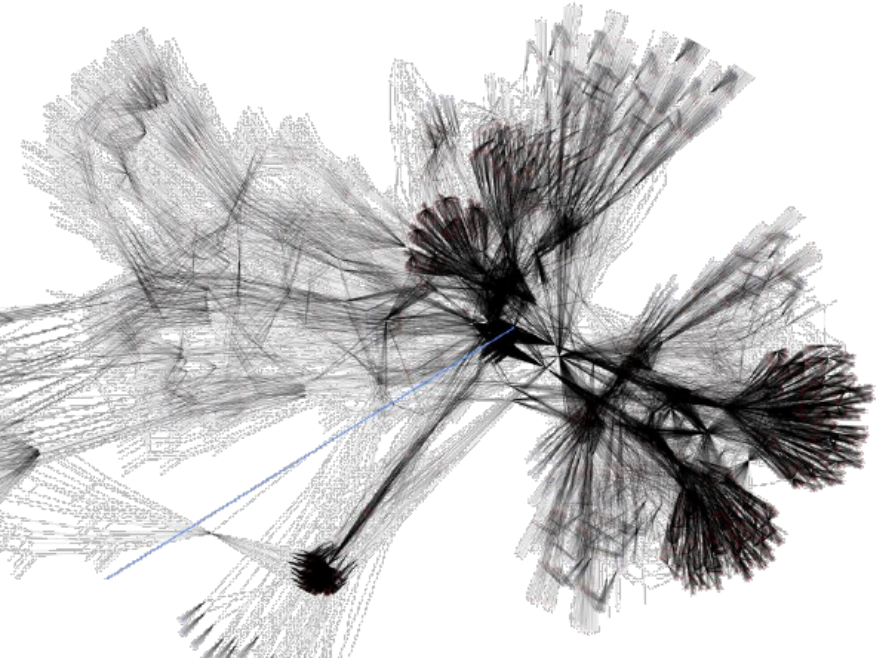


Introduction to statistics for high-energy physics

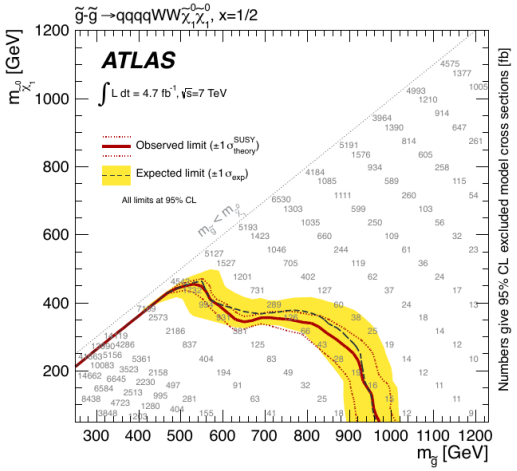
Jeanette Lorenz (LMU Munich)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

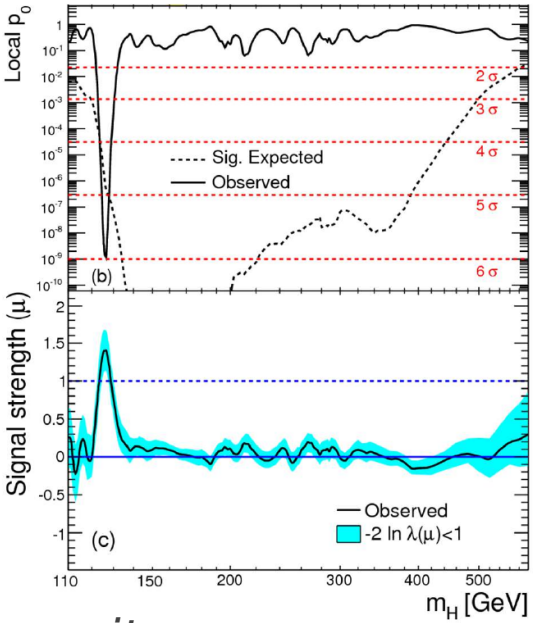
28.2.2019

Basic questions

- Physics questions we want to answer…
  - Is the Higgs boson a SM Higgs boson?
  - What is its production cross section and couplings? → Measurement
  - Is there physics beyond the Standard Model?



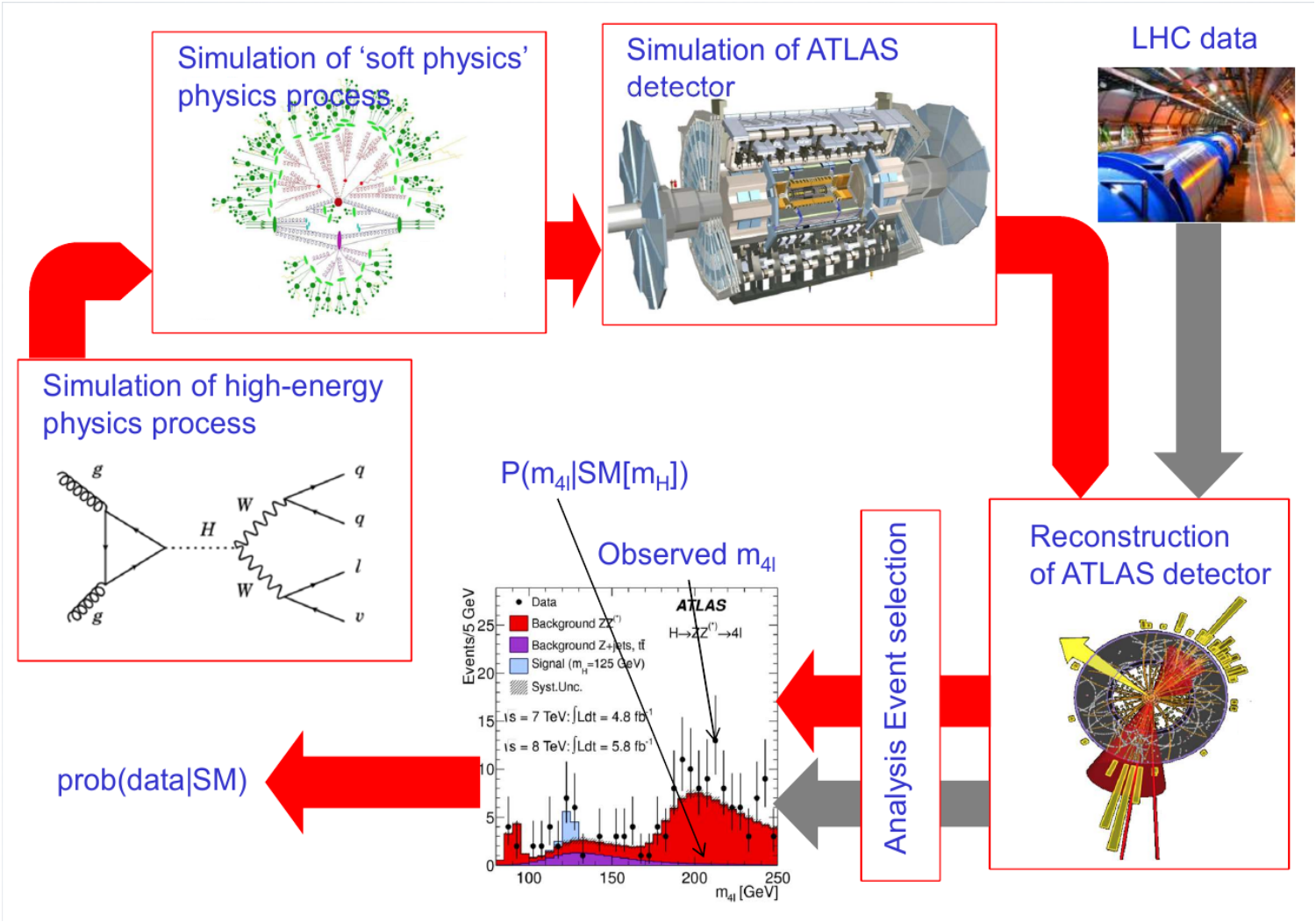
- Enormous efforts in many channels, millions of plots with signal/backgrounds expectations, with systematics and observed data



- How do you conclude on these questions?

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

Workflow in high-energy analyses



W. Verkerke

Example 1: cross-section measurement

[Phys. Lett. B 759 (2016) 601]

e.g. *Measurement of $W^{\pm}$ and Z-boson production cross sections in pp collisions at √s = 13 TeV with the ATLAS detector*

	Z
Signal events	$34865 \pm 187 \pm 7 \pm 3$
Correction C	$0.552^{+0.006}_{-0.005}$
σ^{fid} [nb]	$0.781 \pm 0.004 \pm 0.008 \pm 0.016$
Acceptance A	0.393 ± 0.007
σ^{tot} [nb]	$1.987 \pm 0.011 \pm 0.041 \pm 0.042$

Fiducial cross section calculated by

Signal events

$$\sigma_{fid} = \frac{N_{data} - N_{bkg}}{C_{fid} \cdot L}$$

$$0.781 \pm 0.004 \pm 0.008 \pm 0.016$$

=> Just counting events, use Gaussian uncertainties (see later), using simple error propagation
=> **simplest case**

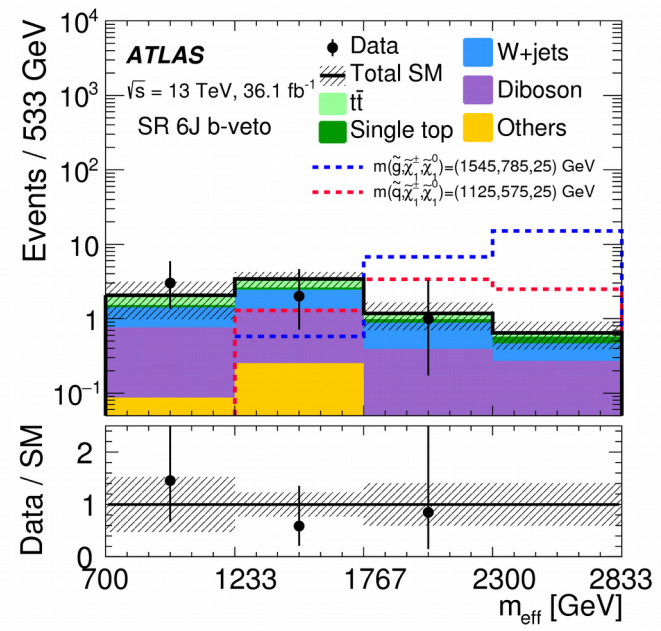
Example 2: multi-bin SUSY search

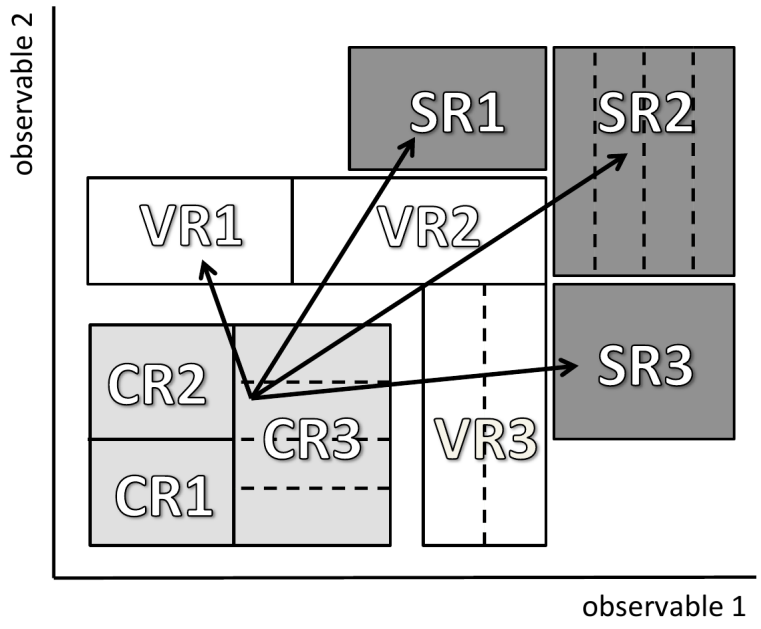
[Phys. Rev. D 96 (2017) 112010]

Most of the SUSY analyses use control regions to constrain backgrounds, some also binned signal regions.

E.g. *Search for squarks and gluinos in events with an isolated lepton, jets and missing transverse momentum at √s = 13 TeV with the ATLAS detector*

Statistical combination of in total 28 bins (signal regions) + 28 bins (control regions), counting experiment in each bin
→ better separation of signal and background by combining information from multiple bins





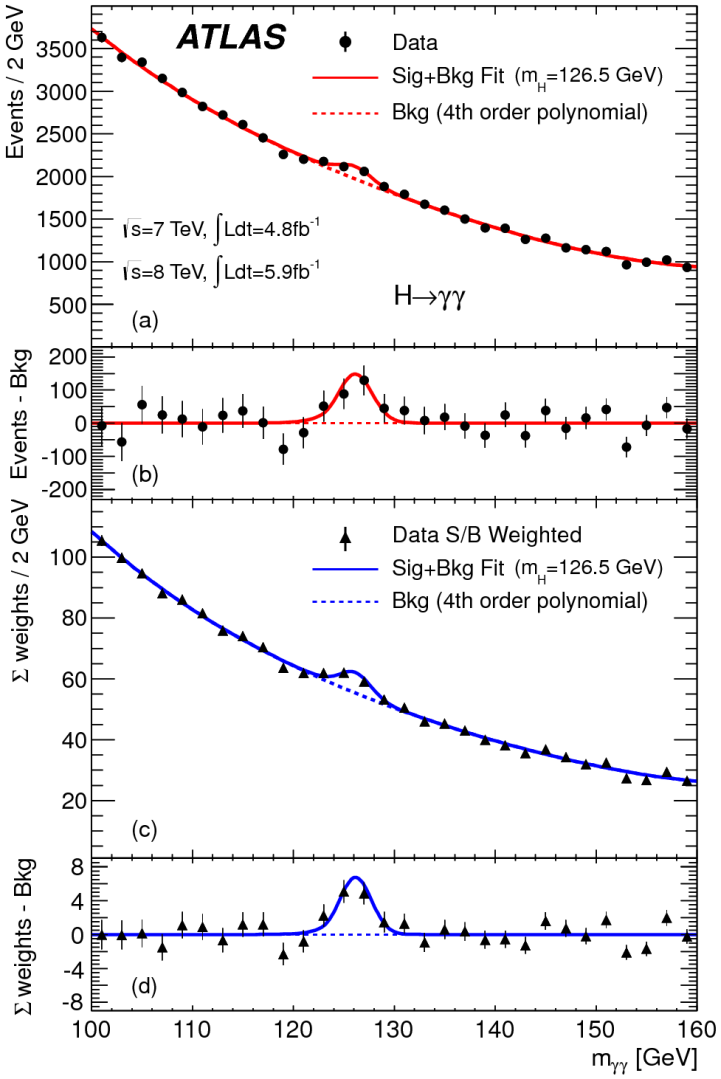
Example 3: Search with continuous background

[Phys. Lett. B 716 (2012) 1-29]

Example: *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*

Background modeled in this case by a fourth-order Bernstein polynomial, so the background is given as a unbinned function.

→ **Unbinned shape analysis**



Outline

- Probability and likelihood
- Discovery case
- Exclusion case
- Statistical tools for high-energy physics

Probability

…following G. Cowan (Statistical Data Analysis), and a bit simplified

Consider a set *S* called the **sample space** consisting of a certain number of elements. To each subset *A* of *S* one assigns a real number *P(A)* called a **probability**, defined by the following three axioms:

(1) For every subset *A* in *S*, *P(A) >= 0*.

(2) For any two subsets *A* and *B* that are disjoint, the probability assigned to the union of *A* and *B* is the sum of the two corresponding probabilities, *P(A U B) = P(A) + P(B)*.

(3) The probability assigned to the sample space is one, *P(S) = 1*.

Interpretation of probability

[again following definitions in G. Cowan 'Statistical Data Analysis']

Following the definition of probability two interpretations are commonly used:

1. Probability as **relative frequency (classical statistics)**

→ *Very common interpretation in data analysis.*
→ *Elements of set S corresponds to possible outcome of a measurement.*
→ *Subset A of S corresponds to the occurrence of any of the outcomes in the subset and the subset is called an event.*

→ $P(A) = \lim_{n \to \infty} \frac{\text{number of occurances of outcome A in n measurments}}{n}$

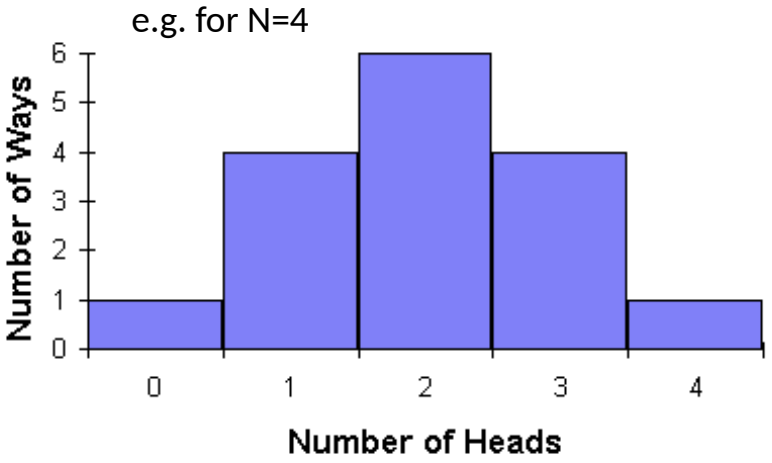
2. **Subjective probability (Bayesian)**

→ *Elements of sample space correspond to hypotheses or propositions, i.e. statements that are either true or false.*
→ *Interpretation as measure of belief:*
P(A) = degree of belief that hypothesis A is true
→ *closely related to Bayes' theorem, e.g. interpretation in particle physics context:*
$P(\text{theory}|\text{data}) \sim P(\text{data}|\text{theory}) \cdot P(\text{theory})$

Example: binominal distributions

Consider throwing a coin – possible outcomes: number or back

e.g. for N=4



Probability to obtain number: p = 0.5
Probability to obtain back: (1-p) = 0.5

Probability to obtain to get *n* times the number in *N* events:

$$f(n;N,p)=\frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}$$

Mean:

$$E[x]=\int_{-\infty}^{\infty} x f(x)dx=\mu$$

In this case:

$$E[n]=\sum_{n=0}^{\infty} n\frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}=Np$$

Variance:

Called **standard deviation**

$$E[(x-E[x])^2]=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx=\sigma^2=V[x]$$

In this case:

V[n] = Np(1-p)

Poisson distribution

For N → ∞ and p very small, but Np = ν equal, where ν is some finite value: The binomial distributions become Poisson distributions in this limit (n is an integer random variable):

$$f(n;\nu)=\frac{\nu^n}{n!}e^{-\nu}$$

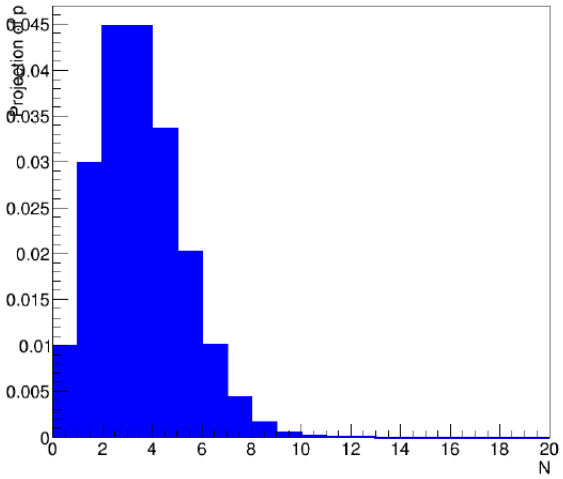
With expectation value:

$$E[n]=\sum_{n=0}^{\infty} n\frac{\nu^n}{n!}e^{-\nu}=\nu$$

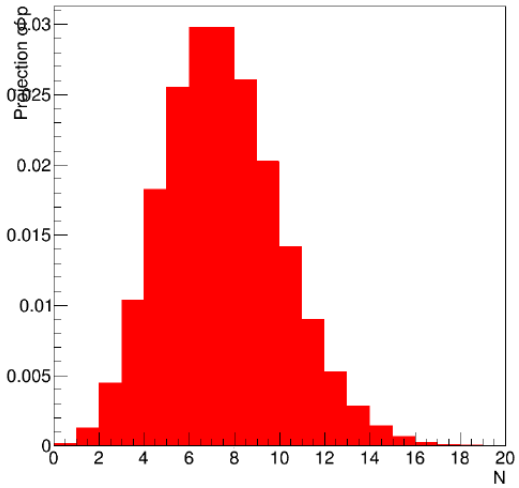
And variance:

$$V[n]=\sum_{n=0}^{\infty}(n-\nu)^2\frac{\nu^n}{n!}e^{-\nu}=\nu$$

μ=3 (“bkg only”)



μ=7 (“bkg+signal”)



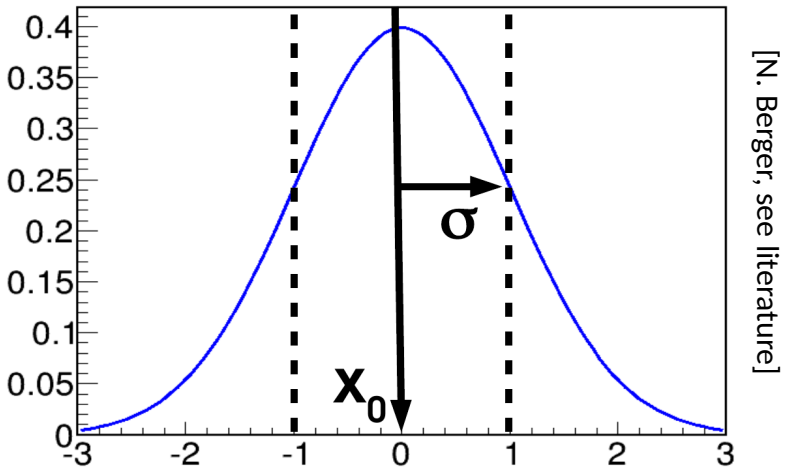
- Corresponds to the usual distribution of events in particle physics.
- Can be treated as a continuous variable x as long as integrated over a range Δx which is large compared to unity.
- For large mean values ν the Poisson variable behaves like a continuous variable following a Gauss distribution → of practical importance

Gaussian distribution

→ another important distribution, in particular for cases of a large number of events

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

With mean μ and variance σ^2



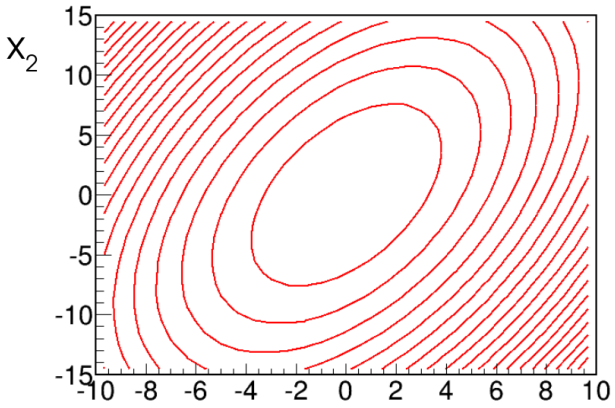
[N. Berger, see literature]

Can be generalized to N dimensions:

$$f(x;\mu,C)=\frac{1}{(2\pi|C|)^{\frac{N}{2}}}e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}$$

With the mean μ and the covariance matrix C, e.g. for two dimensions:

$$C=\begin{bmatrix}\mathrm{Var}(x_1) & \mathrm{Cov}(x_1,x_2)\\ \mathrm{Cov}(x_2,x_1) & \mathrm{Var}(x_2)\end{bmatrix}=\begin{bmatrix}\sigma_{x_1}^2 & \gamma\sigma_{x_1}\sigma_{x_2}\\ \gamma\sigma_{x_1}\sigma_{x_2} & \sigma_{x_2}^2\end{bmatrix}$$



[N. Berger, see literature]

Gaussian quantile

A Gaussian distribution can be transformed into a **standard Gaussian** with μ =0 and σ = 1 by the transformation

Called pull

$$z = \frac{x - \mu}{\sigma}$$

The p.d.f of the standard Gaussian is:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

And the cumulative distribution:

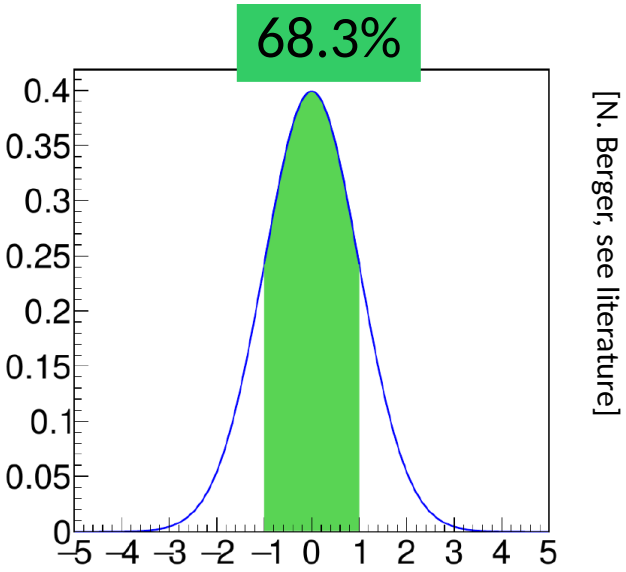
$$\Phi(z) = \int_{-\infty}^{z} \phi(z')\, dz'$$

The cumulative distributions of the original Gaussian F(x) and the standard Gaussian Φ(z) are related by:

$$F(x) = \Phi(z)$$

→ z is often used to express the deviation of a measurement from the mean – we come back to this later when talking about p-values

z	P(|x-μ|>zσ)
1	0.317
2	0.045
3	0.003
5	6 x 10-7



Central limit theorem

Why are Gaussian distributions so important?

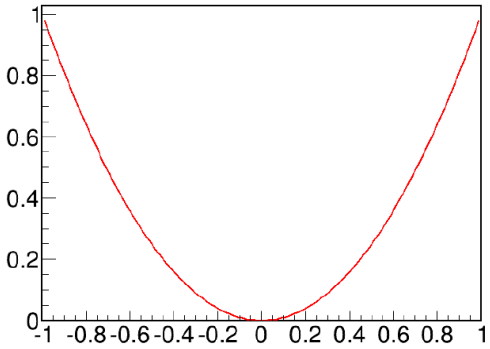
→ **central limit theorem**:

- Sum of n independent continuous random variables x_i with means μ_i and variances σ_i^2 becomes in the limit
  $n \to \infty$ a Gaussian random variable
- Mean: $\mu = \sum_{i=1}^{n} \mu_i$
- Variance: $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$

This means practically, regardless of the original distribution the average of the mean for many measurements is Gaussian → **justification why the statistical error behaves like √N if N is large!**

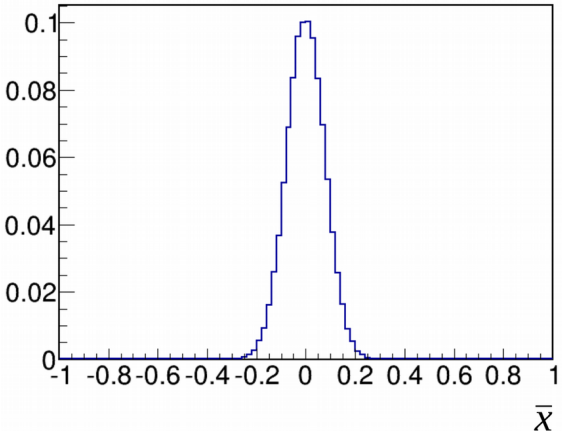
Nice example by N. Berger

Draw random events from a chi2 distribution



n = 100

Repeating many times, the mean will be Gaussian distributed

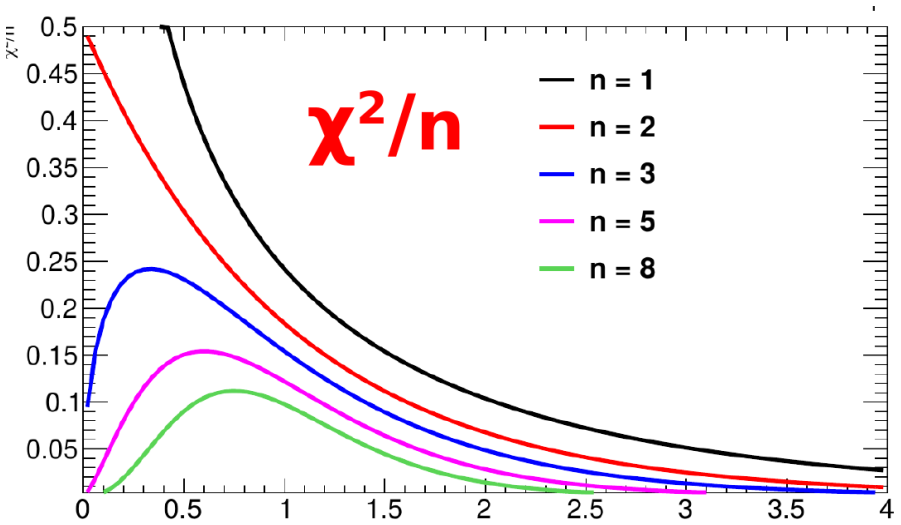


Chi2 test

Considering the sum of N independent squared Gaussian distributed variables:

$$z = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

This is distributed according to a χ^2 distribution for N degrees of freedom.



[N. Berger, see literature]

Important application: Pearson's chi2 test

→ *Typically used to quantify the agreement of two histograms, e.g. observed and expected*

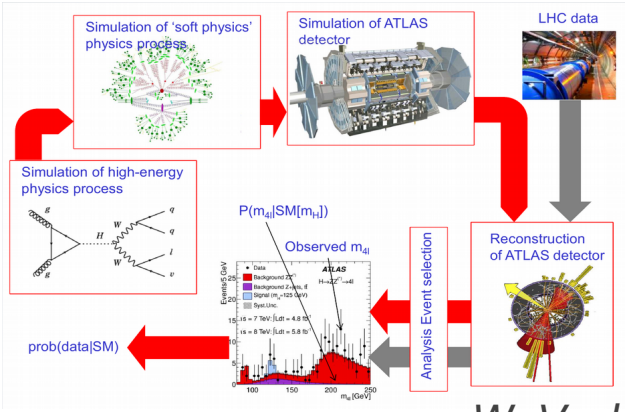
E.g. take a binned histogram of the variable x with observed values $n_1, n_2, \ldots, n_N$ in the bins and if these are Poisson distributed with means $\nu_1, \nu_2, \ldots, \nu_N$

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

And this follows a χ^2 distribution for N degrees of freedom.
$X^2/N <= 1$ is usually considered a good agreement (although there are caveats, see *Cowan*).

General PDF for searches

Learned now which distributions describe a few specific cases (there are of course more cases and more distributions).



W. Verkerke

The goal for every HEP analysis is to build a statistical model that describes the expectation and then to compare this to data.

This statistical model will include two things:

- The **randomness** of data (thus we work with p.d.f.s),
- The **model assumptions** we have (e.g. the knowledge or expectation on how a certain physics process will be distributed or a certain systematic uncertainty is included).

Often the collection of data is described by a Poisson distribution for observing n events (this is the simplest case, we consider more complicated cases in the following slides):

$$f(n;\nu)=\frac{\nu^n}{n!}e^{-\nu}$$

ν is typically composed of a background expectation B and a signal expectation S:

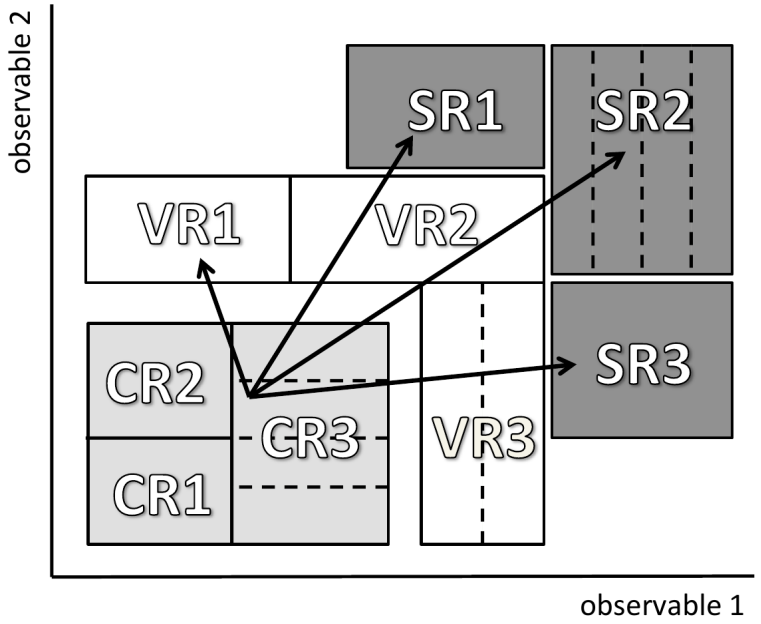
$$f(n;B+\mu S)=\frac{(B+\mu S)^n}{n!}e^{-(B+\mu S)}$$

μ is the signal strength and a priori unknown → parameter of interest (POI)

Binned case

Often analyses consider different search regions:

- **Signal region:** signal-rich region (SR)
- **Control region:** background-rich region (CR), fit simulated backgrounds to data
- **Validation region:** validation of extrapolation (VR)



In this case we have a Poisson distribution for every of these regions or bins (well, at least for CRs and SRs), and the total p.d.f. becomes:

$$f\left(\{n_i\}; B+\mu S\right)=\prod_i^N \frac{\left(B\cdot f_{B,i}+\mu S\cdot f_{S,i}\right)^{n_i}}{n_i!} e^{-\left(B\cdot f_{B,i}+\mu S\cdot f_{S,i}\right)}$$

Where $f_{B,i}$ and $f_{S,i}$ indicate the fractional contribution of the background and the signal in every bin.

Unbinned case

[Phys. Lett. B 716 (2012) 1-29]

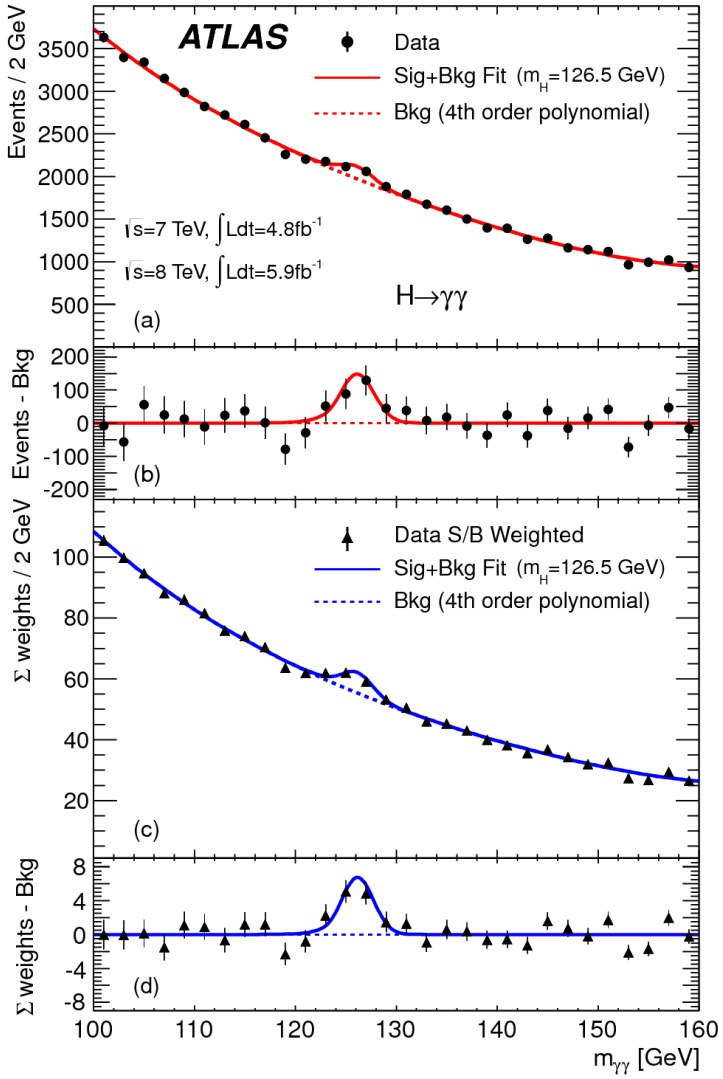
Instead of counting events in every bin and compare this to background and signal predictions in every bin, can also also model background and signal as continuous functions.

The p.d.f. will then be something like e.g.

$$f(m_i; S, B) = \frac{e^{-(S+B)}}{n_{\text{evts}}!} \prod_{i=1}^{n_{\text{evts}}} S P_{\text{sig}}(m_i) + B P_{\text{bkg}}(m_i)$$

Probability to find n_{evts} events

Where m_i runs from 1,..., n_{evts}



Construction of a likelihood

So far we looked at the p.d.f., so f(n; parameters), the probability that a certain data outcome is realized assuming a certain model.

But in HEP we usually have the opposite situation: we have the data recorded and want to know the model describing the data.

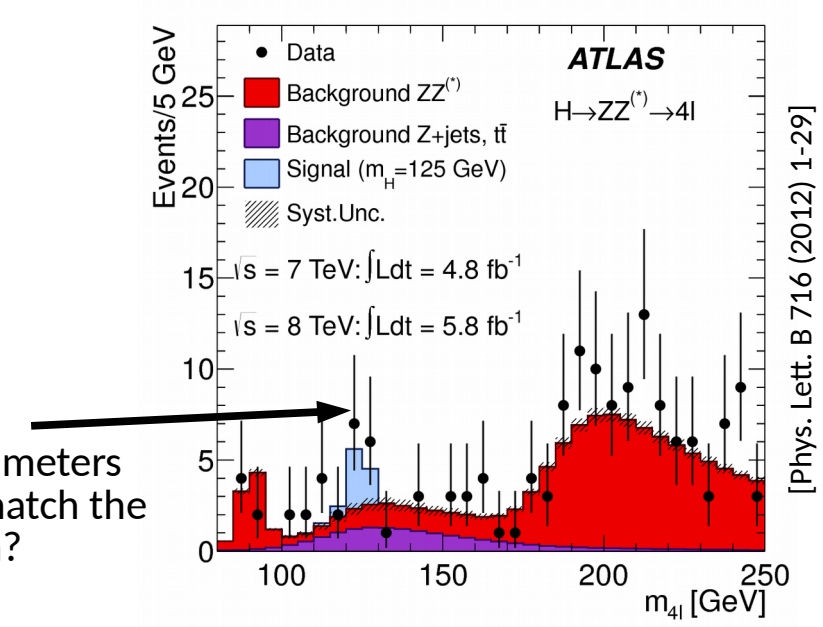
→ *The parameters are the unknowns.*

So instead of using the p.d.f we use the likelihood:

f(n; parameters) = L(parameters)

Data is unknown, parameters are known

Parameters are unknown, data is known



Inclusion of systematic uncertainties

Any model and measurement has uncertainties – statistical, systematic uncertainties and theoretical uncertainties

→ need to include these into the p.d.f. or likelihood
→ p.d.f. for data yields now also depend on systematic uncertainties + additional ***constraint terms*** *for systematic uncertainties from* ***auxiliary measurements***

E.g. for a binned likelihood (see also histfactory documentation):

Includes POIs

Nuisance parameters

$$P(n_{cb}, a_p|\phi_p, \alpha_p, \gamma_b) = \underbrace{\left(\prod_{c\in\text{channels}} \prod_{b\in\text{bins}} \text{Pois}(n_{cb}|\nu_{cb})\right)}_{\text{Poisson terms}} \cdot \underbrace{G(L_0|\lambda, \Delta_L) \cdot \prod_{p\in S+\Gamma} f_b(a_p|\alpha_p)}_{\text{Constraint terms}} \quad (10.2.1)$$

with:

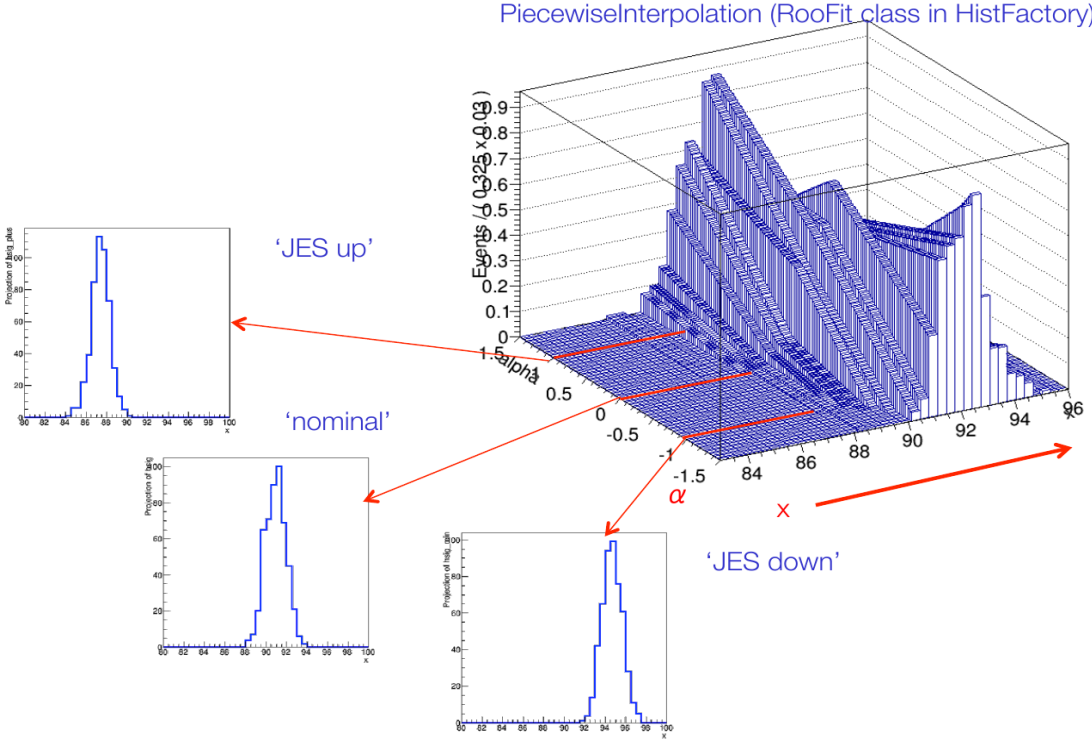
- $c \in$ channels. The channels comprise all regions appearing in the analysis: control, validation and signal regions.
- $b \in$ bins of the histograms used in the construction of the pdf.
- $p \in$ parameters (comprising normalisation parameters and nuisance parameters related to systematic uncertainties, see below).
- $S = \{\alpha_p\}$: the set of all parameters associated to systematic uncertainties with external constraint.
- $\Gamma = \{\gamma_{csb}\}$: the set of all bin-by-bin uncertainties with constraint (detailed below).

Modelling of systematic uncertainties

- 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 (usually obtained by external measurements)
- Interpolation, performed between **+1σ ↔ nominal ↔ -1σ** taken into the model as nuisance parameter

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

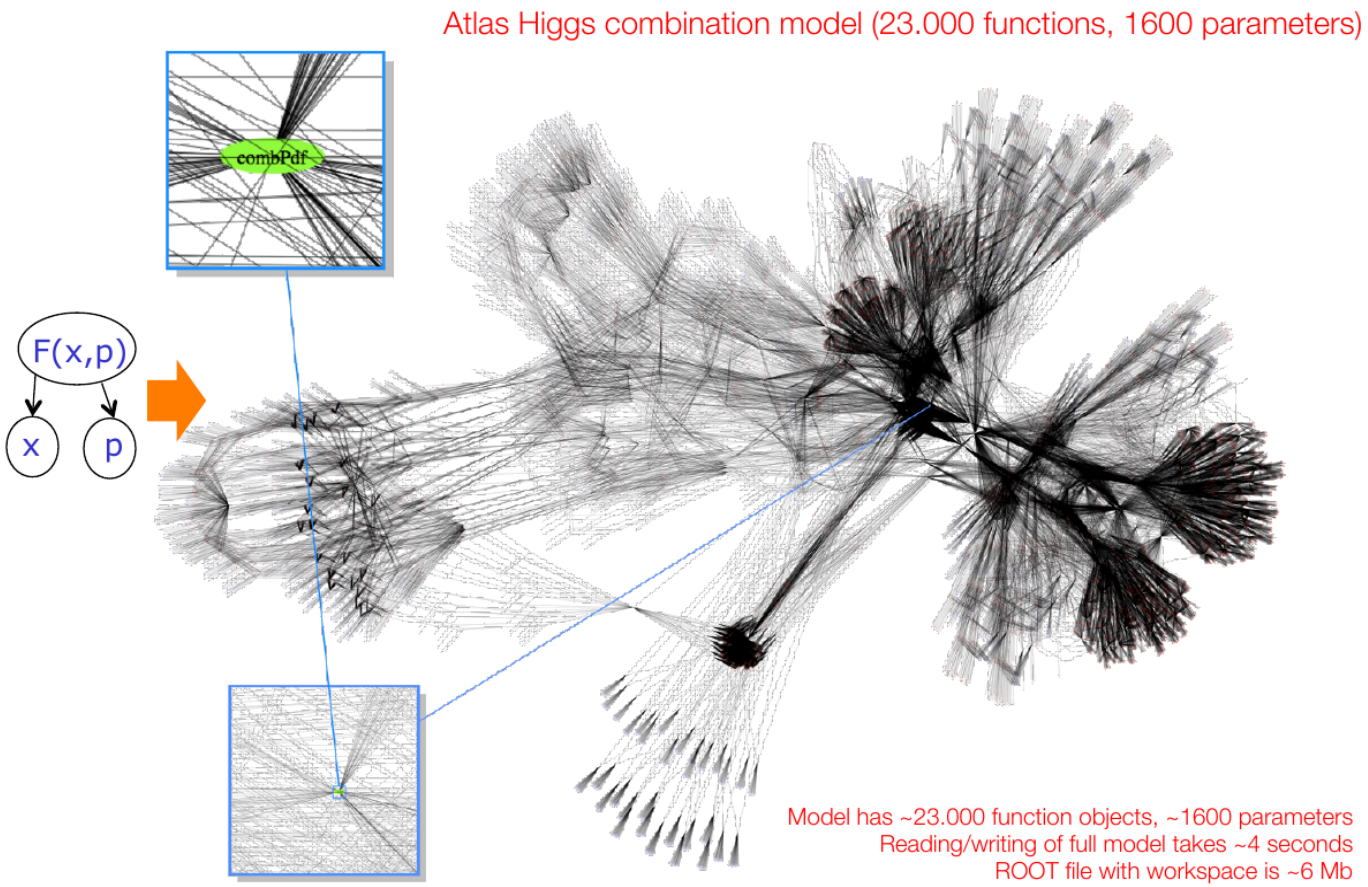
PiecewiseInterpolation (RooFit class in HistFactory)



Example for a complex model

[W. Verkerke at SOS 2014]

The full ATLAS Higgs combination in a single workspace…



Maximum likelihood estimation

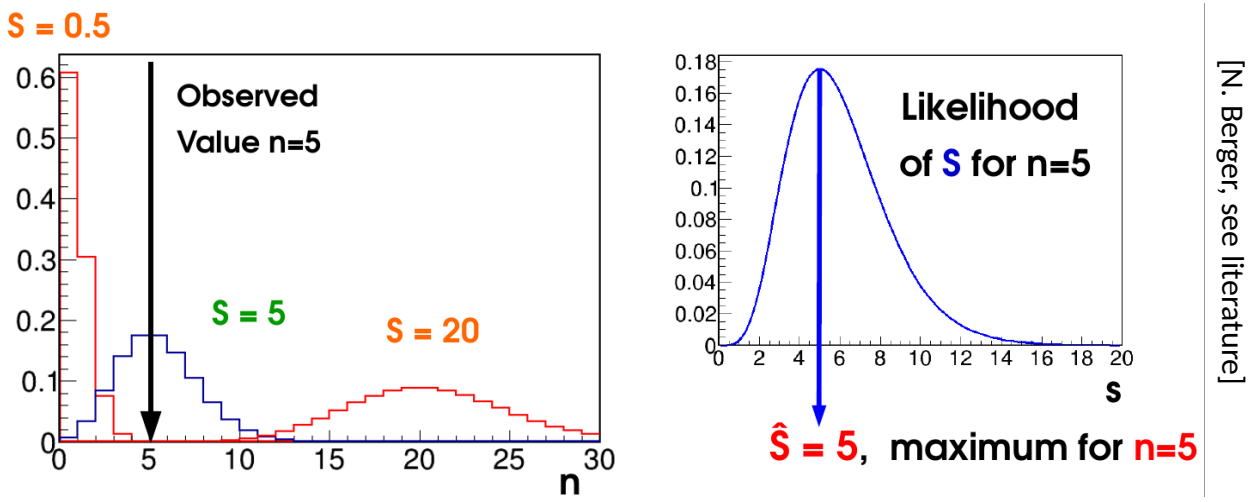
What are the values for the parameters in the likelihood?

→ *Need to estimate them using data which fluctuate.*
→ ***Maximum likelihood estimator (MLE)***

Consider the likelihood L(μ) – find the value μ which **maximizes** L(μ) for the **given data** (which is thus the most likely value for μ given this particular data).

→ maximum likelihood estimator: $\hat{\mu} = \max \mu \text{ for which } \frac{\partial L}{\partial \mu} = 0$

^μ depends on data! - Thus itself an observable and not the true value, but rather 'best guess'.



Example for a binned analysis

Take the likelihood for a binned analysis (without systematic uncertainties):

$$L(\mu S; n_i) = f(n_i; \mu S) = \prod_{i=1}^{N} \mathrm{Pois}(n_i; \mu S f_{S,i} + B f_{B,i})$$

To maximize the likelihood it is usually easier to look at **– log (L)** and to **minimize** this:

$$-2 \log L(\mu S; n_i) = -2 \sum_{i=1}^{N} \log \mathrm{Pois}(n_i; \mu S f_{S,i} + B f_{B,i})$$

Advantage that derivation of a sum, not of a product, also omission of constant terms.

For the special case of Gaussian distributions (instead of Poisson):

$$\lambda_{\mathrm{Gaus}} = \sum_{i=1}^{N} -2 \log \mathrm{G}(n_i; \mu S f_{S,i} + B f_{B,i}, \sigma_i) = \sum_{i=1}^{N} \left(\frac{n_i - (\mu S f_{S,i} + B f_{B,i})}{\sigma_i} \right)^2$$

→ chi2 formula!

Few properties of the MLE:

- For n → ∞ ^μ converges against the real value of μ.
- For large n, ^μ is asymptotically Gaussian distributed.

Profiling of nuisance parameters

[e.g. ATLAS ttH(bb) analysis - Phys. Rev. D 97 (2018) 072016]

Likelihood contains nuisance parameters next to the POIs that are constrained by additional constraint terms:

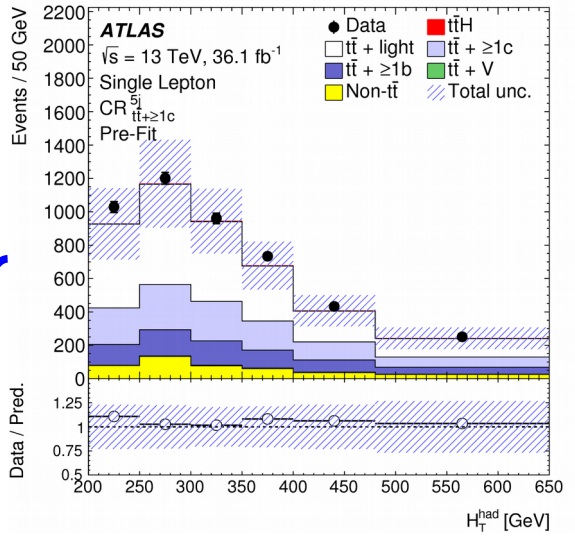
$$L(\mu S, \theta) = \prod_{i=1}^{N} \mathrm{Pois}(n_i; \mu S f_{S,i} + B f_{B,i}) \cdot \prod_{j=1}^{N'} \mathrm{G}(\theta_{\mathrm{obs}}, \theta)$$

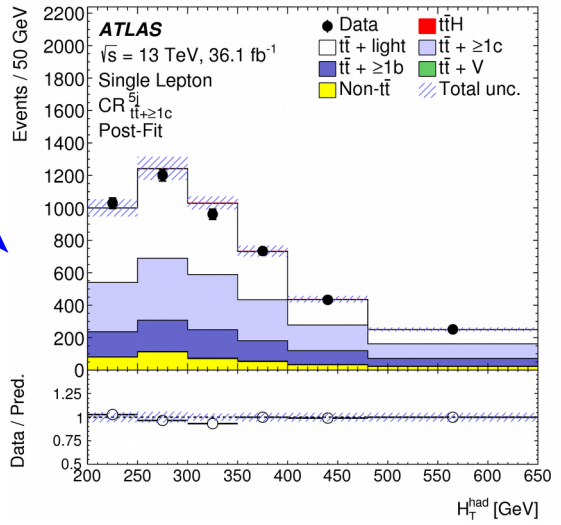
→ *Not only MLE estimate for POIs needed, but also for nuisance parameters θ → ^θ.*

→ best fit values.

→ As the nuisance parameters will propagate as uncertainties to the final results of the analysis, care is needed both in the way the constraint terms are parametrized as also the fit result for nuisance parameters needs to be understood (see next slide).

Reduction of uncertainties in the fit





Pull plots

[e.g. ATLAS ttH(bb) analysis - Phys. Rev. D 97 (2018) 072016]

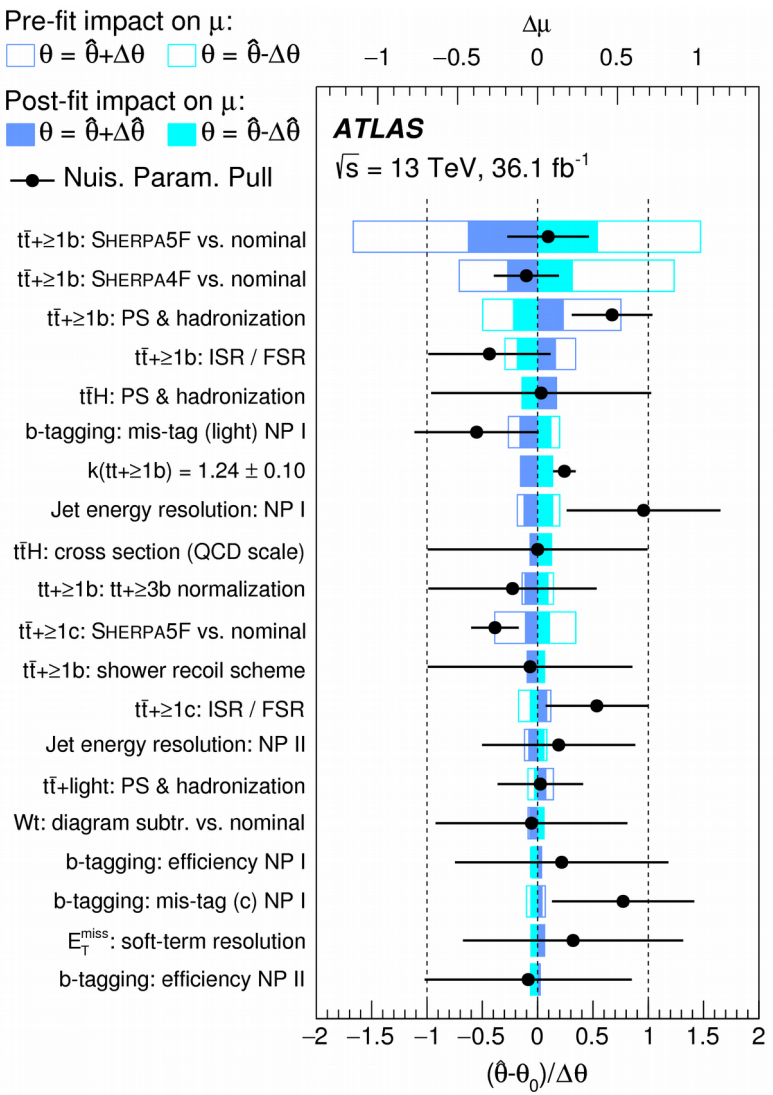
Although systematic uncertainties might have any values, in the likelihood NP typically parametrized as **standard Gaussian →** a systematic uncertainty has a central value and an uncertainty.

Meaning of values for nuisance parameters:

- **NP has central value = 0:** corresponds to unmodified systematic uncertainties, e.g. as derived from MC.
- **NP has an uncertainty of 1 (σ):** the original uncertainty on the NP, as put into the constraint terms.

Different results after fit possible:

- **Central value of NP ≠ 0:** some discrepancy between data and statistical model was absorbed → needs to be understood
- **Uncertainty on NP < 1 (profiling):** the systematic uncertainty was constrained (was reduced) by the data → needs to be understood (opposite direction with uncertainty on NP >1 also possible…)

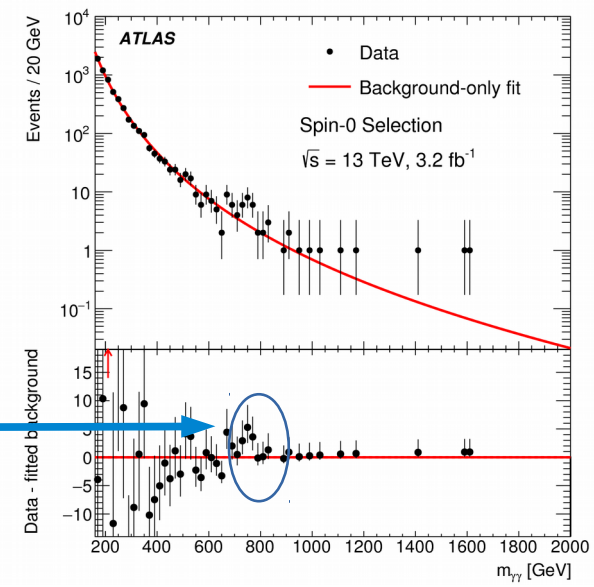


Hypothesis testing

Statistical model built → want to check agreement with data:

- Is there are an (unknown) signal in the data?
  → **discovery p-values**
- Is this model of new physics excluded?
  → **exclusion limits**
- In case of measurements: which is the allowed range for this POI?
  → **confidence intervals**

New particle?



[JHEP 09 (2016) 001]

Overview on different hypothesis tests

Hypothesis: probability that data gets realized for certain model parameters

→ *usually we consider two hypotheses: one that the background model corresponds to data (e.g. H_0), and the other that the background+signal model (e.g. H_1) corresponds to data*

Test: specify a **critical region W**, so that there is only a small probability α that assuming a hypothesis H_0 the data falls into the critical region W: $P(x \text{ in } W | H_0) < \alpha$

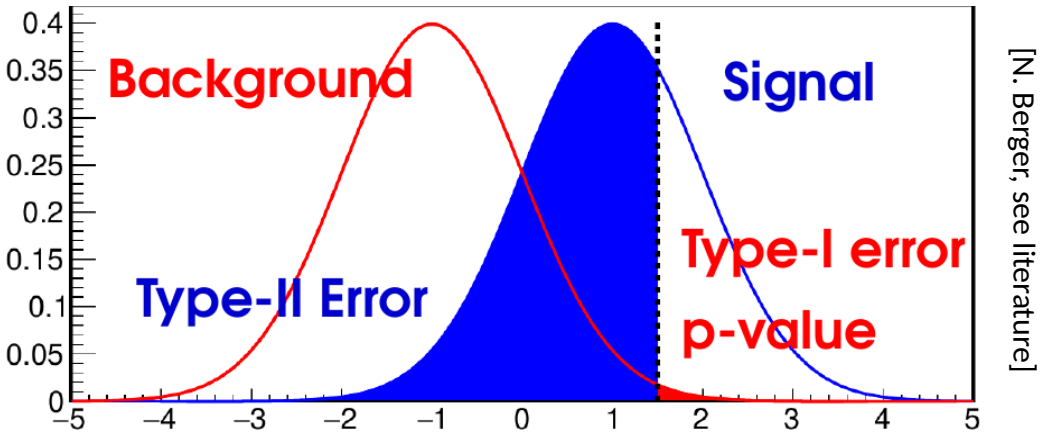
→ *α is the **significance level** of the test*

→ *α needs to be chosen such that the possible errors are minimized:*

Type I error: *H_0 is true, but gets rejected (false discovery claim)*

Type II error: *H_1 is true, but gets rejected (missed discovery!)*

Try to minimize type II error for given level of type I error



Different hypotheses in practice

- **Discovery:** is the data compatible with the background?

  → *try to reject H_0: background model*
  → *p-values and significances*

- **Exclusion:** In case of no excess observed – which signal can be rejected?

  → *H_0 is signal+background, H_1 is only background*
  → *Try to reject H_0*
  → *Upper limits and exclusion limits*

- **Parameter measurements:**

  → *H_0 corresponds to certain values of the parameter μ*
  → *Which parameter values of μ are not rejected at 68% CL level?*
  → *1σ confidence interval for μ*

Trying to reject H_0 in all cases, so be careful, definition of H_0 changes!

Neyman-Pearson lemma

Test statistics:

- Hypothesis depends in general on multiple parameters, e.g. x_1, x_2,..., x_N
- Can map these parameters on a scalar by a function $t(x_1, x_2,..., x_N) = t_{cut}$ → test statistics.
- Transform now all p.d.f.s to be now functions of t, i.e. $f(t;H_0)$ → the distributions are now 1-dimensional.
- The boundary t_{cut} encloses a **critical region** to reject the hypothesis.

Neyman-Pearson lemma:

- Optimal choice of critical region?
- Given by: $t(\boldsymbol{x}) = \frac{P(\boldsymbol{x}|H_1)}{P(\boldsymbol{x}|H_0)} > c$
- Means we always need two hypotheses H_0 and H_1

Discovery test statistics

Is the data consistent with the background-model?

- Two hypotheses: μS=0 and μS≠0
- However, μS≠0 not just one parameter value.

→ Define as test statistics for discovery:

$$t(\mu) = -2\ln\lambda(\mu) = -2\ln\left(\frac{L(\mu S)}{L(\hat{\mu} S)}\right) = -2\ln\left(\frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}\right)$$

And more precisely, as we do not want to reject the background-only hypotheses if we have an underfluctuation in data:

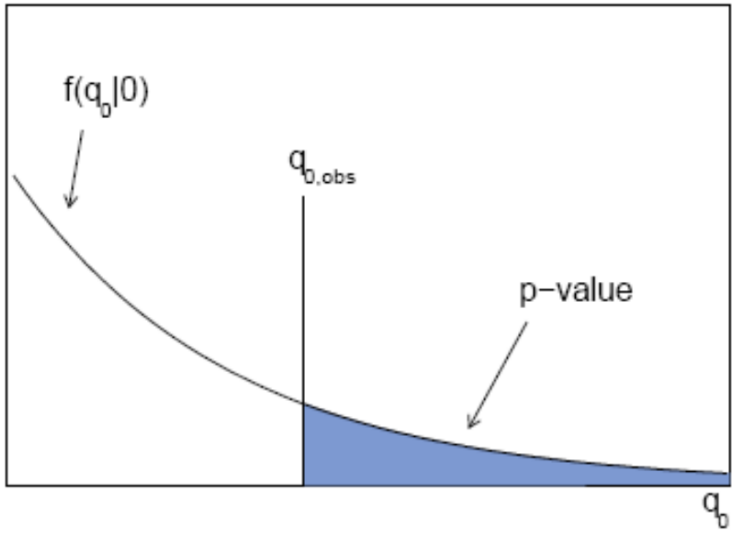
$$q_0 = \begin{cases} t(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

Discovery p-value

Large values of q_0 means that L(^S) and L(S=0) are more and more different/incompatible

→ Define **p-value**:

Probability to obtain observed data, or more extreme, given the hypothesis in future repeated identical experiments



[G. Cowan, see literature]

$$p_0 = \int_{-q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\,dq_0$$

Discovery significance

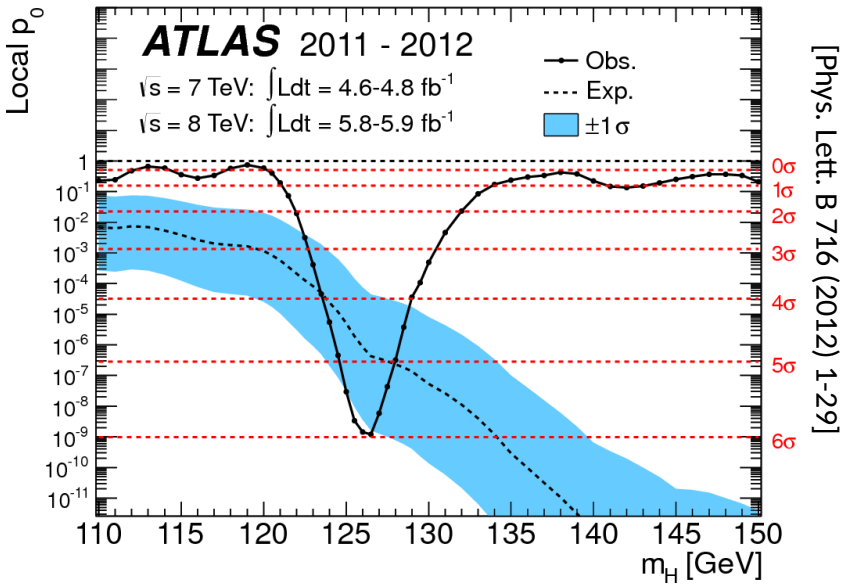
Can transform the p-value in significances – useful as interesting p-values small.

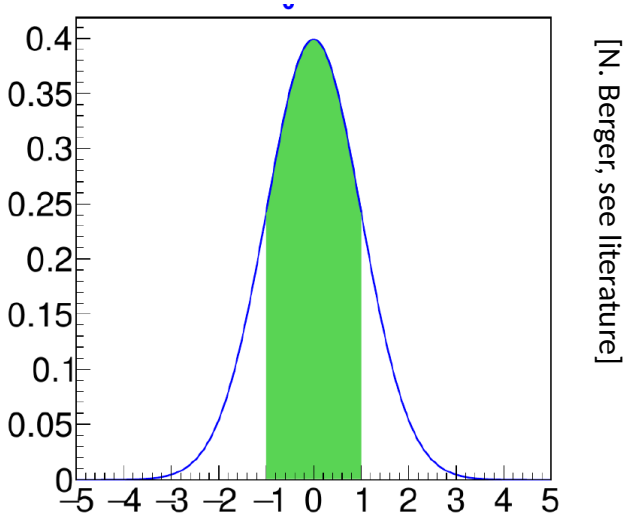
Reminder: Gaussian quantiles

Idea: how many standard deviations σ of a standard Gaussian corresponds the p-value to?

$$\rightarrow \quad Z = \Phi^{-1}(1-p)$$

z	P(|x-μ|>zσ)
1	0.317
2	0.045
3	0.003
5	6 x 10-7





Asymptotic approximation and useful hand formulas

- In certain situations, i.e. in case of a large data statistics, can use the **Wald approximation** (see details on lectures of G. Cowan).

  → *In these cases* $Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$

- For large N, if the process is Gaussian, the significance also takes a simple form:

  → $Z = \sqrt{q_0} = \frac{\hat{\mu} S}{\sqrt{B}}$

- In case of Poisson distributions a bit more complex formula:

  → $Z = \sqrt{2\left((\hat{\mu} S + B)\log\left(1 + \frac{\hat{\mu} S}{B}\right) - \hat{\mu} S\right)}$

- Also note that in case of calculating a discovery sensitivity (so expected), one needs $f(q_0|\mu')$ and not $f(q_0|0)$.

Exclusion test statistics

Is it possible to exclude this signal model with the measured data? What is the upper limit?

- H_0: $\mu S = S_0$ (so a specific signal model), H_1: $\mu S < S_0$

→ Use as test statistics:

$$q_\mu = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

In particular an upward fluctuation should not result in an exclusion of the signal model.

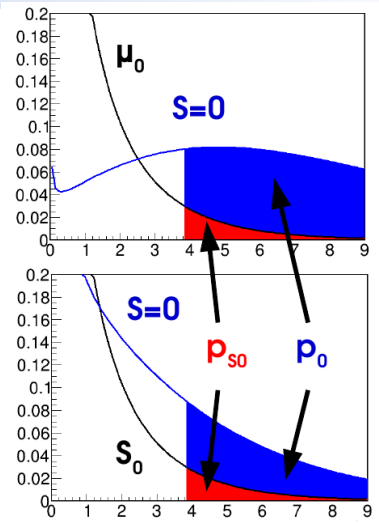
→ Usually we know however, that $\mu S >= 0$, and do not want to allow for negative S. Modified test statistics (similar to q_μ):

$$\tilde{\lambda}(\mu) = \begin{cases} \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0, \\ \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(0, \hat{\boldsymbol{\theta}})} & \hat{\mu} < 0 \, . \end{cases} \qquad \tilde{q}_\mu = \begin{cases} -2 \ln \tilde{\lambda}(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

CLs method

This definition can get problematic if background (B) and background+signal (μS+B) models very similar → in this case can get exclusions of a signal where one would not think to be sensitive to.

Example: Nobs = 2 → $p_{S+B}(\mu S=0) = 0.04$
μS≥0 excluded at 95% C.L. ?

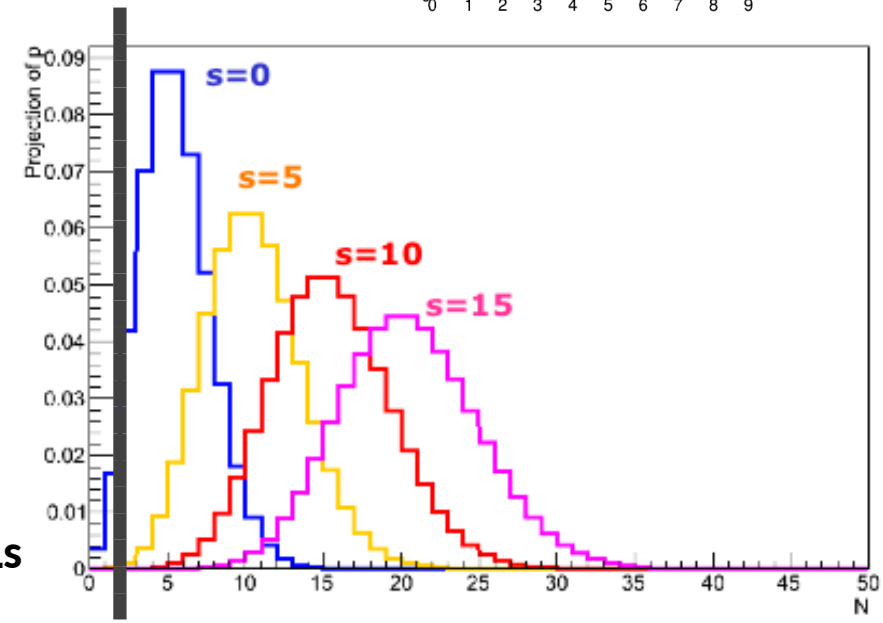


[N. Berger, see literature]

Modified approach to protect against such inference on signal – a little bit ad-hoc, but working nicely:

Instead of requiring $\mathbf{p_{S+B}}$ **<=** 5%, require

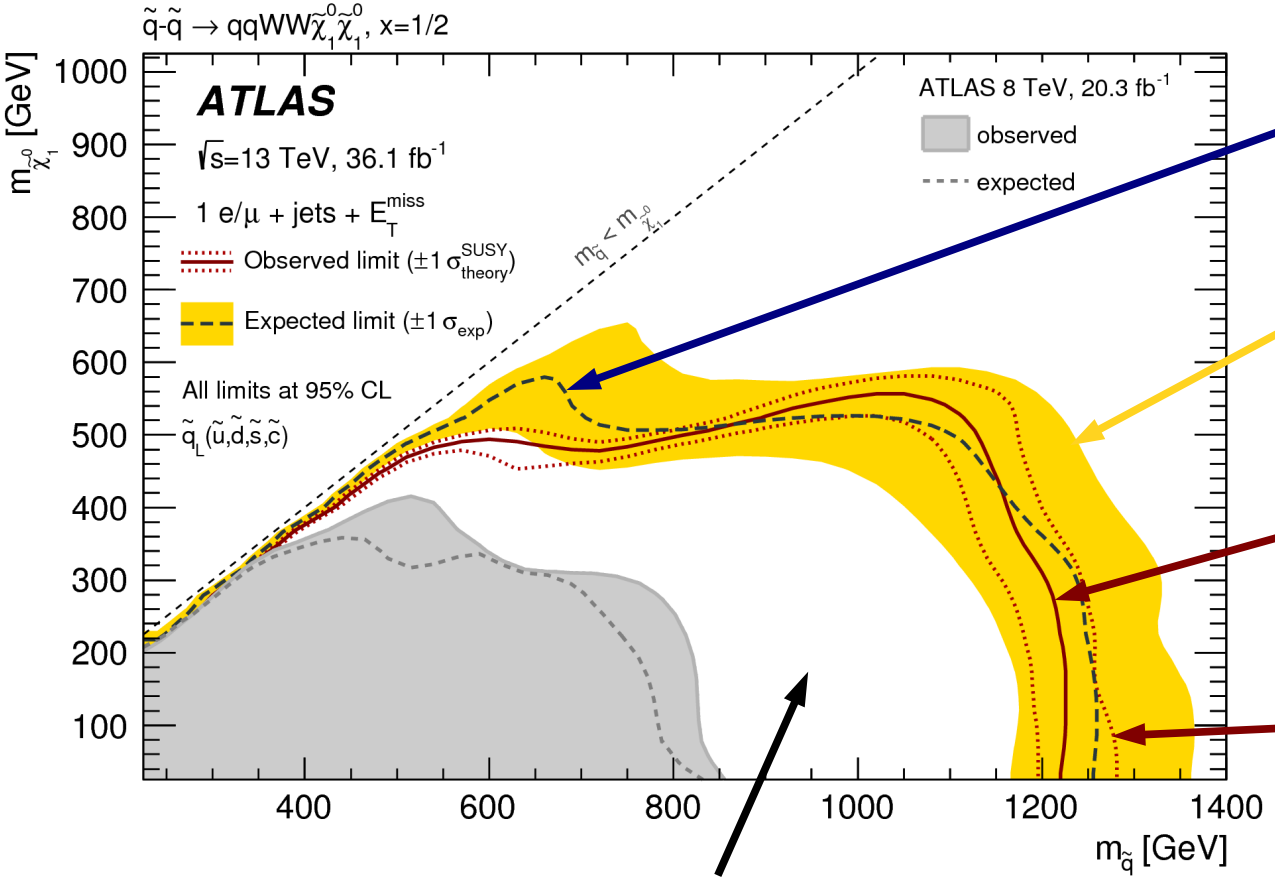
$$\mathrm{CL}_s \equiv \frac{p_{s+b}}{1-p_b} = 5\%$$



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

Example: contour plot

[Phys. Rev. D 96 (2017) 112010]



Every point in this plane corresponds to a signal model – this exact signal model is then tested

Expected limit: Limit if data would correspond to the background model

Uncertainties on the expected limit, depend also on the nuisance parameters

Observed limit, calculated with the data

Uncertainties on the observed limits (varying the signal cross section +- 1 σ)

Confidence intervals

Parameter estimation: what is the allowed range (**confidence interval**) for a parameter, i.e. the interval that contains at a level of x% of the times the true value of the parameter?

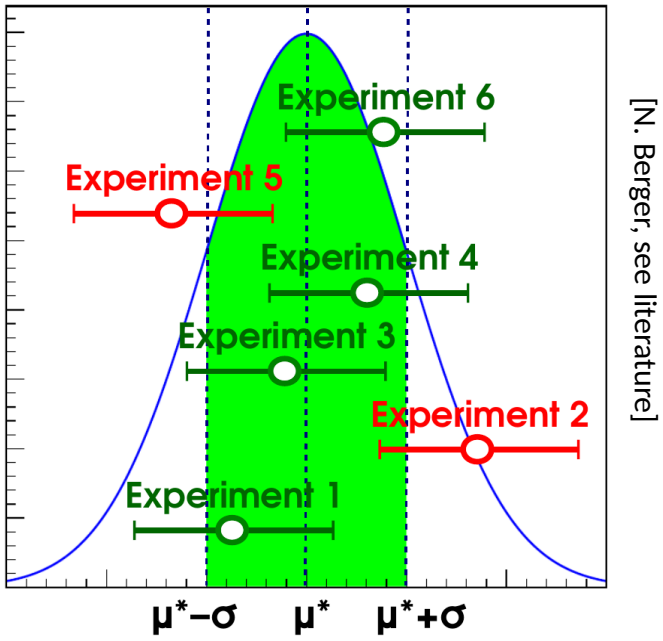
Consider Gaussian case ($\hat{\mu} \sim G(\mu^*, \sigma)$):

$$P(\mu^* - \sigma < \hat{\mu} < \mu^* + \sigma) = 68\,\%$$

$$\Leftrightarrow P(|\hat{\mu} - \mu^*| < \sigma) = 68\,\%$$

$$\Leftrightarrow P(\hat{\mu} - \sigma < \mu^* < \hat{\mu} + \sigma) = 68\,\%$$

→ the interval $[\hat{\mu} - \sigma, \hat{\mu} + \sigma]$ will contain the true value μ^* in 68% of the times

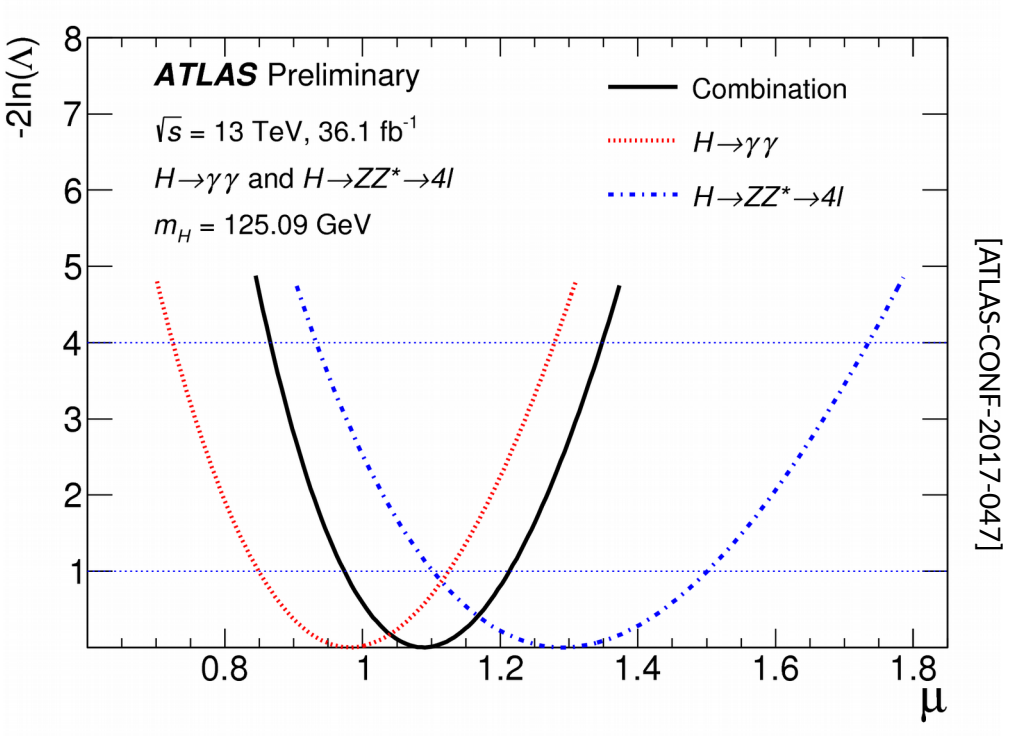


Confidence intervals with likelihoods

- In case of likelihoods test $H(\mu_0)$ use the test statistics $t_{\mu_0} = -2 \ln \frac{L(\mu = \mu_0)}{L(\hat{\mu})}$
- This test needs to be two-sided as the true value can be lower or higher than the observed value.

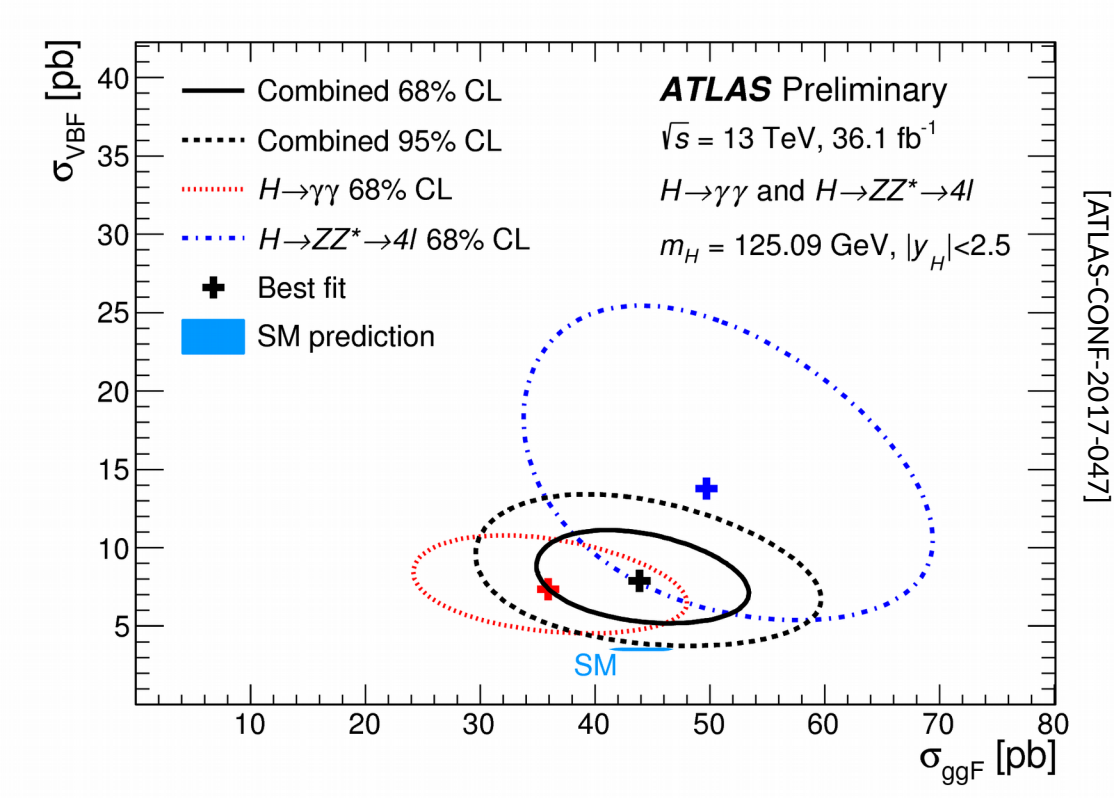
Example 1D:

- Plot t_μ
- Minimum is at $\hat{\mu}$
- +- Z σ uncertainties given by crossings of t_μ with Z^2
- Gaussian case: $[\hat{\mu} - \sigma, \hat{\mu} + \sigma]$ for 68% coverage



Example in 2D

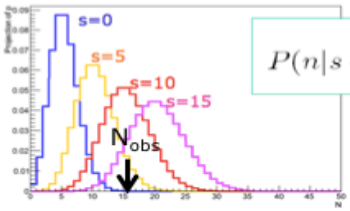
Essentially the same procedure, but now for multiple parameters.



Overview on tools in ROOT

- **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'



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

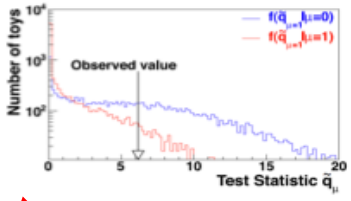
NB: b is a constant in this example

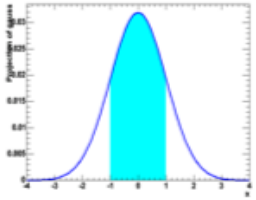
Definition: the Likelihood is P(observed data|theory)

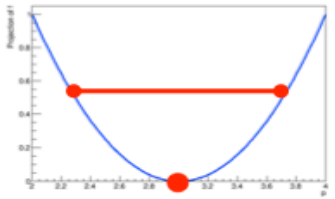
e.g. L(15|s=0)
e.g. L(15|s=10)

Frequentist statistics | Bayesian statistics | Maximum Likelihood

$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}\,|\,\mu)}{L(\vec{N}\,|\,\hat{\mu})} \qquad P(\mu) \propto L(x\,|\,\mu)\cdot\pi(\mu) \qquad \left.\frac{d\ln L(\vec{p})}{d\vec{p}}\right|_{p_i=\hat{p}_i} = 0$$







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

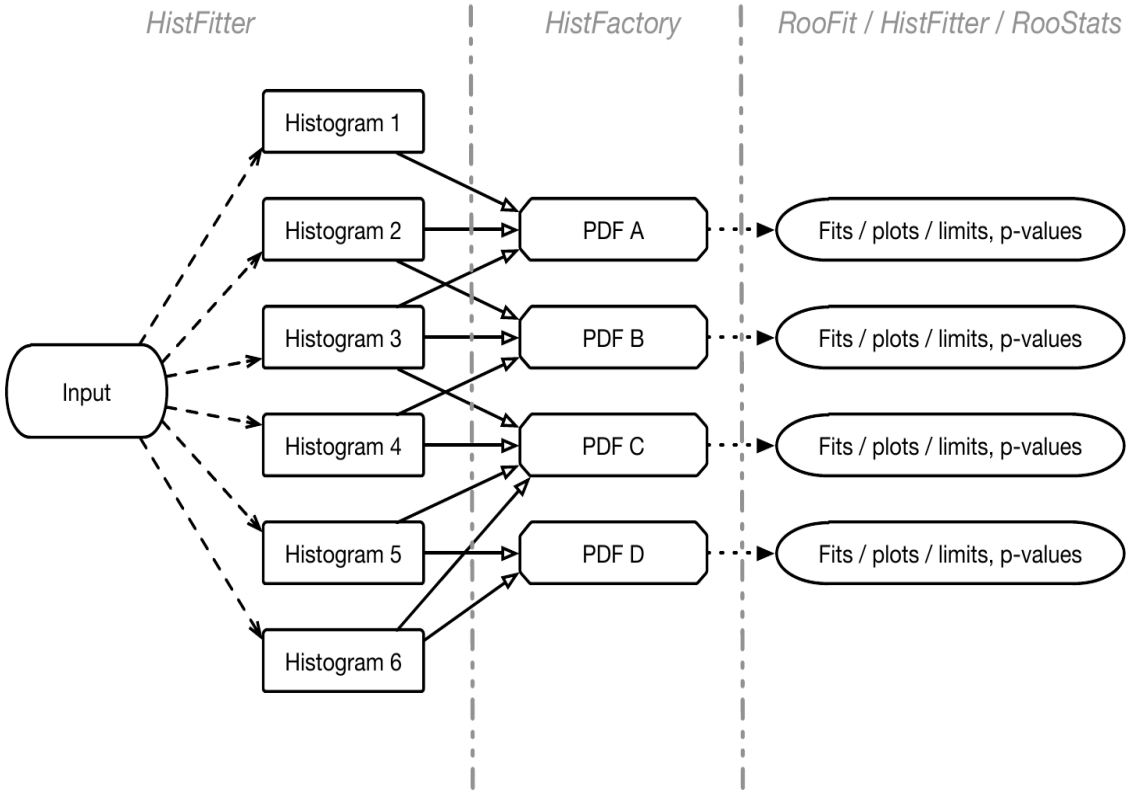
Wouter Verkerke, NIKHEF

7

HistFitter

[Eur.Phys.J. C75 (2015) 153]

The statistical tool/framework HistFitter is built around RooFit/histfactory/RooStats + extends them in key areas → the user can perform a full statistical analysis based on just a user specific configuration file



step-0
Histogram production

step-1
PDF construction
Workspace building

step-2/3
Analysis of models

Summary

Presented a few of the key concepts in statistics for high-energy physics, of course being very brief on many aspects.

→ *Just an appetizer for more detailed lectures (see literature on the next slide).*

In practice, many of the concepts are nicely implemented in statistical tools like RooFit/histfactory/RooStats/HistFitter and others so that you usually do not need to worry about the fine print.

Nevertheless, one should understand the concepts to make sure that the own analysis is solid in terms of the statistical methods used!

Questions?

To read more and references

These slides are largely based on the great lectures by:

- Nicolas Berger, Introduction to statistics for high energy physicists, statistics course in Geneva, 2018, Slides, Slides2 and Slides3
- Glen Cowan, Statistics and Discoveries at the LHC, slides.

There are also few other lectures available:

- Kyle Cranmer's lectures: slides
- Lorenzo Moneta's and Louis Lyons' lectures: slides

And some recommended books:

- G. Cowan, Statistical Data Analysis, Clarendon Press, Oxford, 1998.
- R.J.Barlow, A Guide to the Use of Statistical Methods in the Physical Sciences, John Wiley, 1989;
- F. James, Statistical Methods in Experimental Physics, 2nd ed., World Scientific, 2006; W.T.Eadie et al., North-Holland, 1971;

Backup

Summary of which p.d.f to use when by N. Berger

Summary: How to describe data

Description	Observable	Likelihood
Counting	**n** : measured number of events	**Poisson** $P(n;S,B) = e^{-(S+B)}\frac{(S+B)^n}{n!}$ **S**, **B** : expected signal & background
Binned shape analysis	n_i, i=1..N_{bins} : measured events in each bin.	**Poisson product** $P(n_i;S,B) = \prod_{i=1}^{n_{bins}} e^{-(S f_i^{sig} + B f_i^{bkg})}\frac{(S f_i^{sig} + B f_i^{bkg})^{n_i}}{n_i!}$ **S**, **B** : expected signal & background f^{sig}_i f^{bkg}_i : fraction of sig & bkg in each bin
Unbinned shape analysis	m_i, i=1..n_{evts} : observable value for each event	**Extended Unbinned Likelihood** $P(m_i;S,B) = \frac{e^{-(S+B)}}{n_{evts}!}\prod_{i=1}^{n_{evts}} S\,P_{sig}(m_i) + B\,P_{bkg}(m_i)$ **S**, **B** : expected signal & background P_{sig}, P_{bkg} : PDFs for **m** in signal and bkg.

HistFitter overview

HistFitter: *software framework for statistical data analysis.*

- Built on top of ***HistFactory/RooFit*** (construction of parametric models) and ***RooStats*** (statistical tests of data)
- Consists of a ***Python*** part for configuration and a ***C++*** part for CPU-intensive calculations

HistFitter extends RooFit/HistFactory/RooStats in four key areas:

- Programmable framework:
  Performing complete statistical analyses, using a user-defined configuration file
- Analysis strategy:
  Concepts of analysis control, validation and signal regions deeply woven into the design of HistFitter
- Bookkeeping:
  HistFitter keeps track of numerous data models - including construction and statistical tests of all of them in an organized way
- Presentation and interpretation:
  Easy-to-use tools to present data and interpret results (statistical significances; quality of likelihood fits; tables and plots summarising the results; etc.)

HistFitter used in numerous analyses (e.g. SUSY searches) of the ATLAS Collaboration at the LHC.