DFNAE
DEPARTAMENTO DE FÍSICA NUCLEAR E ALTAS ENERGIAS

Introdução ao RooStat

Sandro Fonseca de Souza

21/09/2021

Referência:

- RooStats twiki page: (link)

Sumário

- Definição do RooStats

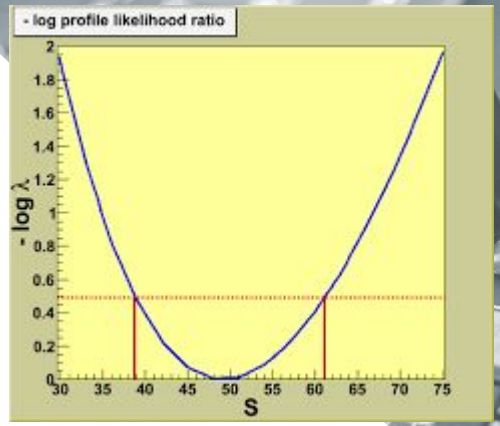
- Aplicativos de estatística e tecnologia do RooStats

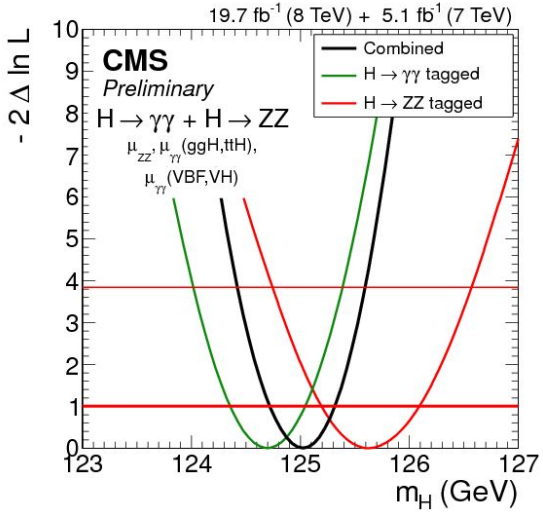
Definição do RooStat

- Projetado para fornecer ferramentas estatísticas avançadas para experimentos do LHC
  - fatorar a descrição do modelo a partir de cálculos estatísticos
  - implementar diferentes técnicas estatísticas
- suporte frequentista, bayesiano ou baseado em probabilidade
- fornecer utilitários para combinação de resultados
- contribuição conjunta entre ATLAS, CMS, RooFit e ROOT

Aplicações estatísticas

- Estimativa pontual (coberta pelo RooFit)
- Estimativa de intervalos confiança (limites superiores e inferiores)
- Testes de hipótese (exemplos de significância de uma descoberta)





Aplicações estatísticas

- Construído sobre o RooFit
  - fornece descrição de modelo genérico (baseado em histogramas) usando modelos não combinados ou parametrizados (*unbinned*) ou combinados (*binned*)
  - possibilita ferramentas para facilitar ferramentas de criação de modelos para combinação de modelos
- Utilizar as principais bibliotecas do ROOT
  - minimização (Minuit), integração numérica e etc.
  - ferramentas adicionais fornecidas quando necessário (por exemplo, MC Markov-Chain)

Tecnologia do RooStats

O HistFactory é uma ferramenta para construir funções parametrizadas de densidade de probabilidade (pdfs) na estrutura RooFit/RooStats com base em histogramas ROOT simples e organizados em um arquivo XML

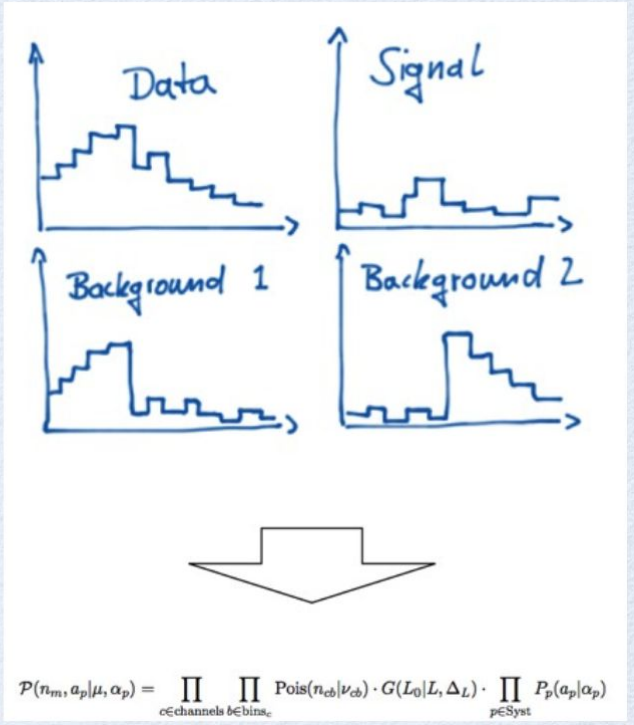
Construindo modelos com Roofit

Mathematical concept		RooFit class
variable	x	RooRealVar
function	$f(x)$	RooAbsReal
PDF	$f(x)$	RooAbsPdf
space point	$\vec{x}$	RooArgSet
integral	$\int_{x_{min}}^{x_{max}} f(x)dx$	RooRealIntegral
list of space points		RooAbsData


```
RooRealVar x("x","x",2,-10,10)
RooRealVar s("s","s",3) ;
RooRealVar m("m","m",0) ;
RooGaussian g("g","g",x,m,s)
```

Gaussian(x,m,s)

Modelos com o HistFactory

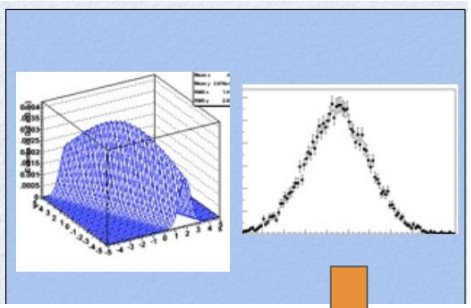


Terminologia

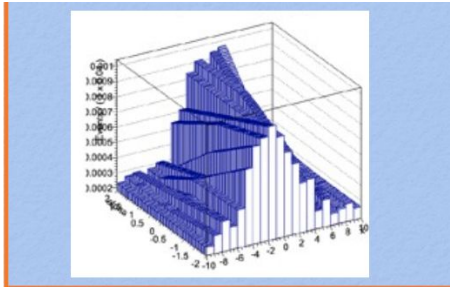
- **observable** (observável) - algo que você mede em um experimento, por exemplo, o momento de uma partícula. Frequentemente, uma função de quantidades medidas, por exemplo, uma massa invariante de várias partículas
- **global observable** or **auxiliary observable** (observável global ou observável auxiliar) - um observável de outra medição, por exemplo, a luminosidade integrada
- **model** (modelo) - um conjunto de funções de densidade de probabilidade (PDFs), que descreve as distribuições dos observáveis ou funções dos observáveis
- **model parameter** (parâmetro do modelo) - qualquer variável em sua expressão da PDF, que não seja observável
- **parameter of interest (POI)** (parâmetro de interesse) - um parâmetro do modelo que você estuda, por exemplo, uma seção transversal de seu processo de sinal
- **nuisance parameter** (parâmetro nuisance) - todos os outros parâmetros do modelo, que não são o seu parâmetro de interesse
- **data** or **dataset** (dados ou conjunto de dados) - um conjunto de valores observáveis, medidos em um experimento ou simulados
- **likelihood** (verossimilhança) - um modelo calculado para um determinado conjunto de dados
- **hypothesis** (hipótese) - um modelo específico, com observáveis especificados, POI, parâmetros nuisance e PDFs anteriores (no caso de inferência bayesiana)
- **prior PDF** - uma densidade de probabilidade para um parâmetro observável ou de modelo, que é conhecido a priori, ou seja, antes de uma medição ser considerada. Este é um conceito exclusivamente bayesiano. Prior não tem significado ou lugar em um tipo de inferência frequentista
- **Bayesian** (Bayesiana) - um tipo de inferência estatística que geralmente produz probabilidade da hipótese frente aos dados. Requer uma prévia
- **frequentist** (frequentista) - um tipo de inferência estatística que geralmente produz probabilidade dos dados frente à hipótese

Exemplo de Análise

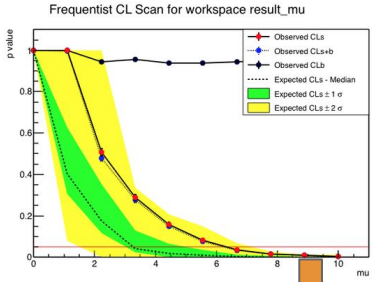
Classe RooWorkspace
Simplifique o empacotamento e o compartilhamento de modelos

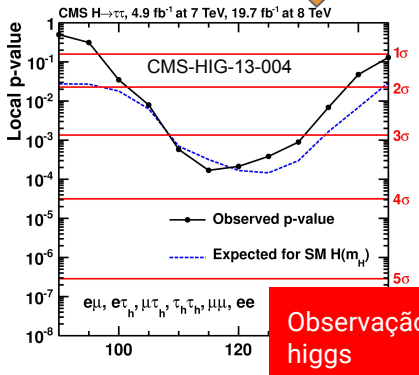


Pacote HistFactory
Construindo modelos a partir de modelos Monte Carlo



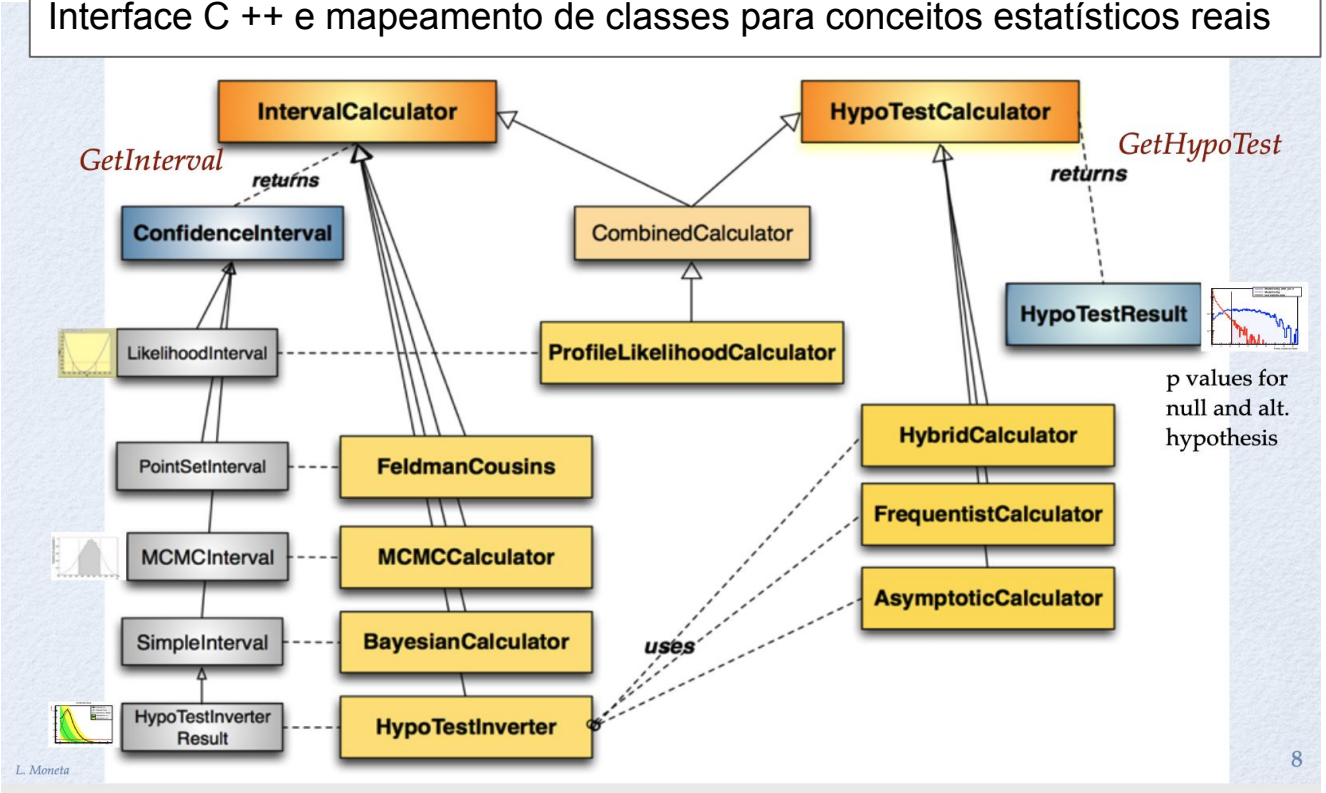
Kit de ferramentas RooStats
Testes estatísticos baseados em verossimilhança de modelos do RooFit





Design do RooStat

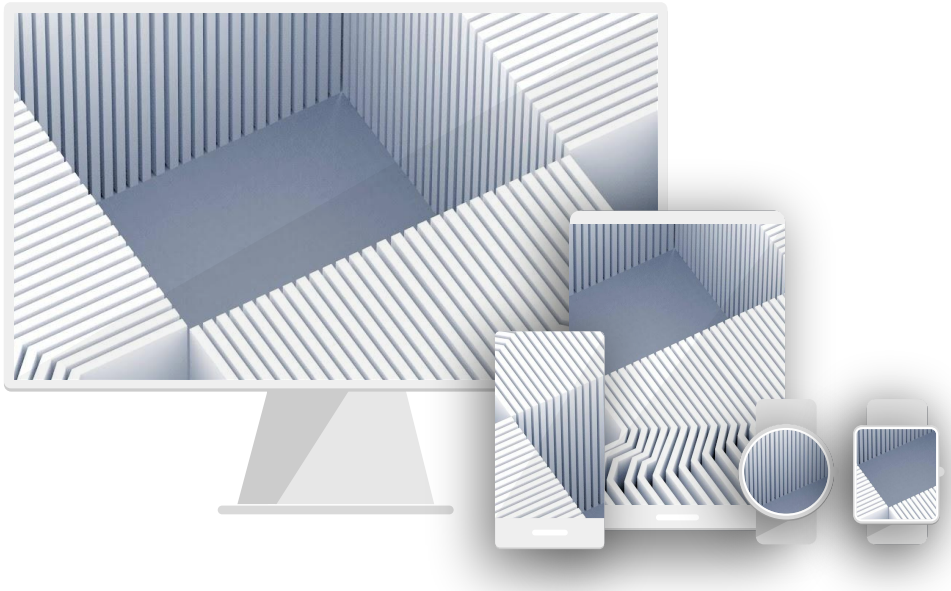
Interface C ++ e mapeamento de classes para conceitos estatísticos reais



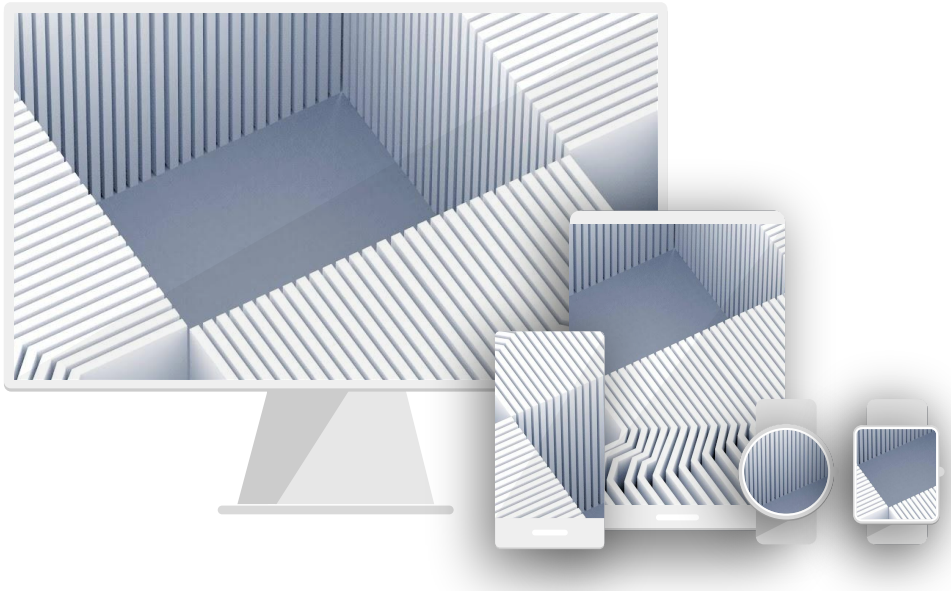
Notebooks

- Ajuste de função para o decaimentos de Higgs em dois fótons (link)
- Significância da descoberta vs massa (link)

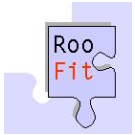
End



backup slides



The Philosophy behind the design the RooFit and RooStat



RooFit and HistFactory

1. Modularity, Generality and Flexibility
2. Construction the likelihood function L(x|p)

Complete description of likelihood model, persistable in ROOT file (RooFit pdf function) Allows full introspection and a-posteriori editing

ROOWorkspace

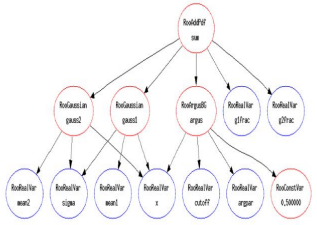
RooStats

Statistical tests on parameter of interest p

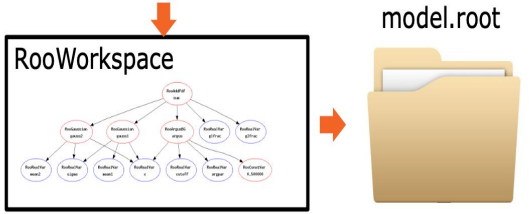
1. Procedure can be Bayesian, Frequentist or Hybrid, but always based on L(x|p)
2. Both steps (construction of the Likelihood func. and statistical test) are conceptually separated and implementation are independent.

RooWorkspace

- Container for all model objects
  - PDF and their parameters uncertainty and their shapes
  - (multiple) data sets
- Maintain complete description of the model
  - can be saved in a ROOT file
  - All information (likelihood function) is available for further analysis

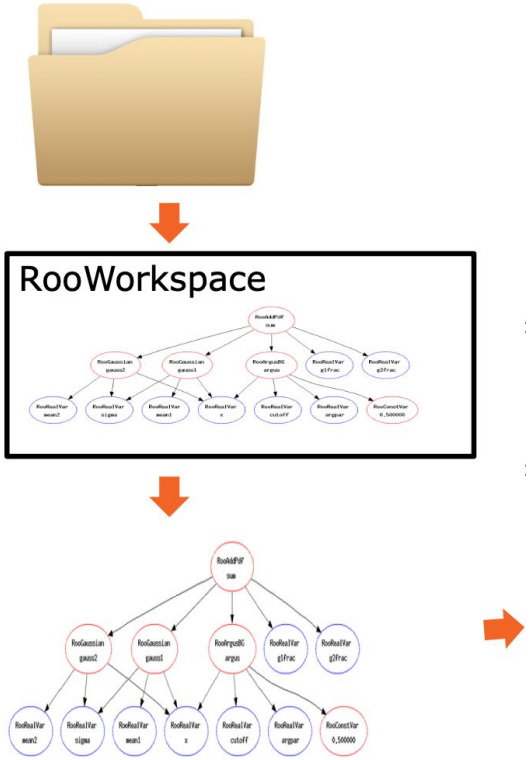


```
RooWorkspace w("w") ;
w.import(sum) ;
w.writeToFile("model.root") ;
```

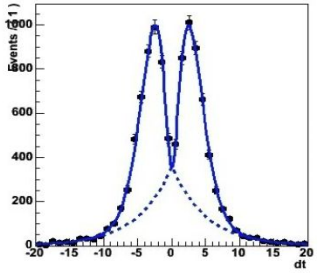


Wouter Verkerke, NIKHEF

RooWorkspace

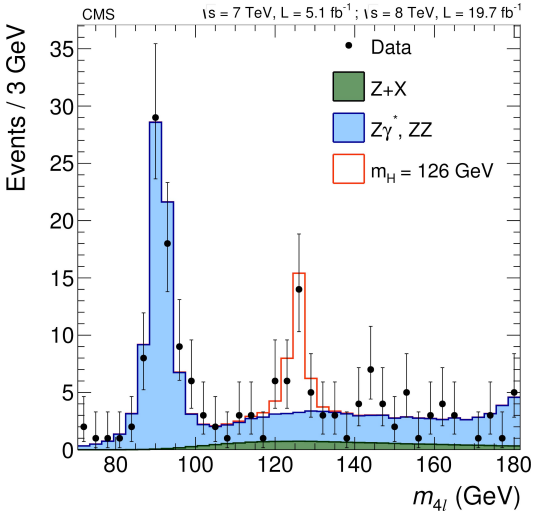



```
// Resurrect model and data
TFile f("model.root") ;
RooWorkspace* w = f.Get("w") ;
RooAbsPdf* model = w->pdf("sum") ;
RooAbsData* data = w->data("xxx") ;

// Use model and data
model->fitTo(*data) ;
RooPlot* frame =
          w->var("dt")->frame() ;
data->plotOn(frame) ;
model->plotOn(frame) ;
```



Exemplo de um modelo de RooWorkspace

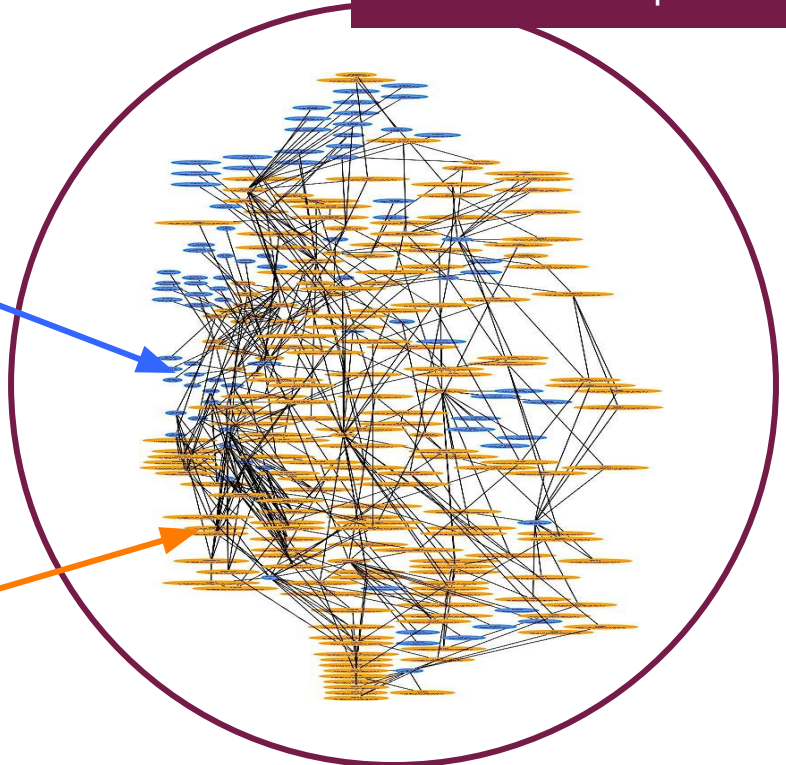
Modelo de probabilidade que descreve a distribuição de massa invariante ZZ incluindo todas as incertezas sistemáticas possíveis



ROOWorkspace

variaveis

objetos de funções



The Benefits of Modularity

RooFit and HistFactory

ROOWorkspace

"Simple fit"
ML Fit with HESSE or MINOS)

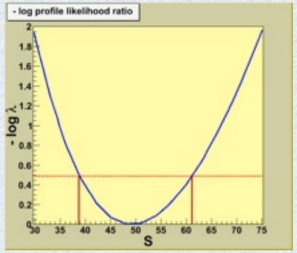
RooStats using Frequentist with toys

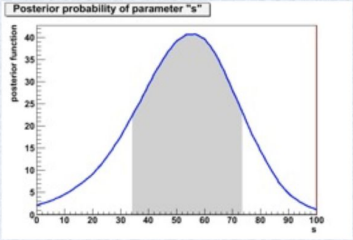
RooStats using Frequentist asymptotic

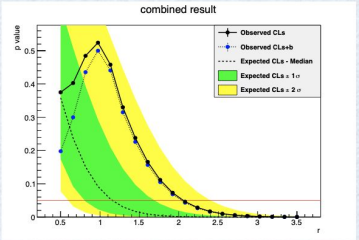
RooStats using Bayesian Markov-Chain Monte Carlo

Métodos do RooStats

- Interval Calculators
  - ProfileLikelihoodCalculator
    - interval estimation using the asymptotic properties of the likelihood function (Minos)
  - BayesianCalculator
    - interval estimation based on Bayes theorem using adaptive numerical integration
  - MCMCCalculator
    - Bayesian calculator using Markov-Chain Monte Carlo
  - HypoTestInverter
    - frequentist interval calculation using hypothesis test
    - can compute CLs limits or Feldman-Cousins interval

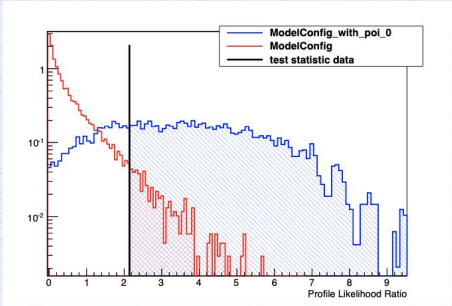


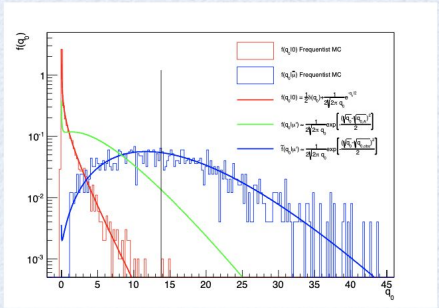




Métodos do RooStats (2)

- Hypothesis Test Calculators
  - FrequentistCalculator
    - frequentist hypothesis tests using pseudo-experiments to determine the test statistics distributions (parametric bootstrap)
  - HybridCalculator
    - same as frequentist calculator by using a bayesian treatment (marginalization) of systematic uncertainties
  - AsymptoticCalculator
    - hypothesis tests using asymptotic likelihood formulae
      - ➡ Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727,EPJC 71 (2011) 1-1

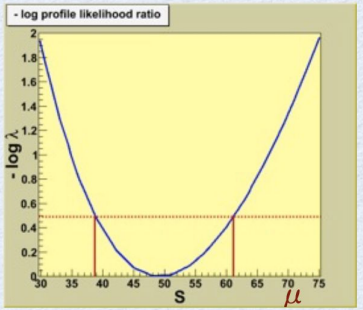




Using RooStats Calculators

- All RooStats calculators require same input:
  - model (described by the ModelConfig class which is linked to a workspace)
  - observed data
- Result is a ConfidenceInterval object or a HypoTestResult object
- Classes for plotting the result are also provided

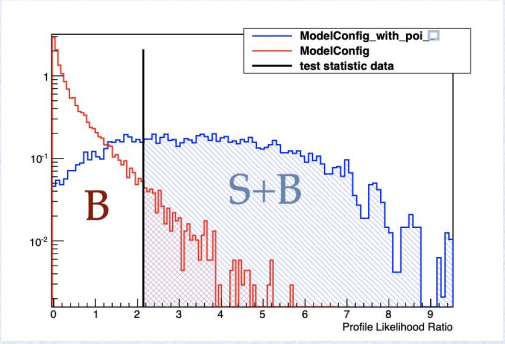
```
// create the class using data and model
ProfileLikelihoodCalculator plc(data, model);

// set the confidence level
plc.SetConfidenceLevel(0.683);

// compute the interval
LikelihoodInterval* interval = plc.GetInterval();

// plot the interval
LikelihoodIntervalPlot plot(interval);
plot.Draw();
```



RooStats Hypothesis Tests

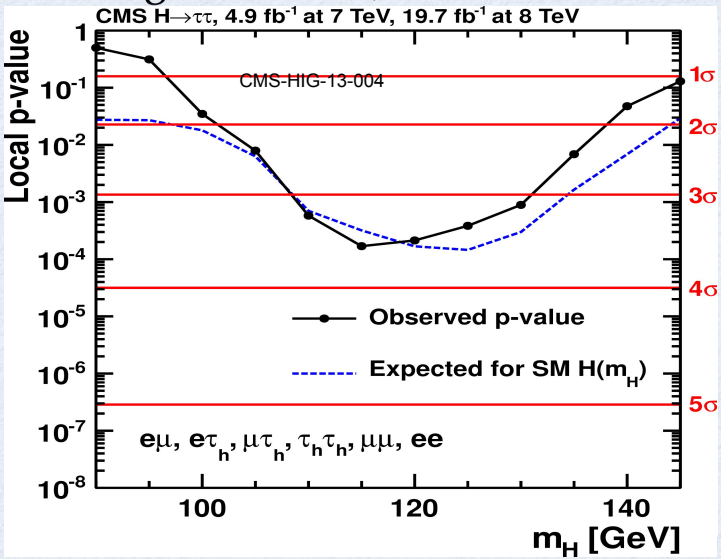
- Define null and alternate model. For discovery test
  - null: Background only model ($\mu = 0$)
  - alternate: Signal + Background model (e.g. $\mu = 1$)
- Select test statistics to use
  - e.g profile likelihood ratio (preferred due to known asymptotic formulae)
- Select type of calculator
  - asymptotic or based on toys
  - treatment of nuisance parameters
- Result is p-value for null (p_0) and alternate models (p_{s+b})

$$\lambda(\mu) = \frac{L(x|\mu, \hat{\hat{\nu}})}{L(x|\hat{\mu}, \hat{\nu})}$$



Exemplo de significância de descoberta

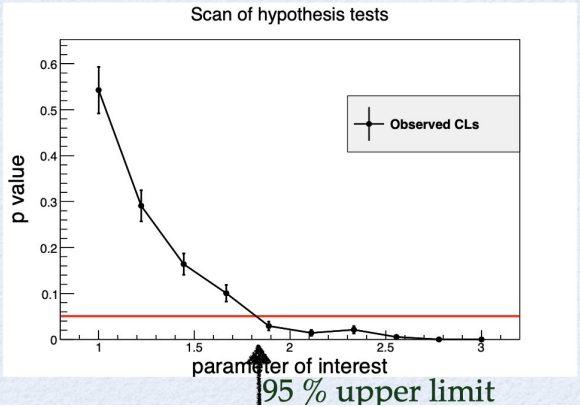
- Performing the tests for different mass hypotheses (*i.e* different signal models):

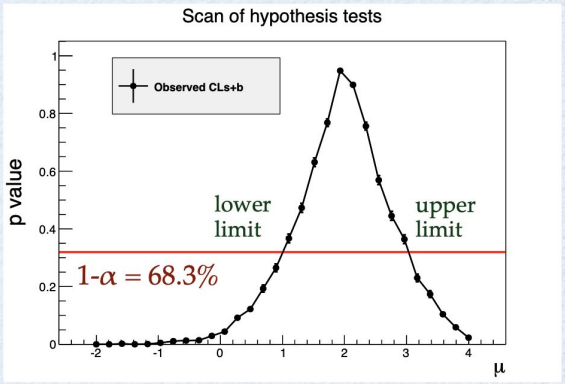


Expected significance is obtained from median of alternate (S+B) model

Inversão do teste de hipótese

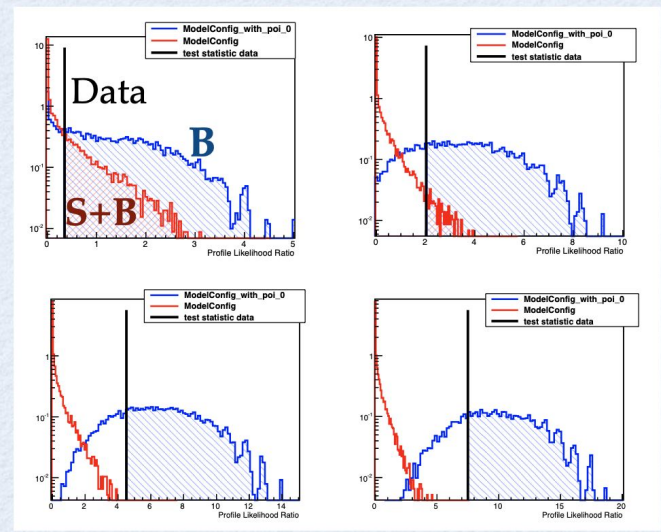
- Perform an hypothesis test at each value of the parameter
- Interval can be derived by inverting the p-value curve, function of the parameter of interest (μ)
  - value of μ which has p-value α (e.g. 0.05), is the upper limit of 1-α confidence interval (e.g. 95%)
    - for upper limits use $CL_s = CL_{s+b}/CL_b$

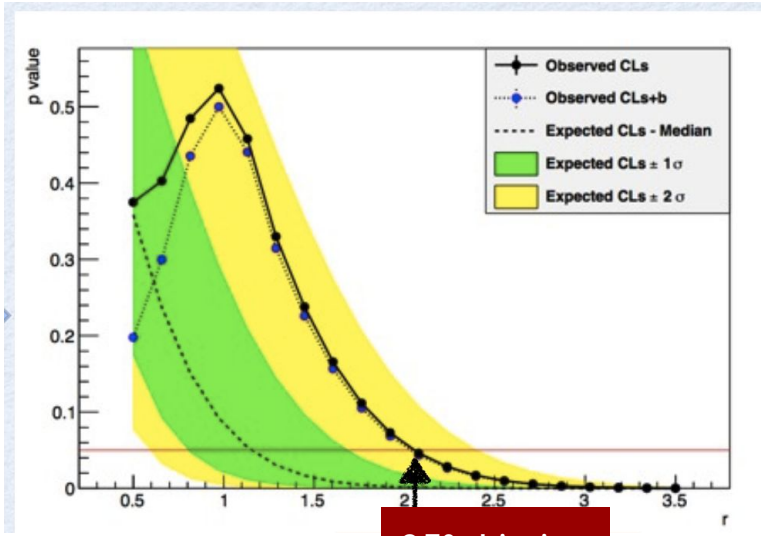




RooStats Hypo Test Inversion

- Can use Frequentist, Hybrid or Asymptotic calculator
- Compute observed, expected limits and bands





95% Limits

Bayesian Analysis in RooStats

- **RooStats** provides classes for
  - marginalize posterior and estimate credible interval

$$P(\mu|x) = \frac{\int L(x|\mu,\nu)\Pi(\mu,\nu)d\nu}{\underbrace{\iint L(x|\mu,\nu)\Pi(\mu,\nu)d\mu d\nu}_{\text{normalisation term}}}$$

posterior probability: $P(\mu|x)$ (POI: μ, data: x)

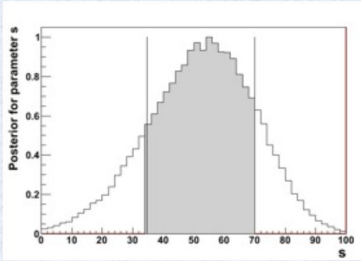
likelihood function: $L(x|\mu,\nu)$

prior probability: $\Pi(\mu,\nu)$

nuisance parameters marginalization: $\int \ldots d\nu$

Bayesian Theorem

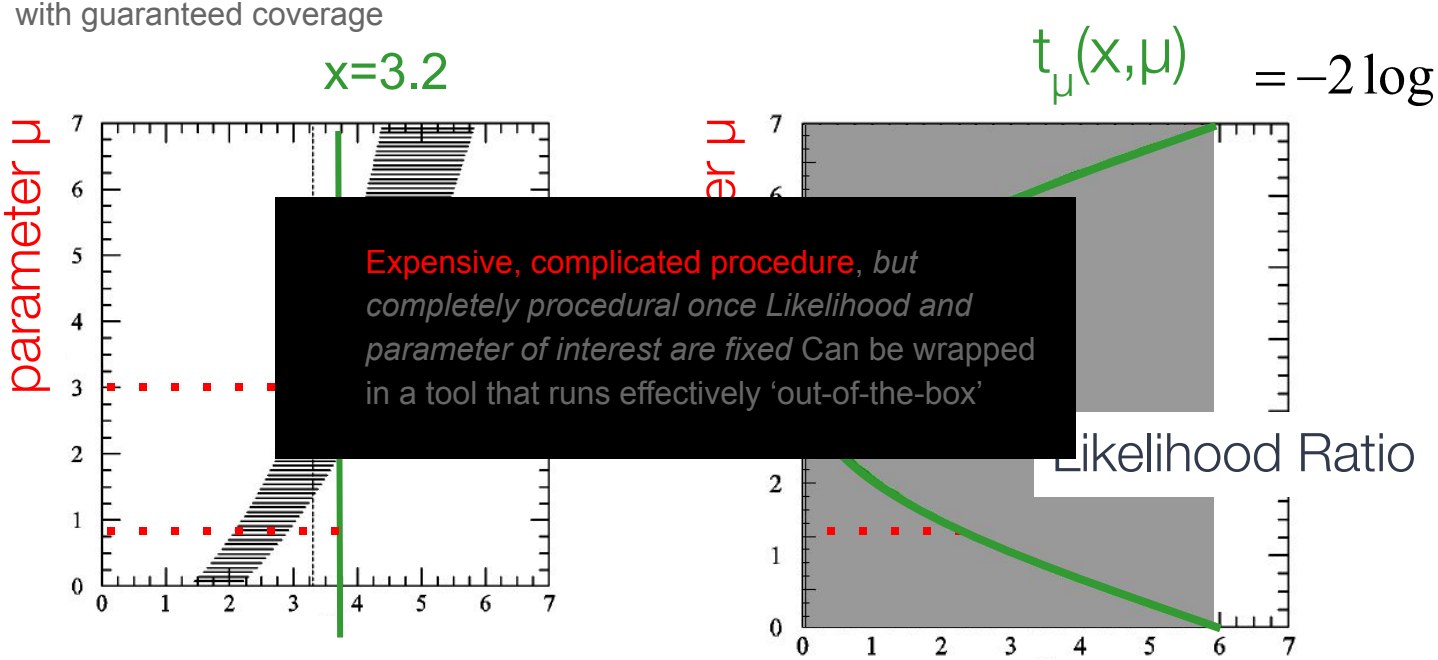
  - support for different integration algorithms:
    - adaptive (numerical)
    - MC integration
    - Markov-Chain
      - can work with models with many parameters (e.g few hundreds)



- Any prior can be given (up to know uniform prior are normally used)
- Working to include Reference priors (least informative and objective)
  - see *L. Demortier, S. Jain, H. B. Prosper, Phys. Rev. D82, 034002, 2010*

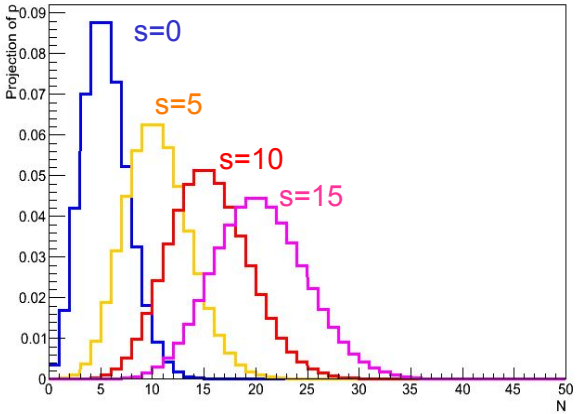
But fundamental techniques can be complicated to execute…

- Example of confidence interval calculation with Neyman construction
  - Need to construct ‘confidence belt’ using toy MC. Intersection observed data with belt defined interval in POI with guaranteed coverage



All experimental results *start* with the formulation of a model

- Examples of HEP physics models being tested
  - **SM with m(top)=172,173,174 GeV:** Measurement top quark mass
  - **SM with/without Higgs boson**: Discovery of Higgs boson
  - **SM with composite fermions/Higgs:** Measurement of Higgs coupling properties
- Via chain of physics simulation, showering MC, detector simulation and analysis software, a physics model is reduced to a statistical model
- A statistical model defines p(data|theory) for all observable outcomes
  - Example of a statistical model for a counting measurement with a known background



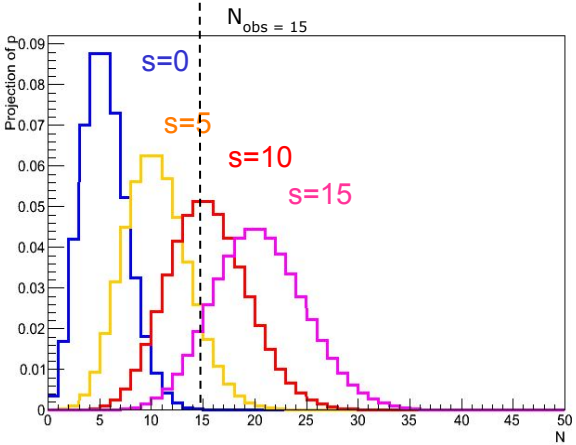
$$P(n|s+b) = \frac{(s+b)^n}{n!}e^{-(s+b)}$$

NB: b is a constant in this example

Definition: the Likelihood is P(observed data|theory)

Everything starts with the likelihood

- **All** fundamental statistical procedures are based on the likelihood function as ‘description of the measurement’



$$P(n|s+b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

NB: b is a constant in this example

Examples:

$$L(N=15|s=0)$$
$$L(N=15|s=10)$$

Definition: the Likelihood is P(observed data|theory)

Frequentist statistics → Confidence interval on s

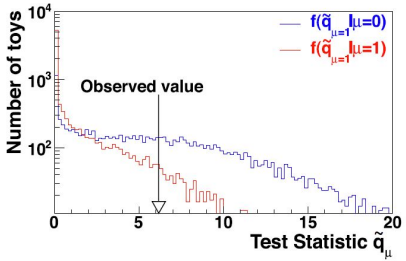
Bayesian statistics → Posterior on s

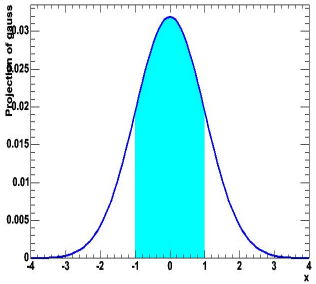
Maximum Likelihood (ML) → s = x ± y

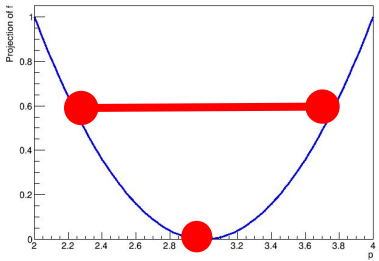
Everything starts with the likelihood

Frequentist statistics | Bayesian statistics | Maximum Likelihood

$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}|\mu)}{L(\vec{N}|\bar{\mu})} \quad P(\mu) \propto L(x|\mu) \times \pi(\mu) \quad \frac{d \ln L(\vec{p})}{d\vec{p}}\Big|_{p_i=\bar{p}_i} = 0$$







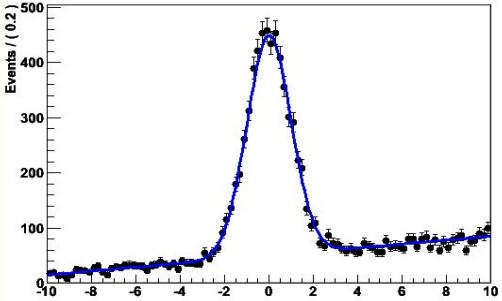
Confidence interval or p value

Posterior on s or Bayes factors

s = x ± y

How is Higgs discovery different from a simples fit?

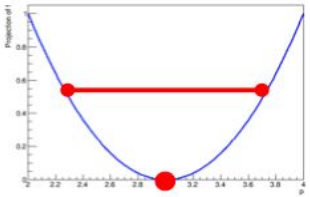
Gaussian + polynomial



ROOT TH1 ROOT TF1

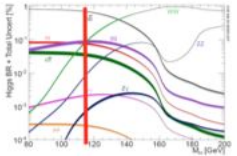
$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta})$$

"inside ROOT"



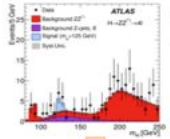
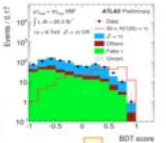
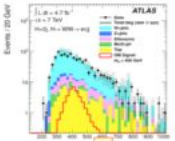
ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

μ = 5.3 ± 1.7



Higgs combination model

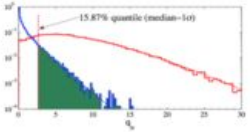
H→ZZ→llll H→ττ H→WW→μνjj

$$L(\vec{N}_{ZZ} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \quad L(\vec{N}_{\tau\tau} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{\tau\tau}, \ldots) \quad L(\vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{WW}, \ldots)$$

$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \cdot \prod Poisson(N^i_{\tau\tau}, \ldots) \cdot \prod Poisson(N^i_{WW}, \ldots) \cdot \ldots$$

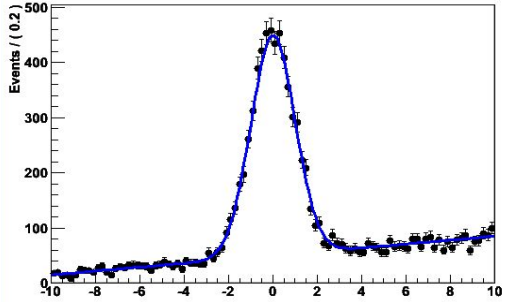
$$\lambda_\mu(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \mu, \hat{\hat{\theta}})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \hat{\mu}, \hat{\theta})}$$



$$p(H_\mu) = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_\mu) d\lambda = \ldots$$

How is Higgs discovery different from a simples fit?

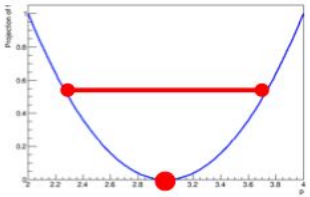
Gaussian + polynomial



ROOT TH1 ROOT TF1

$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta}))$$

"inside ROOT"



ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

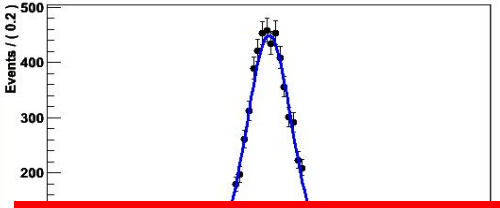
μ = 5.3 ± 1.7

Likelihood Model orders of magnitude more complicated.

- **Describes**
  - **O(100) signal distributions**
  - **O(100) control sample distr.**
  - **O(1000) parameters representing syst. uncertainties**
- **Frequentist confidence interval construction and/or p-value calculation not available as 'ready-to-run' algorithm in ROOT**

How is Higgs discovery different from a simples fit?

Gaussian + polynomial

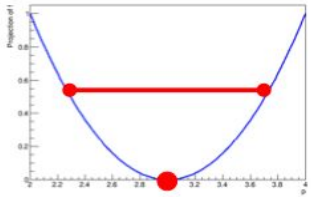


Model Building phase (formulation of L(x|H)

ROOT fit

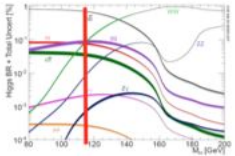
$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta}))$$

"inside ROOT"



ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

μ = 5.3 ± 1.7



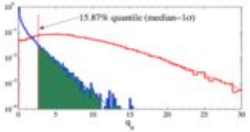
Higgs combination model

H→ZZ→llll H→ττ H→WW→μνjj

$$L(\vec{N}_{ZZ} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \quad L(\vec{N}_{\tau\tau} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{\tau\tau}, \ldots) \quad L(\vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{WW}, \ldots)$$

$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \cdot \prod Poisson(N^i_{\tau\tau}, \ldots) \cdot \prod Poisson(N^i_{WW}, \ldots) \cdot \ldots$$

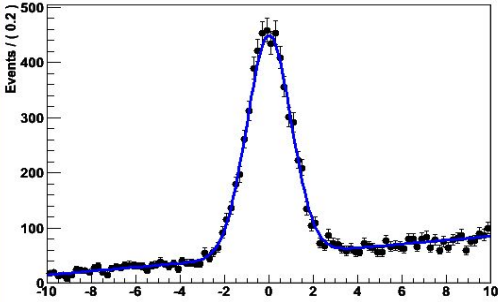
$$\lambda_\mu(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \mu, \hat{\hat{\theta}})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \hat{\mu}, \hat{\theta})}$$



$$p(H_\mu) = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_\mu) d\lambda = \ldots$$

How is Higgs discovery different from a simples fit?

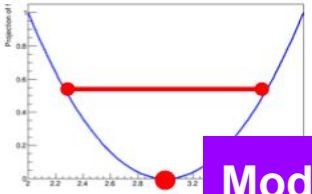
Gaussian + polynomial



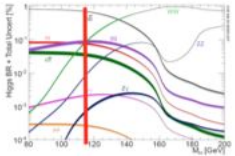
ROOT TH1 ROOT TF1

$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta}))$$

"inside ROOT"

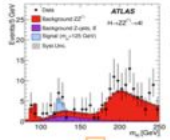
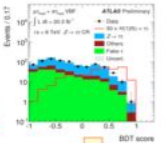
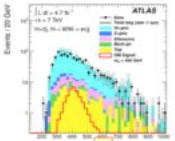


ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)



Higgs combination model

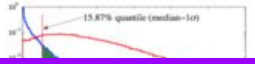
H→ZZ→llll H→ττ H→WW→μνjj

$$L(\vec{N}_{ZZ} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \quad L(\vec{N}_{\tau\tau} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{\tau\tau}, \ldots) \quad L(\vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{WW}, \ldots)$$

$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \cdot \prod Poisson(N^i_{\tau\tau}, \ldots) \cdot \prod Poisson(N^i_{WW}, \ldots) \cdot \ldots$$

$$\lambda_\mu(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \mu, \hat{\hat{\theta}})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \hat{\mu}, \hat{\theta})}$$



$$p(H_\mu) = \int_{\lambda}^{\infty} f(\lambda \mid H_\mu) d\lambda = \ldots$$

Model Usage phase (use L(x|H) to make statement on H)

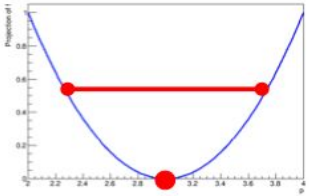
How is Higgs discovery different from a simples fit?

Gaussian + polynomial

ROOT TH1

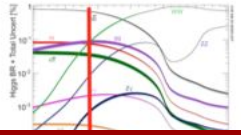
$L(\vec{N} \mid \mu, \vec{\theta}) = \prod$

"inside ROOT"



ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

μ = 5.3 ± 1.7



Higgs combination model

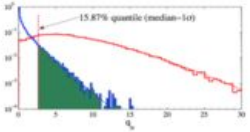
$\prod Poisson(N^i_{WW},\ldots)$

$Poisson(N^i_{WW},\ldots)\cdot\ldots$

Design goal:

Separate **building of Likelihood model** as much as possible from statistical analysis **using the Likelihood model**

- More modular software design
- 'Plug-and-play with statistical techniques
- Factorizes work in collaborative effort

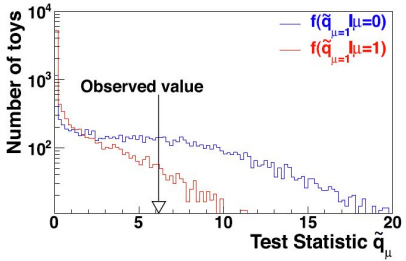


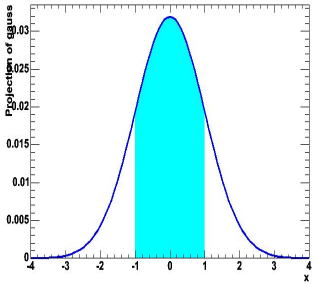
$p(H_\mu) = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_\mu)\, d\lambda = \ldots$

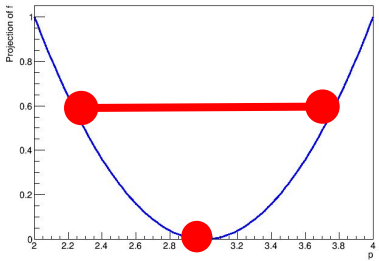
The need for fundamental statistical techniques

Frequentist statistics | Bayesian statistics | Maximum Likelihood

$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}|\mu)}{L(\vec{N}|\bar{\mu})} \quad P(\mu) \propto L(x|\mu) \times \pi(\mu) \quad \frac{d \ln L(\vec{p})}{d\vec{p}}\Big|_{p_i=\bar{p}_i} = 0$$







Confidence interval or p value

Posterior on s or Bayes factors

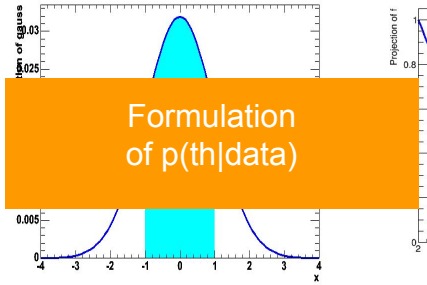
s = x ± y

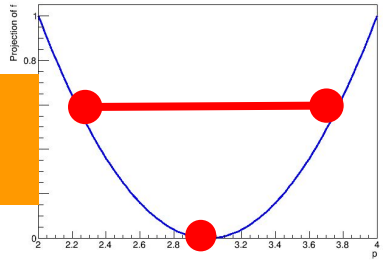
The need for fundamental statistical techniques

Frequentist statistics | Bayesian statistics | Maximum Likelihood

$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}|\mu)}{L(\vec{N}|\bar{\mu})} \quad P(\mu) \propto L(x|\mu) \times \pi(\mu) \quad \frac{d \ln L(\vec{p})}{d\vec{p}}\Big|_{p_i=\bar{p}_i} = 0$$

No assumptions on normal distributions, or asymptotic validity for high statistics





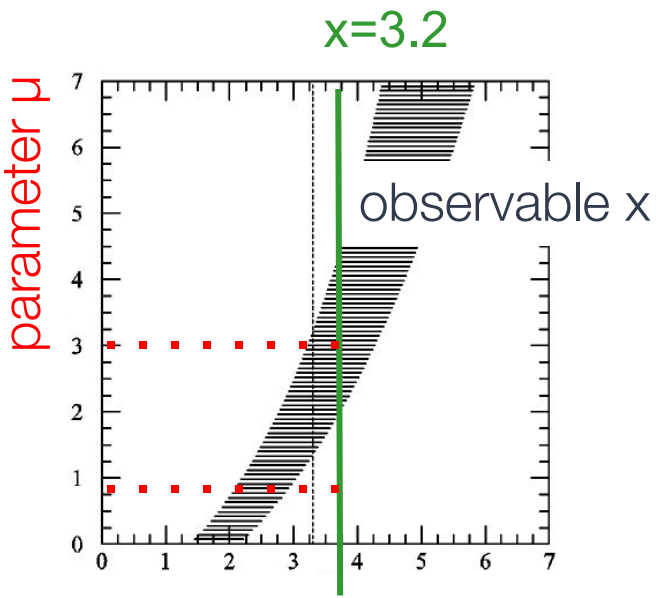
Confidence interval or p value

Posterior on s or Bayes factors

s = x ± y

But fundamental techniques can be complicated to execute…

- Example of confidence interval calculation with Neyman construction
  - Need to construct ‘confidence belt’ using toy MC. Intersection observed data with belt defined interval in POI with guaranteed coverage



$$t_\mu(x,\mu) = -2\log\frac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})}$$