DFNAE
DEPARTAMENTO DE FÍSICA NUCLEAR E ALTAS ENERGIAS

Introduction to RooStats

Sandro Fonseca de Souza

December 22, 2020

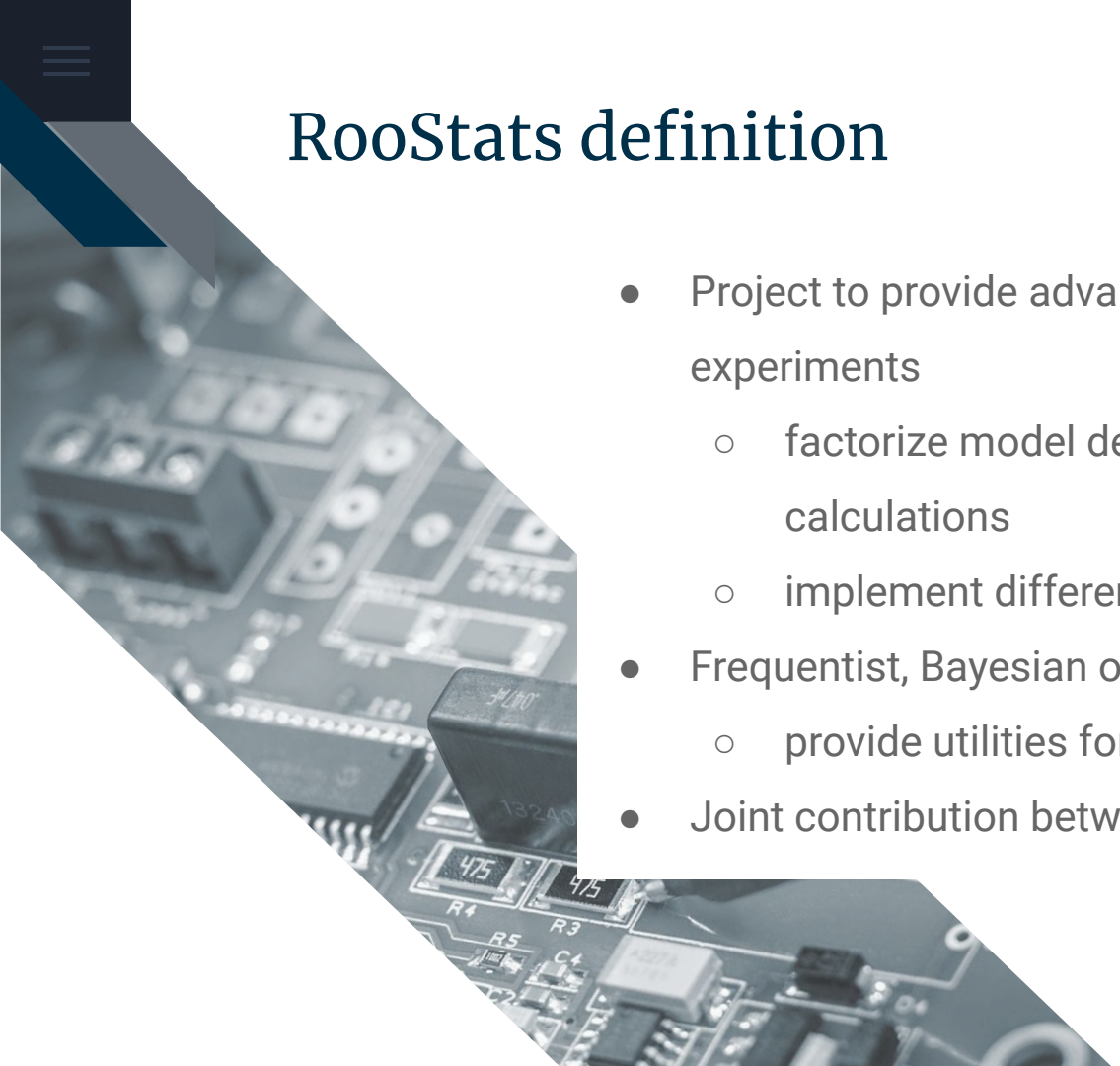
References

- RooStats twiki page: (link)
- RooStats: a framework for advanced statistical analysis (link)
- RooStats tutorial: (link)
- https://twiki.cern.ch/twiki/bin/view/Main/INFNStatRooStats2019
- https://twiki.cern.ch/twiki/bin/view/RooStats/RooStatsTutorialsAugust2012
- https://root.cern/doc/master/group__HistFactory.html
- Statistical analysis tools for the Higgs discovery and beyond

Summary

- RooStats definition
- Statistics applications and RooStats technology
- RooStats Calculators
  - Hypothesis Tests
  - Frequentist interval (CLs)

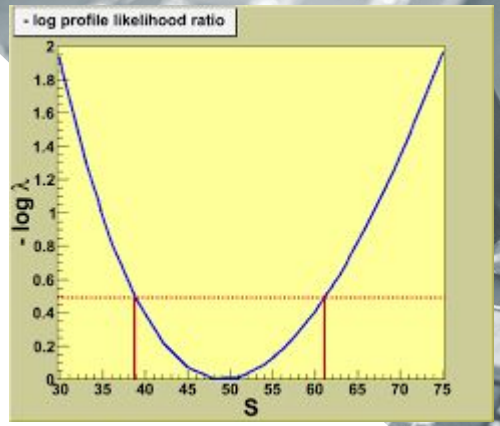
RooStats definition

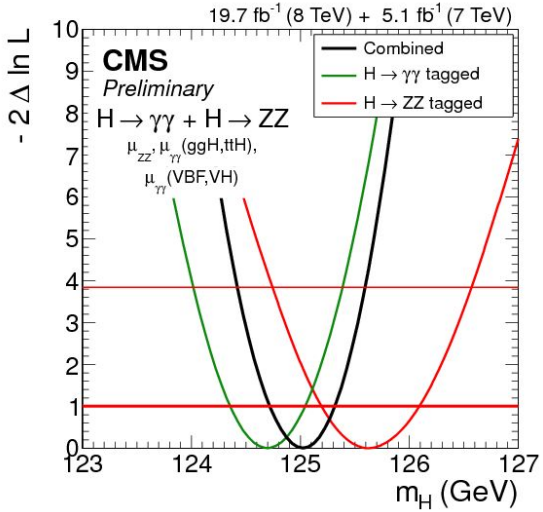


- Project to provide advanced statistical tools for LHC experiments
  - factorize model description from statistical calculations
  - implement different statistics techniques
- Frequentist, Bayesian or Likelihood Based
  - provide utilities for combination of results
- Joint contribution between ATLAS, CMS, RooFit and ROOT

Statistical Applications

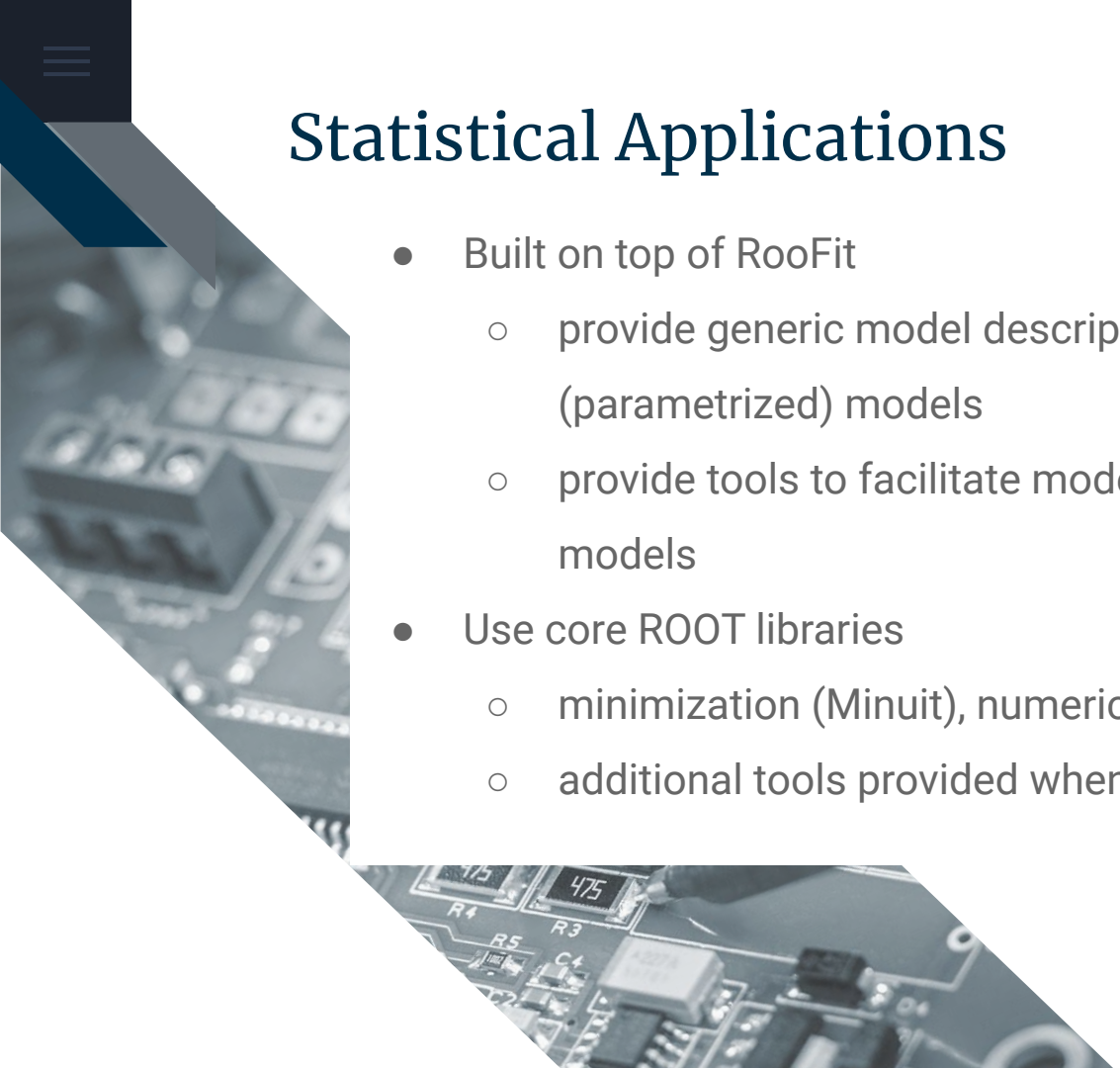
- Point estimation (covered by RooFit)
- Estimation of confidence (credible) intervals (lower/upper limits)
- Hypothesis tests (e.g discovery significance)





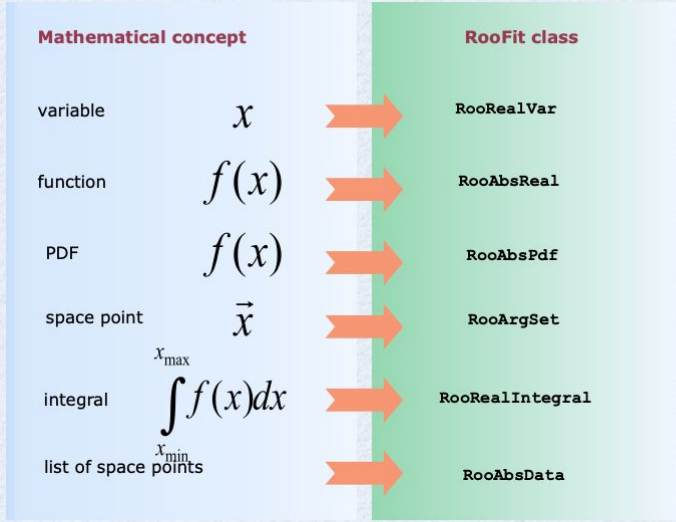
Statistical Applications

- Built on top of RooFit
  - provide generic model description binned (histogram based) unbinned (parametrized) models
  - provide tools to facilitate model creation tools for combination of models
- Use core ROOT libraries
  - minimization (Minuit), numerical integration, etc.
  - additional tools provided when needed (e.g. Markov-Chain MC)



RooStats Technology

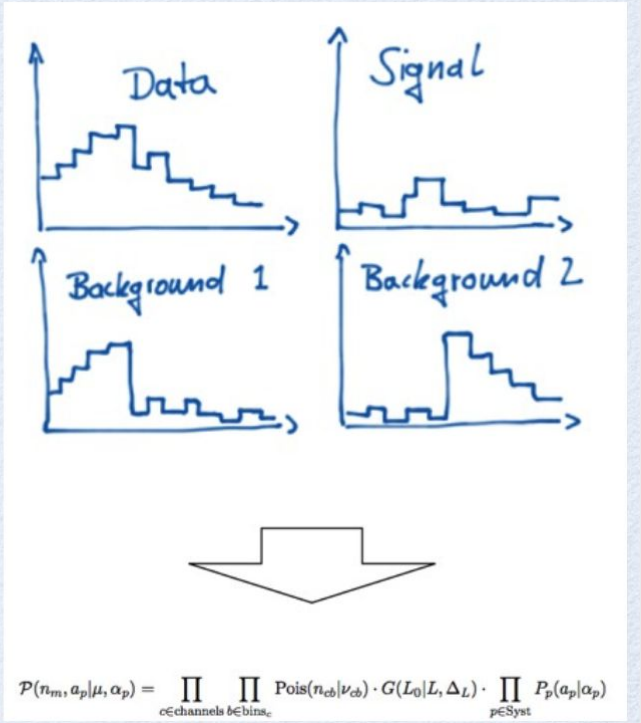
Building models with RooFit




```
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)

Models with HistFactory

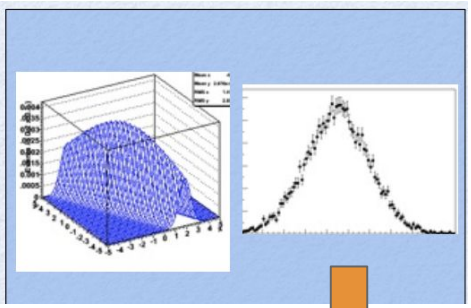


Terminology

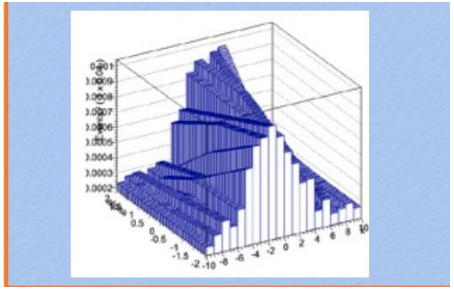
- **observable** - something you measure in an experiment, for example a particle's momentum. Often, a function of measured quantities, for example an invariant mass of several particles
- **global observable** or **auxiliary observable** - an observable from another measurement, for example, the integrated luminosity
- **model** - a set of probability density functions (PDFs), which describe distributions of the observables or functions of observables
- **model parameter** - any variable in your PDF expression, which is not an observable
- **parameter of interest (POI)** - a model parameter that you study, for example a cross section of your signal process
- **nuisance parameter** - every other model parameter, which is not your parameter of interest
- **data** or **dataset** - a set of values of observables, either measured in an experiment, or simulated
- **likelihood** - a model computed for a particular dataset
- **hypothesis** - a particular model, with specified observables, POI, nuisance parameters, and prior PDFs (in case of Bayesian inference)
- **prior PDF** - a probability density for an observable or a model parameter, which is known *a priori*, i.e. before a measurement is considered. This is a Bayesian concept exclusively. Prior has no meaning or place in a frequentist type of inference
- **Bayesian** - a type of statistical inference that usually produces probability of the hypothesis given the data. Requires a prior.
- **frequentist** - a type of statistical inference that usually produces probability of the data given the hypothesis.

RooFit/RooStat at LHC (e.g. Higgs Analysis)

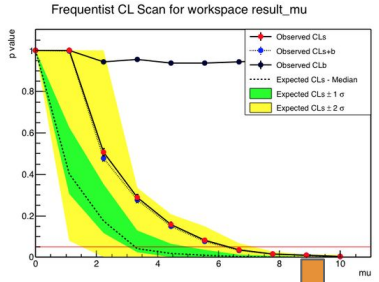
Class RooWorkspace
Simplify packaging and sharing of models

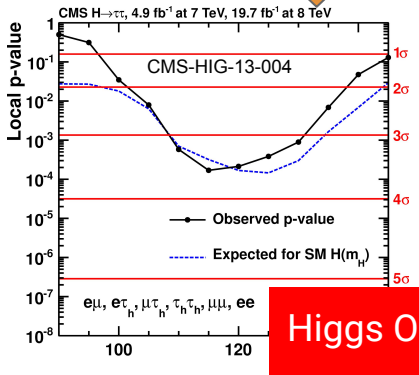


HistFactory package
Constructing models from Monte Carlo templates



RooStats toolkit
Statistical tests based on likelihoods from RooFit models

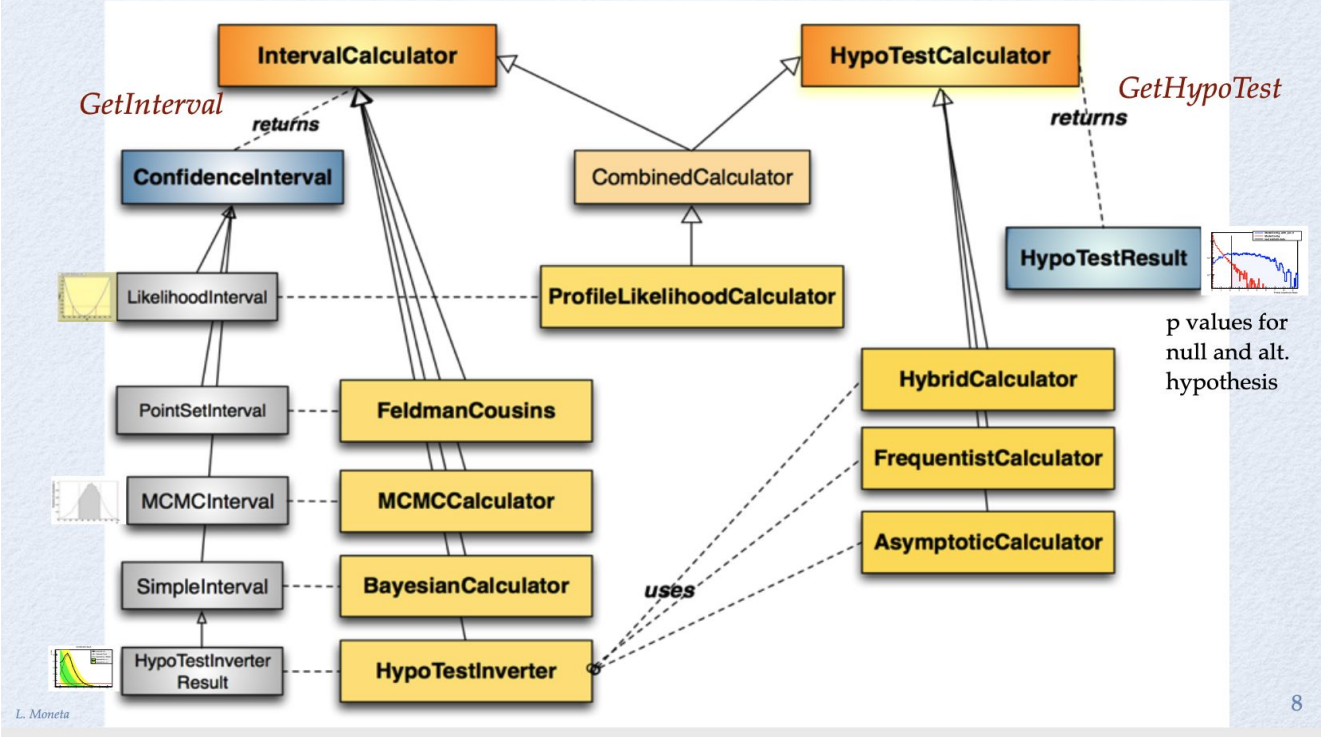




Higgs Observation

RooStats Design

- C++ interfaces and classes mapping to real statistical concepts

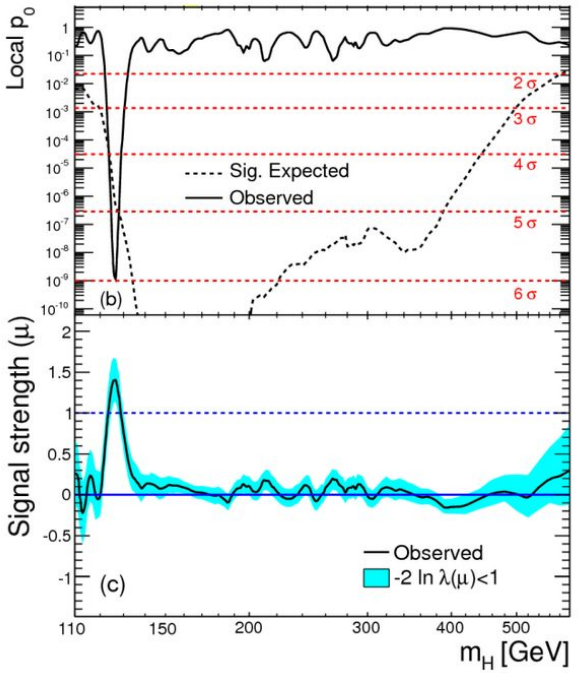


What do you want to know?

- Physics questions we have…
  - Does the (SM) Higgs boson exist?
  - What is its production cross-section?
  - What is its boson mass?
- Statistical tests construct probabilistic statements: p(theo|data), or p(data|theo)
  - Hypothesis testing (discovery)
  - (Confidence) intervals - Measurements & uncertainties

- Result: Decision based on tests

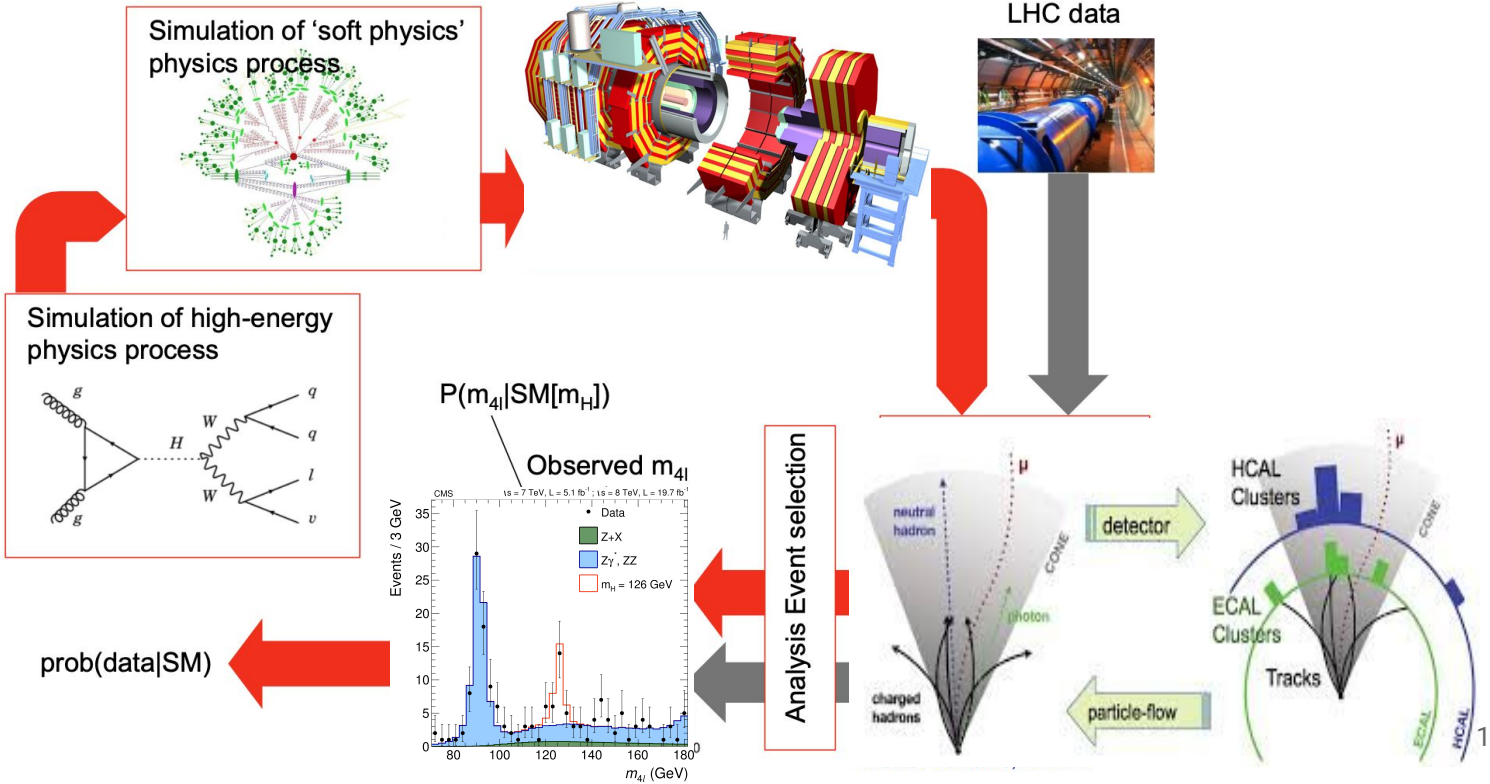
“As a layman I would now say: I think we have it” (Rolf Heuer - CERN DG 2009-2015)



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

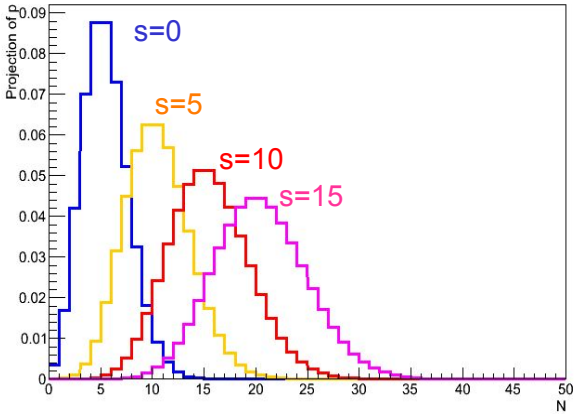
- Examples of HEP physics models being tested
  - **SM with m(top)=172,173,174 GeV:** Measurement top quark mass
  - **SM with/without Higgs boson**: Discovery of Higgs boson
  - **SM with composite fermions/Higgs:** Measurement of Higgs coupling properties
- Via chain of physics simulation, showering 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

The HEP analysis workflow illustrated



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



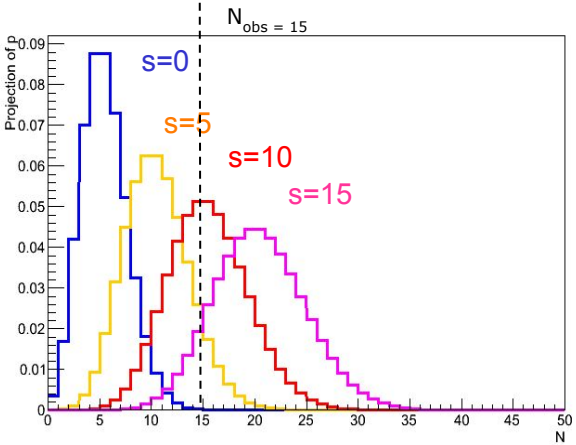
$$P(n|s+b) = \frac{(s+b)^n}{n!}e^{-(s+b)}$$

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

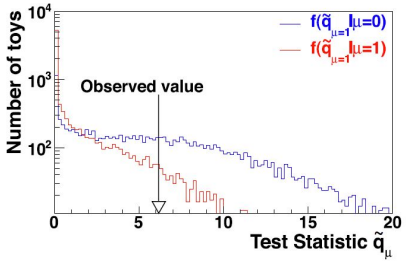
Everything starts with the likelihood

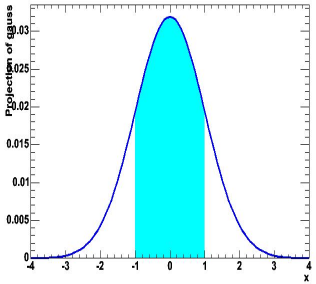
Frequentist statistics

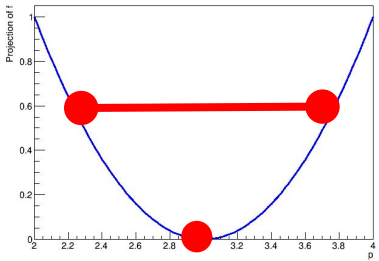
Bayesian statistics

Maximum Likelihood

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







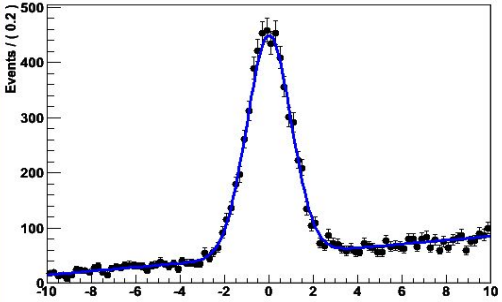
Confidence interval or p value

Posterior on s or Bayes factors

s = x ± y

How is Higgs discovery different from a simples fit?

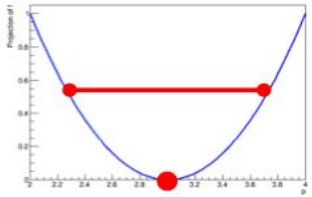
Gaussian + polynomial



ROOT TH1 ROOT TF1

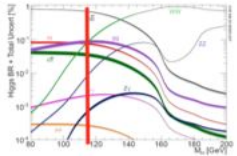
$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta}))$$

"inside ROOT"



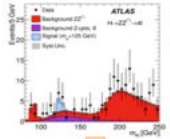
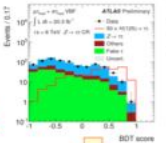
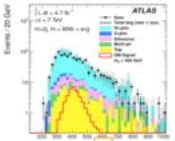
ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

μ = 5.3 ± 1.7



Higgs combination model

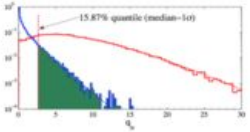
H→ZZ→llll H→ττ H→WW→μνjj

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

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

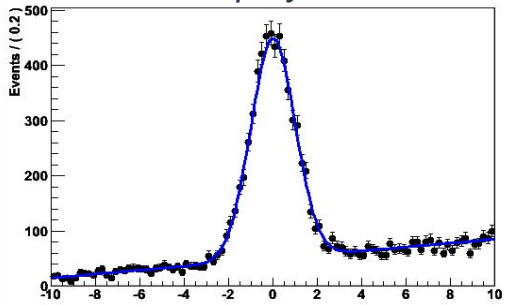
$$\lambda_\mu(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \mu, \hat{\hat{\theta}})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \hat{\mu}, \hat{\theta})}$$



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

How is Higgs discovery different from a simples fit?

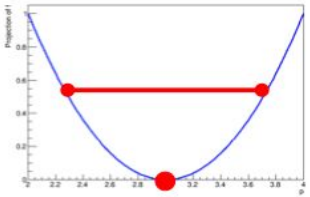
Gaussian + polynomial



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)

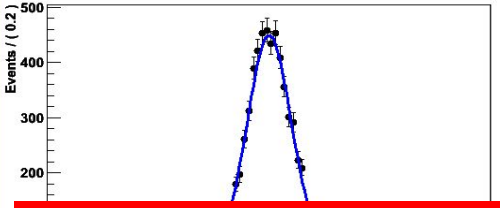
μ = 5.3 ± 1.7

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

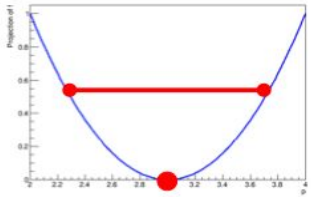


Model Building phase (formulation of L(x|H)

ROOT fit

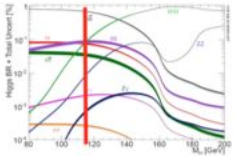
$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta}))$$

"inside ROOT"



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

μ = 5.3 ± 1.7



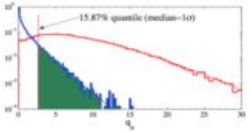
Higgs combination model

H→ZZ→llll H→ττ H→WW→μνjj

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

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

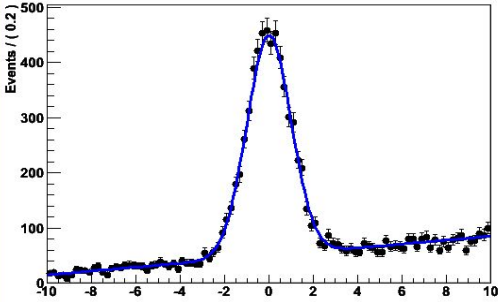
$$\lambda_\mu(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \mu, \hat{\hat{\theta}})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \hat{\mu}, \hat{\theta})}$$



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

How is Higgs discovery different from a simples fit?

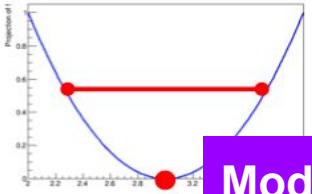
Gaussian + polynomial



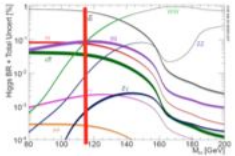
ROOT TH1 ROOT TF1

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

"inside ROOT"

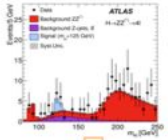
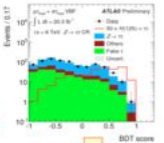
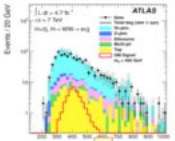


ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)



Higgs combination model

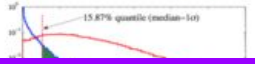
H→ZZ→llll H→ττ H→WW→μνjj

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

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

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



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

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

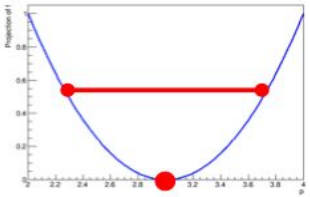
How is Higgs discovery different from a simples fit?

Gaussian + polynomial

ROOT TH1

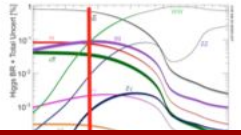
$L(\vec{N} \mid \mu, \vec{\theta}) = \prod$

"inside ROOT"



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

μ = 5.3 ± 1.7



Higgs combination model

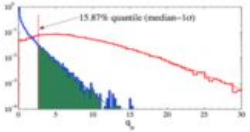
$\prod Poisson(N^i_{WW},\ldots)$

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

Design goal:

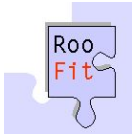
Separate **building of Likelihood model** as much as possible from statistical analysis **using the Likelihood model**

- 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 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

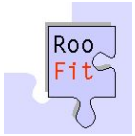
ROOWorkspace

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.

The Benefits of Modularity



RooFit and HistFactory

ROOWorkspace

"Simple fit"
ML Fit with HESSE or MINOS)

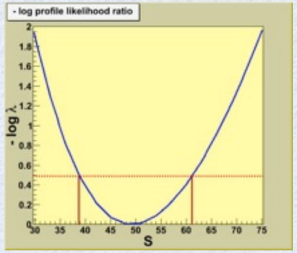
RooStats using Frequentist with toys

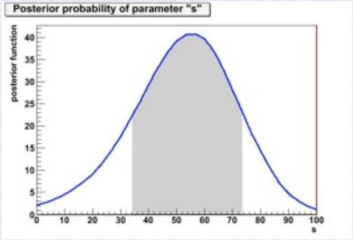
RooStats using Frequentist asymptotic

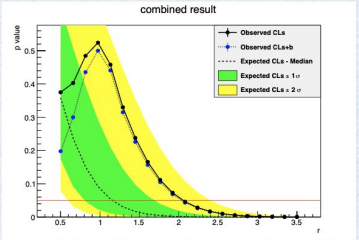
RooStats using Bayesian Markov-Chain Monte Carlo

RooStats Calculators

- 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

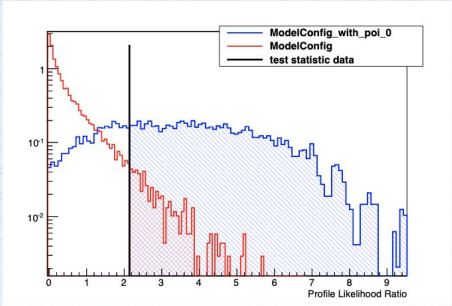


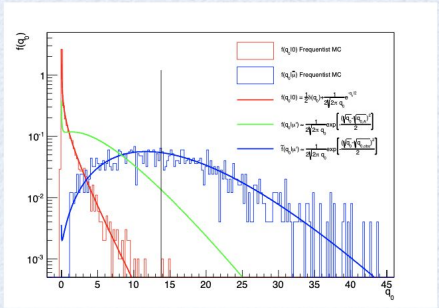




RooStats Calculators (2)

- 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

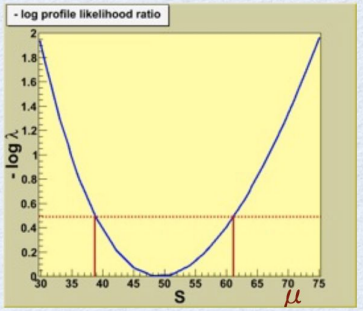




Using RooStats Calculators

- 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

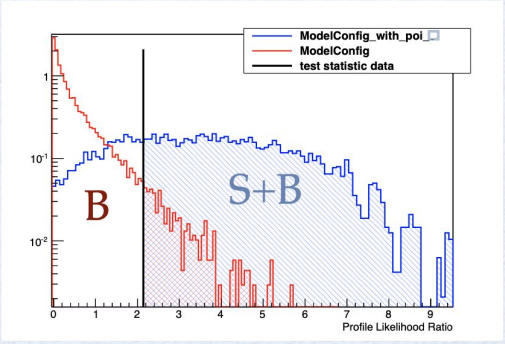
```
// create the class using data and model
ProfileLikelihoodCalculator plc(data, model);

// 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

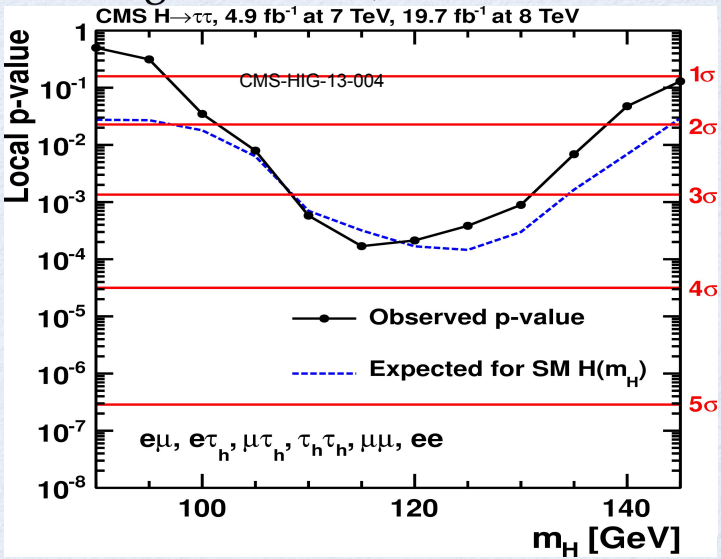
- Define null and alternate model. For discovery test
  - null: Background only model ($\mu = 0$)
  - alternate: Signal + Background model (e.g. $\mu = 1$)
- Select test statistics to use
  - e.g profile likelihood ratio (preferred due to known asymptotic formulae)
- Select type of calculator
  - asymptotic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

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



Example: Discovery Significance

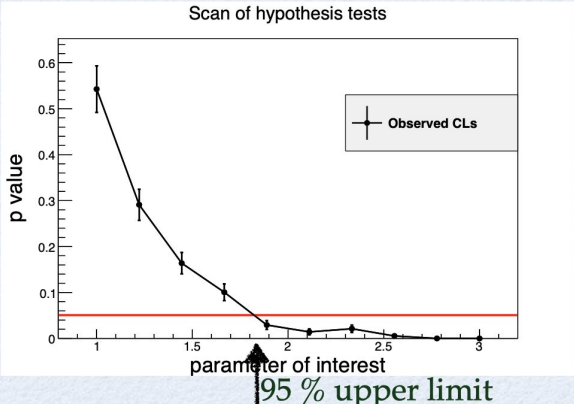
- Performing the tests for different mass hypotheses (*i.e* different signal models):

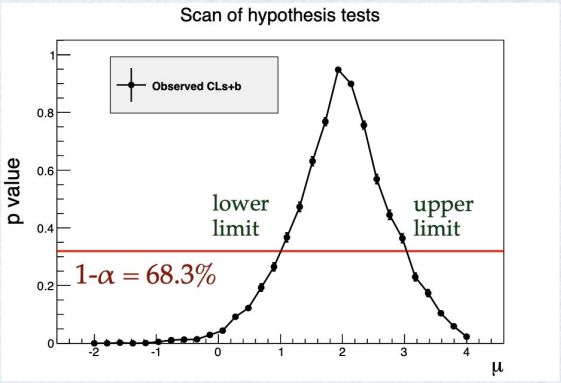


Expected significance is obtained from median of alternate (S+B) model

Hypothesis test Inversion

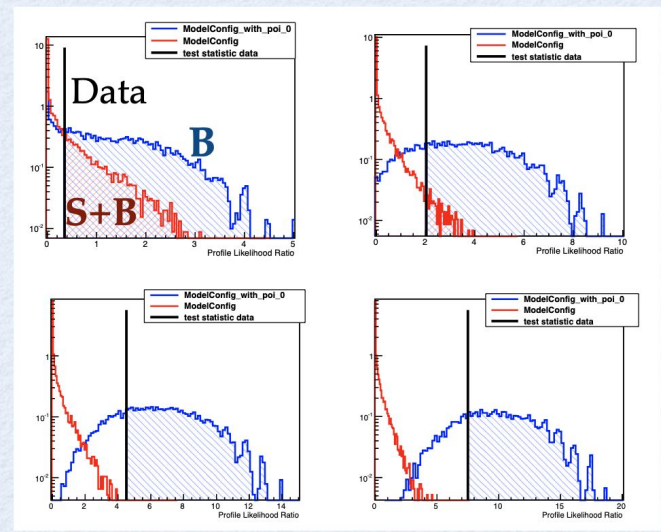
- Perform an hypothesis test at each value of the parameter
- Interval can be derived by inverting the p-value curve, function of the parameter of interest (μ)
  - value of μ which has p-value α (e.g. 0.05), is the upper limit of 1-α confidence interval (e.g. 95%)
    - for upper limits use $CL_s = CL_{s+b}/CL_b$

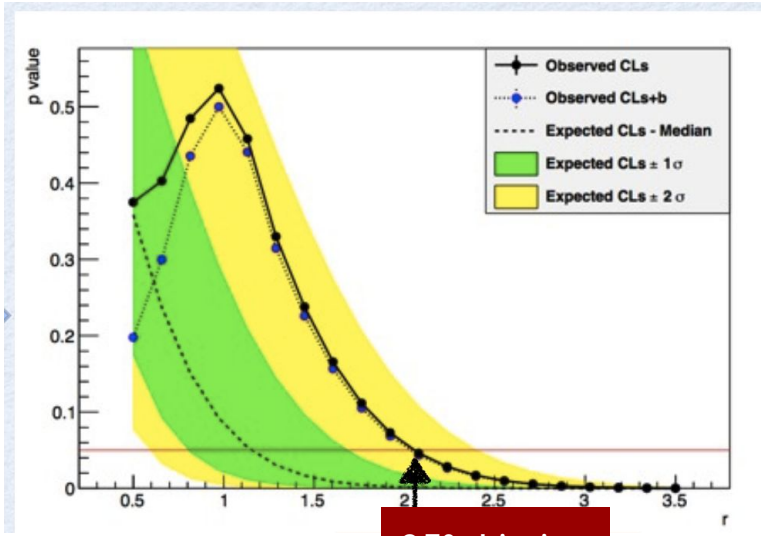




RooStats Hypo Test Inversion

- 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)

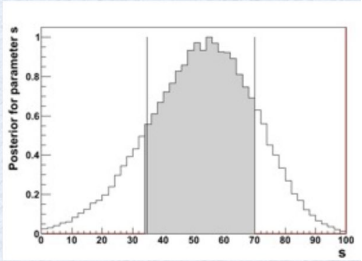
likelihood function: $L(x|\mu,\nu)$

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

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

Bayesian Theorem

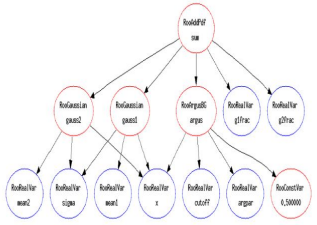
  - support for different integration algorithms:
    - adaptive (numerical)
    - MC integration
    - Markov-Chain
      - can work with models with many parameters (e.g few hundreds)



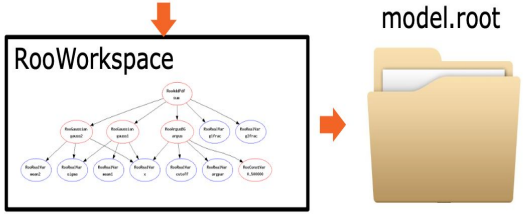
- Any prior can be given (up to know uniform prior are normally used)
- Working to include Reference priors (least informative and objective)
  - see *L. Demortier, S. Jain, H. B. Prosper, Phys. Rev. D82, 034002, 2010*

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

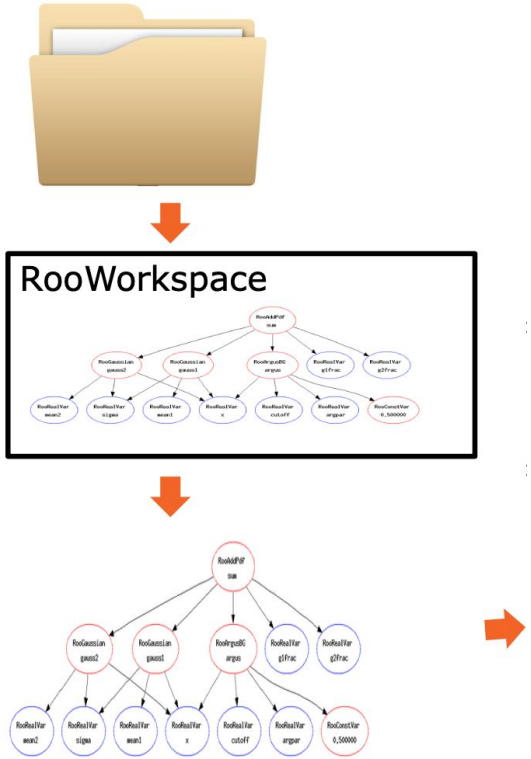


```
RooWorkspace w("w") ;
w.import(sum) ;
w.writeToFile("model.root") ;
```

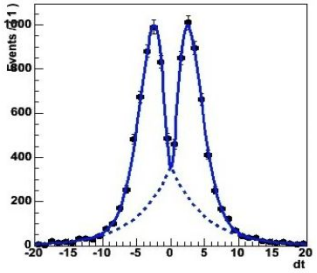


Wouter Verkerke, NIKHEF

RooWorkspace



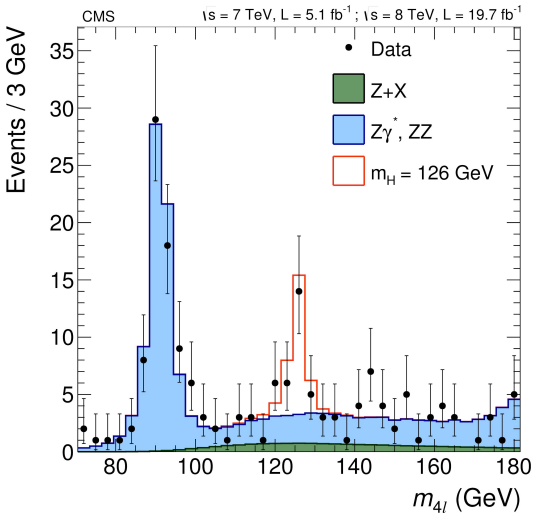

```
// Resurrect model and data
TFile f("model.root") ;
RooWorkspace* w = f.Get("w") ;
RooAbsPdf* model = w->pdf("sum") ;
RooAbsData* data = w->data("xxx") ;

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



Courtesy Wouter Verkerke, NIKHEF

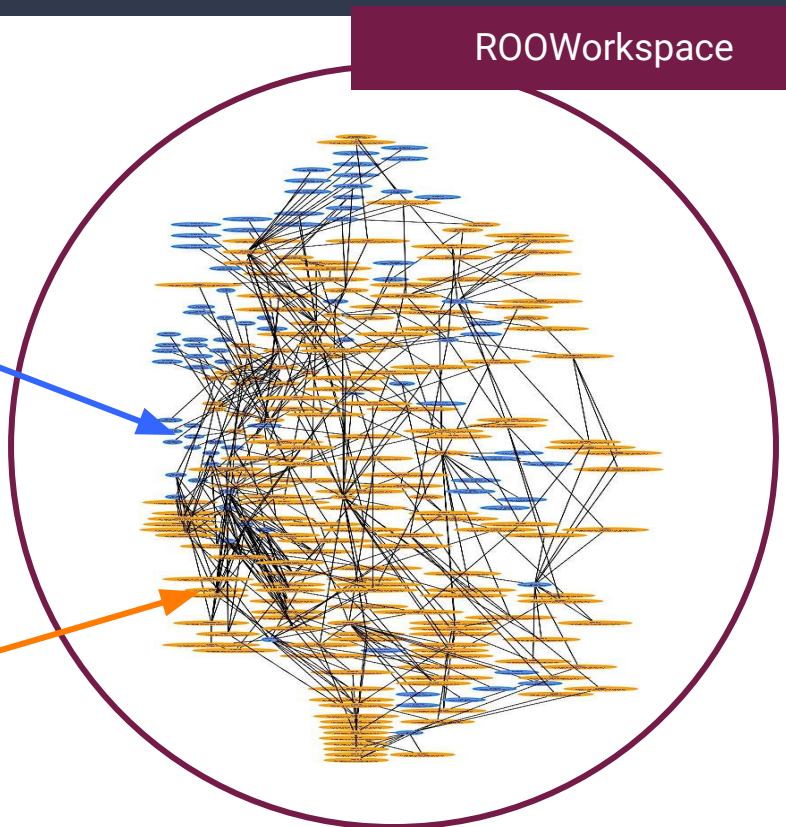
Example RooWorkspace component model for realistic Higgs analysis

Likelihood model describing the ZZ invariant mass distribution *including all possible systematic uncertainties*



variables

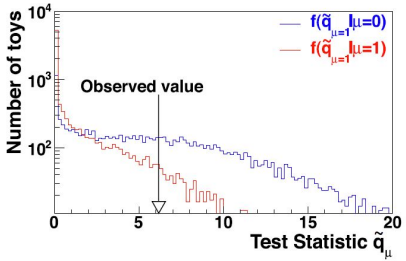
function objects

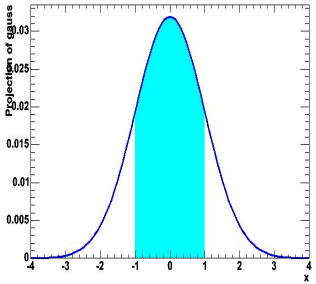


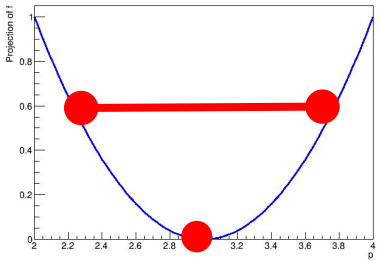
The need for fundamental statistical techniques

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

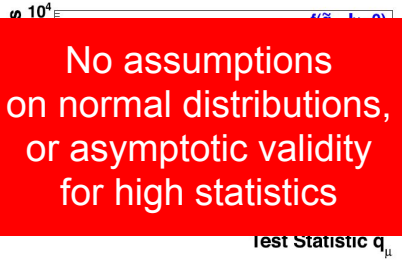
s = x ± y

The need for fundamental statistical techniques

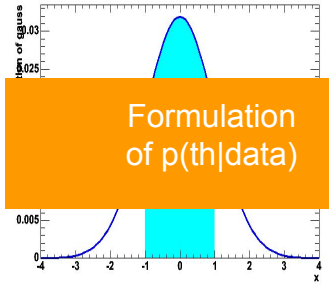
Frequentist statistics | Bayesian statistics | Maximum Likelihood

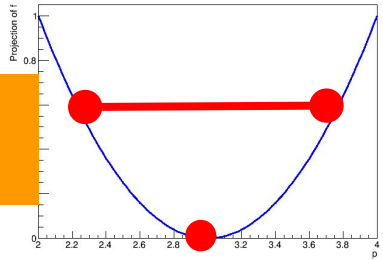
$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}|\mu)}{L(\vec{N}|\bar{\mu})} \quad P(\mu) \propto L(x|\mu) \times \pi(\mu) \quad \frac{d \ln L(\vec{p})}{d\vec{p}}\Big|_{p_i=\bar{p}_i} = 0$$

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



Formulation of p(th|data)





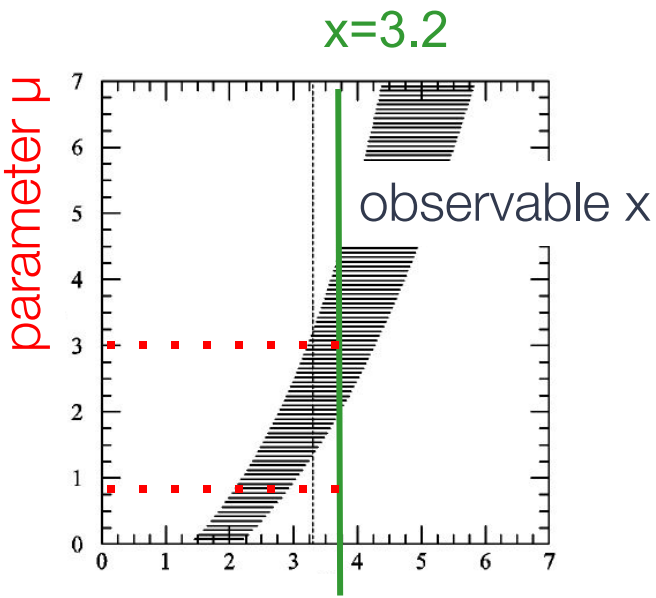
Confidence interval or p value

Posterior on s or Bayes factors

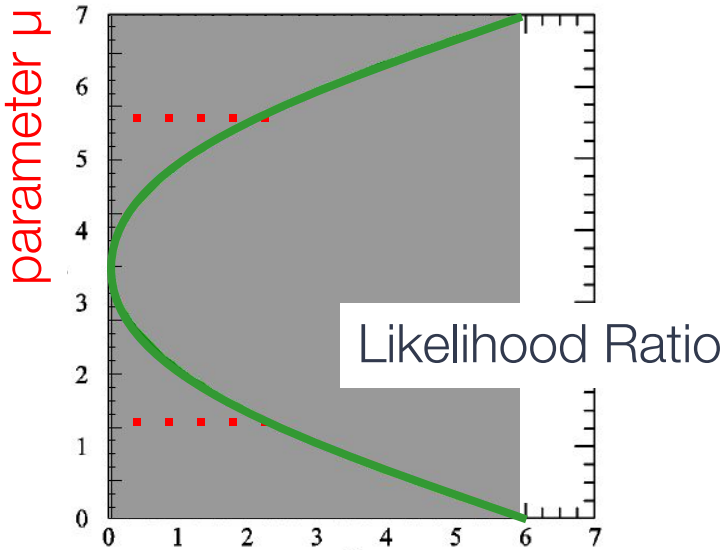
s = x ± y

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

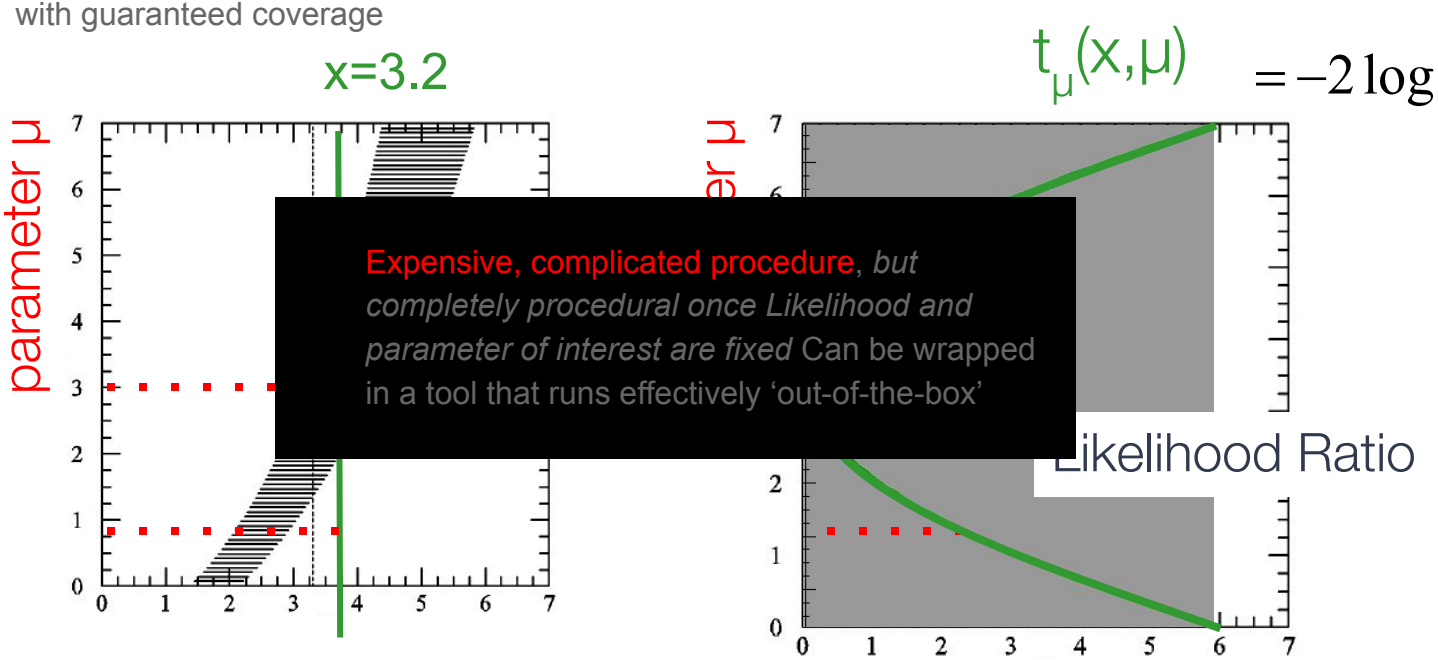


$t_\mu(x,\mu) = -2\log\frac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})}$



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



Expensive, complicated procedure, *but completely procedural once Likelihood and parameter of interest are fixed* Can be wrapped in a tool that runs effectively ‘out-of-the-box’

Examples

1. RooFit example on fitting the 2-photon invariant mass to determine the number of Higgs signal events. (link) (github)
2. Create the RooStats::ModelConfig- How to add the ModelConfig class for the H->gg workspace we have created for the RooFit exercise (link) (github)
3. Discovery Significance vs Mass to make a p0 plot a scan of the discovery significance vs mass for the H->gg case (link) (github)

End