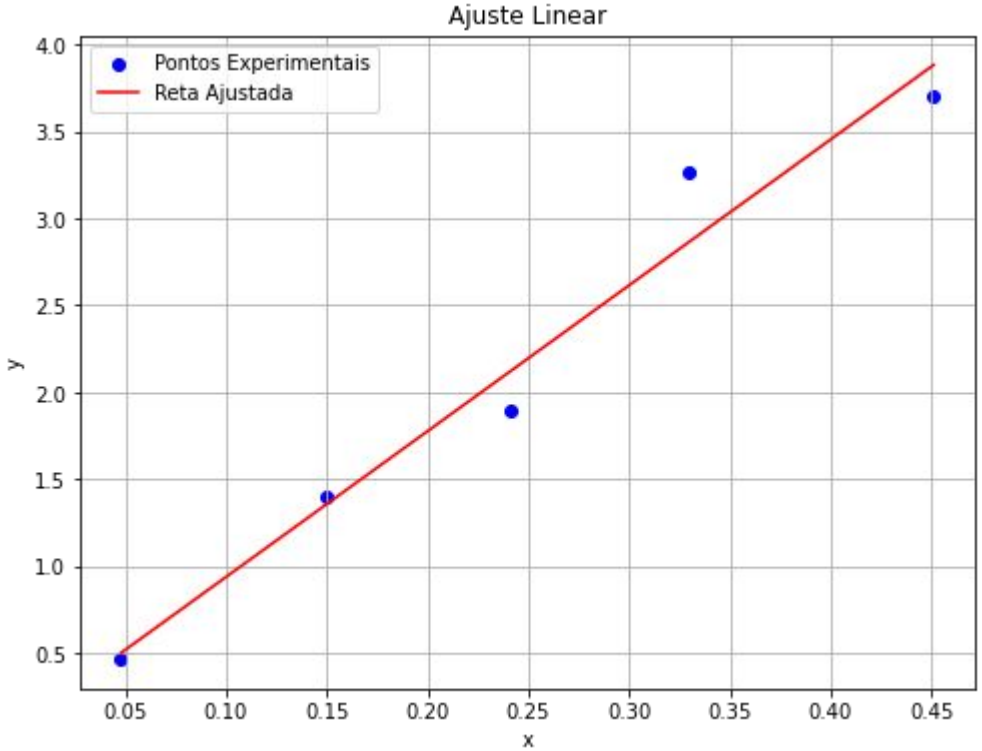
$$y'(x) = m \cdot x + b$$

$$m = \frac{M_{xy}}{M_{xx}}$$

$$b = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i\right)$$

$$M_{xy} = \sum_{i=1}^{N} x_i.y_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i\right)$$

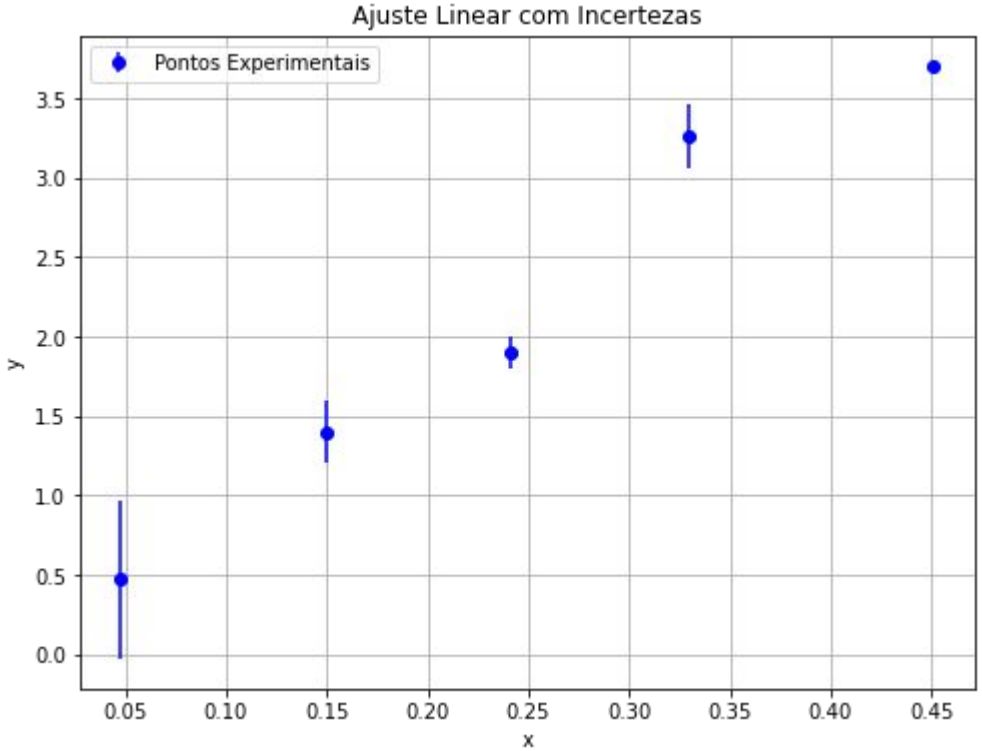
$$M_{xx} = \sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2$$



Métodos estatísticos básicos: **Método dos Mínimos Quadrados**

- Podemos também considerar a incerteza em cada medida

- **E aplicar o MMQ ponderado**

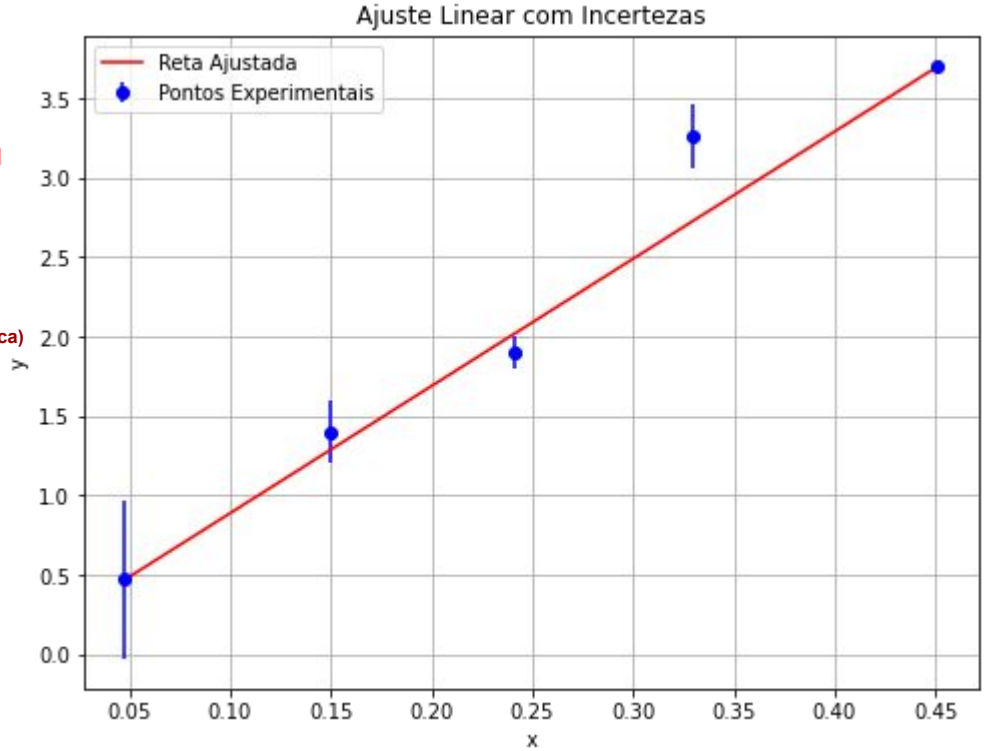


Métodos estatísticos básicos: **Método dos Mínimos Quadrados**

- Podemos também considerar a incerteza em cada medida.

- **Exercício 1: Deduza as equações para o ajuste linear para o caso em que as incertezas em y são diferentes para cada ponto.**

(Sugestão: seção F.2 do livro Estimativas e Erros em Experimentos de Física)



Probabilidade e Distribuições

Probabilidade

- A probabilidade fornece a base matemática para interpretar, analisar e modelar dados em física de altas energias.

- É essencial para lidar com as incertezas inerentes aos experimentos e para desenvolver modelos teóricos que possam ser testados e verificados empiricamente.

Probabilidade e Distribuições

Probabilidade

- **Probabilidade à *priori***: refere-se à probabilidade atribuída a um evento ou hipótese antes de se considerar quaisquer dados ou evidências adicionais. É uma forma de representar nosso conhecimento ou crenças iniciais sobre o evento com base em informações prévias ou suposições antes de analisar os dados.

- **Probabilidade à *posteriori***: conceito central na estatística bayesiana e refere-se à probabilidade de uma hipótese ou parâmetro após ter considerado as evidências ou dados disponíveis. É uma atualização da probabilidade a priori com base nas novas informações fornecidas pelos dados observados.

Probabilidade e Distribuições

Probabilidade

processo aleatório

- evento (i) $\Leftrightarrow$ ocorrência de uma face ou valor i
- freqüência (n_i) $\Leftrightarrow$ número de ocorrências da face i

- probabilidade a posteriori $\rightarrow p_i = \frac{n_i}{N}$ (experimental)
- probabilidade a priori $\rightarrow p_i = \frac{1}{6}$ (teórica)

Mais aplicações na aula de métodos a MC

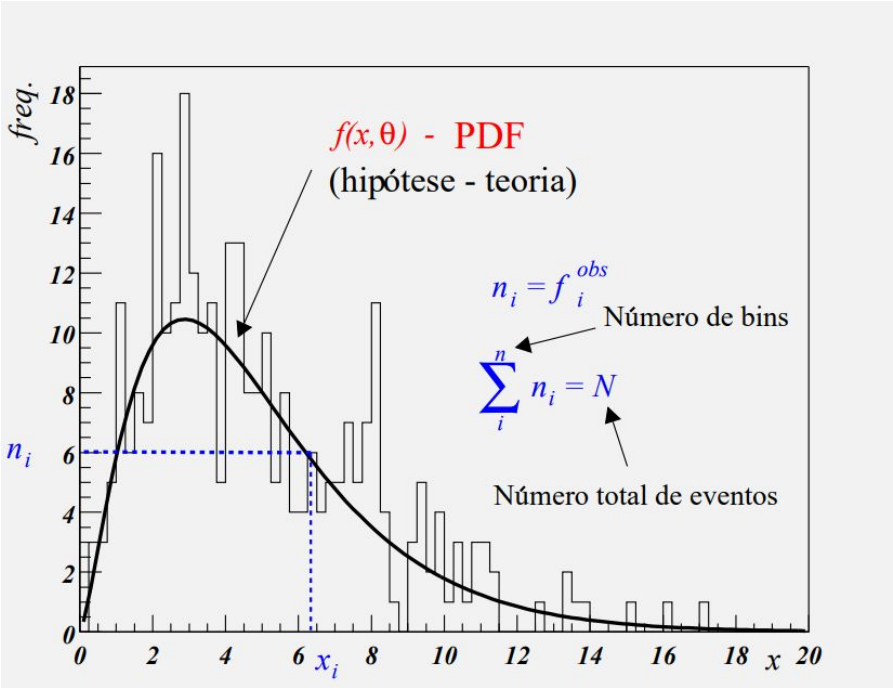
Probabilidade e Distribuições

Distribuições de probabilidade

- **Função Densidade de Probabilidade** (pdf) ou Distribuição de Probabilidade: é a função que fornece a probabilidade de se observar um valor x de uma determinada variável contínua dentro de um intervalo infinitesimal |x,x+dx|
  - Em física de altas energias, as medições experimentais frequentemente seguem distribuições contínuas. A pdf descreve como os dados se distribuem ao longo dos possíveis valores.
  - Permite ajustar modelos teóricos aos dados experimentais, ajudando a testar a validade das teorias e a entender a física subjacente.

Probabilidade e Distribuições

Distribuições de probabilidade



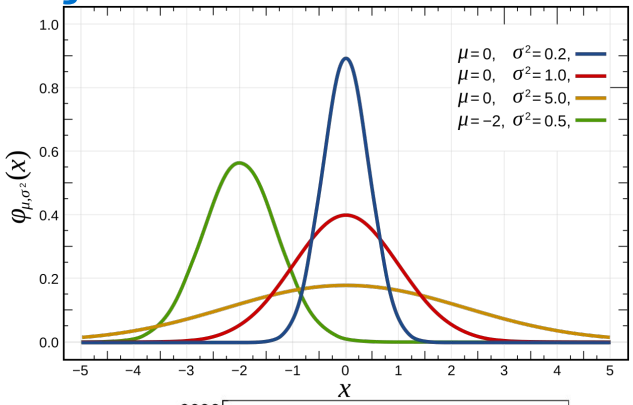
Mais aplicações na aula de métodos a MC

Probabilidade e Distribuições: Distribuições básicas

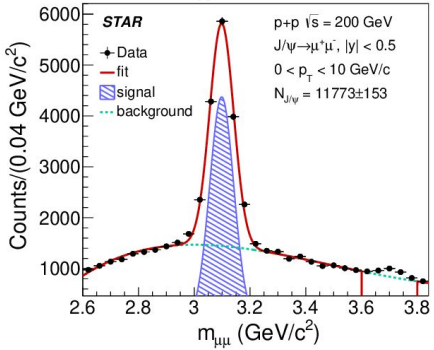
Distribuições Gaussiana

- A Distribuição de Gauss, também conhecida como distribuição normal, é uma distribuição contínua de probabilidade que é simétrica em torno de sua média.
- É caracterizada por dois parâmetros: a média (μ) e a variância (σ²).

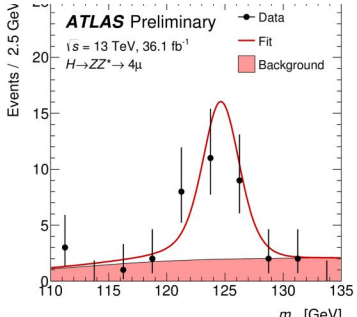
$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma_x^2}}$$



- Modelagem de Erros e Incertezas: A distribuição normal é frequentemente usada para modelar erros e incertezas nas medições experimentais devido ao Teorema Central do Limite, que afirma que a soma de muitas variáveis aleatórias independentes e identicamente distribuídas tende a seguir uma distribuição normal.



- Ajuste de Modelos e Estimação de Parâmetros: Muitos modelos teóricos e ajustes experimentais assumem que as variáveis seguem uma distribuição normal. Isso simplifica a estimação de parâmetros e a realização de inferências estatísticas.
  - Um exemplo é a estimativa da massa das partículas



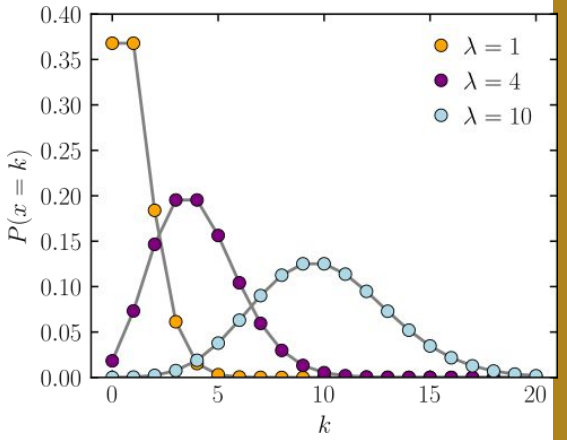
Probabilidade e Distribuições: Distribuições básicas

Distribuições de Poisson

- é uma distribuição discreta de probabilidade que descreve o número de eventos que ocorrem em um intervalo fixo de tempo ou espaço, dado que esses eventos ocorrem com uma taxa média constante e independentemente uns dos outros.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- onde λ é a taxa média de eventos (ou a média da distribuição) e k é o número de eventos observados.
- É ideal para modelar a contagem de eventos raros ou pouco frequentes em física de altas energias, como o número de partículas detectadas em um detector.



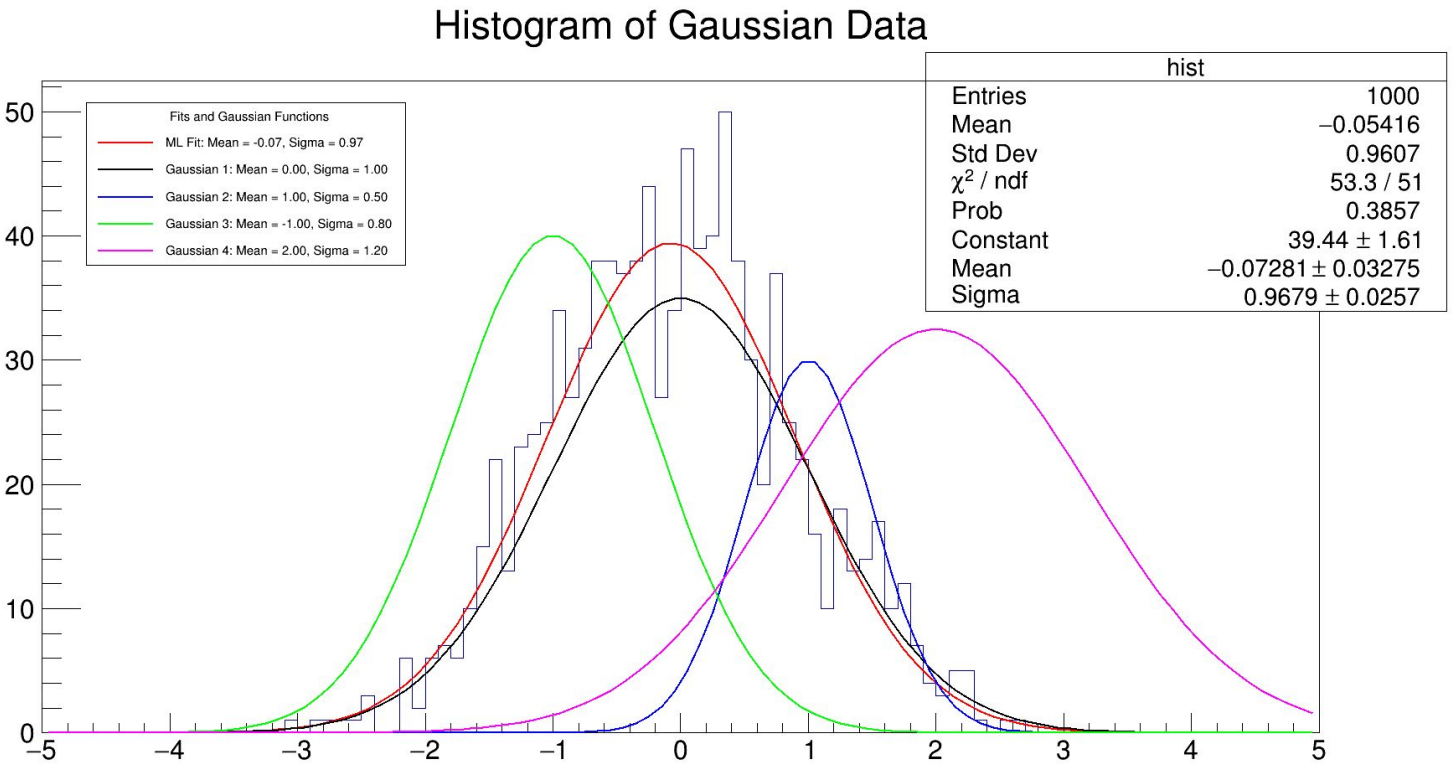
Ajuste de Função

- O Método dos Mínimos Quadrados é uma ferramenta útil e é amplamente utilizado para ajustar funções lineares.
- Em Física de Altas Energias, muitas vezes encontramos distribuições mais complexas.
- O Método da Máxima Verossimilhança, por outro lado, permite modelar uma ampla variedade de distribuições de probabilidade, tornando-o mais flexível e adequado para a análise de dados em física de altas energias.
- Método da Máxima Verossimilhança nos diz que:
  - A função que melhor se ajusta é aquela que maximiza a probabilidade de observar os dados fornecidos, ou seja, a função cuja verossimilhança é maximizada.

Ajuste de Função

- Aplicações comuns da Máxima Verossimilhança em FAE incluem:
  - Ajuste de curvas de eficiência: Determinando a eficiência de detectores em função da energia das partículas.
  - Estimação de parâmetros de modelos teóricos: Extraindo informações sobre as propriedades de partículas e interações.
  - Construção de histogramas e gráficos de contorno: Visualizando a distribuição de eventos em um espaço multidimensional.
- Softwares como ROOT e RooFit fazem uso desse método para os seus ajustes de funções.
  - Nas próximas aulas abordaremos esses ajustes.

Ajuste de Função



Teste do χ^2

Os dados observados são consistentes com o modelo proposto?

- O teste do χ^2 é uma ferramenta estatística usada para avaliar a adequação de um modelo aos dados observados.
- Ele mede a discrepância entre os dados observados e os valores esperados de um modelo, ajudando a verificar a qualidade do ajuste.
- É calculado por:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- onde O_i são os valores observados e E_i são os valores esperados.

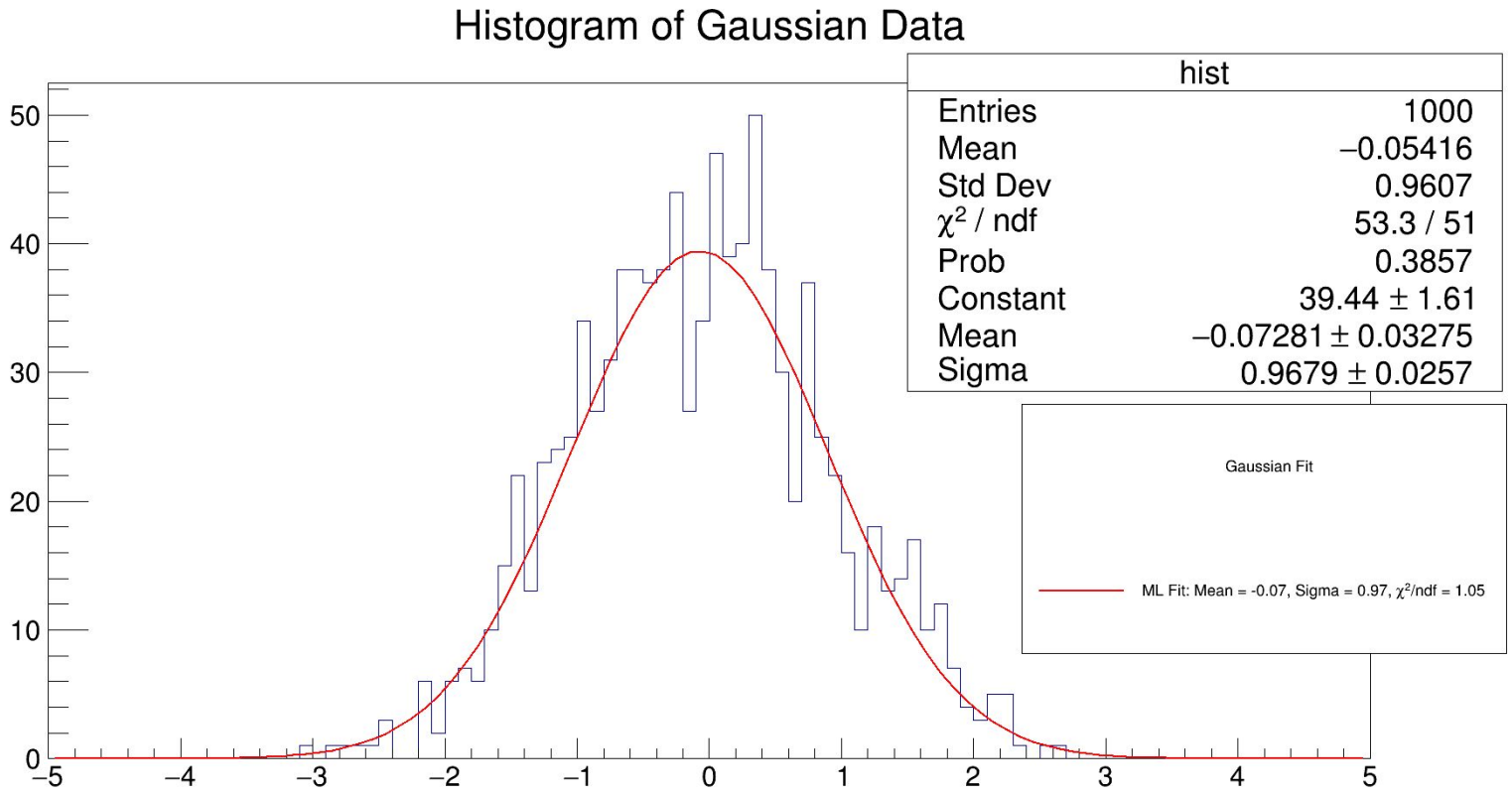
Teste do χ^2

- O melhor ajuste é obtido quando o valor de χ^2 dividido pelo número de graus de liberdade (ndf) se aproxima de 1.

- Graus de liberdade referem-se ao número de valores independentes que podem variar em uma análise.

- Em um ajuste de curva, os graus de liberdade são geralmente dados pelo número de pontos de dados menos o número de parâmetros ajustados.

Teste do χ^2



Exercícios

Exercício 2: A seção de choque de um processo definido como sinal pode ser obtida experimentalmente pela equação:

$$\sigma = \frac{N_{Total} - N_{background}}{\mathcal{L}}$$

onde N_{Total} é o número de eventos observados; $N_{background}$ é o número de eventos de fundo esperados e L é a luminosidade integrada. Considerando para este caso que o número total de eventos observados foi 2567, o número de eventos de fundo esperado é 1223.5, e a luminosidade integrada de $25fb^{-1}$ com uma incerteza sistemática de 10%. Calcule o valor da seção de choque propagando separadamente as incertezas estatísticas (lembrando que aqui se trata de distribuição de Poisson) e sistemáticas.

Exercício 3: A seleção de eventos em uma análise em Física de Altas Energias por busca de acoplamentos anômalos previu um número de eventos de fundo de 0.07 eventos após todos os cortes. Ao olhar para os dados, também após todos os cortes, se observou zero eventos. Diferentes modelos para acoplamentos anômalos previram diferentes números de eventos a serem observados, desde 0.09 até 35 eventos. Considerando que a contagem de eventos se dá através de uma pdf de Poisson, calcule até quantos eventos esperados podemos excluir com essa análise com 95% de C.L. (ver a tese "Study of WW Central Exclusive Production in the semileptonic channel with tagged protons at CMS detector" cap. 4.4)

Exercício 4: Mostre que a melhor função que se adequada aos dados é quando χ^2/ndf -> 1. (ver Vuolo 12.6)

Referências