Introduction to statistics / statistical tools

Max Baak (CERN), Geert-Jan Besjes (Nijmegen/Nikhef), Jeanette Lorenz (LMU/Excellence Cluster Universe), Sophio Pataraia (Bergische Universitaet Wuppertal)

+ many other people

30 March 2015

HistFitter

Overview

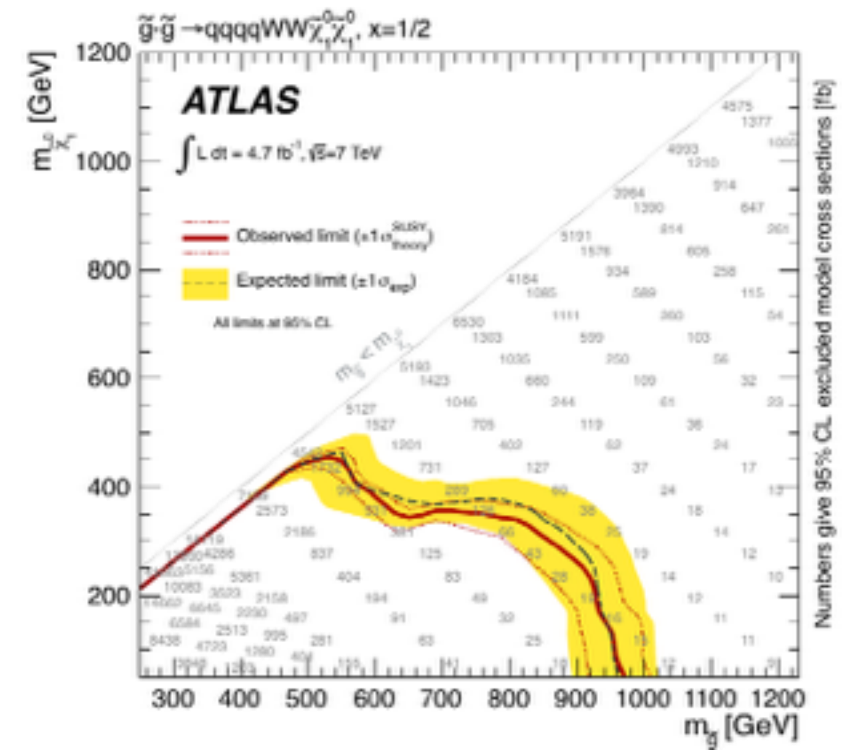
- Introduction to statistics (short)
- Introduction to statistical analysis (RooFit, RooStats, HistFactory)

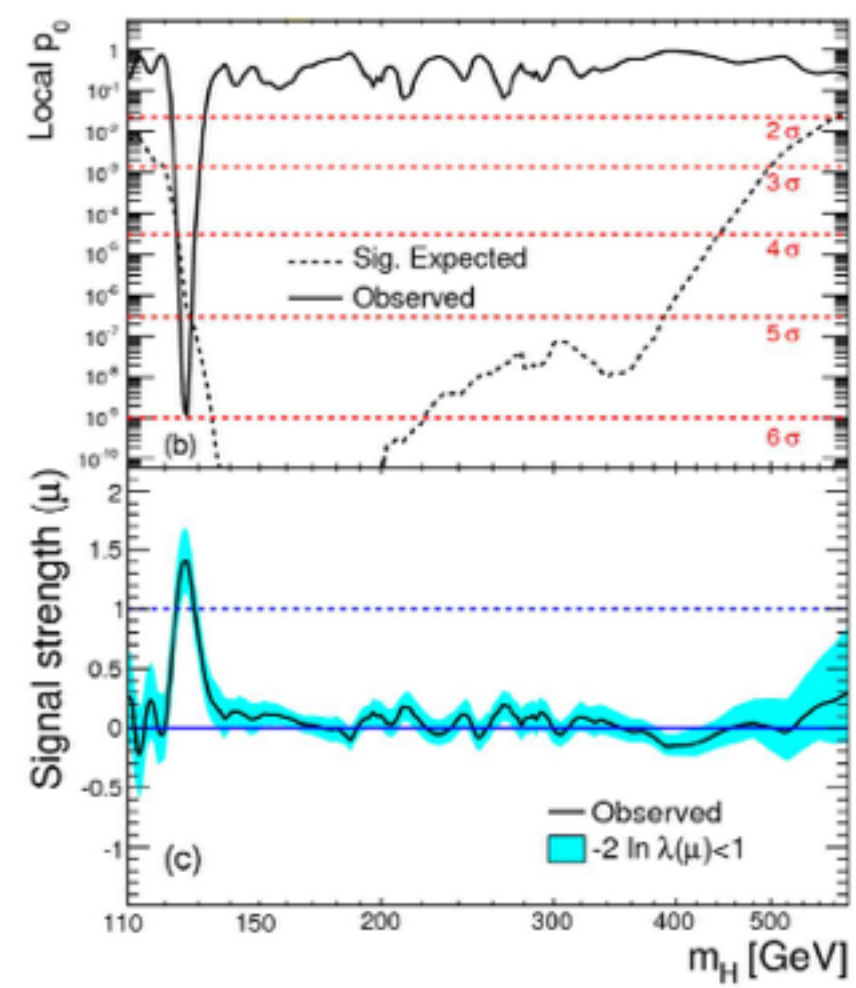
Introduction to HEP statistics

Largely borrowed from lectures/slides by W. Verkerke

Basic questions

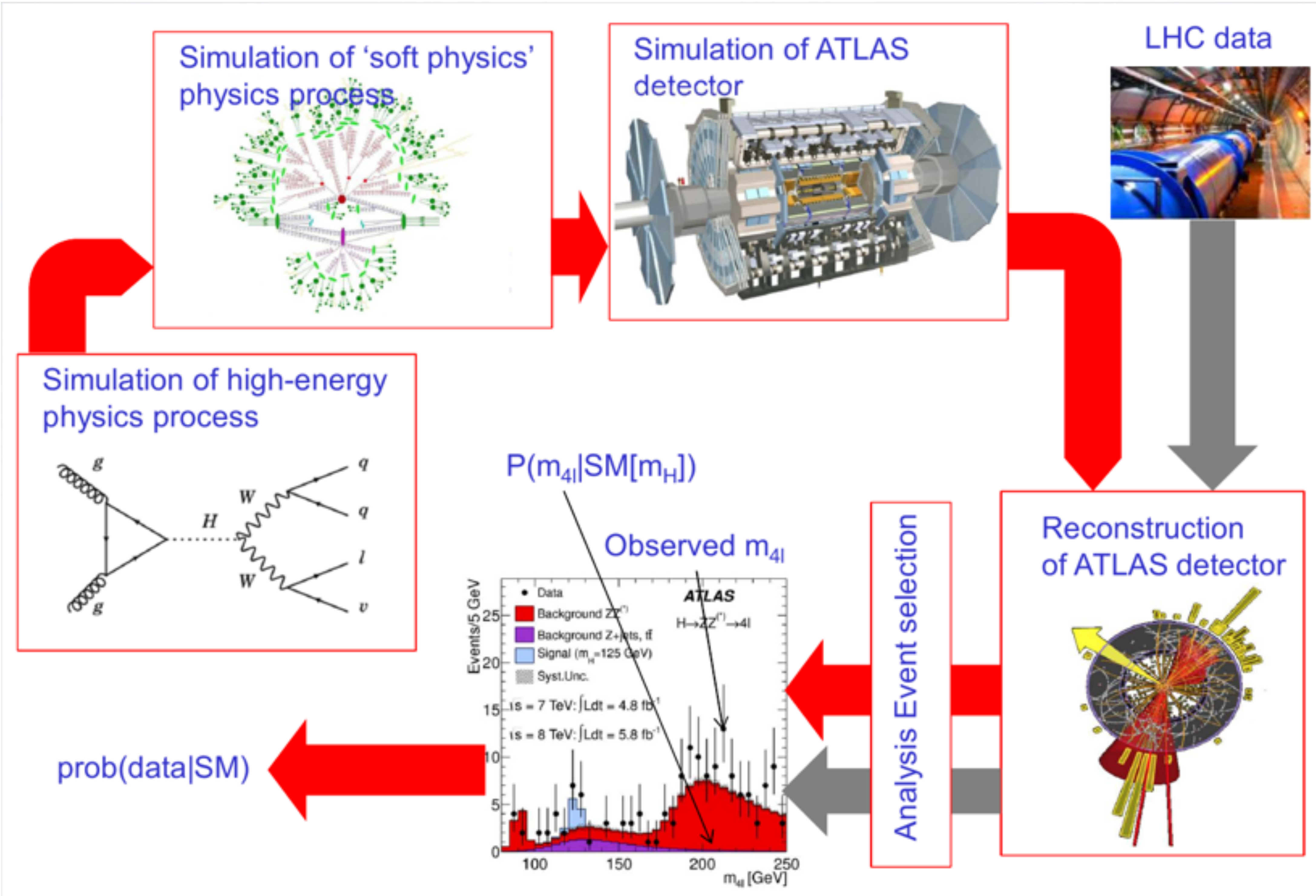
- Physics questions we want to answer…
  - Is the new discovered particle a 'vanilla' Higgs boson?
  - What is its production cross section and couplings?
  - Is there any SUSY in ATLAS data?
    - If not, what models do not agree with data?
- Enormous efforts in many channels, millions of plots with signal/backgrounds expectations, with systematics and observed data
- How do you conclude on these questions?
- Statistical tests construct probabilistic statements/models on P(theory|data) or P(data|theory)
  - Likelihood fits
  - Systematics/uncertainties
  - Hypothesis testing
  - Setting limits …
- Result: decisions based on these tests





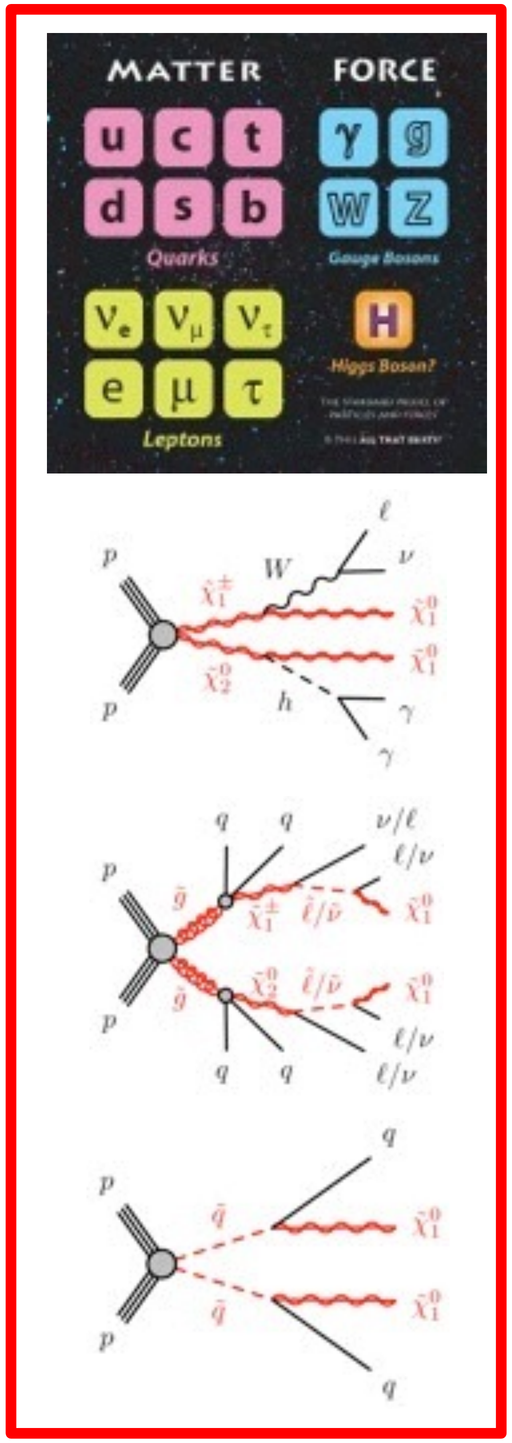
As a layman I would now say, I think we have it

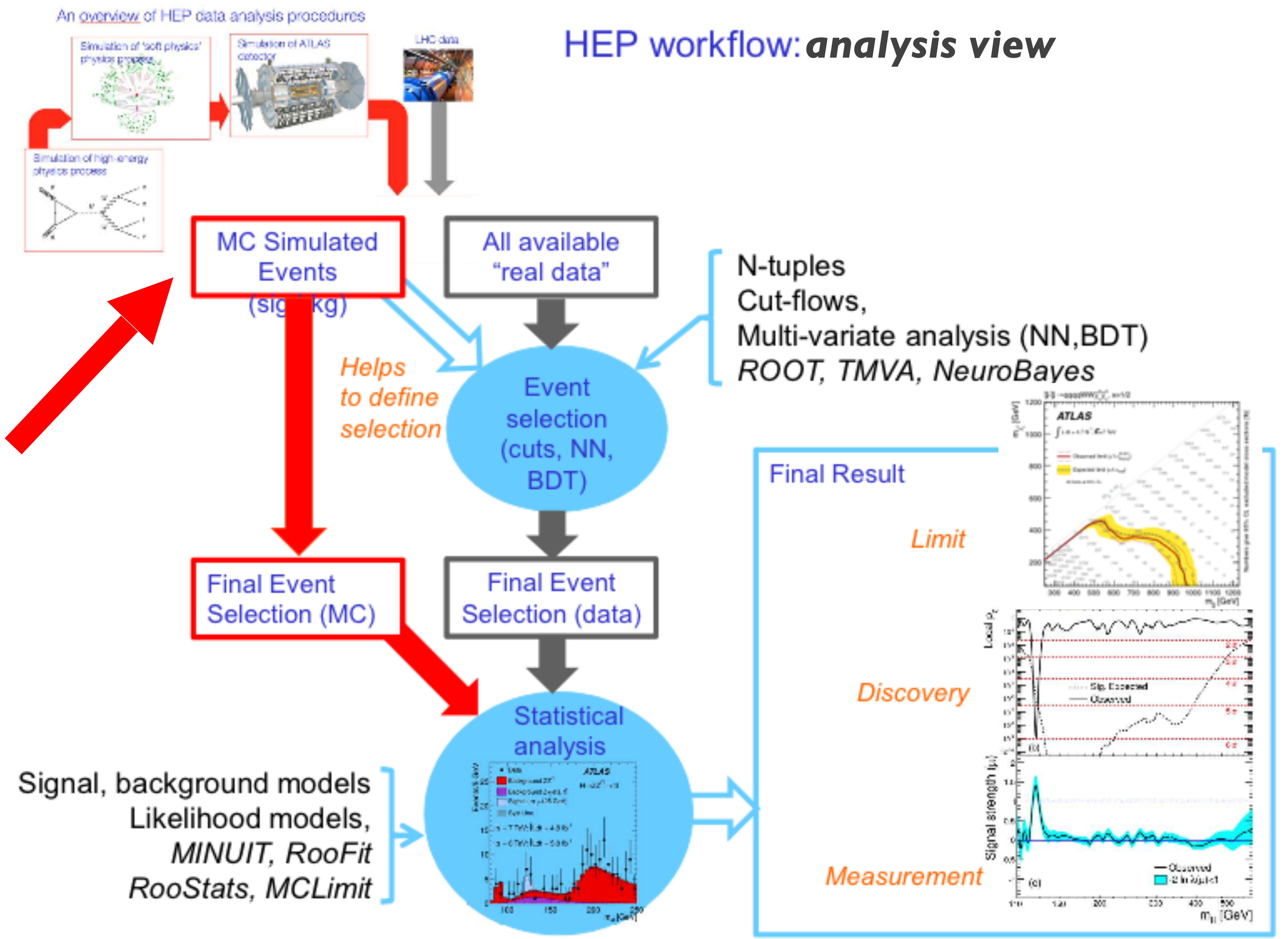
HEP workflow



W. Verkerke

HEP data analysis





W. Verkerke

- HEP Data Analysis is (should be) for a large part the reduction of a physics theory(s) to a statistical model
- Statistical/probability model: Given a measurement ***x*** (eg N events), what is the probability to observe each possible ***x***, under the hypothesis that the physics theory is true?

Simple statistical example

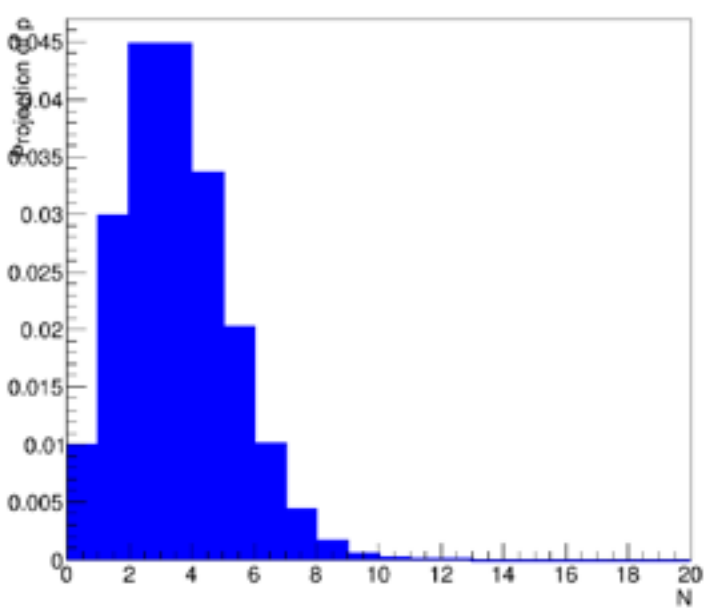
- Central concept in statistics is the ‘probability model’ : assigns a probability to each possible experimental outcome

- <u>Example</u>: a HEP counting experiment
  - Count number of events in your signal region (SR) in your data (specific lumi): Poisson distribution
  - Given the *expected(MC)* event count, the probability model is fully specified

$$P(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$

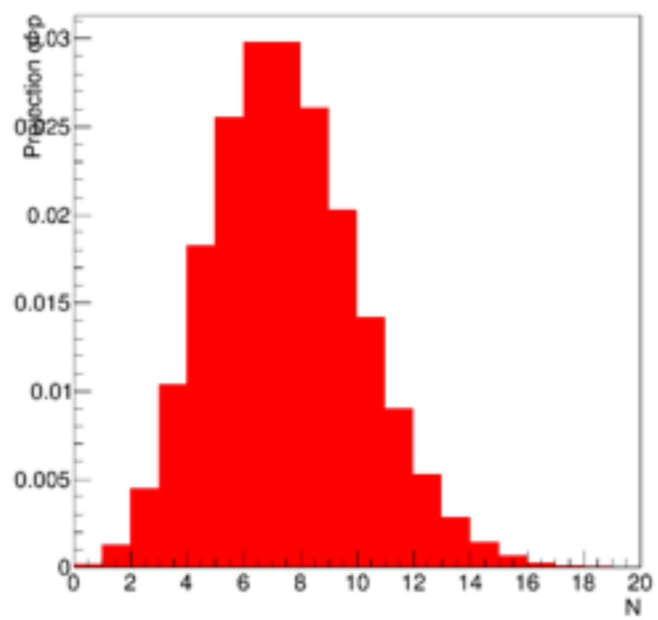
PROBABILITY

μ=3 (“bkg only”)

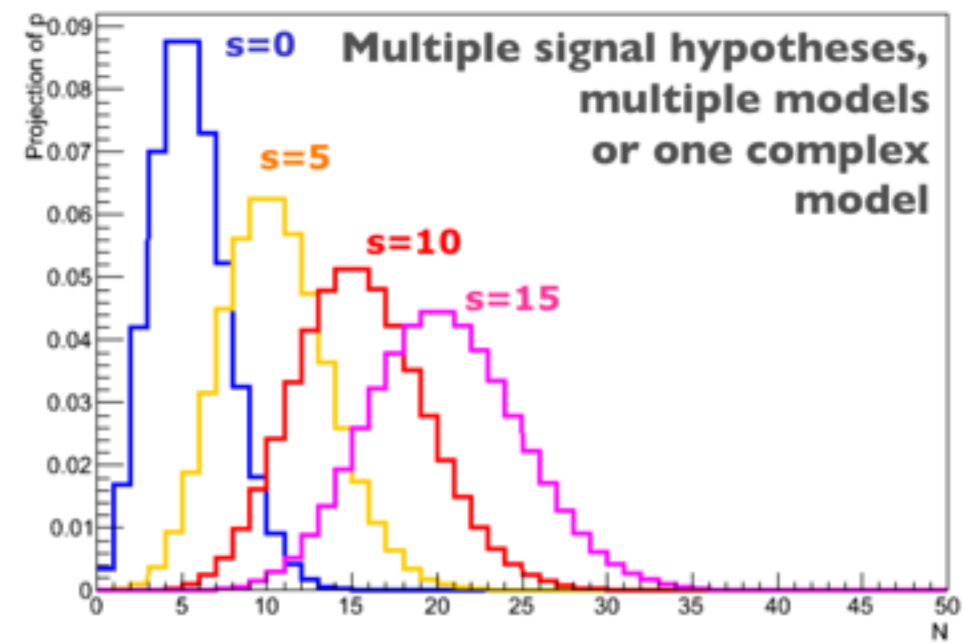


Poisson(N| b)

μ=7 (“bkg+signal”)



Poisson(N| s + b)



Poisson(N| s + b)

- Suppose we measure N = 7 events (Nobs), then can calculate the probability
- P(Nobs|hypothesis) is called **<u>LIKELIHOOD</u>** **- L(Nobs|b), L(Nobs|s+b), L(observed data|theory)**

p(Nobs|b) = 2.2% **p(Nobs|s+b) = 14.9%**

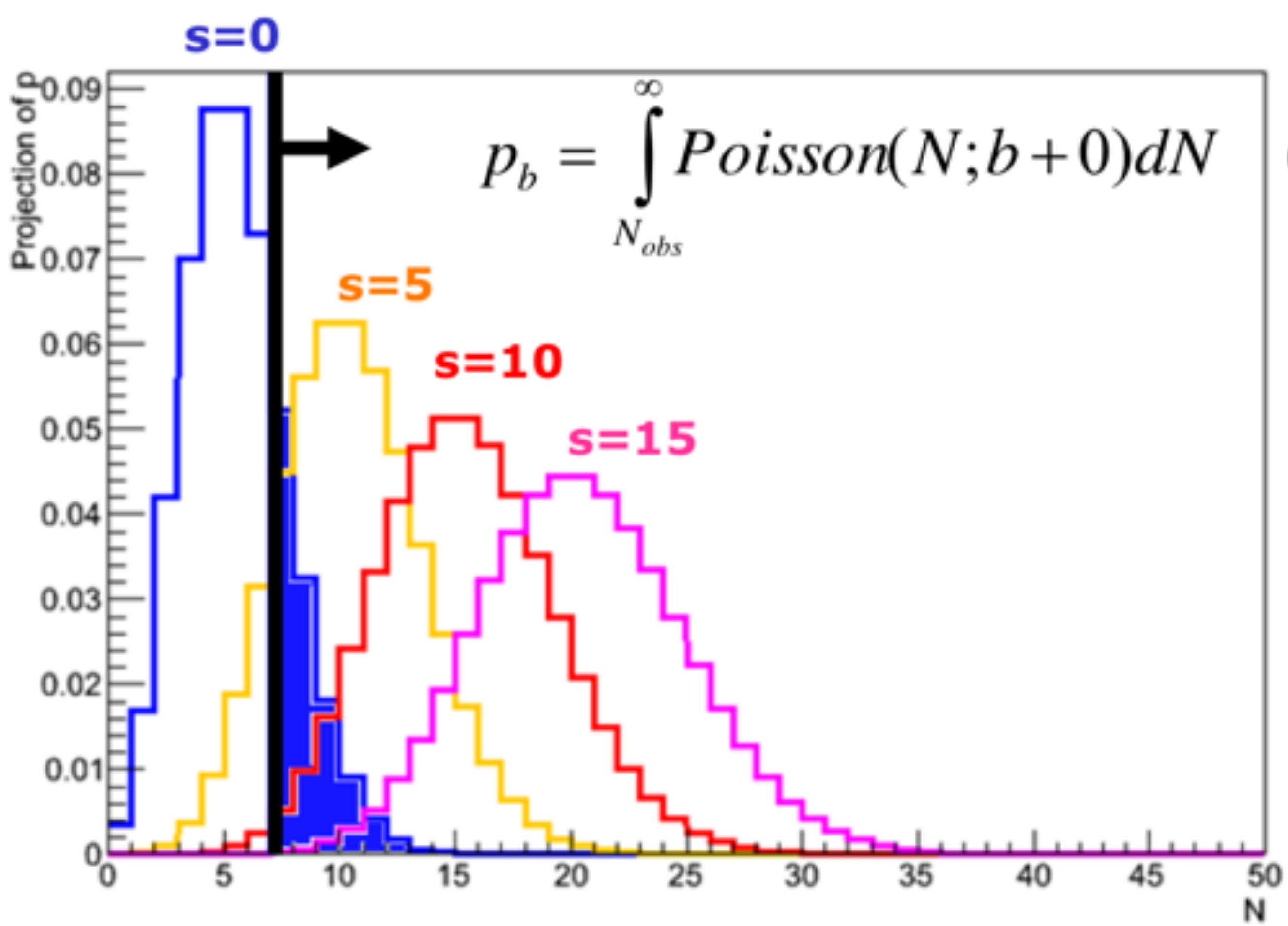
  - Data is more likely under s+b hypothesis than bkg-only

W. Verkerke

p-value

- **P-VALUE:** probability to obtain observed data, or more extreme, given the hypothesis in future repeated identical experiments

- For our example from previous page:
  - <u>For the bkg-only hypothesis:</u> $\mathbf{p_b}$ = Fraction of future measurements with N=Nobs (or larger) if s=0



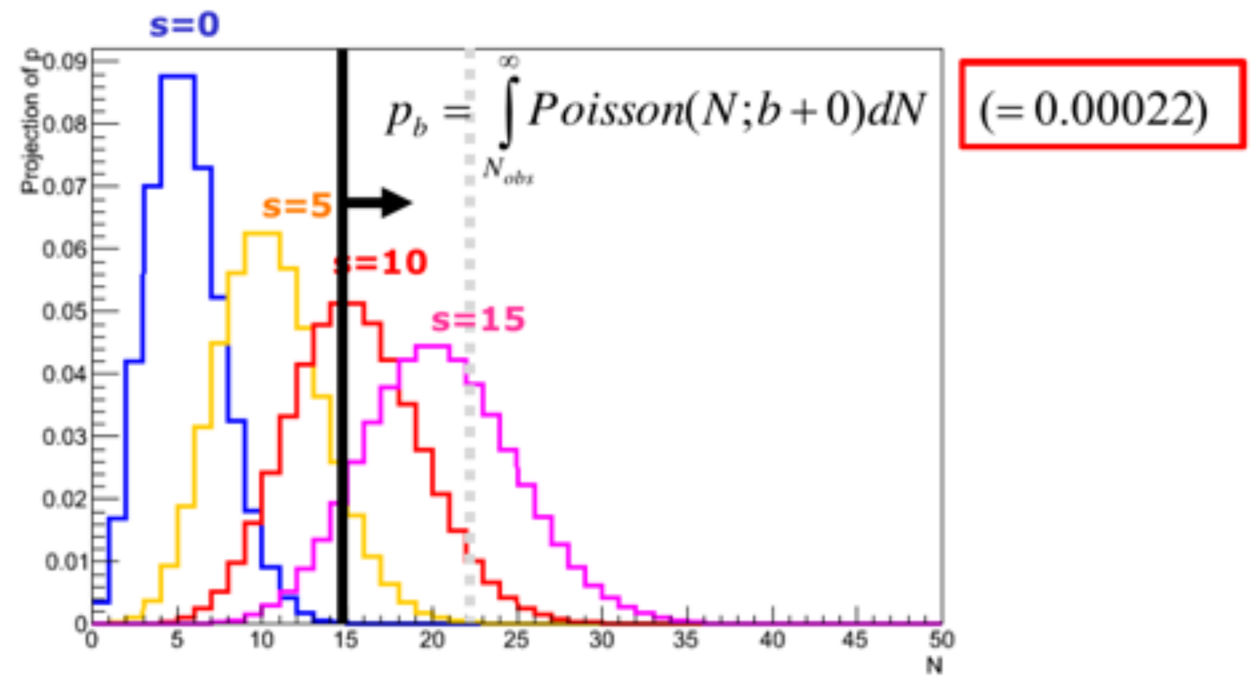
$$p_b = \int_{N_{obs}}^{\infty} Poisson(N; b+0)dN \quad (=0.23)$$

- Frequentist p-values (apologies to Bayesians) -- see links later

Excess over background

- **p_b** or p-values of background hypothesis is used to quantify 'discovery'
- 'discovery' = excess of events over background expectation

- One more example:
  - Nobs=15 for same model, what is **p_b**?



- Results customarily expressed as odds of a Gaussian fluctuation with equal p-value: **significance, Zn, z-value**

- Nobs = 15 → Zn = 3.5σ

- Nobs = 22 → **Zn = 5σ**
  or **p_b < 2.87 ×10^{-7}**

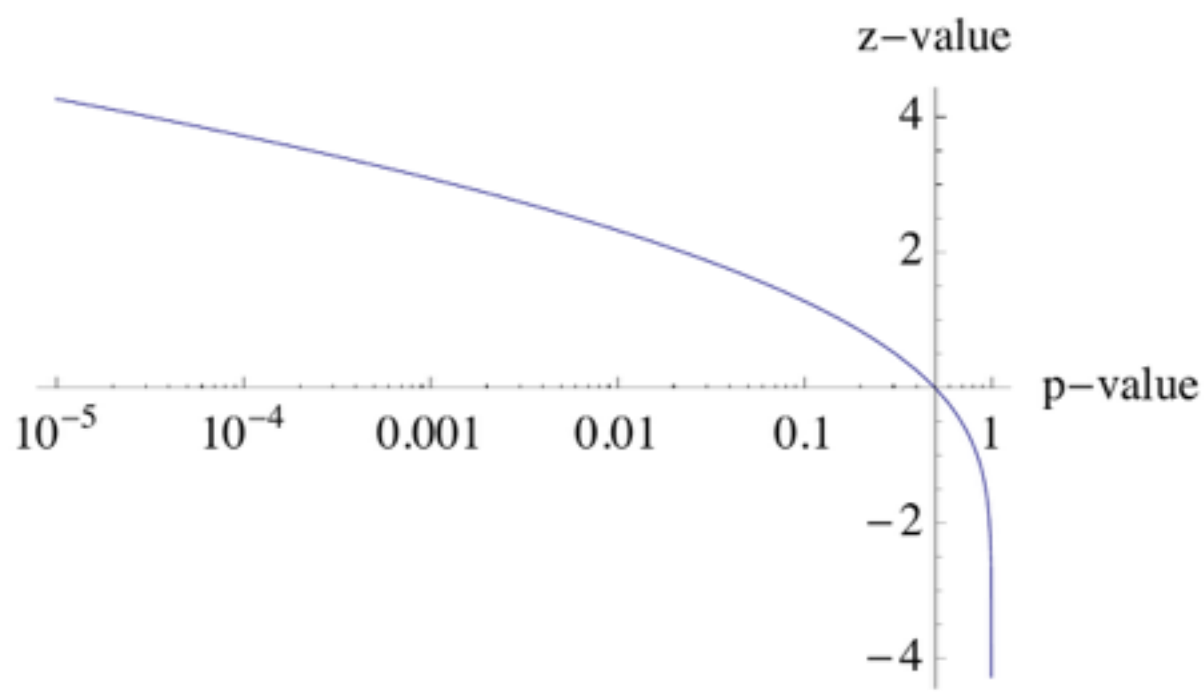


Fig. 1. Relationship between *p*-value and *z*-value.

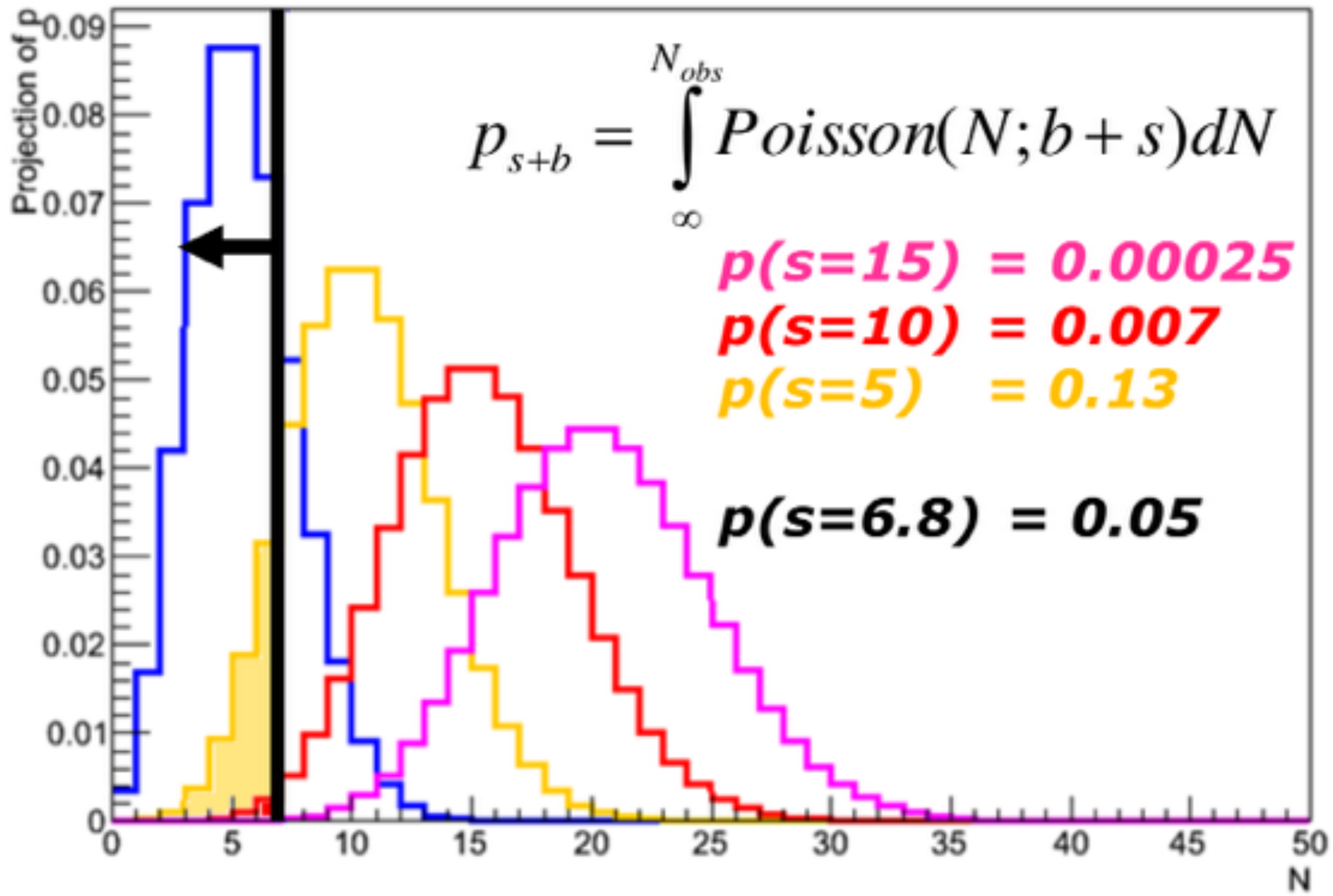
z-value =

`sqrt(2.) * TMath :: ErfInverse(1. − 2. * pvalue)`

$$p\text{-value} = \int_{z\text{-value}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx,$$

Upper limits

- Can also define p-value for s+b hypothesis $\mathbf{p_{s+b}}$
  - Note convention change: integration range in $\mathbf{p_{s+b}}$ is flipped



- Convention: express result as value (upper limit) of **s** for which $\mathbf{p_{s+b}}$ = 5%
  or excluded at 95% confidence level (95% C.L.)
- Our example:
  - s>6.8 is excluded at 95% C.L.

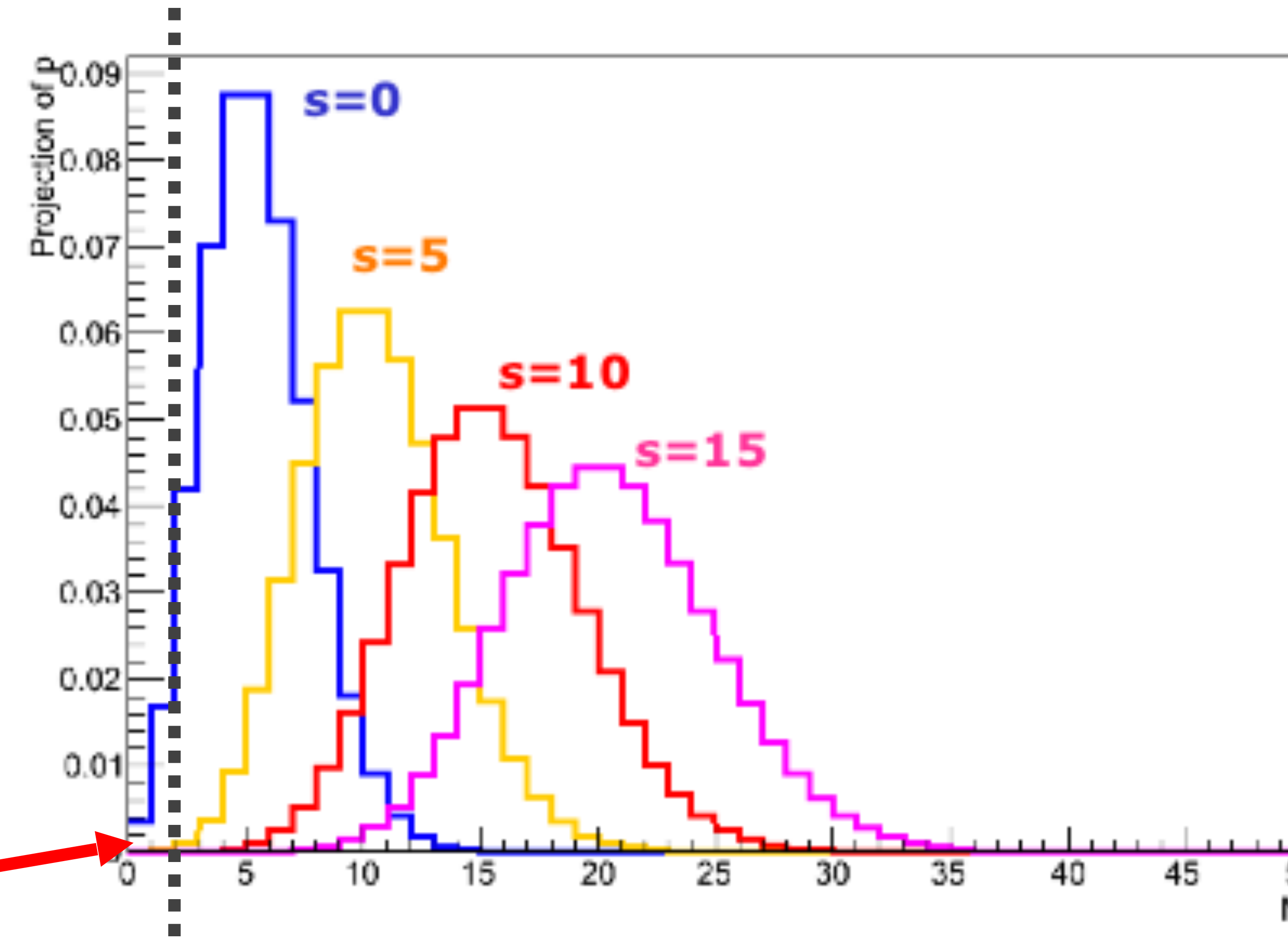
Modified Upper limits : CLs

- Interpretation of $\mathbf{p_{s+b}}$ in terms of inference on signal only is problematic
  - Since $\mathbf{p_{s+b}}$ quantifies consistency with data of signal + background
  - Problem apparent when observed data has downward fluctuation wrt background expectation

- Example: Nobs = 2 → $\mathbf{p_{s+b}(s=0) = 0.04}$
  - **s≥0 excluded at 95% C.L. ???**

- Modified approach to protect against such inference on signal (LHC convention):
  - Instead of requiring $\mathbf{p_{s+b}}$ = 5%, require

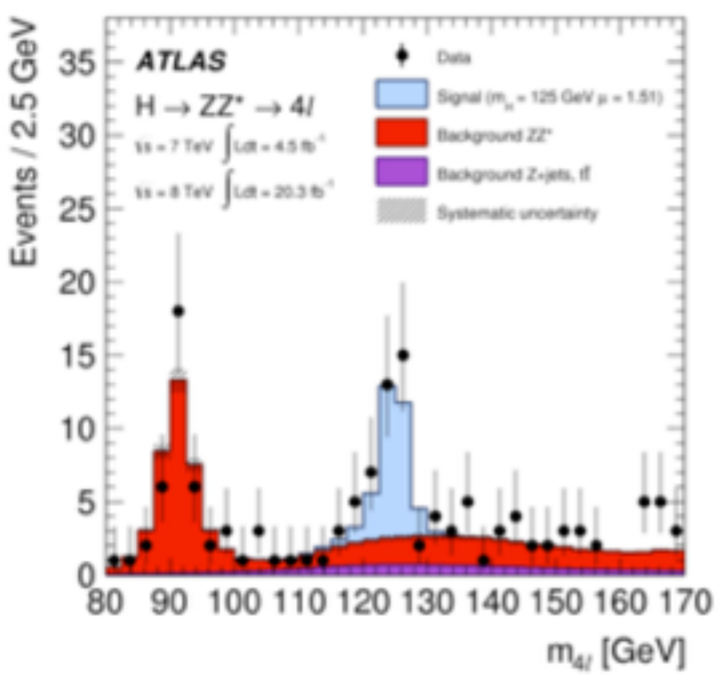
$$\mathrm{CL}_s \equiv \frac{p_{s+b}}{\boxed{1 - p_b}} = 5\%$$

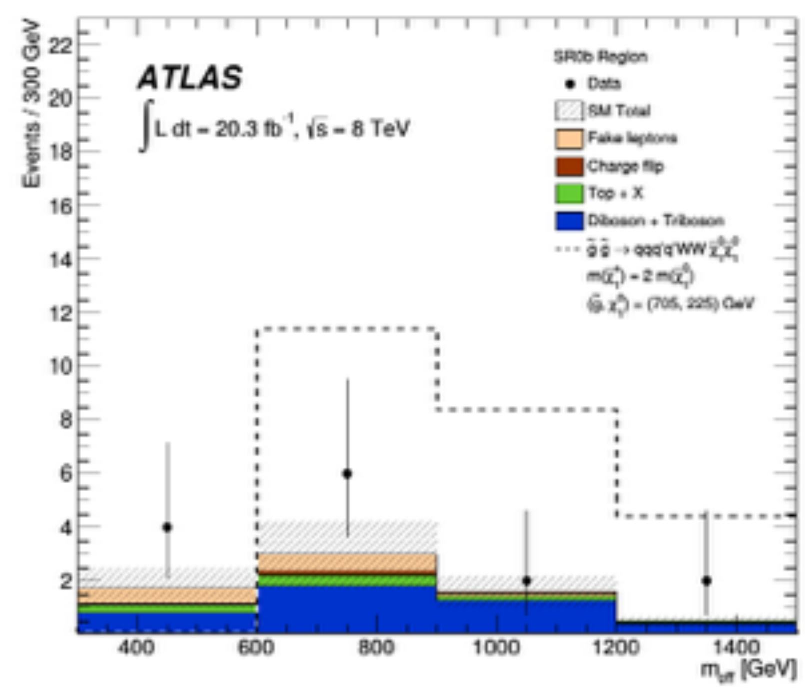


- Example: Nobs = 2 → **s>3.4 excluded at 95% CLs**
- For large Nobs effect on limit is small as $\mathbf{p_b}$ → 0

More complex examples

- Typical analysis is not a simple counting experiment
  - Many intrinsic uncertainties on signal and bkg
  - Result is a distribution, not a single number
    - SUSY searches: discovery is cut&count, but many exclusion limits are shape-fits/multi-bin





- Any result can be converted into a single number by constructing a ***test statistic***
  - A test statistic compresses all signal-to-background discrimination power into one number
  - Most powerful discriminators are
    Likelihood Ratios
    (Neyman-Pearson lemma)
  - $\boldsymbol{q_\mu}$ is a common test statistic
    (LHC convention)

$$q_\mu = -2\ln \frac{L(data\,|\,\mu)}{L(data\,|\,\hat{\mu})}$$

Likelihood ratio test statistic

- **Signal strength μ** = signal rate / nominal signal rate (also know as $\mu_{\mathbf{SIG}}$)
  - Bkg-only hypothesis: **μ** = 0
  - Bkg + signal hypo: **μ** = 1
  - Bkg + 2 X signal hypo: **μ** = 2
- Likelihood with nominal signal strength (**μ** = 1)

'likelihood assuming nominal signal strength'

$$q_1 = -2\ln\frac{L(data\,|\,\mu=1)}{L(data\,|\,\hat{\mu})}$$

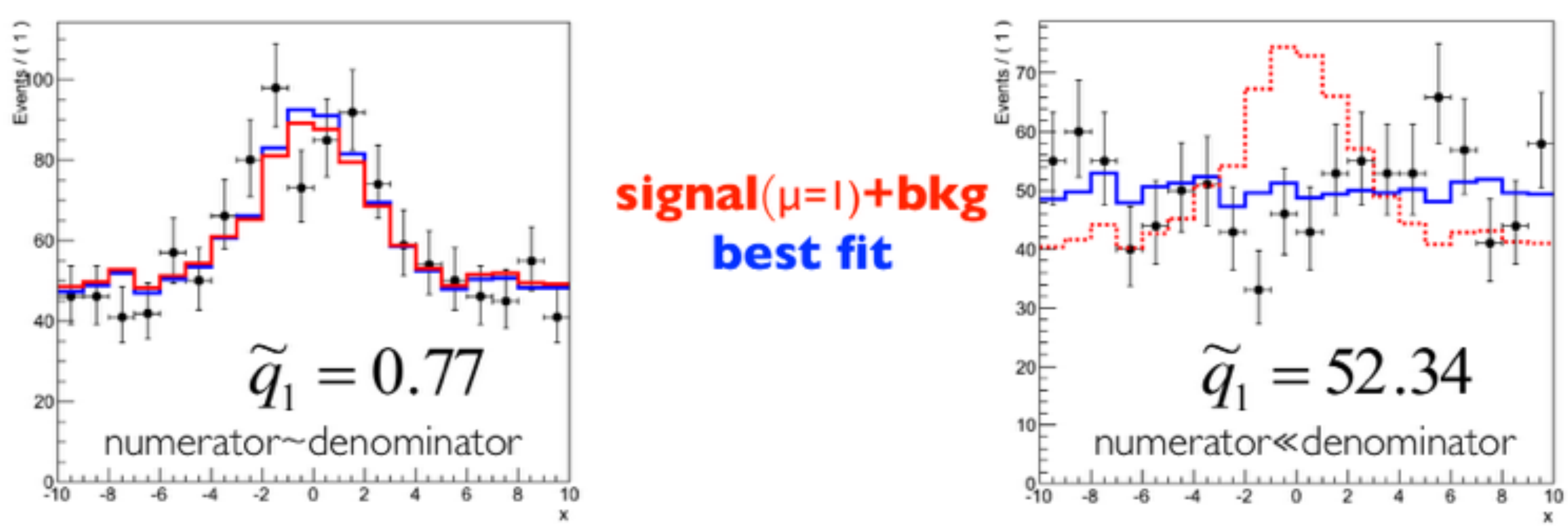
$\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

- **Example:** simple s + b model with no uncertainties

*On signal-like data q1 is **small***

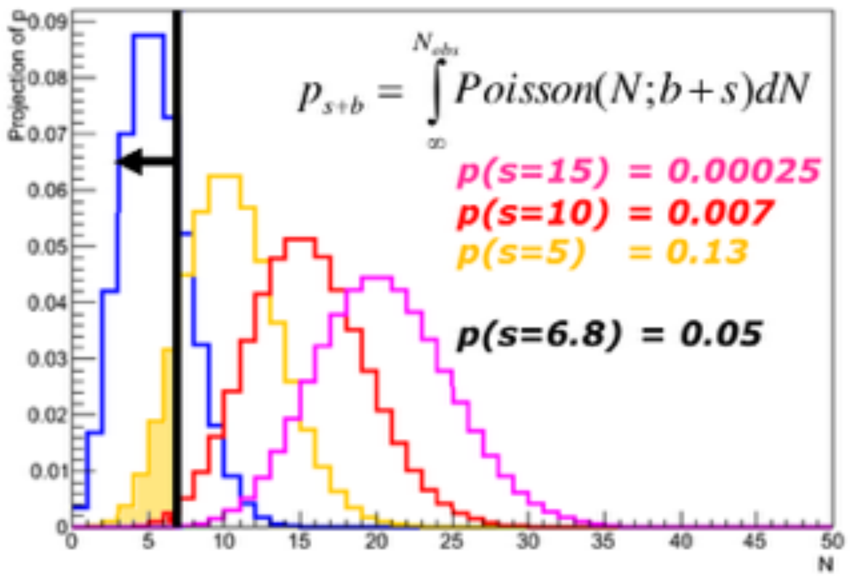
*On background-like data q1 is **large***

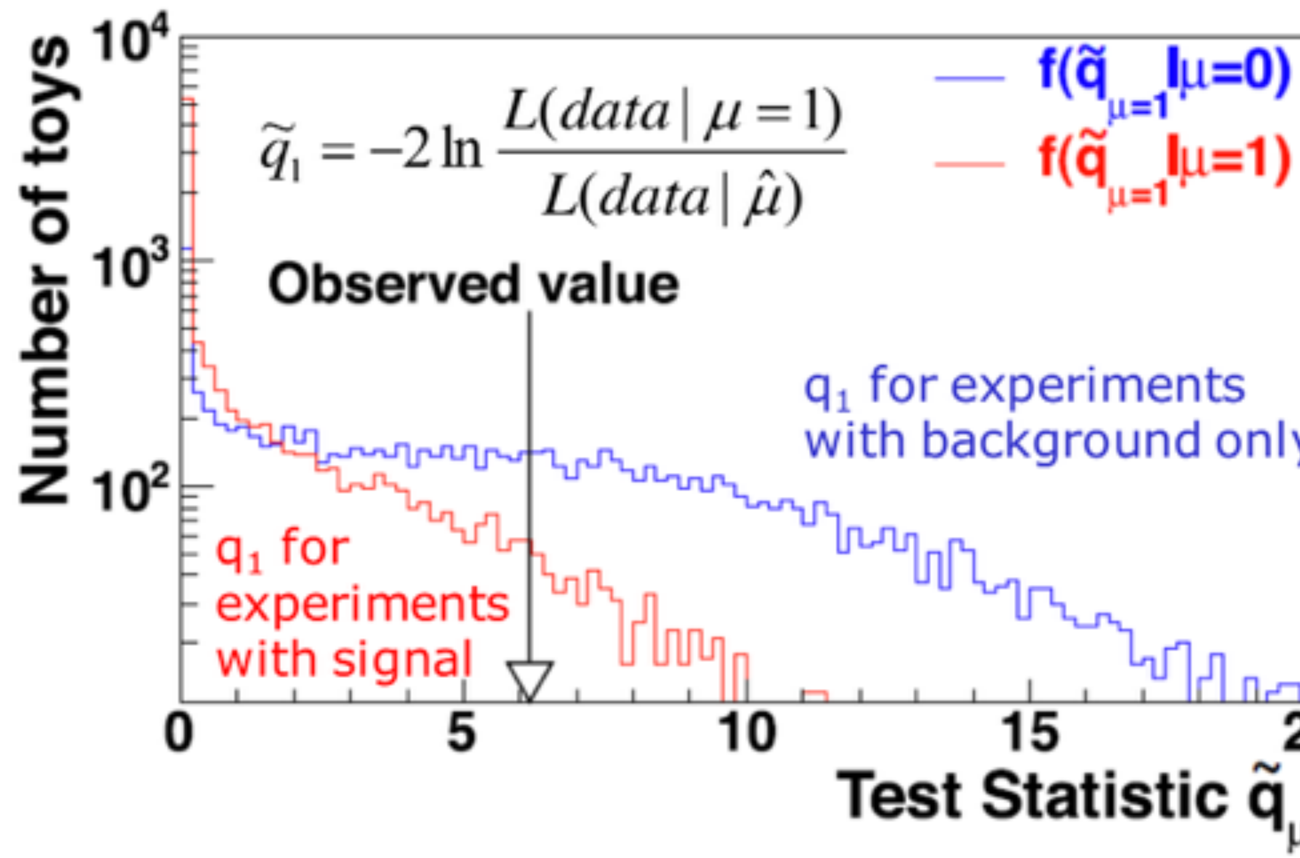


Distribution of test statistic

- Value of q_1 on data is now the *measurement*
- Distribution of q_1 is **not** calculable → But can be obtained from pseudo-experiments (toys)
  - Generate a large number of pseudo-experiments with a given value of **μ**, calculate q_1 for each, plot distribution

Note analogy to Poisson counting example





- From q_{obs} and these test statistic distributions, $f(q_\mu)$, can then set limits or calculate discovery significance similar to what was shown for Poisson example

- Typically CPU-intensive to run many toy-experiments → approximate with **asymptotic formulae**, aka **asimov data** (only works in cases when Nobs≳10, see links for details)

Systematic uncertainties

- Typically HEP models will have uncertainties: experimental (JES,trigger eff.) or theoretical (Q,σ)

$$L(data \mid \mu) \rightarrow L(data \mid \mu, \vec{\theta}) \qquad L(data \mid \mu, \theta) = Poisson(N_i \mid \mu \cdot s_i(\theta) + b_i(\theta)) \cdot p(\tilde{\theta}, \theta)$$

- Models w/ uncertainties, described by additional parameters θ that describe effect of uncert.
- Likelihood includes *auxiliary measurement* terms that constrain the nuisance parameters θ
  - Auxiliary measurement given by performance group (jet perf.) or theory variations (renorm. scale up/down)

- Likewise uncertainties quantified by nuisance parameters are incorporated into test statistic using **Profile Likelihood Ratio**

$$q_\mu = -2 \ln \frac{L(data \mid \mu)}{L(data \mid \hat{\mu})} \longrightarrow \tilde{q}_\mu = -2 \ln \frac{L(data \mid \mu, \hat{\hat{\theta}}_\mu)}{L(data \mid \hat{\mu}, \hat{\theta})}$$

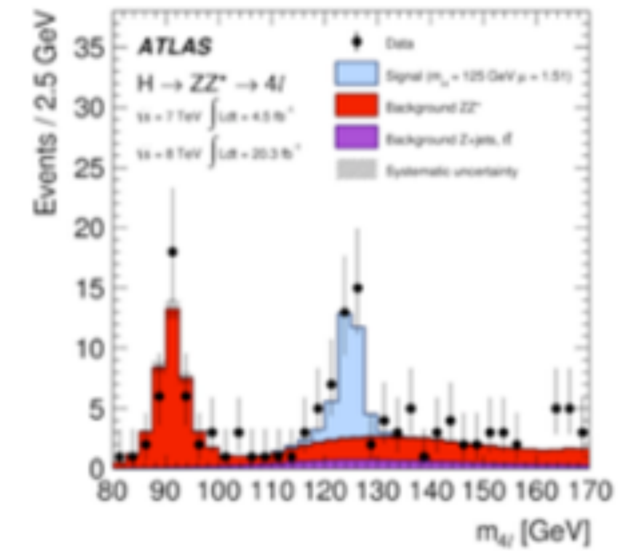
'likelihood of best fit for a given fixed value of μ'

'likelihood of best fit'

(with a constraint $0 \leq \hat{\mu} \leq \mu$)

Overview for a search

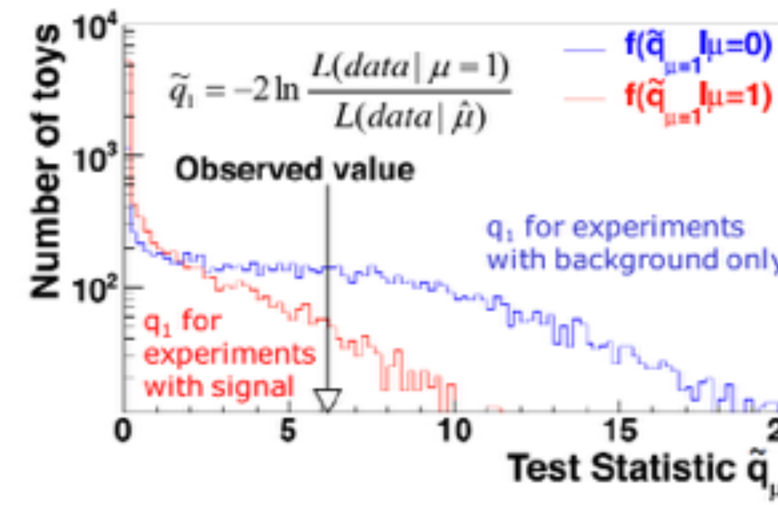
- Take Higgs search as example, and put it all together



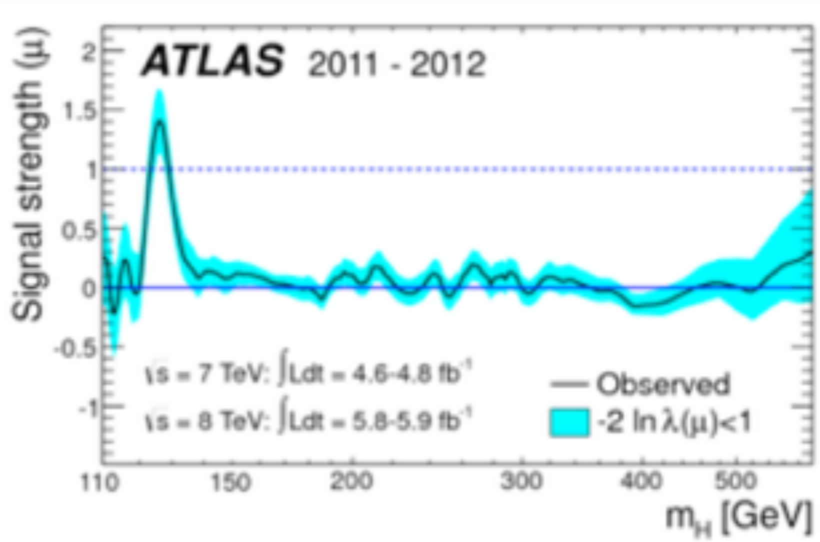
- Result from data is a distribution (eg m(4l))
- Model signal and background by PDF (probability density function) for a given Higgs mass hypothesis

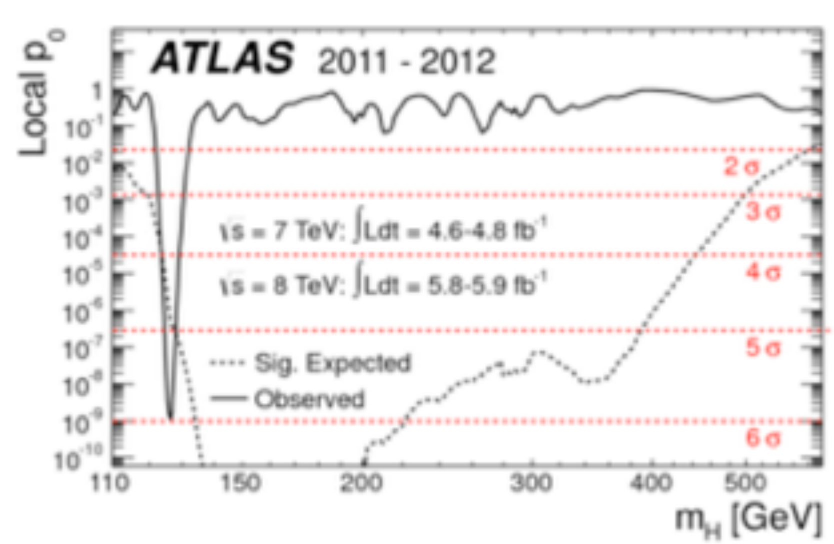
- Construct likelihood(s) by joining data and model(s)
- Construct test statistic q_μ from likelihoods

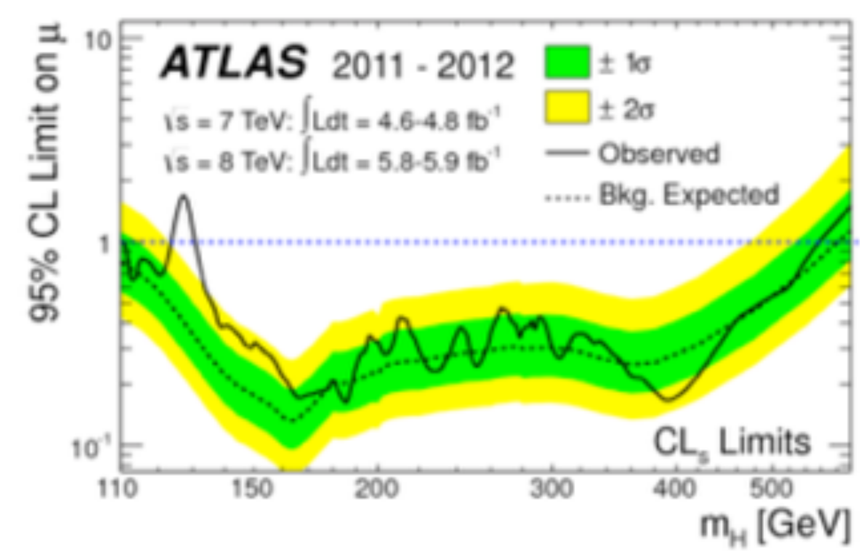
$$\tilde{q}_\mu(m_H) = -2\ln\frac{L(data\,|\,\mu, m_H \hat{\hat{\theta}}_\mu)}{L(data\,|\,\hat{\mu}, m_H \hat{\theta})}$$



- Obtain expected distributions of q_μ
- Determine discovery p_0 and signal exclusion limit
- Repeat for each assumed m_H







Links

- Statistics lectures (CERN school, 2014, W. Verkerke):
  - Part-1: https://indico.cern.ch/event/287744/contribution/7/material/slides/0.pdf
  - Part-2: https://indico.cern.ch/event/287744/contribution/11/material/slides/1.pdf
  - Part-3: https://indico.cern.ch/event/287744/contribution/14/material/slides/0.pdf

- Plotting the Differences Between Data and Expectation, G. Choudalakis, D. Casadei http://arxiv.org/abs/1111.2062

- CLs: https://twiki.cern.ch/twiki/pub/AtlasProtected/StatisticsTools/CLsInfo.pdf

[28] A. Read, Presentation of search results: the CL s technique, Journal of Physics G: Nuclear and Particle Physics 28 (10) (2002) 2693.

[29] G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur.Phys.J. C71 (2011) 1554. arXiv:1007.1727, doi:10.1140/epjc/s10052-011-1554-0.

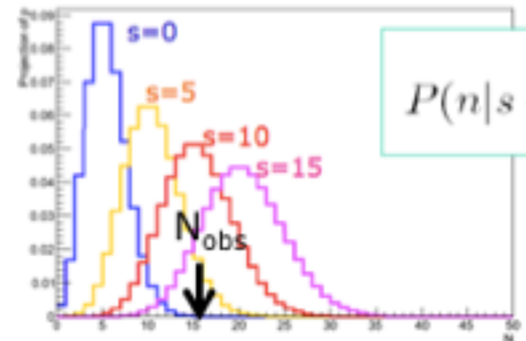
[30] S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Statist. 9 (1938) 60–62.

Introduction to statistics tools

Largely borrowed from lectures/slides by W. Verkerke

LIKELIHOOD, LIKELIHOOD, LIKELIHOOD...

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



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

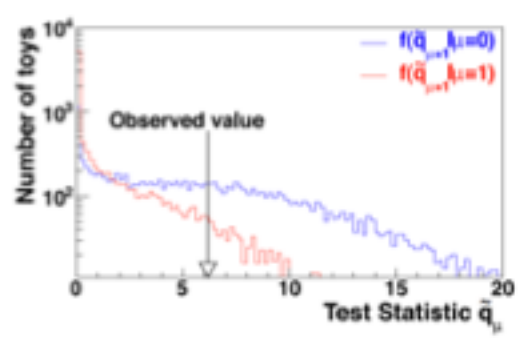
NB: b is a constant in this example

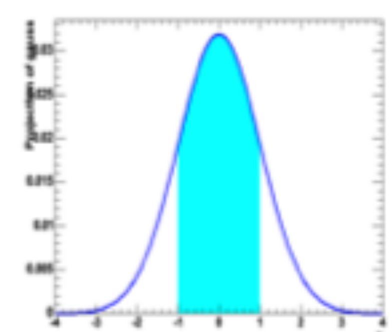
Definition: the Likelihood is P(observed data|theory)

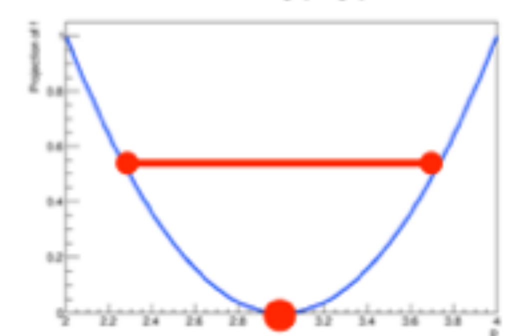
e.g. L(15|s=0)

e.g. L(15|s=10)

Frequentist statistics	Bayesian statistics	Maximum Likelihood
$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}\,\vert\,\mu)}{L(\vec{N}\,\vert\,\hat{\mu})}$	$P(\mu) \propto L(x\,\vert\,\mu) \cdot \pi(\mu)$	$\left.\frac{d \ln L(\vec{p})}{d\vec{p}}\right\vert_{p_i=\hat{p}_i} = 0$







Confidence interval or p-value	**Posterior on s or Bayes factor**	**s = x ± y**

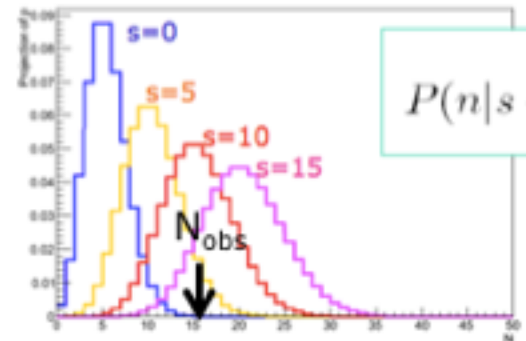
Wouter Verkerke, NIKHEF

Modular software design

- **RooFit:** tool/language for building probability models: datasets, likelihoods, minimization, toy data, visualization
- **HistFactory:** tool to construct binned template models of arbitrary complexity using classes of physics concepts: channel/region, sample, uncertainties
  Builds a RooFit stat. model from HistFactory physics model
- **RooWorkspace:** persistent RooFit object to transport a likelihood, containing model/data. Completely factorizes process of building and using likelihood functions.

- **RooStats:** tool/suite to calculate intervals and perform hypothesis tests using a variety of statistical techniques; easy to use with RooWorkspace

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

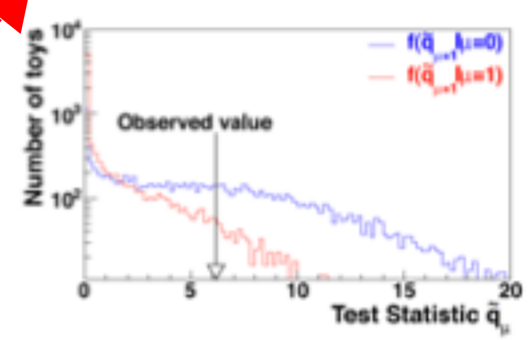
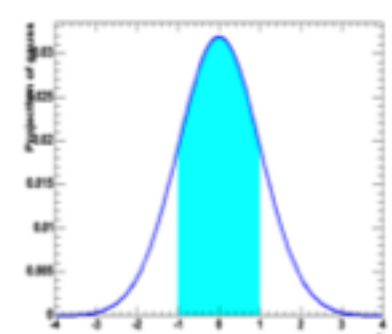
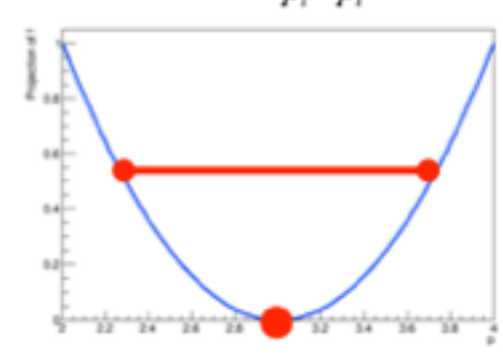


$$P(n|s+b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

NB: b is a constant in this example

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

e.g. L(15|s=0)
e.g. L(15|s=10)

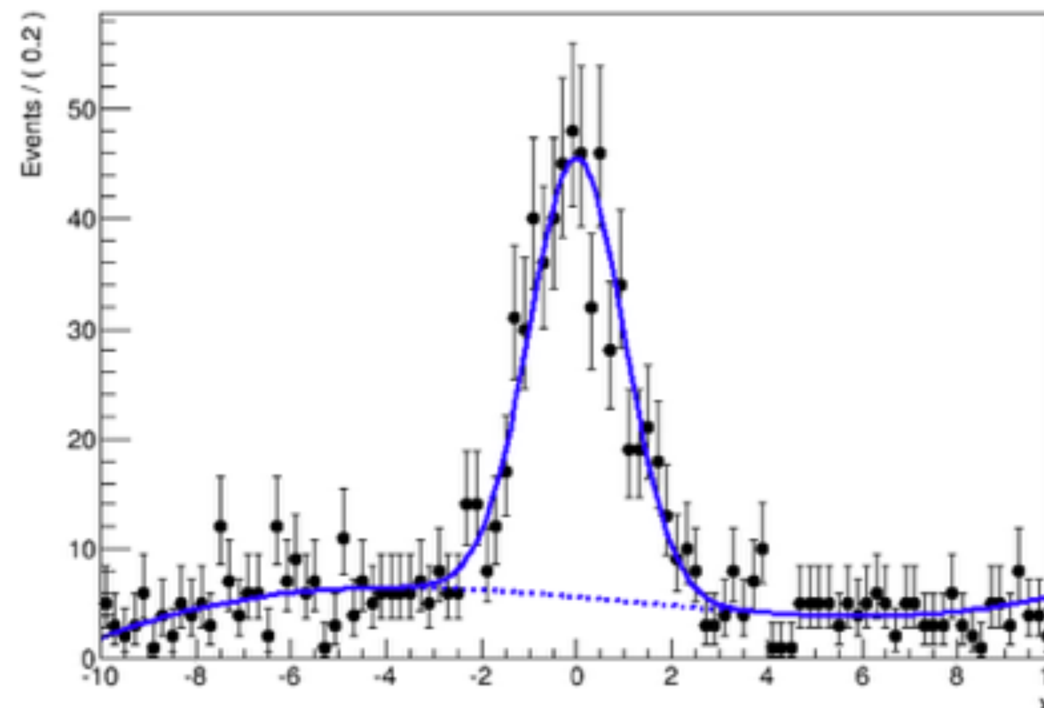
Frequentist statistics	Bayesian statistics	Maximum Likelihood
$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}\mid\mu)}{L(\vec{N}\mid\hat{\mu})}$	$P(\mu) \propto L(x\mid\mu)\cdot\pi(\mu)$	$\left.\frac{d\ln L(\vec{p})}{d\vec{p}}\right\vert_{p_i=\hat{p}_i} = 0$
		
Confidence interval or p-value	Posterior on s or Bayes factor	s = x ± y

Wouter Verkerke, NIKHEF

RooFit

- <u>Focus:</u> coding a probability density function PDF : how do you formulate a PDF in ROOT?

- Simple example: gauss (signal) + polynomial (bkg)
- Quickly becomes complicated: multidimensional, unbinned fits, non-trivial functions, non-analytic functions
- <u>Core design philosophy:</u>
  mathematical objects represented as C++ objects

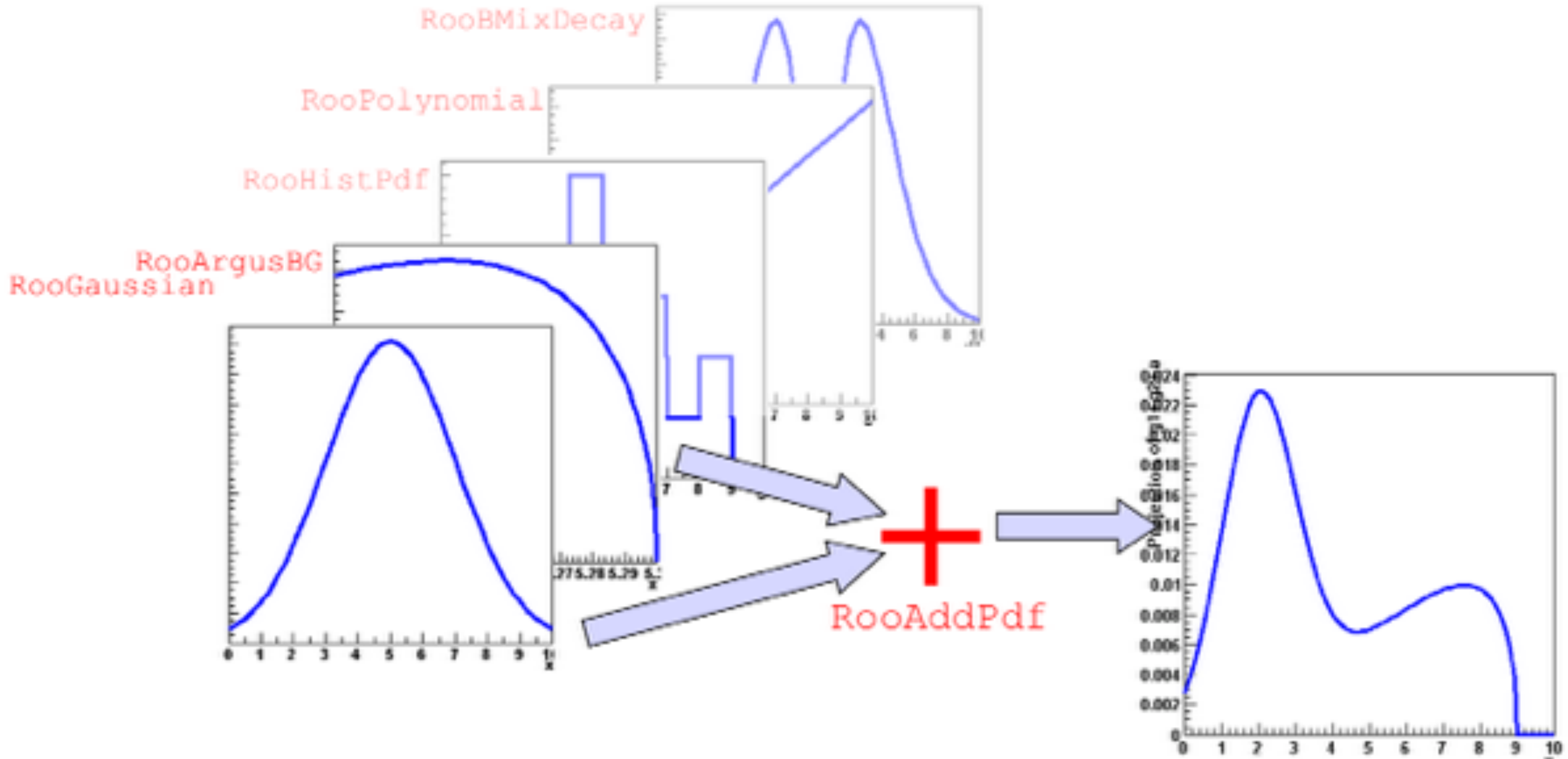


Mathematical concept		RooFit class
variable	x	RooRealVar
function	$f(x)$	RooAbsReal
PDF	$f(x)$	RooAbsPdf
space point	$\vec{x}$	RooArgSet
integral	$\int_{x_{min}}^{x_{max}} f(x)dx$	RooRealIntegral
list of space points		RooAbsData

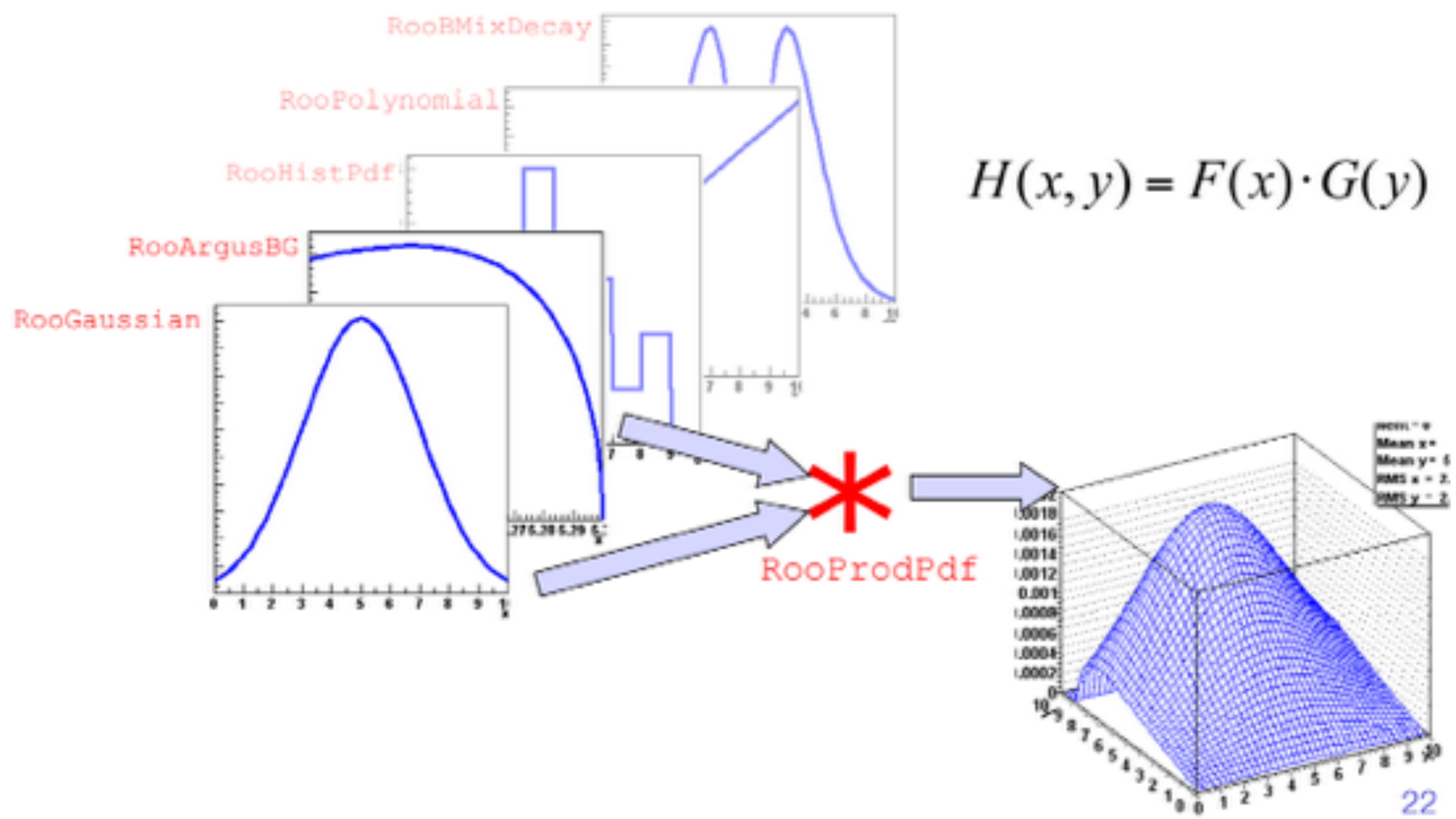
RooFit - model building

- Easy to use standard components to build more complex/realistic models

- Addition



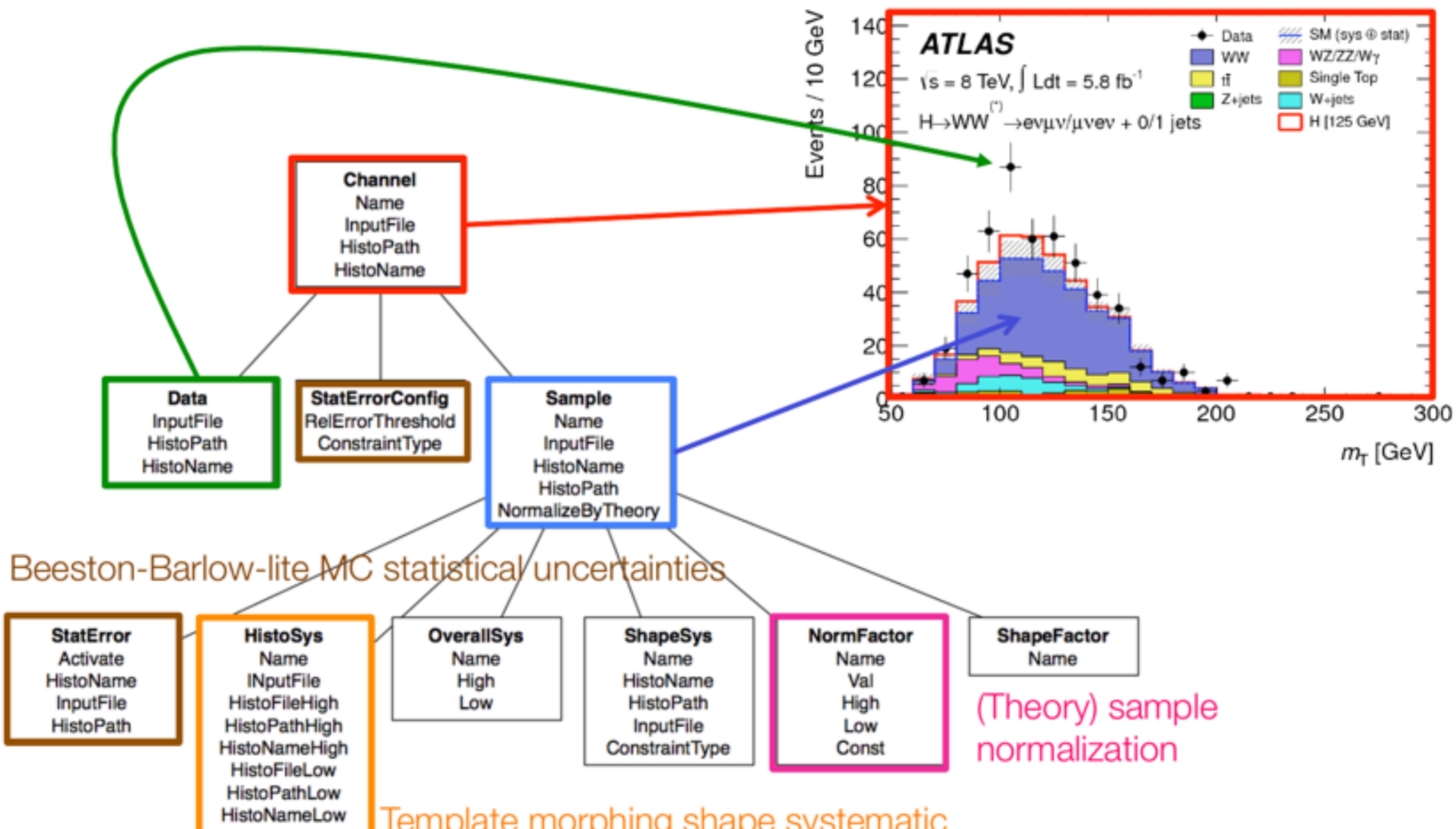
- Product (multi-dimensional)



$$H(x, y) = F(x) \cdot G(y)$$

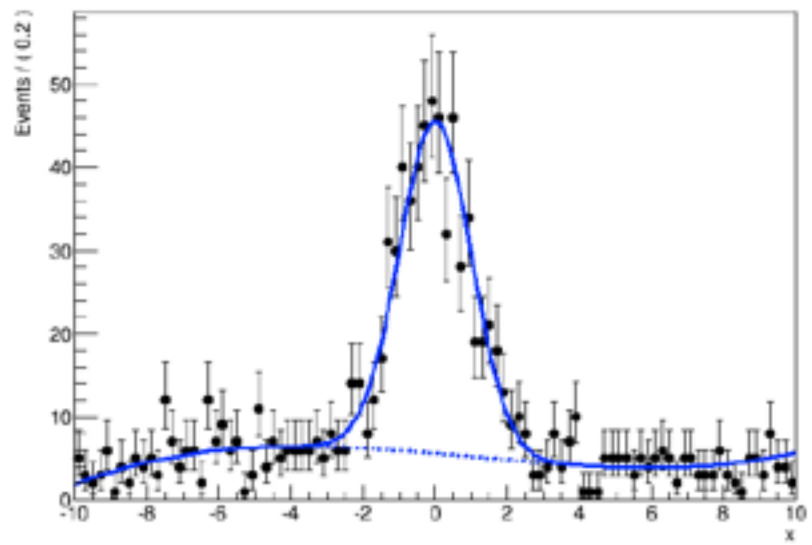
HistFactory

- Structured building of complex models based on binned templates (histograms)
- Classes of physics concepts:
  - Channel = region of phase space
    - One or more channels are combined to form a measurement
  - Sample = physics process: either data-driven or described by Monte Carlo (MC) simulation
  - Systematics = intrinsic uncertainty on your model



Systematics : nuisance parameters

- Empirical modeling of your model is easy to do, but expect some hard questions
  - Gaussian for signal + polynomial for background



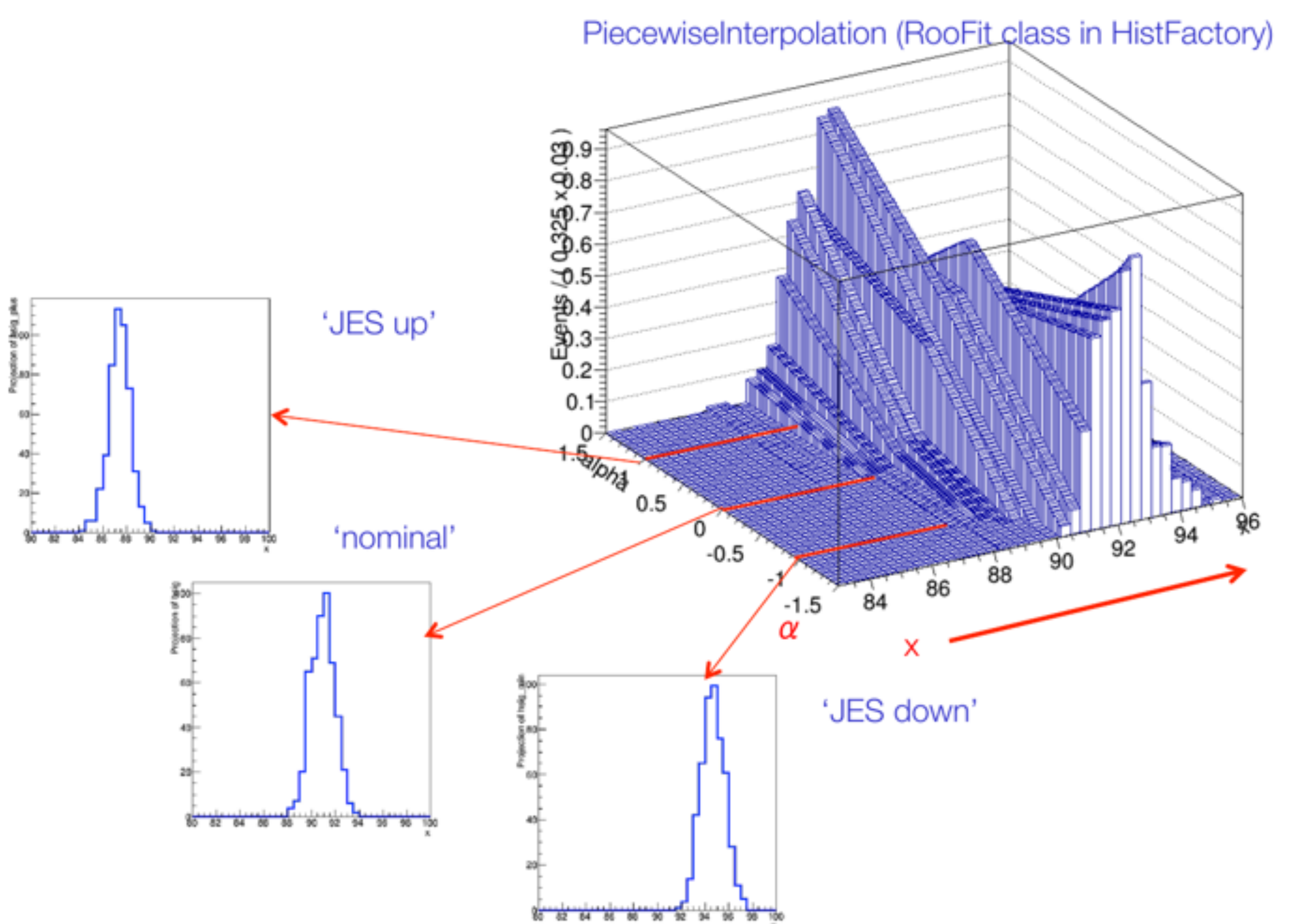
$$L(x \mid f, m, \sigma, a_0, a_1, a_2) = fG(x, m, \sigma) + (1 - f)Poly(x, a_0, a_1, a_2)$$

- Is your model correct?
  - Is the true signal distribution captured by a Gaussian?
- Is your model flexible enough?
  - Why use 4th order polynomial and not 6th order?
- How do your model parameters connect to known detector/theory uncertainties for your distribution?
  - What conceptual uncertainty does what parameter represent? And are all conceptual uncertainties represented?

Systematics modeling - interpolation

- A common solution is to introduce degrees of freedom in model that describe specific systematic/uncertainty!
- The +1/-1 σ variations sampled from MC simulation are compared to nominal MC response
    - (corrected/checked/double-checked to data by Perf. Groups)
- Interpolation, performed between **+1σ** ↔ **nominal** ↔ **-1σ** taken into the model as nuisance parameter

$$L(data \mid \mu,\theta) = Poisson(N_i \mid \mu \cdot s_i(\theta) + b_i(\theta)) \cdot p(\tilde{\theta},\theta)$$



RooWorkspace

- Complete description of likelihood persistable in a ROOT file
- Factorizes building and using likelihood functions
  - In setup, team member, place and time

- Construct RooFit model **sum** and persist to ROOT file

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

- Pass file to your colleague

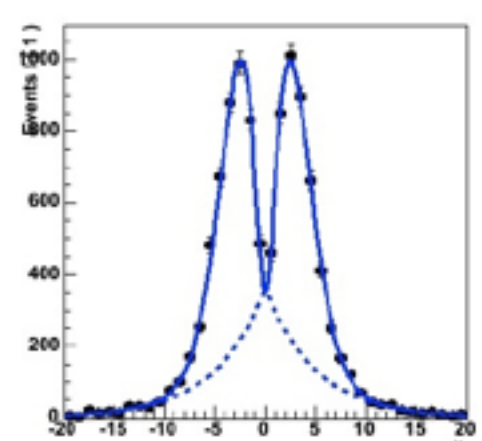
model.root

- Colleague resurrects likelihood, runs fit and produces plots

```
// 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) ;
```



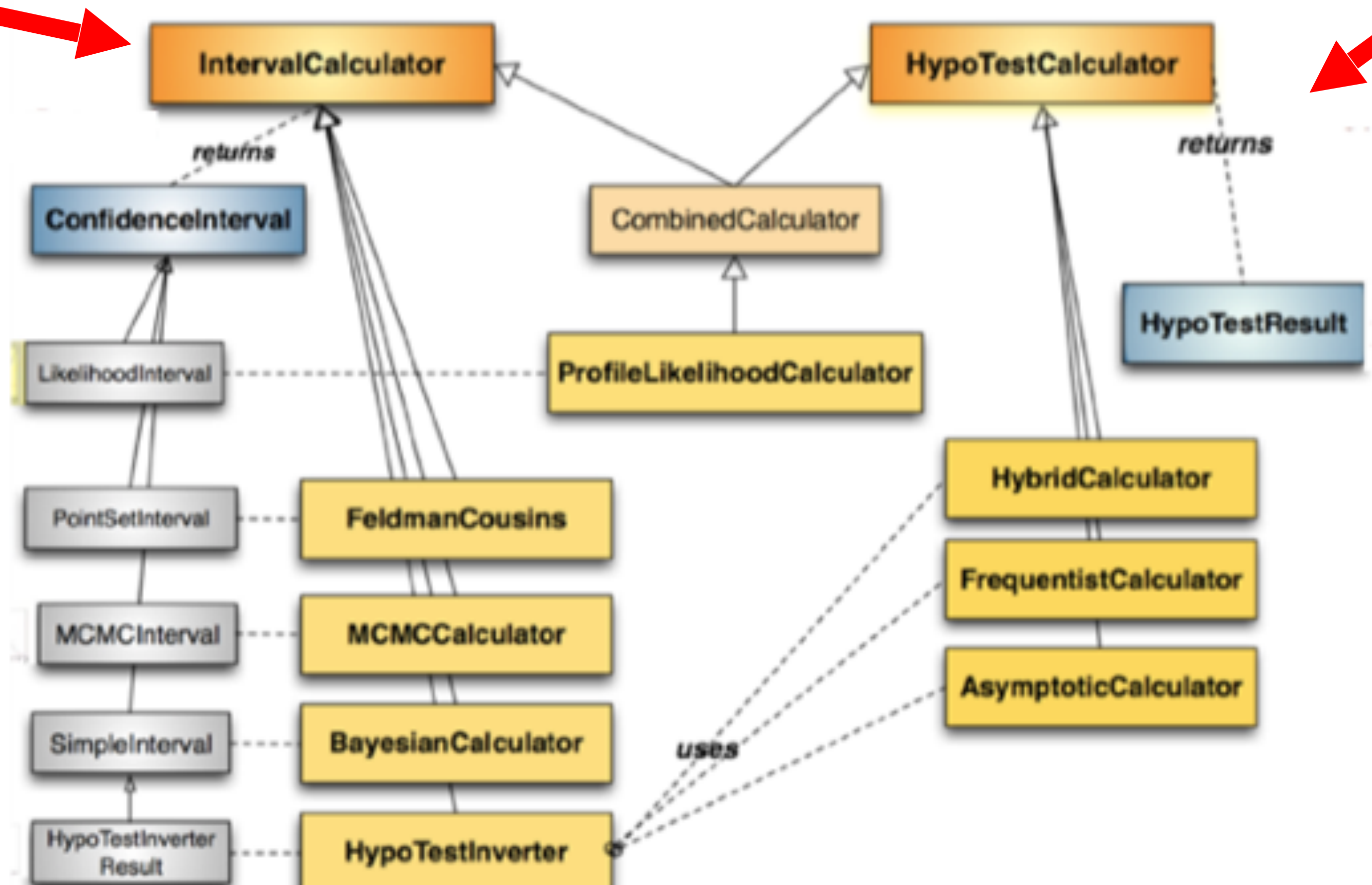
RooStats

- RooFit/HistFactory give tools to construct (complex) probability density functions
- RooWorkspace makes it possible to decouple statistical test tools from model contruction
- **RooStats** project/tools suite delivers a series of tools that can calculate intervals and perform hypothesis tests using a variety of statistical techniques
  - Frequentist/Bayesian/Likelihood-based methods (confidence/credible interval, hypothesis tests)

Confidence intervals: $[\theta_-, \theta_+]$,or θ<X at 95% C.L.
Hypothesis testing: → p(data| θ=0) = 1.10^{-7}

- Abstract interface for procedure to calculate a confidence interval
- Abstract interface for result: *s>6.8 is at 95% C.L.*

RooStats class structure



- Abstract interface for hypothesis tester to calculate a p-value
- Concrete result class: <u>*HypoTestResult*</u>
  $p_b = 0.023$
  $CL_S = 0.00513$

Overview

- **Step-0:** define signal/control/validation regions
  - Input TTrees (derived from xAOD), histograms, numbers
- **Step-1:** Construct PDF and the likelihood function

  RooFit or HistFactory + RooFit
  - Result from data is a distribution
  - Model signal and background by PDF (prob. density func.)
  - Construct likelihood(s) by joining data and model(s)
- ↓
- **RooWorkspace**
- ↓
- **Step-2:** Statistical tests on parameter of interest μ

  RooStats
  - Construct test statistic q_μ from likelihoods
  - Obtain expected distributions of q_μ for various μ values
  - Determine discovery p_0 and signal exclusion limit
- **Step-3:** Repeat for each model (assumed value m_H)

HistFitter

- adds steps-0 and 3
- allows full analysis chain from simple configuration file

Links

- RooFit overview (2004): http://www.nikhef.nl/~verkerke/talks/chep03/chep2003_v4.pdf
- ATLAS Statistics Forum page on Stat. Tools: https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/StatisticsTools
- RooFit/RooStats at ACAT 2014: https://indico.cern.ch/event/258092/session/0/contribution/140/material/slides/1.pdf
- Higgs Combination procedure/explanation of CLs observed/expected and error bands: http://cds.cern.ch/record/1375842
- HistFactory documentation:
  https://cdsweb.cern.ch/record/1456844/
  https://twiki.cern.ch/twiki/bin/view/RooStats/HistFactory

[23] K. Cranmer, G. Lewis, L. Moneta, A. Shibata, W. Verkerke, HistFactory: A tool for creating statistical models for use with RooFit and RooStats, CERN-OPEN-2012-016.

[24] L. Moneta, K. Belasco, K. S. Cranmer, S. Kreiss, A. Lazzaro, et al., The RooStats Project, PoS ACAT2010 (2010) 057. arXiv:1009.1003.

[25] W. Verkerke, D. P. Kirkby, The RooFit toolkit for data modeling, eConf C0303241 (2003) MOLT007. arXiv:physics/0306116.

[26] R. Brun, F. Rademakers, ROOT: An object oriented data analysis framework, Nucl.Instrum.Meth. A389 (1997) 81–86. doi:10.1016/S0168-9002(97)00048-X.

[27] I. Antcheva, M. Ballintijn, B. Bellenot, M. Biskup, R. Brun, et al., ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization, Comput.Phys.Commun. 182 (2011) 1384–1385. doi:10.1016/j.cpc.2011.02.008.

Backup