

Statistics and Data Analysis (HEP at LHC)

- ☐ Reminders from lecture 1 and 2
- ☐ Expected results and toys
 - Pseudo-experiments and Asimov datasets
 - Dealing with non-asymptotic situations
- ☐ Profiling
- ☐ Look-Elsewhere Effect
- ☐ Bayesian method
- ☐ Presentation of results

Slides extracted from N. Berger lectures at CERN Summer School 2019

Reminders from Lecture 1: Statistical Model

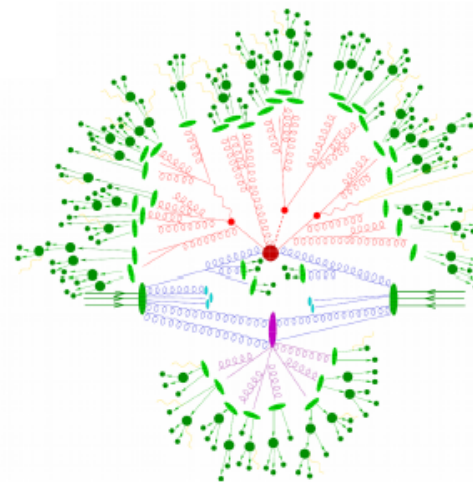
Goal:

Describe the random process by which the data was obtained.

→ Build a **Statistical Model**

Ingredients:

1. **Statistical description** of the random aspects
⇒ **Probability distributions**
2. **Assumptions** on the underlying statistical processes (physics, etc.)
→ Uncertainties on the assumptions themselves: **systematic uncertainties**

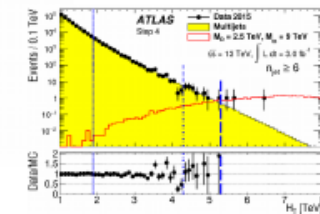


Hard scattering

Decays

Detector response

Reconstruction



"Systematic uncertainty is, in any statistical inference procedure, the uncertainty due to the incomplete knowledge of the probability distribution of the observables.

G. Punzi, *What is systematics?*

Statistical results can only be as accurate as the model itself !

Reminders from Lecture 1: Statistical Model

Physics measurement data are produced through **random processes**,
Need to be described using a statistical model:

Description	Observable	Likelihood
Counting	n	Poisson $P(n; S, B) = e^{-(S+B)} \frac{(S+B)^n}{n!}$
Binned shape analysis	$n_i, i=1..N_{\text{bins}}$	Poisson product $P(n_i; S, B) = \prod_{i=1}^{n_{\text{bins}}} e^{-(S f_i^{\text{sig}} + B f_i^{\text{bkg}})} \frac{(S f_i^{\text{sig}} + B f_i^{\text{bkg}})^{n_i}}{n_i!}$
Unbinned shape analysis	$m_i, i=1..n_{\text{evts}}$	Extended Unbinned Likelihood $P(m_i; S, B) = \frac{e^{-(S+B)}}{n_{\text{evts}}!} \prod_{i=1}^{n_{\text{evts}}} S P_{\text{sig}}(m_i) + B P_{\text{bkg}}(m_i)$

Model can include multiple **categories**, each with a separate description
Includes **parameters of interest** (POIs) but also **nuisance parameters** (NPs)

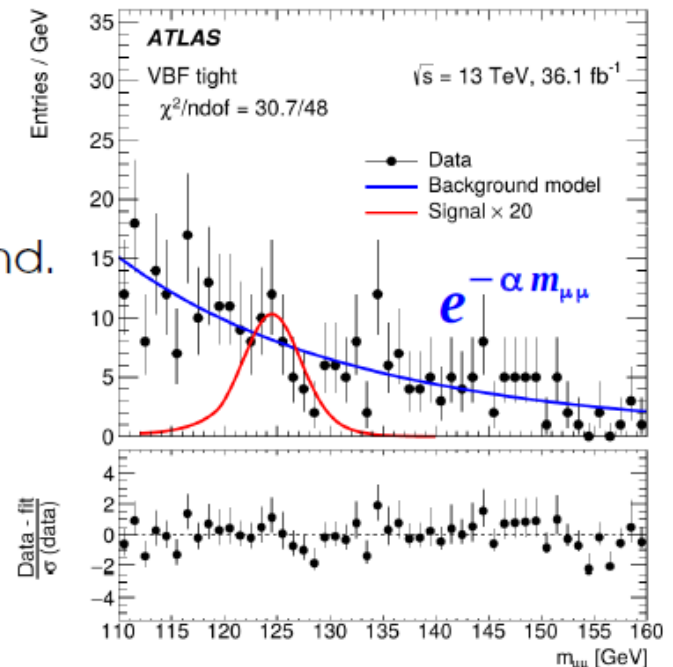
Model Parameters

Model typically includes:

- **Parameters of interest** (POIs) : what we want to measure
→ $S, \sigma \times B, m_W, \dots$
- **Nuisance parameters** (NPs) : other parameters needed to define the model
→ B
→ For binned data, f_{sig}_i, f_{bkg}_i
→ For unbinned data, parameters needed to define P_{bkg}
e.g. exponential slope α of $H \rightarrow \mu\mu$ background.

NPs must be either

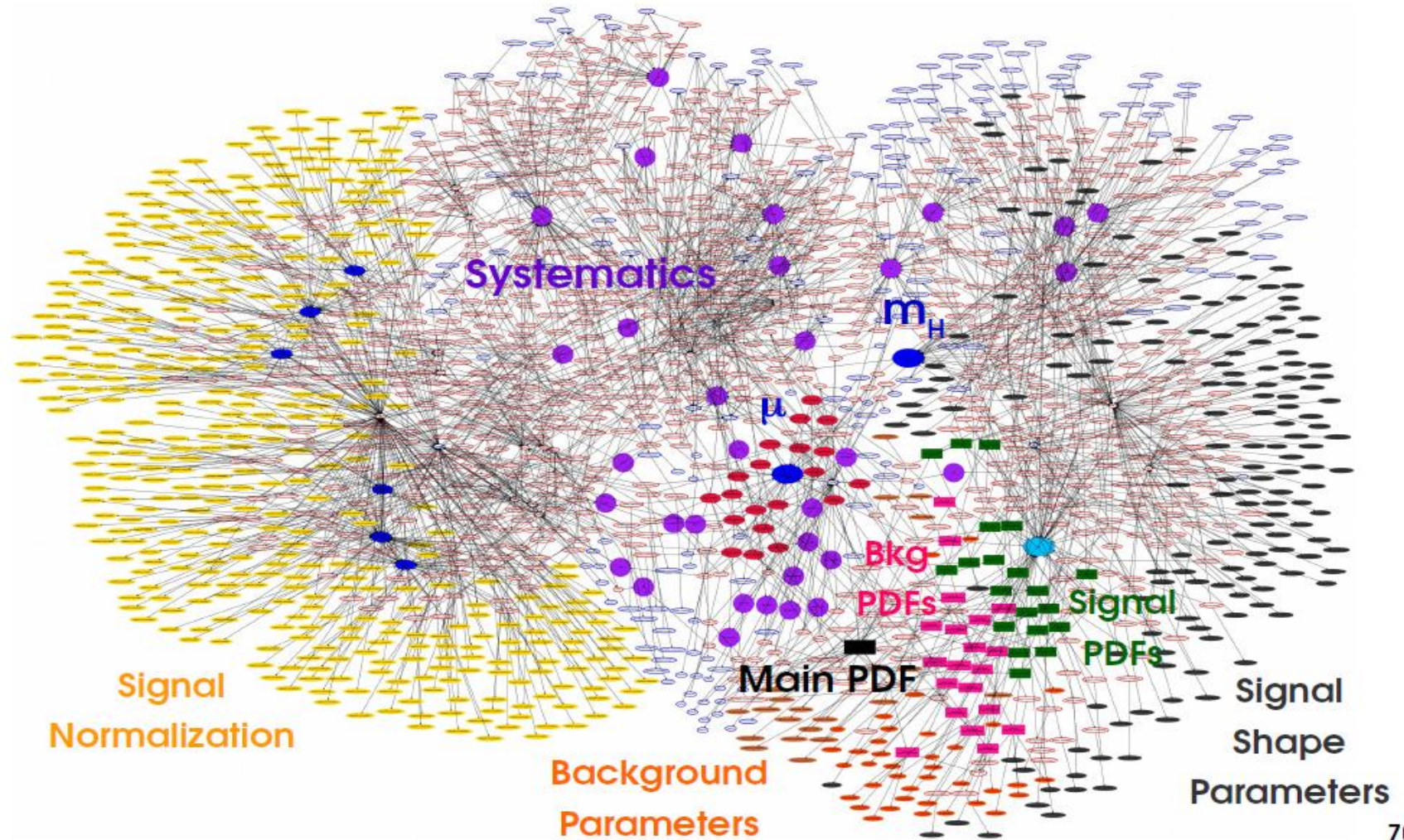
- **known a priori** (possibly within systematics) or
- **constrained by the data** (e.g. in sidebands)



Phys. Rev. Lett. 119 (2017) 051802

Model Example

$H \rightarrow \gamma\gamma$ Discovery Analysis



Statistical Results as Hypothesis Tests

Usual HEP results can be recast in terms of **hypothesis testing**:

- **Discovery**: Is the data compatible with background-only ?
 - H_0 : only background is present
 - How well can we **reject H_0** ? → **p-value (significance)**
- **Upper limits**: no excess observed – how small must the signal be ?
 - $H_0(S)$: B + some signal S
 - How small can we make S, and still reject $H_0(S)$ at 95% C.L. (p-value=5%) ?
- **Parameter measurement**
 - $H_0(\mu)$: some parameter value μ
 - What values μ are not rejected at 68% C.L. (p=32%) ?
 - ⇒ **1σ confidence interval on μ**

In all cases, H_0 : **null hypothesis** – what we are trying to disprove

Test Statistics for Discovery

Discovery :

- H_0 : background only ($S = 0$) against
- H_1 : presence of a signal ($S \neq 0$)



→ For H_1 , any $S \neq 0$ is possible, which to use ? **The one preferred by the data, $\hat{S}$.**

⇒ Use LR $\frac{L(S=0)}{L(\hat{S})}$

→ In fact use the **test statistic** $t_0 = -2 \log \frac{L(S=0)}{L(\hat{S})}$

→ t_0 is computed from the observed data – fit to data to get $\hat{S}$.

→ t_0 **always** ≥ 0 , $t_0 = 0$ reached for $\hat{S} = 0$.

→ t_0 measures the relative *likelihood* of H_1 vs. H_0 in data:

Large values of $t_0 \Leftrightarrow$ large observed S

Discovery p-value

Large values of $t_0 = -2 \log \frac{L(S=0)}{L(\hat{S})}$

⇒ large observed $\hat{S}$

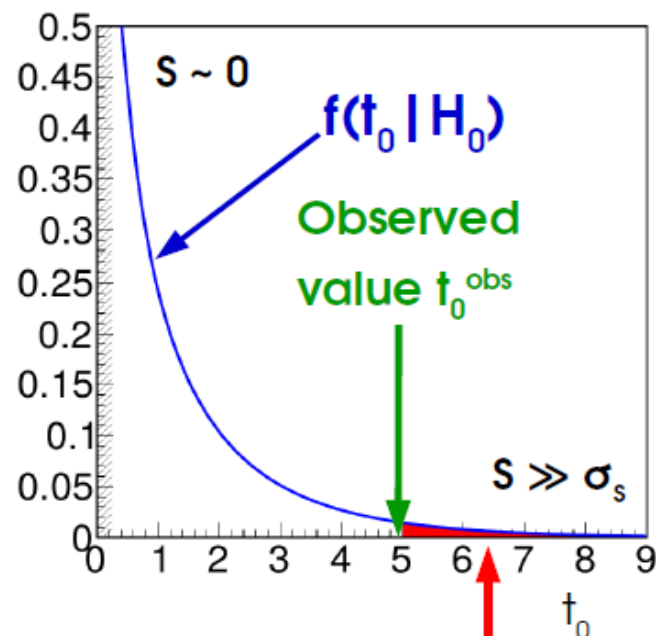
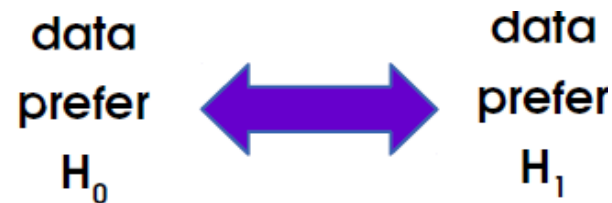
⇒ $H_0(S=0)$ **disfavored** compared to $H_1(S \neq 0)$.

How large t_0 before we can exclude H_0 ?
(and claim a discovery!)

p-value : Fraction of outcomes that are **at least as H_1 -like** (signal-like) **as data**, when **H_0 is true** (no signal present).

→ Smaller p-value ⇒ Stronger case for discovery

→ Compute from distribution $f(t_0 | H_0)$ of t_0 if H_0 is true:



$$p_0 = \int_{t_0^{\text{obs}}}^{\infty} f(t_0 | H_0) dt_0$$

Reminder: Wilk's Theorem

Cowan, Cranmer, Gross & Vitells
Eur.Phys.J.C71:1554,2011

Consider $t_{S_0} = -2 \log \frac{L(S=S_0)}{L(\hat{S})}$

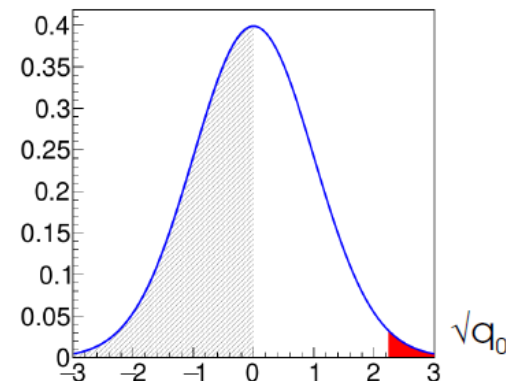
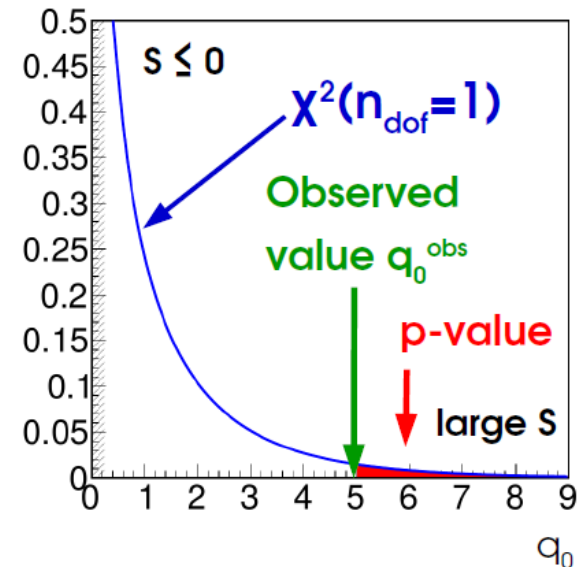
→ Assume **Gaussian regime** (e.g. large n_{evts} , Central-limit theorem) : then:

Wilk's Theorem: t_{S_0} is distributed as a χ^2
under $H_{S_0}(S=S_0)$:

$$f(t_{S_0} | S=S_0) = f_{\chi^2(n_{\text{dof}}=1)}(t_{S_0})$$

⇒ The significance is:

$$Z = \sqrt{q_0}$$



Asymptotic Approximation

Cowan, Cranmer, Gross & Vitells
Eur.Phys.J.C71:1554,2011

→ Assume **Gaussian regime for $\hat{S}$** (e.g. large n_{evts}) ⇒ Central-limit theorem :

⇒ **t_0 is distributed as a χ^2** under the hypothesis H_0

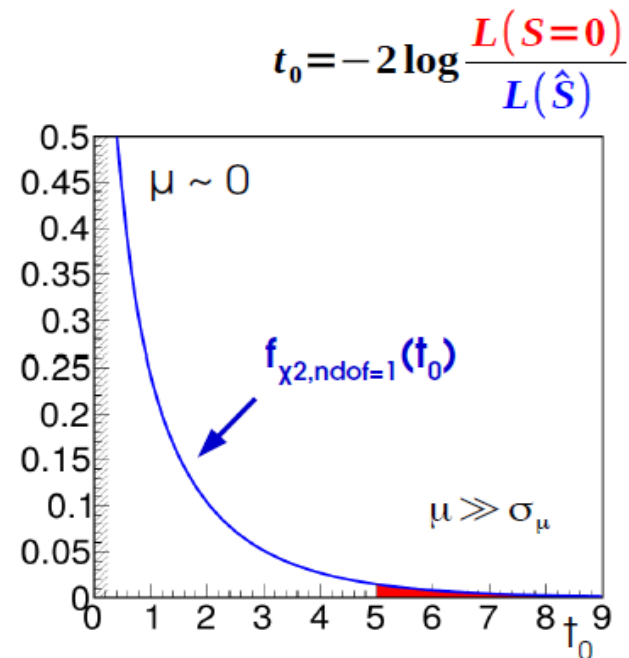
$$f(t_0 | H_0) = f_{\chi^2(n_{\text{dof}}=1)}(t_0)$$

In particular, significance:

$$Z = \sqrt{t_0}$$

By definition,
 $t_0 \sim \chi^2 \Rightarrow \sqrt{t_0} \sim G(0,1)$

Typically works well for event counts $O(5)$
and above (5 already “large”...)



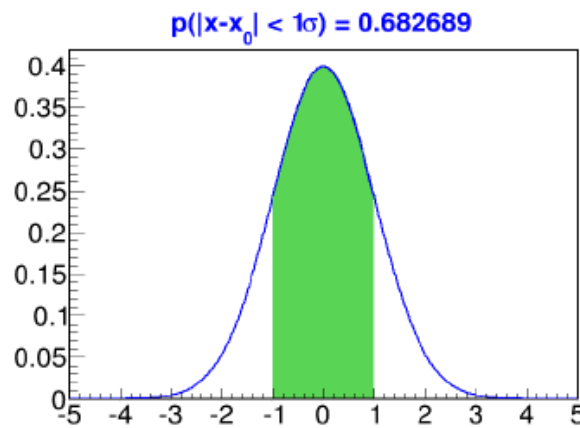
The 1-line “proof” : asymptotically L and S are Gaussian, so

$$L(S) = \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}}{\sigma}\right)^2\right] \Rightarrow t_0 = \left(\frac{\hat{S}}{\sigma}\right)^2 \Rightarrow t_0 \sim \chi^2(n_{\text{dof}}=1) \text{ since } \hat{S} \sim G(0, \sigma)$$

Discovery significance

Interesting p-values are quite small
 $\Rightarrow$ express in terms of Gaussian quantiles

$\rightarrow$ **Significance Z**



$$p_0 = 1 - \int_{-Z}^{+Z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

$$= 1 - 2 \Phi(Z)$$

$$\Phi(Z) = \int_{-\infty}^Z G(u; 0, 1) du$$

Z	p-value
1	0.32
2	0.045
3	0.003
5	6×10^{-7}

In ROOT:

$p_0 \rightarrow Z \ (\Phi) : \text{ROOT::Math::gaussian_quantile_c}$

$Z \rightarrow p_0 \ (\Phi^{-1}) : \text{ROOT::Math::gaussian_cdf_c}$

$\Rightarrow$ How small is small enough ?

$\rightarrow$ Conventionally, discovery for $p_0 = 6 \times 10^{-7} \Leftrightarrow Z = 5\sigma$

Takeaways: Discovery Significance

Given a statistical model $P(\text{data}; \mu)$, define likelihood $\mathbf{L}(\mu) = P(\text{data}; \mu)$

To estimate a parameter, use the value $\hat{\mu}$ that maximizes $L(\mu)$.

To decide between hypotheses H_0 and H_1 , use the **likelihood ratio** $\frac{L(H_0)}{L(H_1)}$

To test for **discovery**, use $q_0 = -2 \log \frac{L(S=0)}{L(\hat{S})} \quad \hat{S} \geq 0$

For large enough datasets ($n > 5$), $\mathbf{Z} = \sqrt{q_0}$

For a **Gaussian** measurement, $\mathbf{Z} = \frac{\hat{S}}{\sqrt{B}}$

For a **Poisson** measurement, $\mathbf{Z} = \sqrt{2 \left[(\hat{S} + B) \log \left(1 + \frac{\hat{S}}{B} \right) - \hat{S} \right]}$

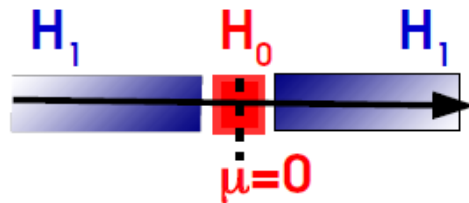
Hypothesis testing: One-Sided vs Two-Sided

If $\hat{S} < 0$, is it a *discovery* ? (does reject the $S=0$ hypothesis...)

Usual assumption : only $\hat{S} > 0$ is a *bona fide* signal

⇒ Change statistic so that $\hat{S} < 0 \Rightarrow t_0 = 0$ (perfect agreement with H_0 , as for $\hat{S} = 0$)

Two-sided

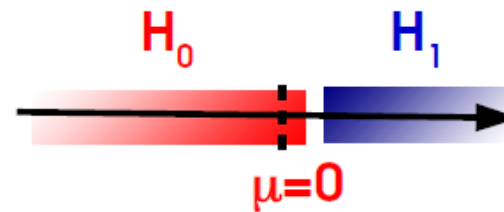


$$t_0 = -2 \log \frac{L(S=0)}{L(\hat{S})}$$

$$Z = \Phi^{-1}\left(1 - \frac{p_0}{2}\right)$$

By convention, factor 2
in p-values for a given Z

One-sided



$$q_0 = \begin{cases} -2 \log \frac{L(S=0)}{L(\hat{S})} & \hat{S} \geq 0 \\ 0 & \hat{S} < 0 \end{cases}$$

$$Z = \Phi^{-1}(1 - p_0)$$

⇒ Same Z in both cases
for a given signal S

p_0	Z	p_0
0.32	1	0.16
0.003	3	0.0015
6×10^{-7}	5	3×10^{-7}

One-Sided Asymptotics

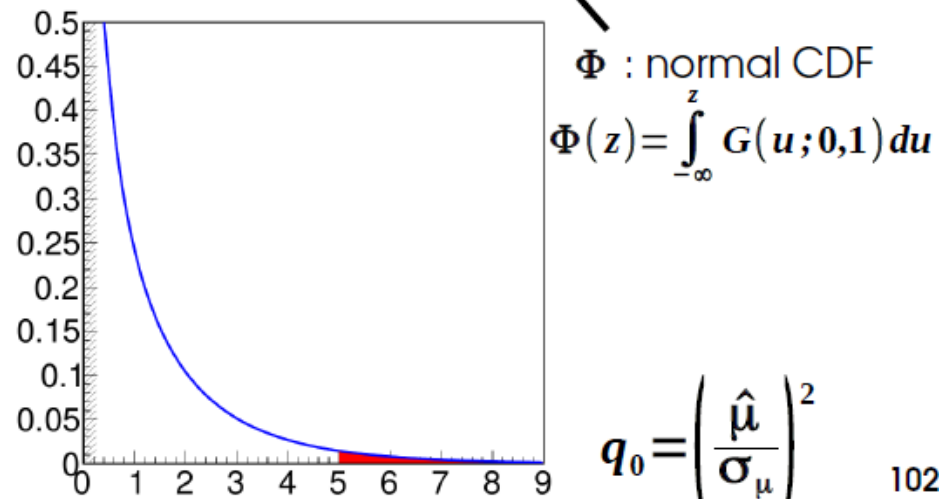
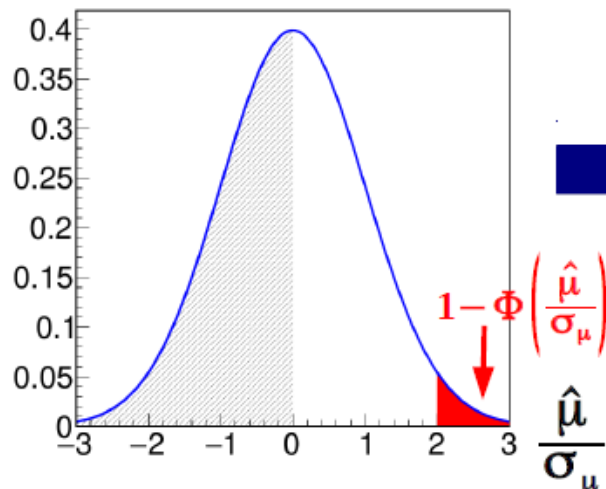
→ One-sided test:



$$q_0 = \begin{cases} -2 \log \frac{L(S=0)}{L(\hat{S})} & \hat{S} \geq 0 \\ 0 & \hat{S} < 0 \end{cases}$$

Asymptotics: “half- χ^2 ” distribution: $f(q_0 | S=0) = \frac{1}{2} \delta(q_0) + \frac{1}{2} f_{\chi^2(n_{\text{dof}}=1)}(q_0)$

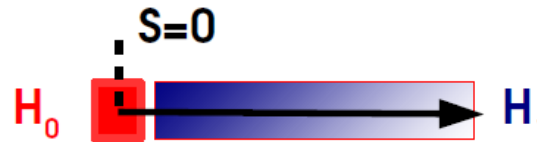
Discovery p-value: $p_0 = 1 - \Phi(\sqrt{q_0})$ Significance: $Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$



Test Statistic for Limit-Setting

Discovery :

- $H_0 : S = 0$
- $H_1 : S > 0$



$$q_0 = -2 \log \frac{L(S=0)}{L(\hat{S})}$$

Compare

← Likelihood of H_0 ($\hat{S} > 0$)

← Likelihood of H_1

Limit-setting

- $H_0 : S = S_0$
- $H_1 : S < S_0$



$$q_{S_0} = -2 \log \frac{L(S=S_0)}{L(\hat{S})}$$

Compare

← Likelihood of H_0 ($\hat{S} < S_0$)

← Likelihood of H_1

Same as q_0 :

→ large values $\Rightarrow$ good rejection of H_0 .

$\Rightarrow$ Can compute p-value from q_{S_0} .

Takeaways: Limits & Intervals

Limits : use LR-based test statistic:

$$q_{S_0} = -2 \log \frac{L(S=S_0)}{L(\hat{S})} \quad \hat{S} \leq S_0$$

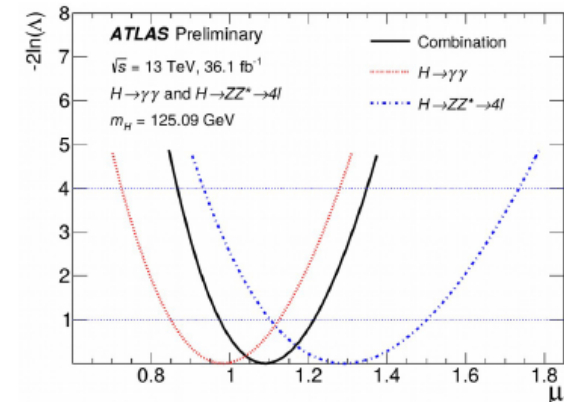
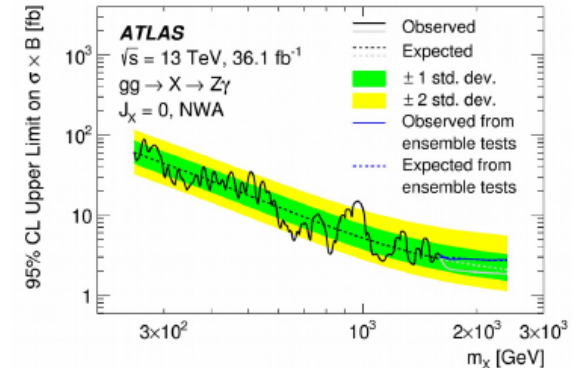
→ Use **CL_s procedure** to avoid negative limits

Poisson regime, $n=0$: $S_{\text{up}} = 3 \text{ events}$

Confidence intervals: use $t_{\mu_0} = -2 \log \frac{L(\mu=\mu_0)}{L(\hat{\mu})}$

→ 1D: crossings with $t_{\mu_0} = Z^2$ for $\pm Z\sigma$ intervals

Gaussian regime: $\mu = \hat{\mu} \pm \sigma_{\text{Gauss}}$ for a 1σ interval



Generating Pseudo-data

Model describes the distribution of the observable: $P(\text{data}; \text{parameters})$

⇒ Possible outcomes of the experiment, for given parameter values

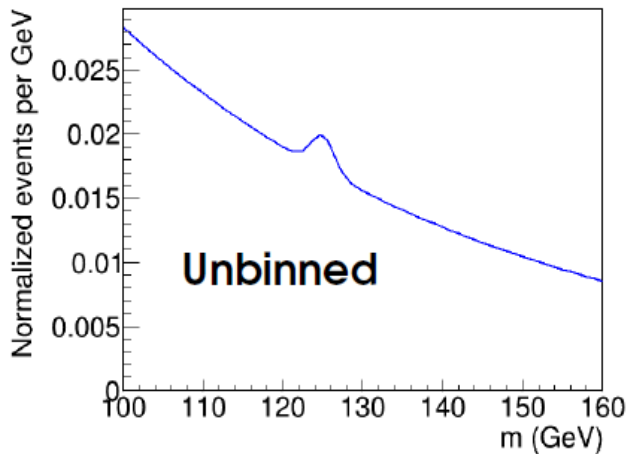
Can draw random events according to PDF : **generate** *pseudo-data*

$$P(\lambda=5)$$

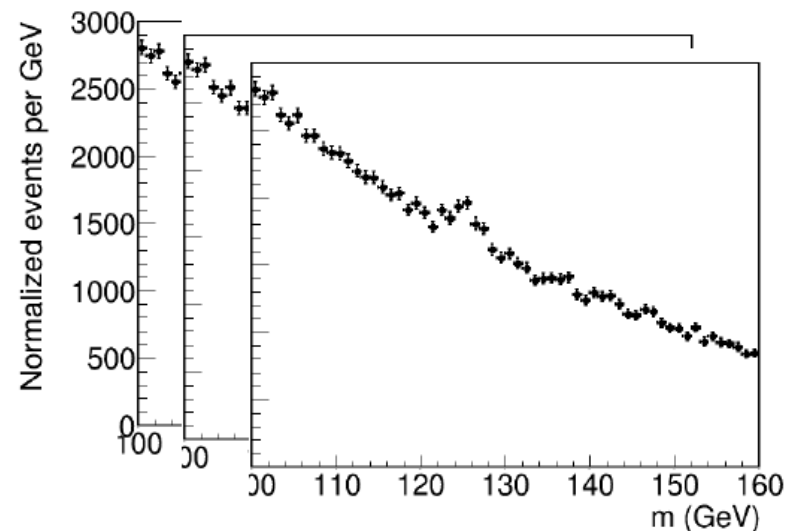


2, 5, 3, 7, 4, 9,

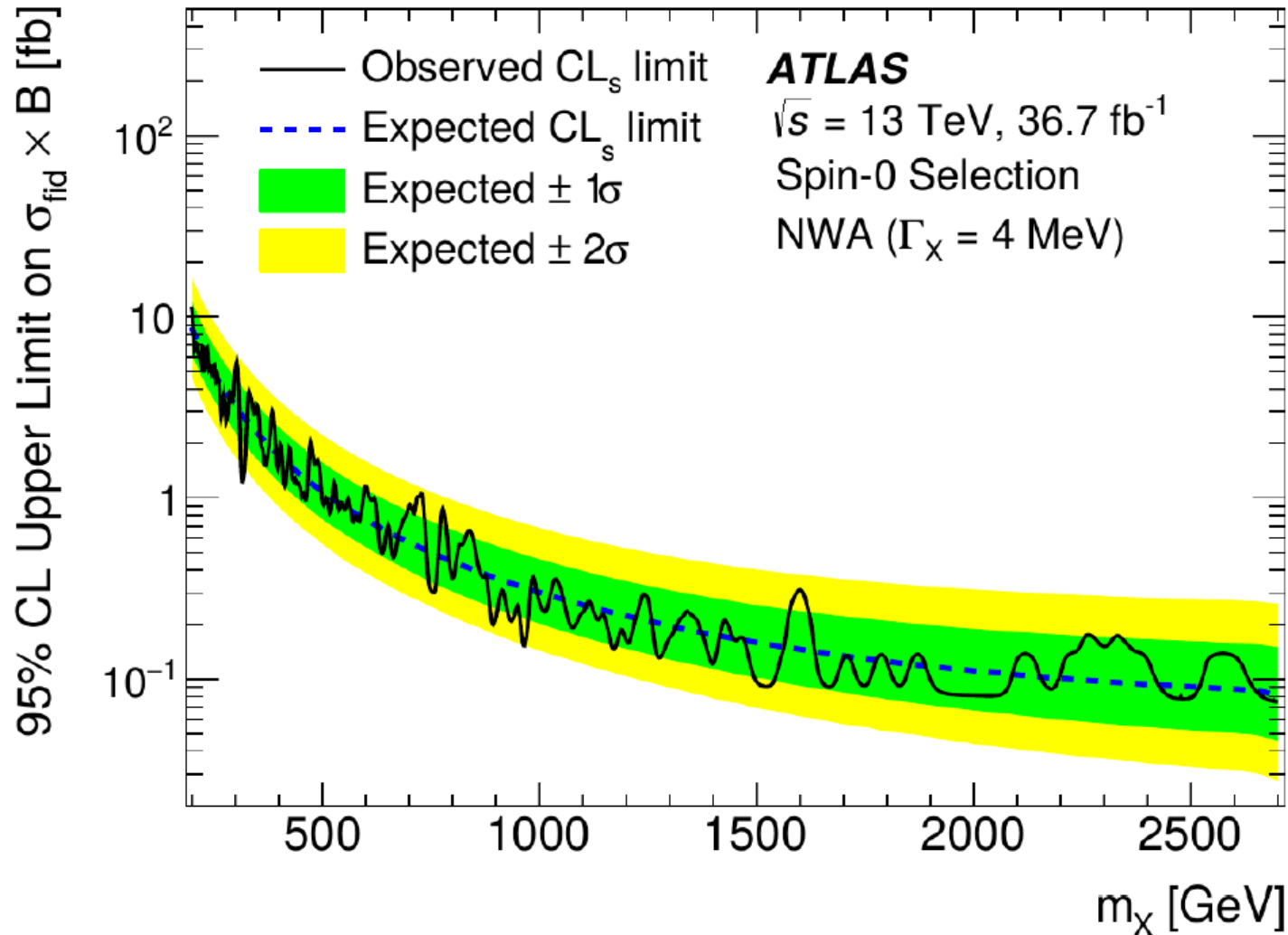
Each entry = separate "experiment"



Generate



Expected results



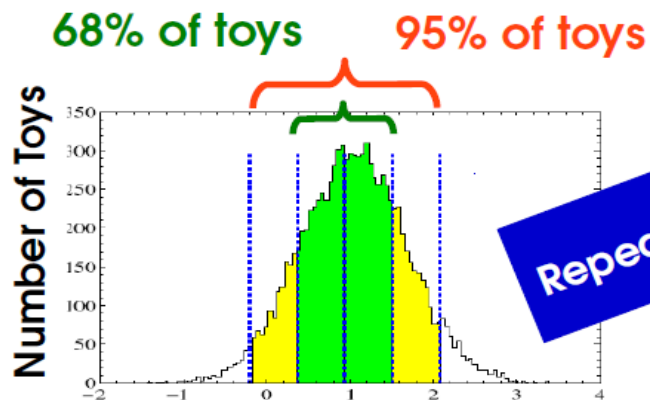
Expected limits: Toys

Expected results: median outcome under a given hypothesis
→ usually B-only for searches, but other choices possible.

Two main ways to compute:

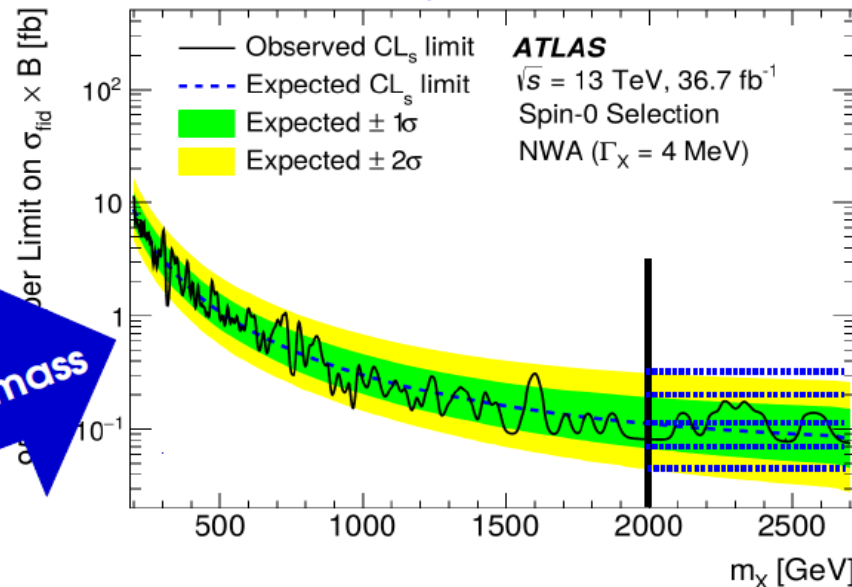
→ **Pseudo-experiments (toys):**

- Generate a pseudo-dataset in B-only hypothesis
- Compute limit
- Repeat and histogram the results
- Central value = median, bands based on quantiles



Repeat for each mass

Phys. Lett. B 775 (2017) 105



Eur.Phys.J.C71:1554,2011

Computed limit

CL_s: Gaussian Bands

Usual Gaussian counting example with known B:
95% CL_s upper limit on S:

$$S_{\text{up}} = \hat{S} + \left[\Phi^{-1} \left(1 - 0.05 \Phi(\hat{S}/\sigma_s) \right) \right] \sigma_s \quad \text{with} \quad \sigma_s = \sqrt{B}$$

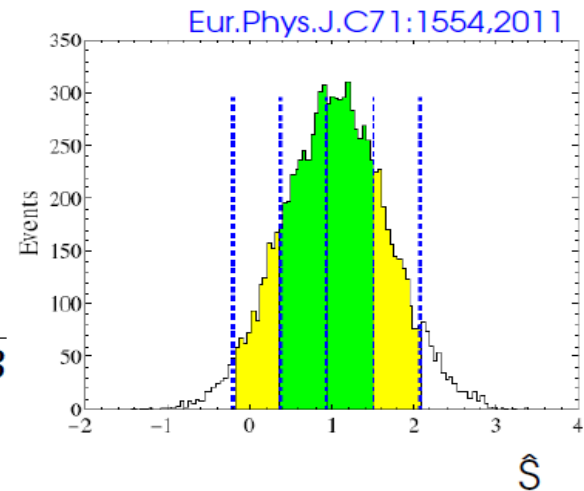
Compute expected bands for S=0:

→ **Asimov dataset** $\Leftrightarrow \hat{S} = 0$:

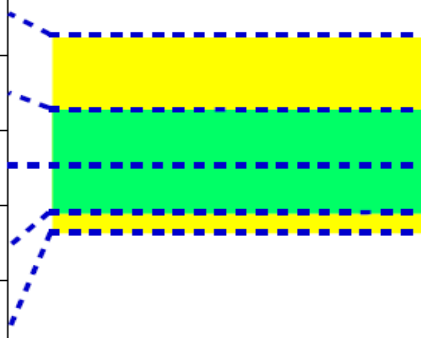
$$S_{\text{up,exp}}^0 = 1.96 \sigma_s$$

→ **$\pm n \sigma$ bands**:

$$S_{\text{up,exp}}^{\pm n} = \left(\pm n + \left[1 - \Phi^{-1} \left(0.05 \Phi(\mp n) \right) \right] \right) \sigma_s$$



n	$S_{\text{exp}}^{\pm n} / \sqrt{B}$
+2	3.66
+1	2.72
0	1.96
-1	1.41
-2	1.05



CLs :

- Positive bands somewhat reduced,
- Negative ones more so

Band width from $\sigma_{S,A}^2 = \frac{S^2}{q_s(\text{Asimov})}$
depends on S, for non-Gaussian cases, different values for each band...

Expected limits: Asimov Datasets

Expected results: median outcome under a given hypothesis

→ usually B-only by convention, but other choices possible.

Two main ways to compute:

Strictly speaking, Asimov dataset if
 $\hat{\mathbf{X}} = \mathbf{X}_0$ for all parameters $\mathbf{X}$,
where $\mathbf{X}_0$ is the generation value

→ Asimov Datasets

- Generate a “perfect dataset” – e.g. for binned data, set bin contents carefully, no fluctuations.

- Gives the median result immediately:

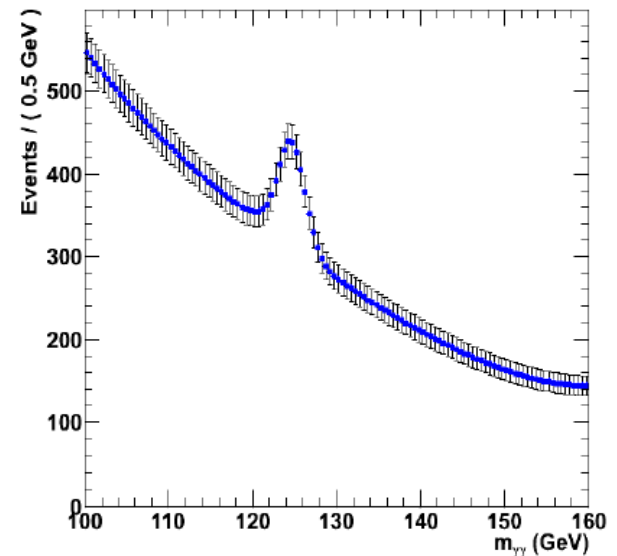
median(toy results) ↔ result(median dataset)

- Get bands from asymptotic formulas:
Band width

$$\sigma_{S_0, A}^2 = \frac{S_0^2}{q_{S_0}(\text{Asimov})}$$

⊕ Much faster (1 “toy”)

⊖ Relies on Gaussian approximation



Beyond Asymptotics: Toys

CMS-PAS-HIG-11-022

Asymptotics usually work well, but break down in some cases – e.g. **small event counts**.

Solution: generate *pseudo data* (**toys**) using the PDF, under the tested hypothesis

→ Also randomize the observable

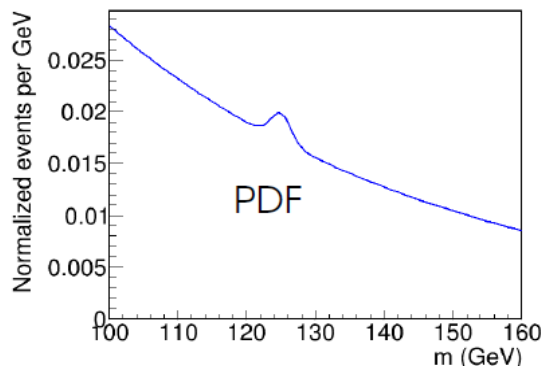
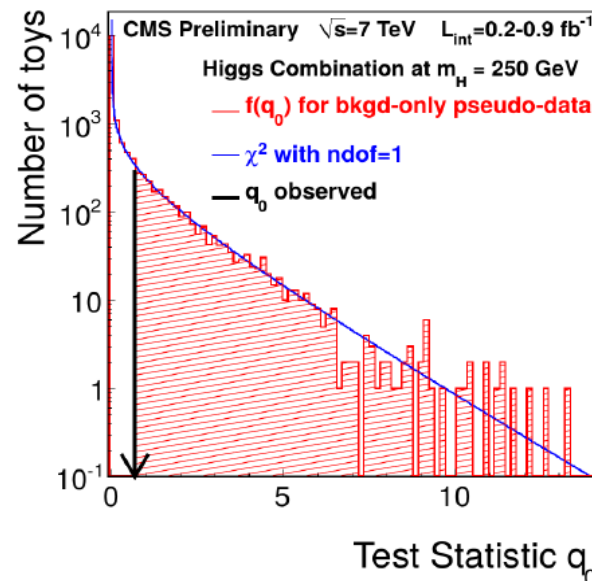
(θ^{obs}) of each auxiliary experiment: $G(\theta^{\text{obs}}; \theta, \sigma_{\text{syst}})$

→ Samples the true distribution of the PLR

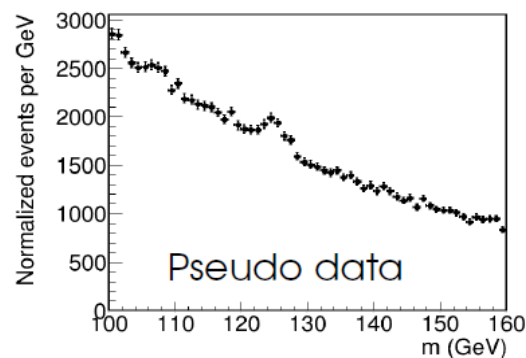
⇒ Integrate above observed PLR to get the p-value

→ Precision limited by number of generated toys,

Small p-values (5σ : $p \sim 10^{-7}$!) ⇒ **large toy samples**



$p(\text{data} | x)$



Repeat N_{toys} times



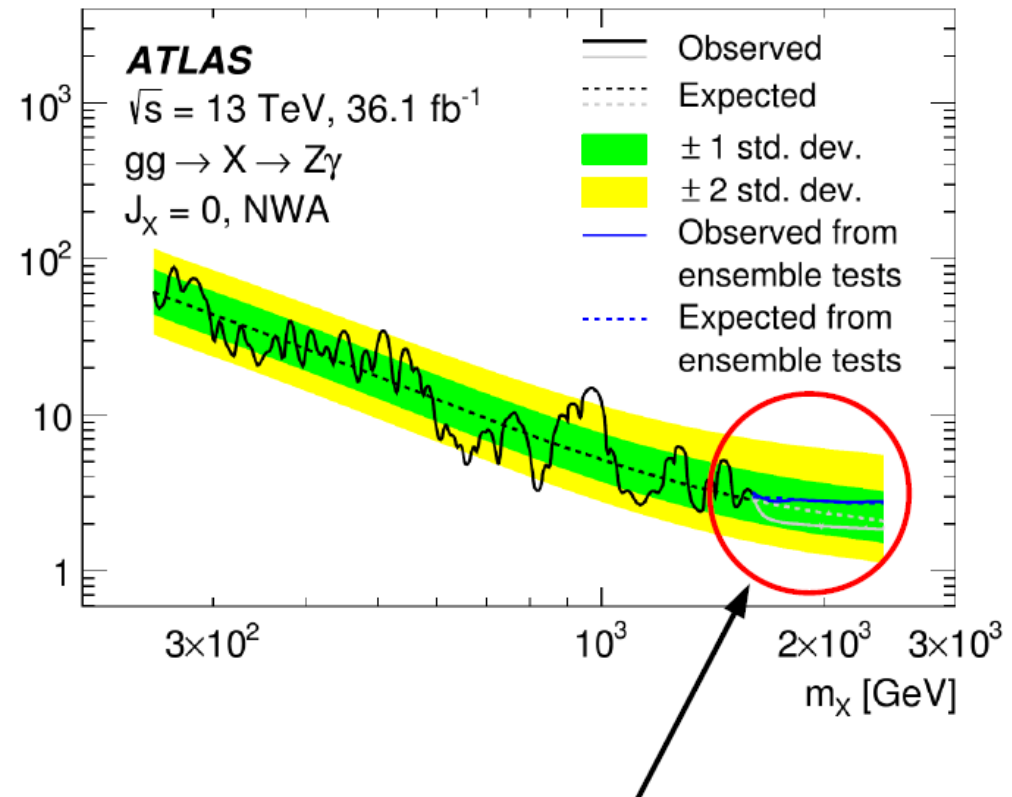
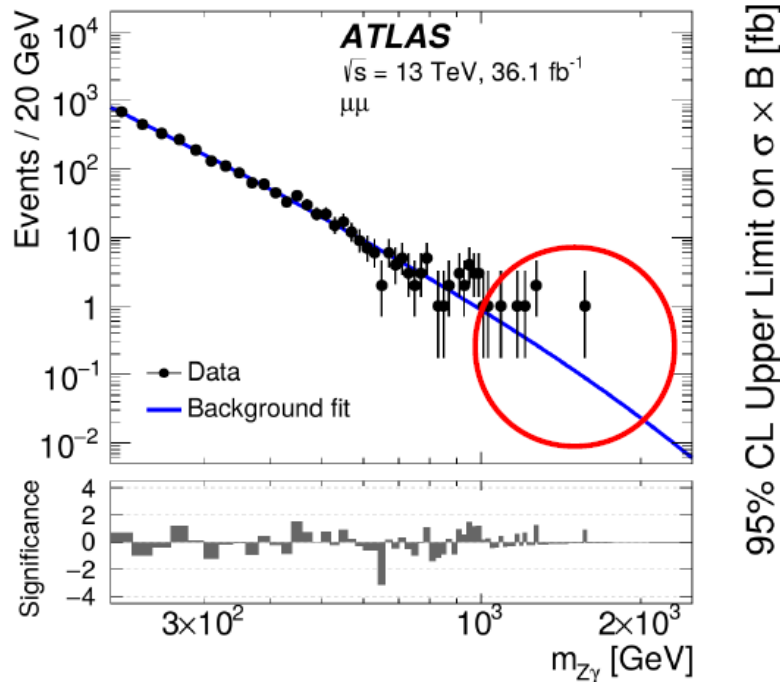
q_0

Toys: Example

ATLAS $X \rightarrow Z\gamma$ Search: covers $200 \text{ GeV} < m_X < 2.5 \text{ TeV}$

JHEP 10 (2017) 112

→ for $m_X > 1.6 \text{ TeV}$, low event counts ⇒ derive results from toys



Asymptotic results (in gray) give optimistic result compared to toys (in blue)

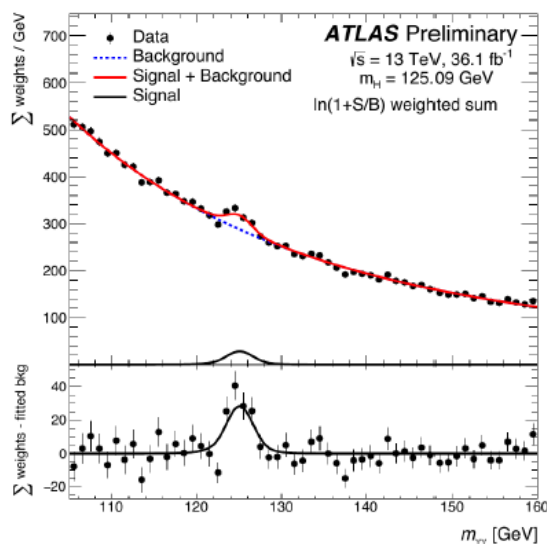
Remarks

Short answer: The high-signal, low-background experiments have been done already (although a surprise would be welcome...)

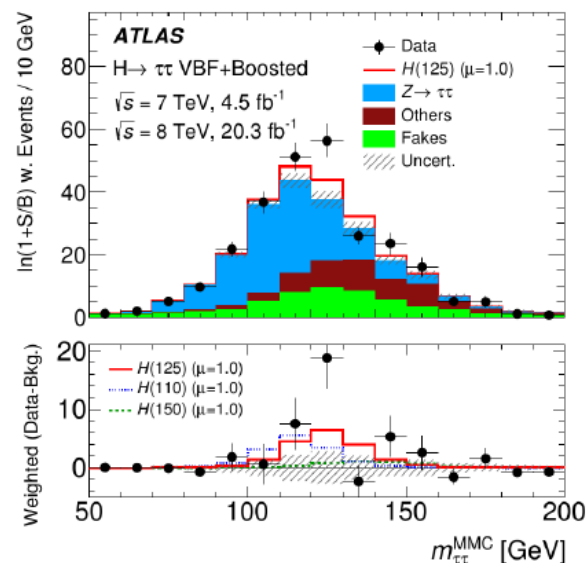
e.g. at LHC:

- **High background levels**, need precise modeling
- **Large systematics**, need to be described accurately
- **Small signals**: need optimal use of available information :
 - **Shape analyses** instead of counting
 - **Categories** to isolated signal-enriched regions

ATLAS-CONF-2017-045



JHEP 12 (2017) 024

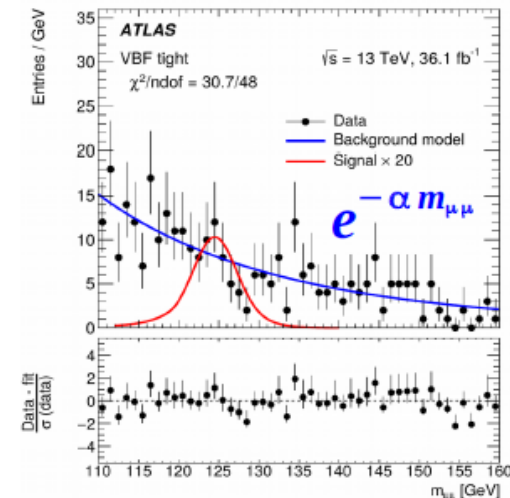


Nuisances and Systematics

Phys. Rev. Lett. 119 (2017) 051802

Likelihood typically includes

- **Parameters of interest** (POIs) : $\mu, \sigma \times B, m_W, \dots$
- **Nuisance parameters** (NPs) : other parameters needed to define the model
 → Ideally, **constrained by data** like the POI
 e.g. shape of $H \rightarrow \mu\mu$ continuum bkg



What about systematics ?

= what we don't know about the random process

⇒ **Parameterize using additional NPs**

→ By definition, **not constrained by the data**

⇒ Cannot be free, or would spoil the measurement
 (lumi free ⇒ no $\sigma \times B$ measurement!)

⇒ **Introduce a constraint in the likelihood:**

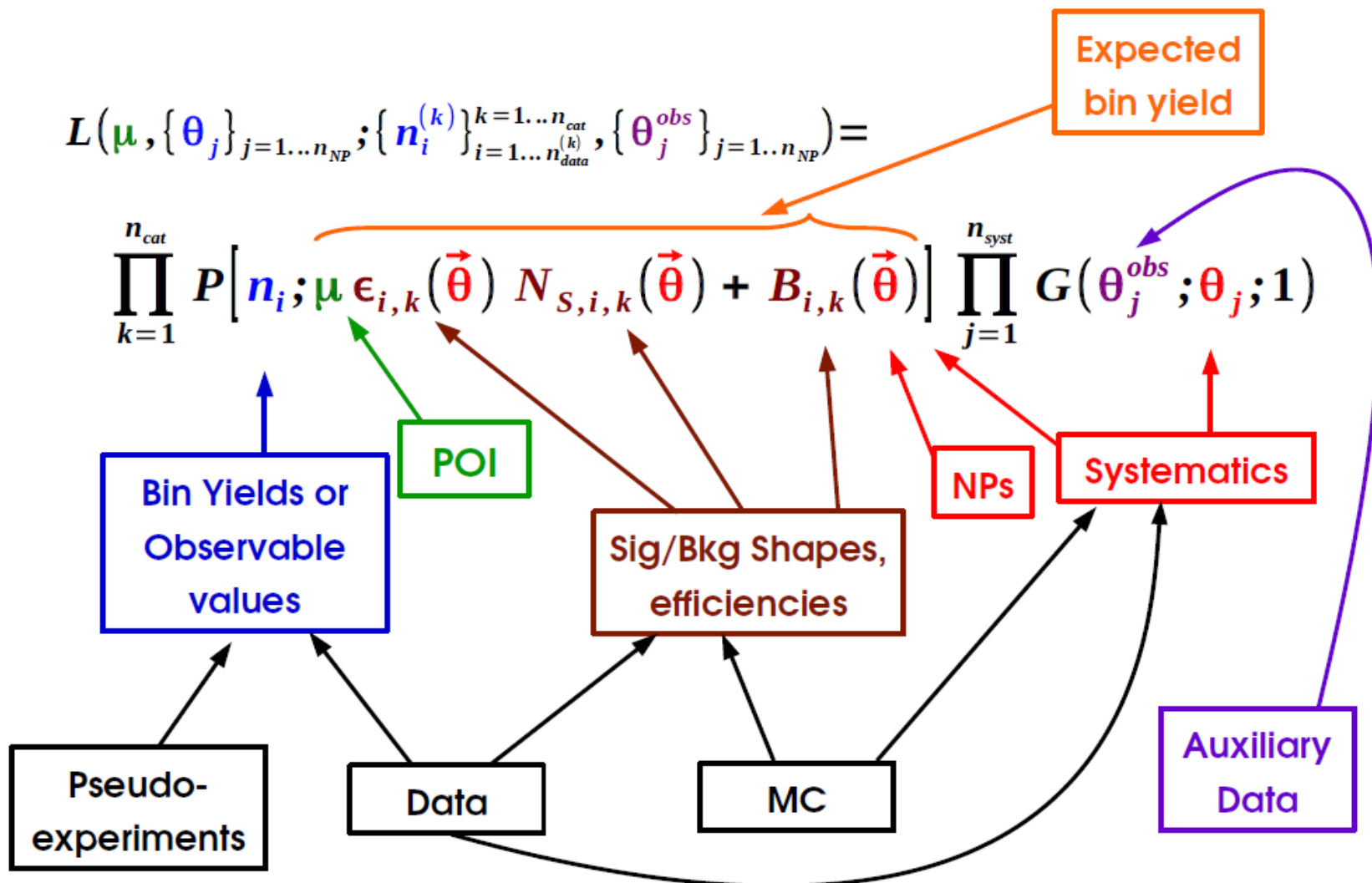
"Systematic uncertainty is, in any statistical inference procedure, the uncertainty due to the incomplete knowledge of the probability distribution of the observables.

G. Punzi, *What is systematics ?*

$$L(\underbrace{\mu}_{\text{POI}}, \underbrace{\theta}_{\text{Systematics NP}}; \text{data}) = \underbrace{L_{\text{measurement}}(\mu, \theta; \text{data})}_{\text{Measurement Likelihood}} \underbrace{C(\theta)}_{\text{NP Constraint term}}$$

⇒ penalty for $\theta \neq \theta^{\text{nominal}}$

Likelihood, the full version (binned case)



Frequentist Constraints

Prototype: NP measured in a separate *auxiliary experiment*

e.g. luminosity measurement

→ Build the combined likelihood of the main+auxiliary measurements

$$L(\mu, \theta; \text{data}) = L_{\text{main}}(\mu, \theta; \text{main data}) L_{\text{aux}}(\theta; \text{aux. data})$$

Independent
measurements:
⇒ just a product

Gaussian form often used by default: $L_{\text{aux}}(\theta; \text{aux. data}) = G(\theta^{\text{obs}}; \theta, \sigma_{\text{syst}})$

In the combined likelihood, **systematic NPs are constrained**

→ now same as other NPs: **all uncertainties statistical in nature**

→ Often no clear setup for auxiliary measurements

e.g. theory uncertainties on missing HO terms from scale variations

→ **Implemented in the same way nevertheless** (“pseudo-measurement”)

Wilks' Theorem

The likelihood usually has NPs:

- **Systematics**
- Parameters fitted in data

→ What values to use when defining the hypotheses ? → $H(\mu=0, \theta=?)$

Answer: let the data choose ⇒ use the best-fit values (*Profiling*)

⇒ **Profile Likelihood Ratio** (PLR)

$$t_{\mu_0} = -2 \log \frac{L(\mu = \mu_0, \hat{\hat{\theta}}_{\mu_0})}{L(\hat{\mu}, \hat{\theta})}$$

$\hat{\hat{\theta}}_{\mu_0}$ best-fit value for $\mu = \mu_0$ (conditional MLE)
 $\hat{\theta}$ overall best-fit value (unconditional MLE)

Wilks' Theorem: PLR also follows a χ^2 ! $f(t_{\mu_0} | \mu = \mu_0) = f_{\chi^2(n_{dof}=1)}(t_{\mu_0})$
also with NPs present

→ Profiling “builds in” the effect of the NPs

⇒ Can treat the PLR as a **function of the POI only**

Systematics implementation

Prototype: NP measured in a separate *auxiliary experiment*

e.g. luminosity measurement

→ Build the combined likelihood of the main+auxiliary measurements

$$L(\mu, \theta; \text{data}) = L_{\text{main}}(\mu, \theta; \text{main data}) L_{\text{aux}}(\theta; \text{aux. data})$$

Independent
measurements:
⇒ just a product

Gaussian form often used by default: $L_{\text{aux}}(\theta; \text{aux. data}) = G(\theta^{\text{obs}}; \theta, \sigma_{\text{syst}})$

→ Often no clear setup for auxiliary measurements

e.g. theory uncertainties on missing HO terms from scale variations

→ **Implemented in the same way nevertheless** (“pseudo-measurement”)

Gaussian Profiling

Gaussian counting with systematic on background: $n = S + B + \theta$:

→ $n_{\text{obs}} \sim \mathbf{G}(S + B + \theta, \sigma_{\text{stat}})$

→ constraint $\mathbf{G}(\theta, \sigma_{\text{syst}})$ on θ

$$L(n; S, \theta) = G(n; S + B + \theta, \sigma_{\text{stat}}) G(\theta_{\text{obs}} = 0; \theta, \sigma_{\text{syst}})$$

Then: **MLE:** $\hat{S} = n - B, \hat{\theta} = 0$

$$\text{Conditional MLE: } \hat{\theta}(S) = \frac{\sigma_{\text{syst}}^2}{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2} (n - S - B)$$

$$\text{PLR: } \lambda(S) = \left(\frac{S + B - n}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right)^2$$

$$\sigma_{\mu} = \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

⇒ Statistical and systematic uncertainties add in quadrature as expected

Executive summary:

- **Systematic = NP with an external constraint** (auxiliary measurement)
- Profiling systematics includes their effect into the total uncertainty
- No special treatment for systematics: treated like any other NP, automatically accounted for through profiling.
- Guaranteed to work only as long as everything is Gaussian, but typically robust against non-Gaussian behavior.

Profiling example

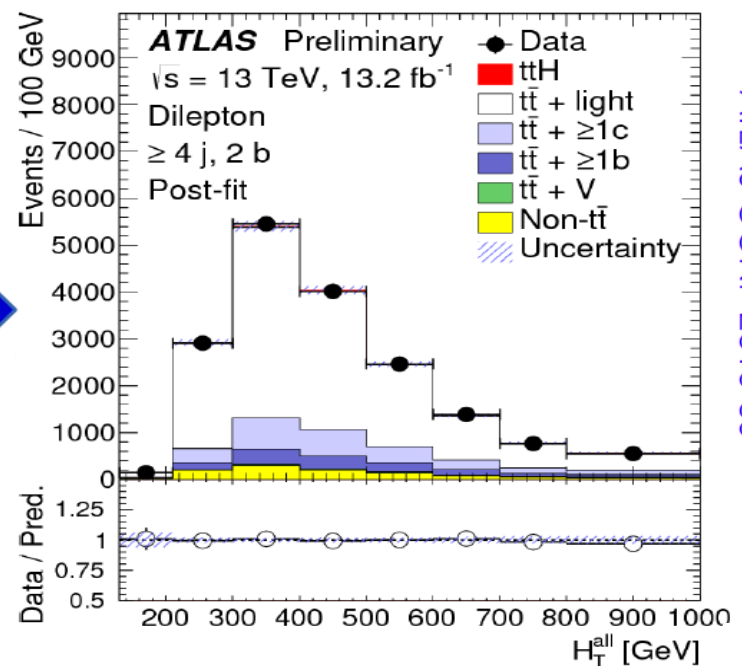
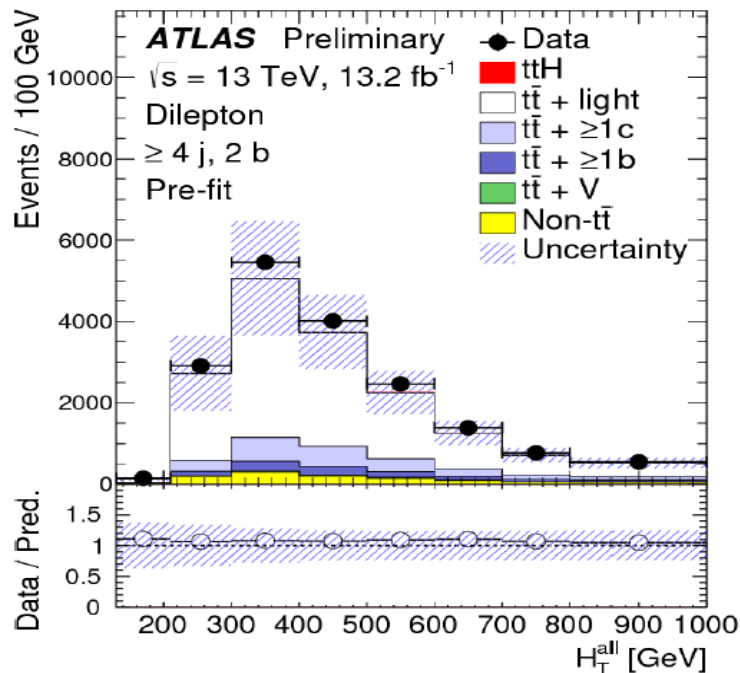
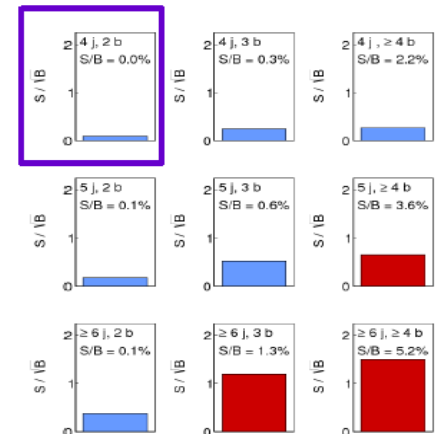
$t\bar{t}H \rightarrow b\bar{b}$

Analysis uses low-S/B categories to constrain backgrounds.

→ **Reduction in large uncertainties on $t\bar{t}$ bkg**

→ **Propagates to the high-S/B categories** through the statistical modeling

⇒ **Care needed in the propagation** (e.g. different kinematic regimes)

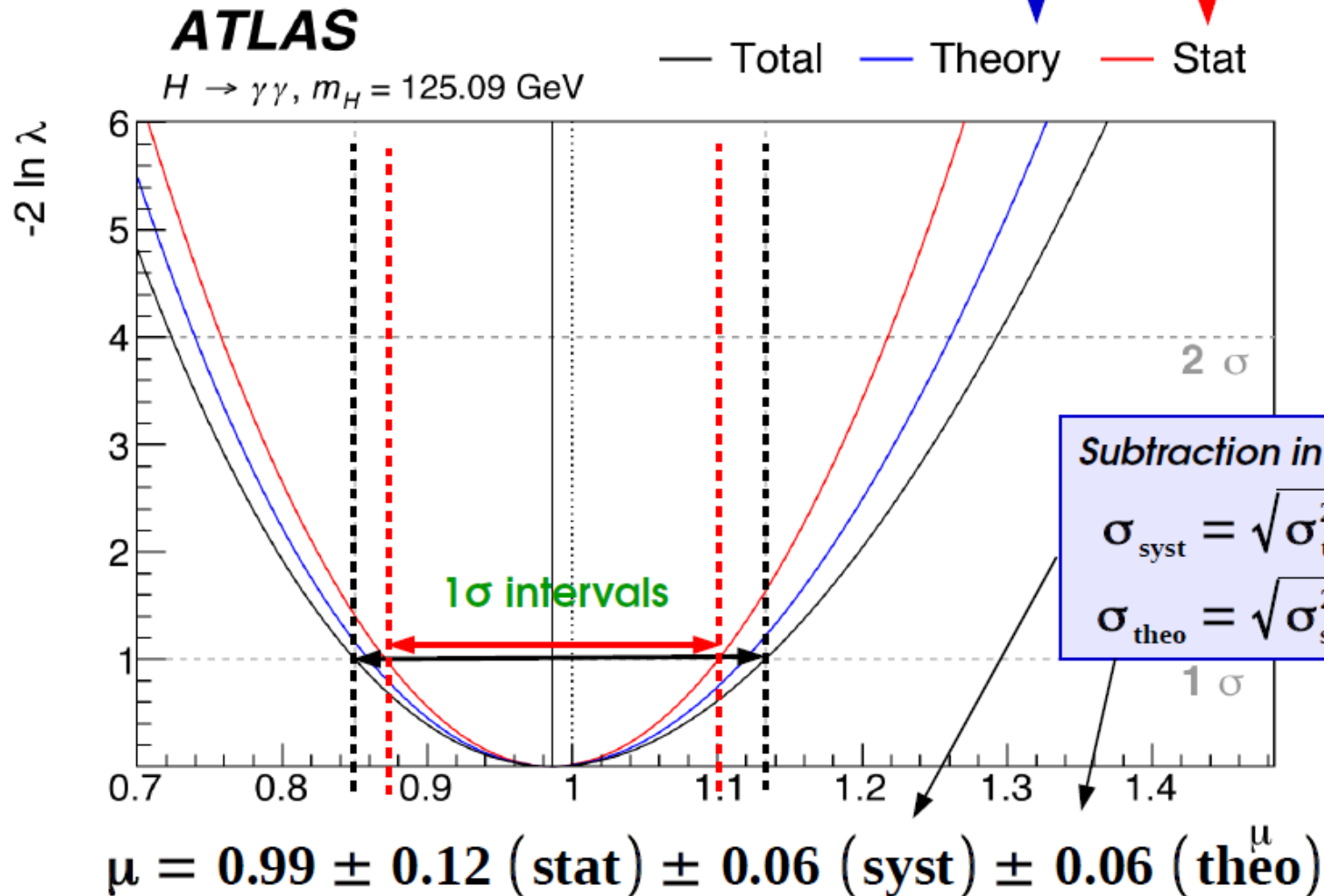


ATLAS-CONF-2016-08

Uncertainty decomposition

All systematics NPs fixed to 0 : statistical uncertainty only

exp. syst. NPs fixed to 0 : stat+theory uncertainty



Pull/Impact plots

Systematics are described by NPs included in the fit. Nominally:

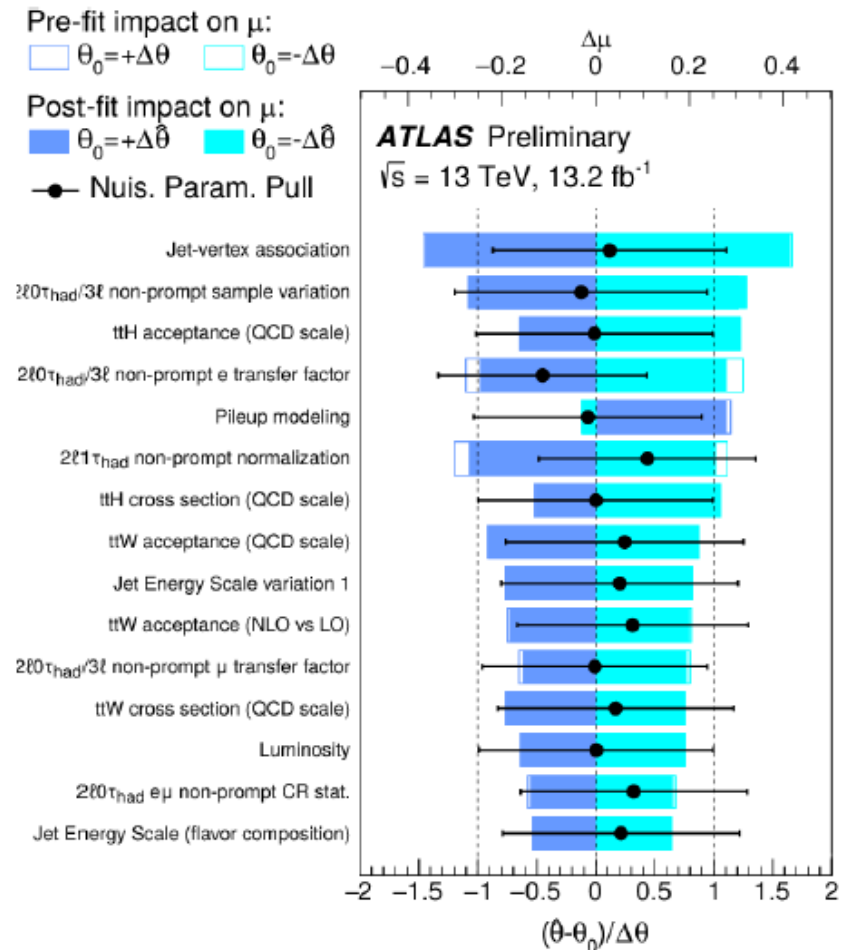
- **NP central value = 0** : corresponds to the pre-fit expectation (usually MC)
- **NP uncertainty = 1** : since NPs normalized to the value of the syst. :

$$N = N_0 (1 + \sigma_{\text{syst}} \theta), \theta \sim G(0, 1)$$

Fit results provide information on impact of the systematic on the result:

- **If central value $\neq 0$** : some data feature absorbed by nonzero value $\Rightarrow$ Need investigation if large pull
- **If uncertainty < 1** : systematic is constrained by the data $\Rightarrow$ Needs checking if this legitimate or a modeling issue
- **Impact on result** of $\pm 1\sigma$ shift of NP

ATLAS-CONF-2016-058



Pull/Impact plots

Systematics are described by NPs included in the fit. Nominally:

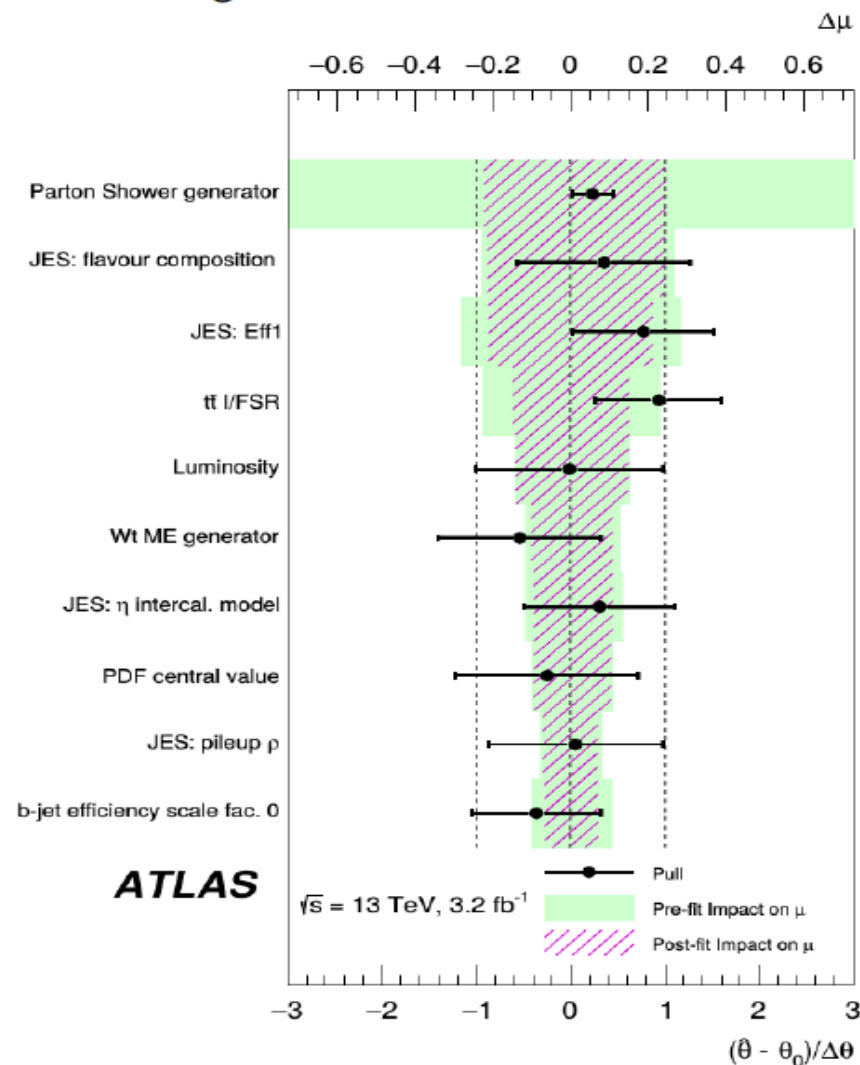
- **NP central value = 0** : corresponds to the pre-fit expectation (usually MC)
- **NP uncertainty = 1** : since NPs normalized to the value of the syst. :

$$N = N_0 (1 + \sigma_{\text{syst}} \theta), \theta \sim G(0, 1)$$

Fit results provide information on impact of the systematic on the result:

- **If central value $\neq 0$** : some data feature absorbed by nonzero value
 $\Rightarrow$ Need investigation if large pull
- **If uncertainty < 1** : systematic is constrained by the data
 $\Rightarrow$ Needs checking if this legitimate or a modeling issue
- **Impact on result** of $\pm 1\sigma$ shift of NP

13 TeV single- t XS ([arXiv:1612.07231](https://arxiv.org/abs/1612.07231))



Takeaways

Systematics: uncertainties on the **form of the statistical model**

(as opposed to the uncertainties encoded in the model itself)

→ Implemented using additional nuisance parameters in the model

→ Constrained by adding **auxiliary measurements** (sometimes fictitious ones) to the model – usually represented by a single Gaussian for each NP.

$$L(\mu, \theta; \text{data}) = L_{\text{main}}(\mu, \theta; \text{main data}) G(\theta^{\text{obs}}, \theta, 1)$$

⇒ **Systematics treated in the same way as statistical uncertainties**, although we still keep track of **systematics NPs** for bookkeeping purposes

Profiling: when testing a hypothesis, use the best-fit values of the nuisance parameters: **profile likelihood ratio**.

$$\frac{L(\mu = \mu_0, \hat{\theta}_{\mu_0})}{L(\hat{\mu}, \hat{\theta})}$$

Wilks' Theorem: the PLR has the same asymptotic properties as the LR without systematics: can profile out NPs and just deal with POIs.

→ NPs still show up in the PLR as increased uncertainties – Gaussian case:

$$\sigma_{\text{total}} = \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

Profiling can have unintended effects – need to carefully check behavior

Summary on Statistical Results Computation

Methods provide:

→ **Optimal use of information from the data under general hypotheses**

→ **Arbitrarily complex/realistic models (up to computing constraints...)**

→ **No Gaussian assumptions in the measurements**

Still often assume Gaussian behavior of PLR – but weaker assumption and can be lifted with toys

Systematics treated as auxiliary measurements – modeling can be tailored as needed

→ **Single PLR-based framework for all usual classes of measurements**

Discovery testing

Upper limits on signal yields

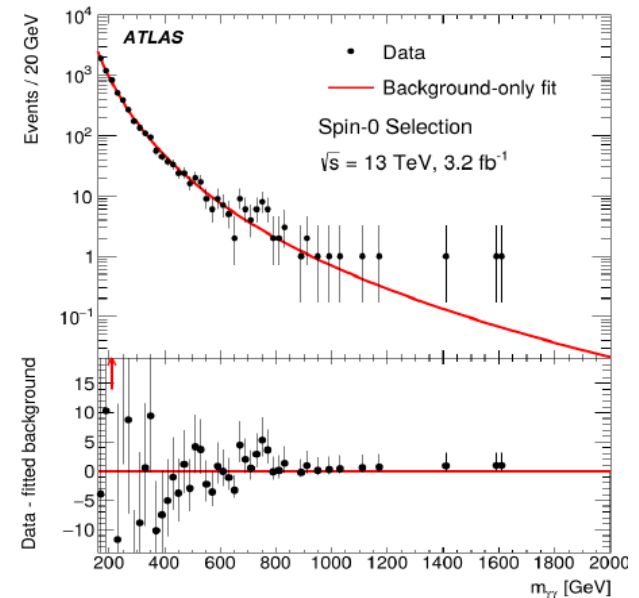
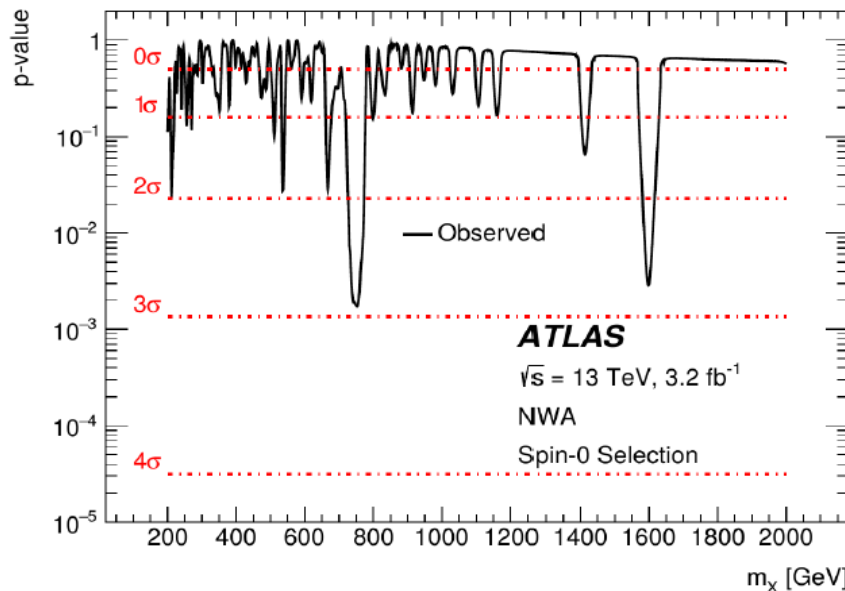
Parameter estimation

Look-Elsewhere effect

Sometimes, unknown parameters in signal model
e.g. p-values as a function of m_X

⇒ Effectively: **multiple, simultaneous searches**

→ If e.g. small resolution and large scan range,
many independent experiments



→ More likely to find an excess
anywhere in the range, rather
than in a **predefined** location
⇒ **Look-elsewhere effect** (LEE)

Global Significance

Probability for a fluctuation **anywhere** in the range → **Global** p-value.
 at a given location → **Local** p-value

$$\begin{array}{ccc}
 \boxed{\text{Global p-value}} & \xrightarrow{\quad} & p_{\text{global}} = 1 - (1 - p_{\text{local}})^N \approx N p_{\text{local}} \xleftarrow{\quad} \boxed{\text{Local p-value}} \\
 & & \uparrow \\
 & & \boxed{\text{Trial factor}}
 \end{array}$$

→ $p_{\text{global}} > p_{\text{local}} \Rightarrow Z_{\text{global}} < Z_{\text{local}}$ – global fluctuation more likely ⇒ less significant

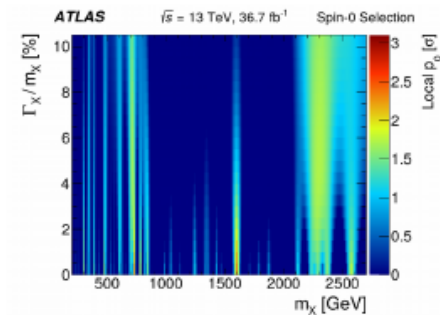
Trial factor :  **naïvely** = # of independent intervals: $N_{\text{trials}} = N_{\text{indep}} = \frac{\text{scan range}}{\text{peak width}}$??
 However this is usually **wrong** – more on this later

For searches over a parameter range, p_{global} **is the relevant p-value**

→ Depends on the scanned parameter ranges

e.g. $X \rightarrow \gamma\gamma$: $200 < m_X < 2000 \text{ GeV}$, $0 < \Gamma_X < 10\% m_X$.

→ However what comes out of the usual asymptotic formulas is p_{local} .

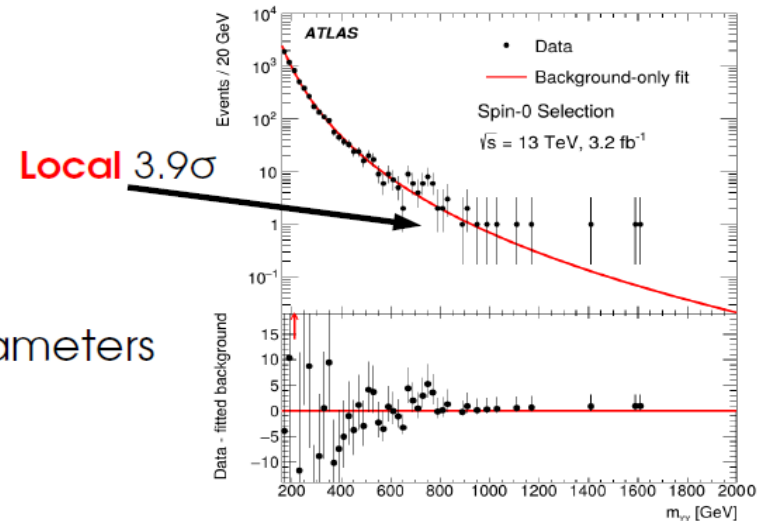


How to compute p_{global} ? → **Toys** (brute force) or **asymptotic formulas**.

Global Significance from Toys

Principle: repeat the analysis in toy data:

- generate pseudo-dataset
- perform the search, scanning over parameters as in the data
- report the largest significance found
- repeat many times



⇒ The frequency at which a given Z_0 is found **is** the global p-value

e.g. **$X \rightarrow \gamma\gamma$ Search:** $Z_{\text{local}} = 3.9\sigma$ ($\Rightarrow p_{\text{local}} \sim 5 \cdot 10^{-5}$),

→ However we are scanning $200 < m_X < 2000 \text{ GeV}$ and $0 < \Gamma_X < 10\% m_X$!

→ Toys : find such an excess **2%** of the time somewhere in the range

⇒ $p_{\text{global}} \sim 2 \cdot 10^{-2}$, $Z_{\text{global}} = 2.1\sigma$ Less exciting, and better indication of true Z!

⊕ **Exact treatment**

⊖ **CPU-intensive** especially for large Z (need $\sim O(100)/p_{\text{global}}$ toys)

Global Significance from Asymptotics

Principle: approximate the global p-value in the asymptotic limit

→ reference paper: **Gross & Vitells, EPJ.C70:525-530,2010**

Asymptotic trials factor (1 POI):

$$N_{\text{trials}} = 1 + \sqrt{\frac{\pi}{2}} N_{\text{indep}} Z_{\text{local}}$$

$N_{\text{indep}} = \frac{\text{scan range}}{\text{peak width}}$

→ Trials factor is **not just** N_{indep} ,
also depends on Z_{local} !

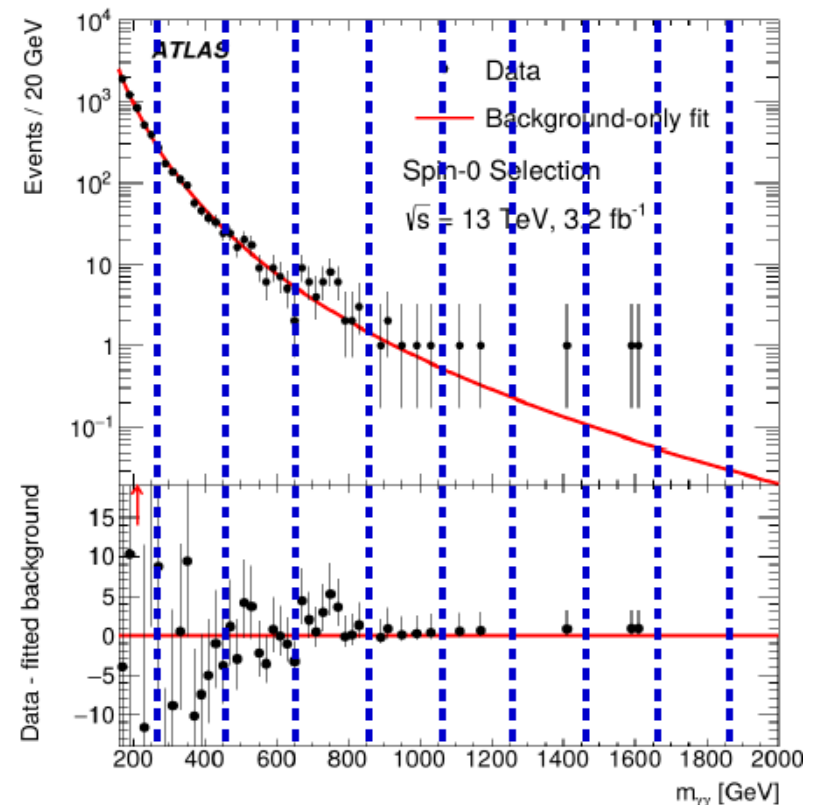
Why ?

→ slice scan range into N_{indep} regions
of size $\sim$ peak width
→ search for a peak in each region

⇒ Indeed gives $N_{\text{trials}} = N_{\text{indep}}$.

However this misses peaks sitting on
edges between regions

⇒ true N_{trials} is **>** N_{indep} !



Global Significance from Asymptotics

Principle: approximate the global p-value in the asymptotic limit

→ reference paper: **Gross & Vitells, EPJ.C70:525-530,2010**

Asymptotic trials factor (1 POI):

$$N_{\text{trials}} = 1 + \sqrt{\frac{\pi}{2}} N_{\text{indep}} Z_{\text{local}}$$

$N_{\text{indep}} = \frac{\text{scan range}}{\text{peak width}}$

→ Trials factor is **not just** N_{indep} ,
also depends on Z_{local} !

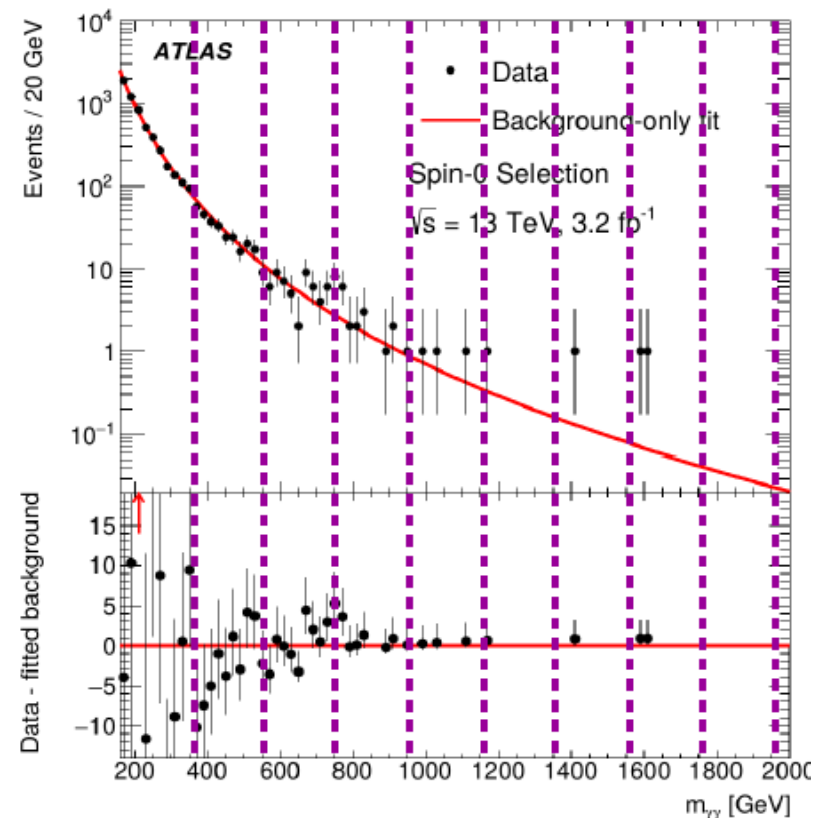
Why ?

→ slice scan range into N_{indep} regions
of size $\sim$ peak width
→ search for a peak in each region

⇒ Indeed gives $N_{\text{trials}} = N_{\text{indep}}$.

However this misses peaks sitting on
edges between regions

⇒ true N_{trials} is **>** N_{indep} !



Illustrative Example (1)

Test on a simple example: generate toys with

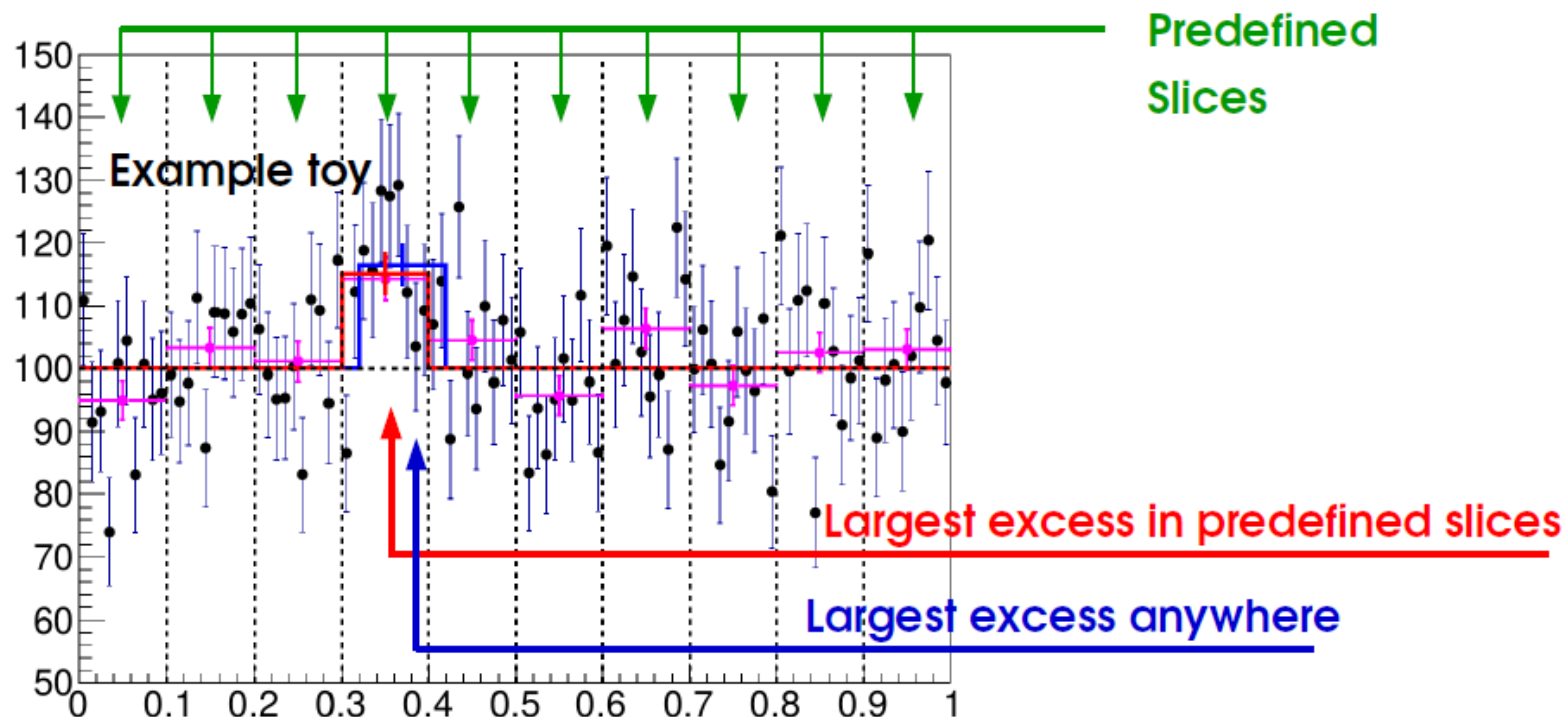
→ flat background (100 events/bin)

→ count events in a fixed-size sliding window, look for excesses

Two configurations:

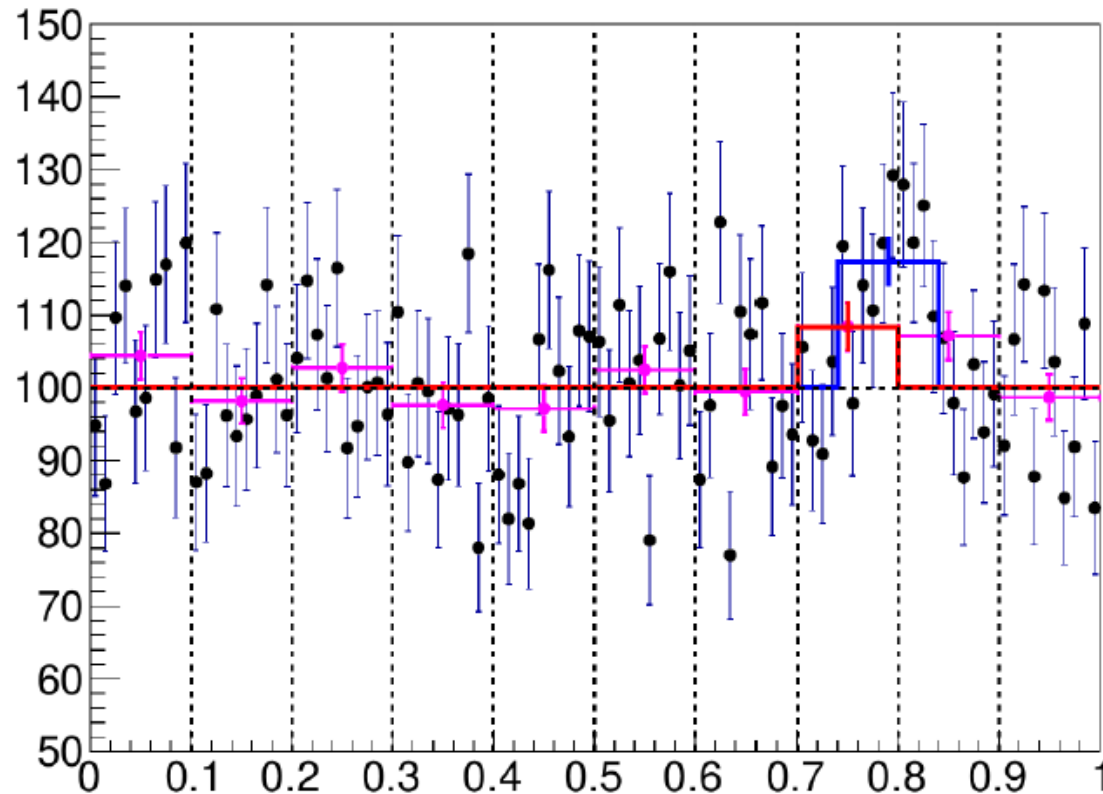
1. Look only in 10 slices of the full spectrum

2. Look in any window of same size as above, anywhere in the spectrum



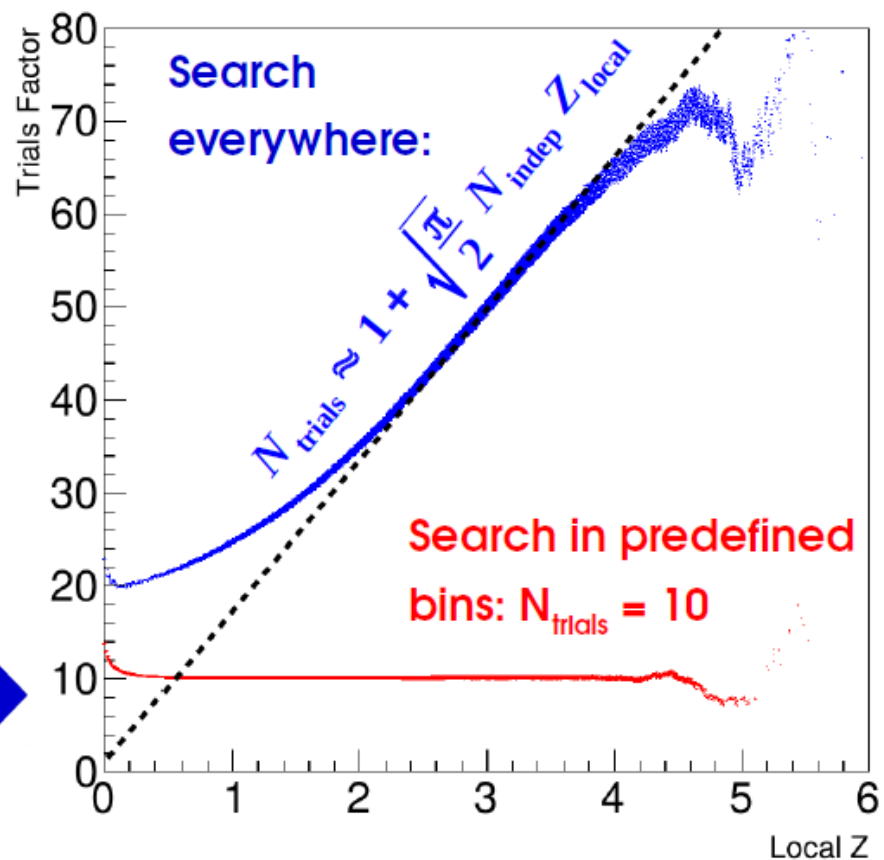
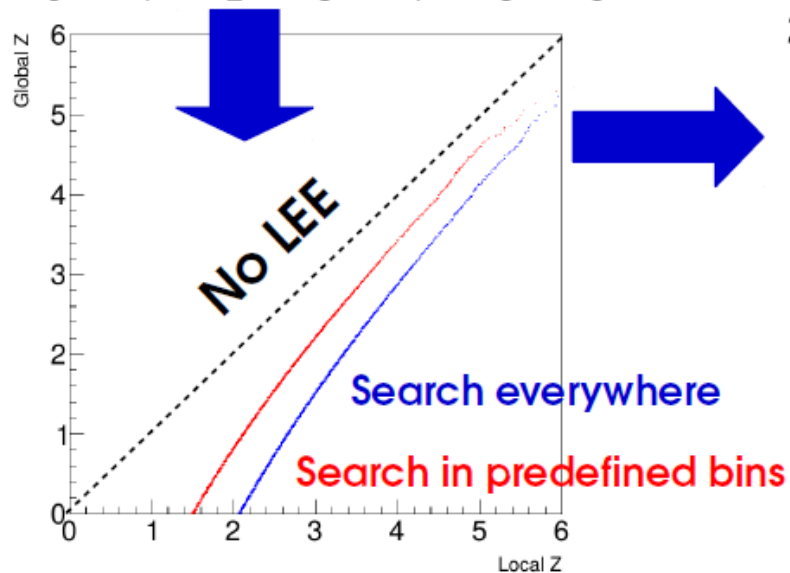
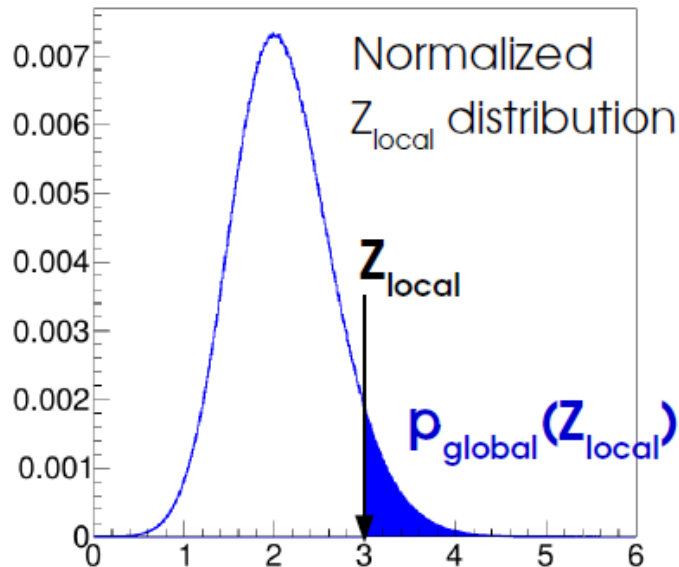
Illustrative Example (2)

Very different results if the excess is **near a boundary** :



1. Look only in 10 slices of the full spectrum
2. Look in any window of same size as above, anywhere in the spectrum

Illustrative Example (3)



Searching everywhere gives the extra Z_{local} dependence

Z_{Global} Asymptotics Extrapolation

Asymptotic trials factor (1 POI): $N_{\text{trials}} = 1 + \sqrt{\frac{\pi}{2}} N_{\text{indep}} Z_{\text{local}}$

How to get N_{indep} ? Usually work with a slightly different formula:

$$N_{\text{trials}} = 1 + \frac{1}{p_{\text{local}}} \langle N_{\text{up}}(Z_{\text{test}}) \rangle e^{\frac{Z_{\text{local}}^2 - Z_{\text{test}}^2}{2}}$$


 Number of excesses with $Z > Z_{\text{test}}$

→ **Get N_{up} From toys ? but high $Z_{\text{local}} \Rightarrow$ many toys needed**

⇒ **calibrate for small Z_{test} , apply result to higher Z_{local}**

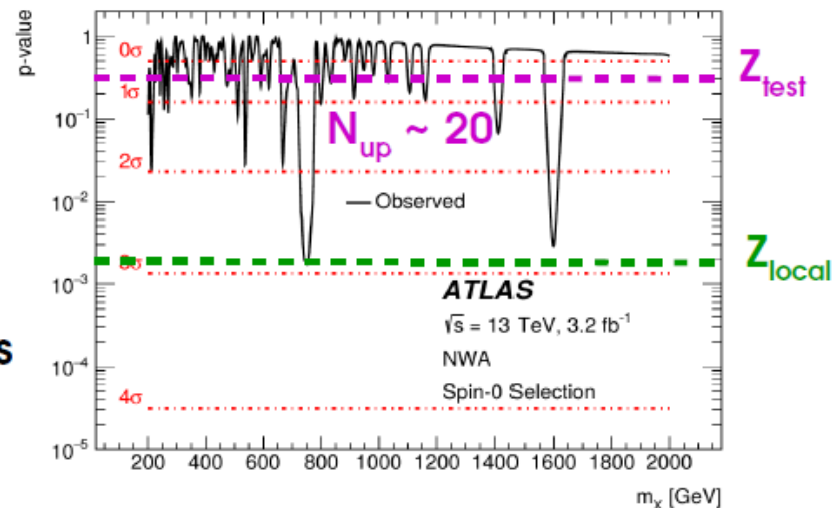
Can choose arbitrarily small Z_{test}

⇒ many excesses

⇒ can measure N_{up} in data (1 “toy”)

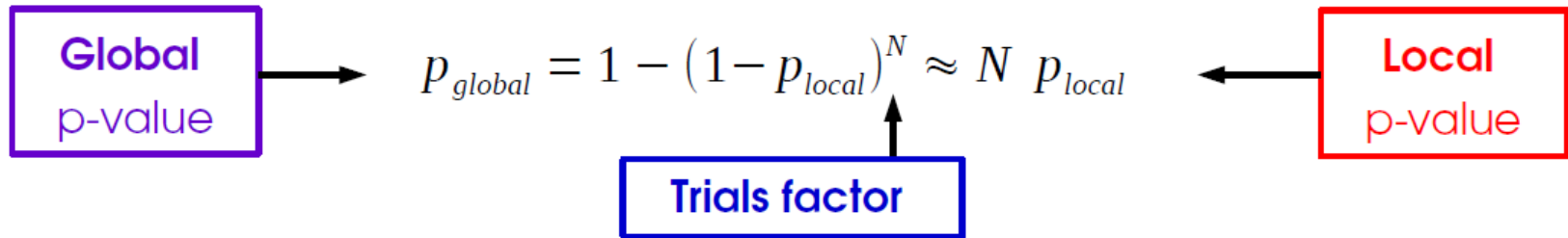
Can also measure $\langle N_{\text{up}} \rangle$ in multiple toys

if large stat uncertainty from
too few excesses



Trials factor

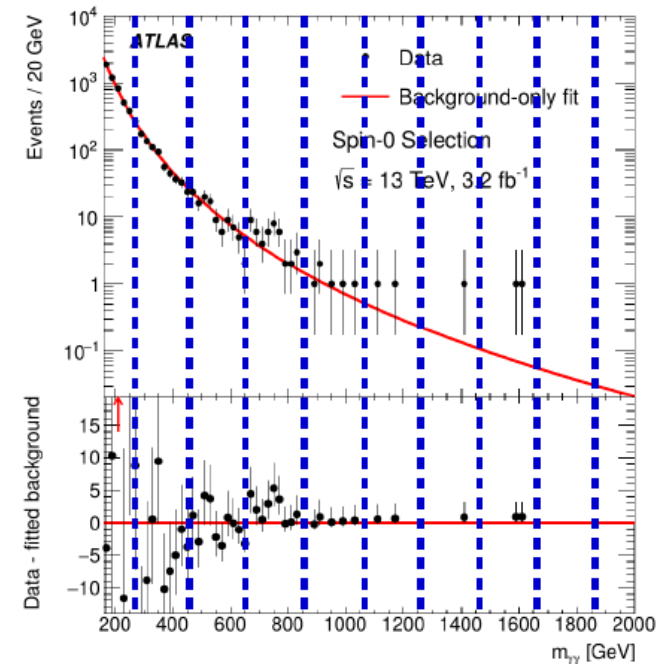
Trials factor N = # of independent searches:



Naively, one could expect

$$N_{\text{trials}} = ?? N_{\text{indep}} = \frac{\text{scan range}}{\text{peak width}}$$

However this is usually **wrong** !



Frequentist vs. Bayesian

All methods described so far are **frequentist**

- Probabilities (p-values) refer to outcomes if the experiment were **repeated identically many times**
- Parameters value are **fixed but unknown**

• Probabilities apply to measurements:

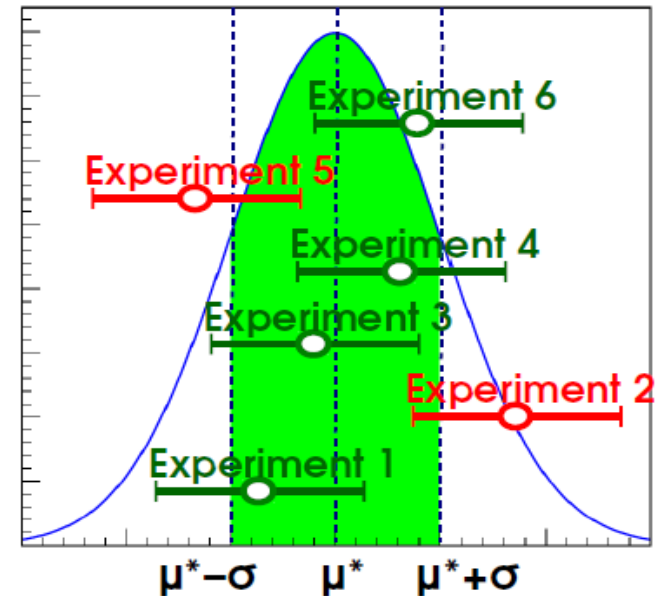
→ “ $m_H = 125.09 \pm 0.24 \text{ GeV}$ ” :

→ i.e. $[125.09 - 0.24 ; 125.09 + 0.24] \text{ GeV}$ has $p=68\%$ to contain **the** true m_H .

→ if we repeated the experiment many times, we would get different intervals, 68% of which would contain the true m_H .

→ “**5 σ Higgs discovery**”

- if there is really no Higgs, such fluctuations observed in 3.10^{-7} of experiments



Not exactly the crucial question – what we would really like to know is

What is the probability that the excess we see is a fluctuation

→ we want **P(no Higgs | data)** – but all we have is **P(data | no Higgs)**

Frequentist vs. Bayesian

Can use **Bayes' theorem** to address this:

$$P(\mu|data) = \frac{P(data|\mu)}{P(data)} P(\mu)$$

same as in the frequentist formalism (=likelihood)

Prior Probability

irrelevant normalization factor

Can compute $P(\mu | data)$, **if we provide $P(\mu)$**

→ Implicitly, we have now made μ into a random variable

- Is m_H , or the presence of H(125), randomly chosen ?
- In fact, different definition of p: **degree of belief**, not from frequencies.
- $P(\mu)$ **Prior degree of belief** – critical ingredient in the computation

Compared to frequentist PLR:

- ⊕ answers the “right” question
- ⊖ answer depends on the prior

“Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone.” - **Louis Lyons**

Bayesian methods

Probability distribution (= likelihood) : same form as frequentist case, but

P(θ) constraints now **priors for the systematics NPs**, $P(\theta)$

not auxiliary measurements $P(\theta^{\text{mes}}; \theta)$

⊕ Simply integrate them out, no need for profiling: $P(\mu) = \int P(\mu, \theta) d\theta$

→ Use probability distribution $P(\mu)$ directly for limits, credibility intervals

e.g. define 68% CL (“Credibility Level”) interval (A, B) by: $\int_A^B P(\mu) d\mu = 68\%$

⊖ No simple way to test for discovery

⊖ Integration over NPs can be CPU-intensive

Priors : most analyses still using flat priors in the analysis variable(s)

⇒ **Parameterization-dependent**: if flat in $\sigma \times B$, then not flat in $\kappa \dots$

→ Can use the Jeffreys’ or reference priors, but difficult in practice

Frequentist-Bayesian Hybrid methods (“Cousins-Highland”)

- Integrate out NPs as in Bayesian measurements
- Once only POIs left, Use $P(\text{data} | \mu)$ in a frequentist way
 - “Bayesian NPs, frequentist POIs”
- Some use in Run 1, now phased out in favor of frequentist PLR.

Frequentists method: CL_s computation

Gaussian counting with systematic on background: $n = S + B + \sigma_{\text{syst}} \theta$

$$L(n; S, \theta) = G(n; S + B + \sigma_{\text{syst}} \theta, \sigma_{\text{stat}}) G(\theta_{\text{obs}} = 0; \theta, 1)$$

$$\text{MLE: } \hat{S} = n - B$$

$$\text{Conditional MLE: } \hat{\theta}(S) = \frac{\sigma_{\text{syst}}}{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2} (n - S - B)$$

$$\text{PLR: } \lambda(S) = \left(\frac{S + B - n}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right)^2$$

Gaussian $\Rightarrow$ from previous studies, CL_s limit is

$$\text{CL}_s: S_{\text{up}}^{\text{CL}_s} = n - B + \left[\Phi^{-1} \left(1 - 0.05 \Phi \left(\frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right) \right] \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

Bayesian method: Bayesian limit

Gaussian counting with systematic on background: $n = S + B + \sigma_{\text{syst}} \theta$

$$P(n | S, \theta) = G(n; S + B + \sigma_{\text{syst}} \theta, \sigma_{\text{stat}}) G(\theta | 0, 1)$$

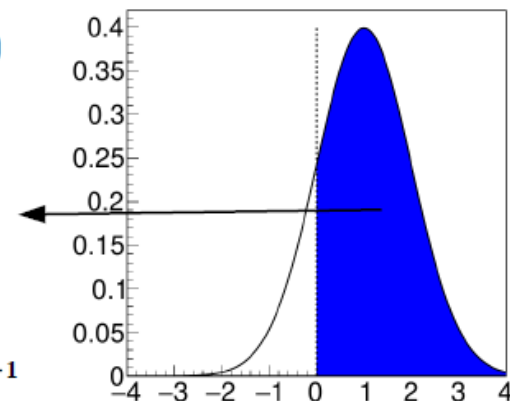
Bayesian: $G(\theta)$ is actually a **prior** on $\theta \Rightarrow$ perform integral (**marginalization**)

$$P(n | S) = G(S; n - B, \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}) \quad \text{same effect as profiling!}$$

Need $P(S | n) \Rightarrow$ a prior for S – take flat PDF over $S > 0$

$\Rightarrow$ Truncate Gaussian at $S=0$: $P(S | n) = P(n | S) P(S)$

$$P(S | n) = G(S; n - B, \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}) \left[\Phi \left(\frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right]^{-1}$$



Bayesian Limit:

$$\int_{S_{\text{up}}}^{\infty} P(S | n) dS = 5\% = \left[1 - \Phi \left(\frac{S_{\text{up}} - (n - B)}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right] \left[\Phi \left(\frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right]^{-1}$$

$$S_{\text{up}}^{\text{Bayes}} = n - B + \left[\Phi^{-1} \left(1 - 0.05 \Phi \left(\frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right) \right] \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

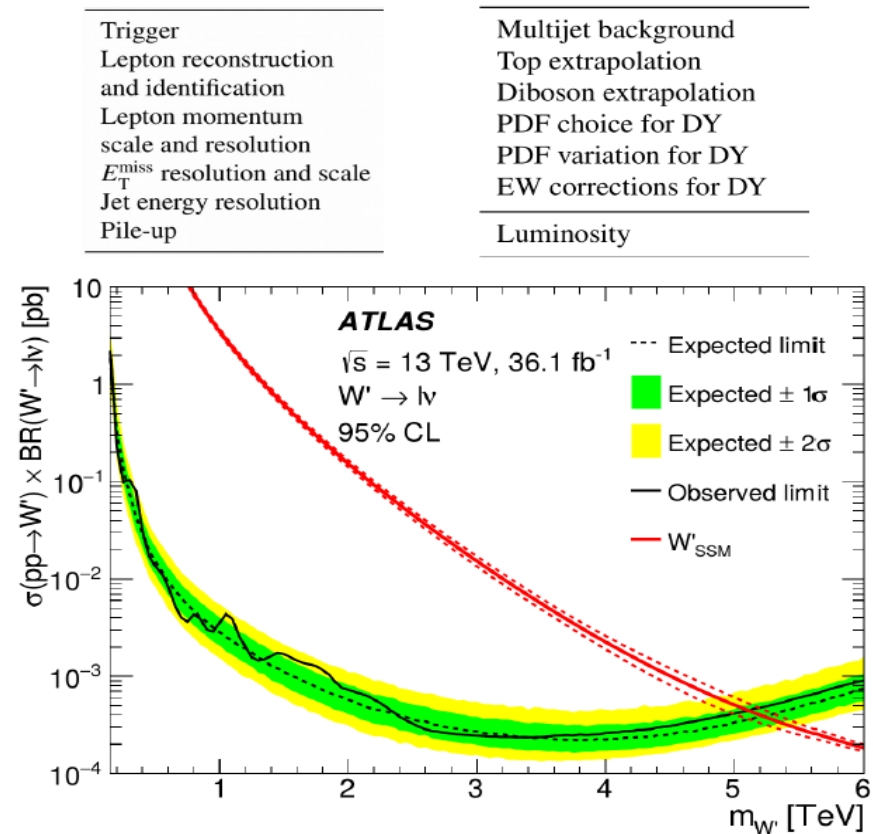
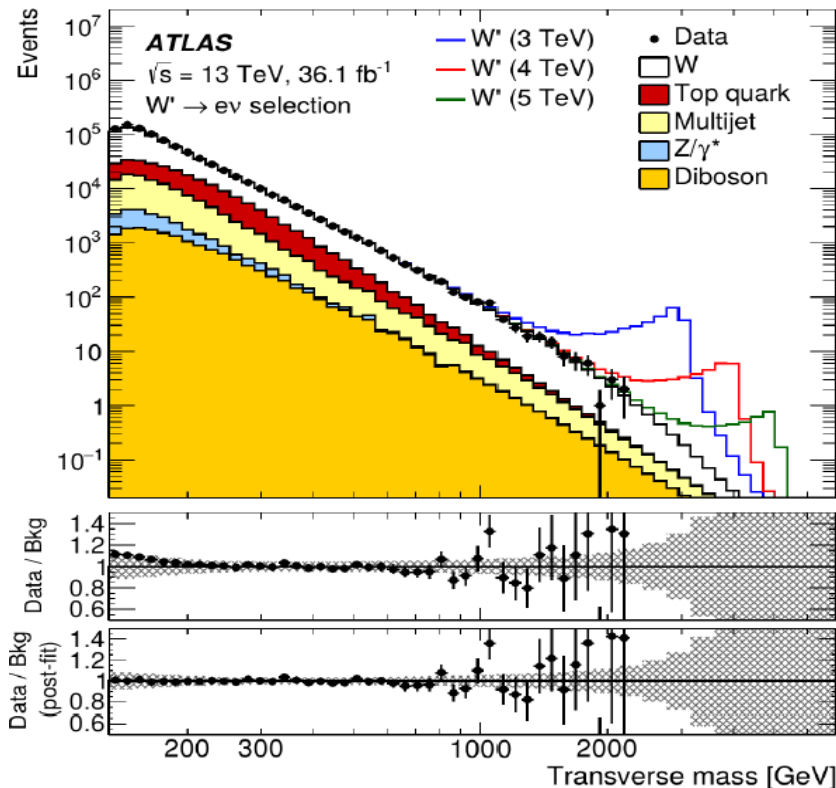
same result as CLs!

Bayesian methods

Example: $W' \rightarrow l\nu$ Search

arXiv:1706.04786

- **POI:** $W' \sigma \times B$ → use **flat prior** over $[0, +\infty[$.
- **NPs:** syst on **signal ϵ** (6 NPs), **bkg** (6), **lumi** (1) → integrate over Gaussian priors



48

Why 5σ ?

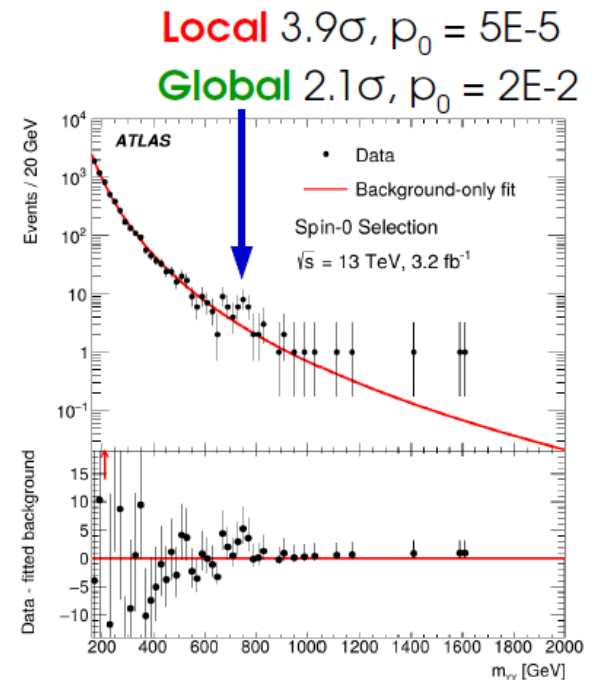
One-sided discovery: $5\sigma \Leftrightarrow p_0 = 3 \cdot 10^{-7} \Leftrightarrow 1 \text{ chance in } 3.5\text{M}$

→ Overly conservative ?

→ Do we even control such small probabilities ?

Reasons for sticking with 5σ (from Louis Lyons):

- **LEE** : searches typically cover multiple independent regions
⇒ Global p-value is the relevant one
 $N_{\text{trials}} \sim 1000$: local $5\sigma \Leftrightarrow \mathcal{O}(10^{-4})$ more reasonable
- **Mismodeled systematics**: factor 2 error in syst-dominated analysis ⇒ factor 2 error on Z...
- **History**: 3σ and 4σ excesses do occur regularly, for the reasons above
- **“Subconscious Bayes Factor”** : p-value should be at least as small as the subjective $p(S)$:



Extraordinary claims require extraordinary evidence

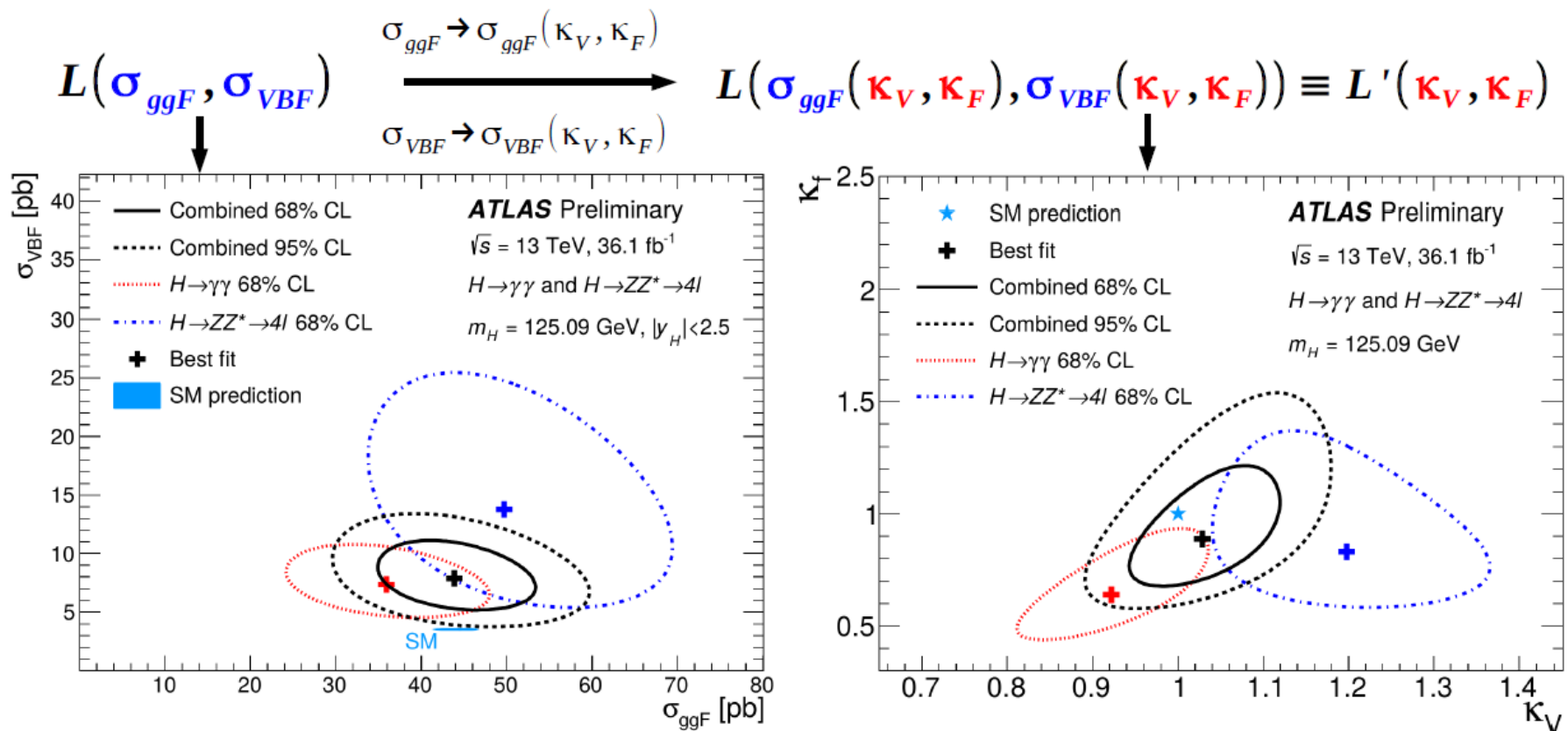
⇒ Stay with 5σ ...

Reparametrisation

Start with basic measurement in terms of e.g. $\sigma \times \mathbf{B}$

→ How to measure derived quantities (couplings, parameters in some theory model, etc.) ? → **just reparameterize the likelihood:**

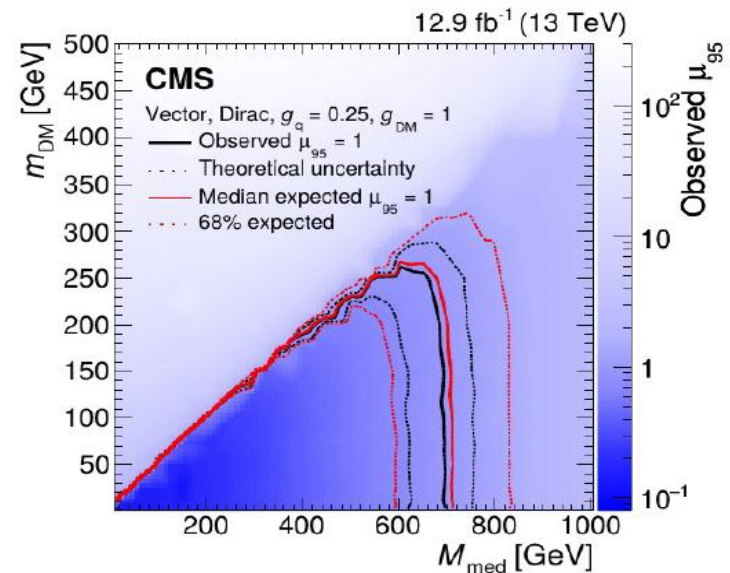
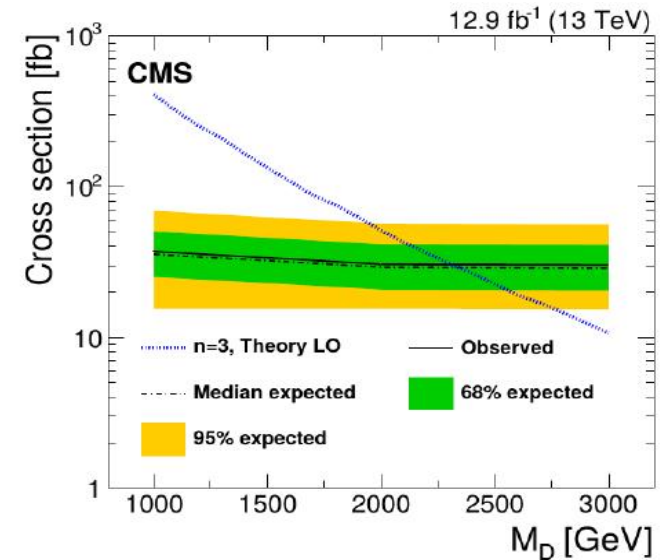
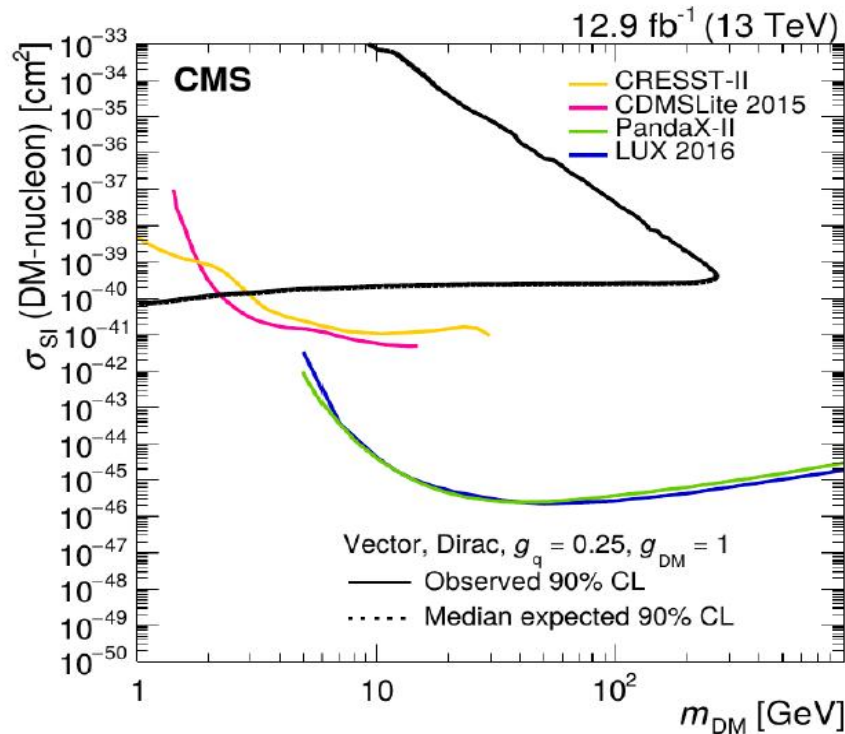
e.g. Higgs couplings: σ_{ggF} , σ_{VBF} sensitive to Higgs coupling modifiers κ_V , κ_F .



Reparametrization: Limits

Reparameterization: Limits

CMS Run 2 Monophoton Search: measured N_s in a counting experiment reparameterized according to various DM models

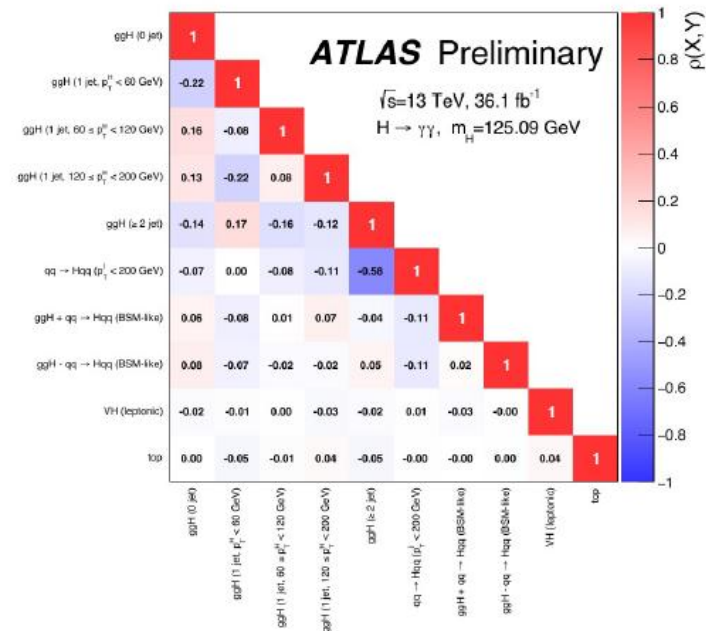
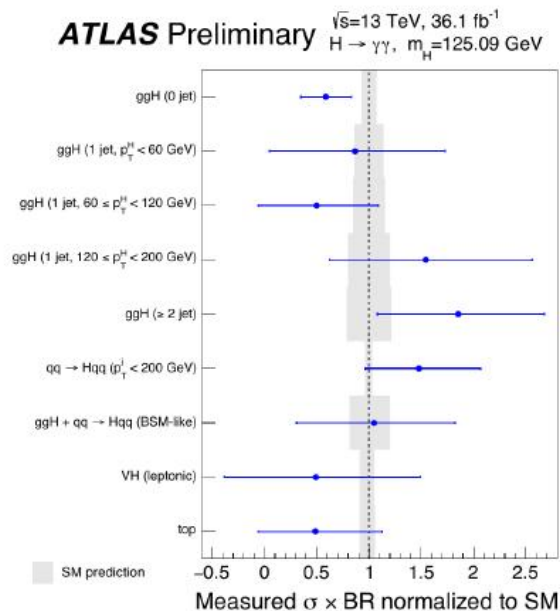


Presentation of results

→ Cannot test every model : need to make enough information public so that others (theorists) are able to do it independently

⇒ **Gaussian case**: sufficient to provide measurements + covariance matrix

→ For example using the [HEPData](#) repository.



Non-Gaussian case: no simple method

Conclusions

- Significant evolution in the statistical methods used in HEP
- Variety of methods, adapted to various situations and target results
- Allow to
 - model the statistical process with high precision in difficult situations (large systematics, small signals)
 - make optimal use of available information
- Implemented in standard RooFit/RooStat toolkits within the ROOT framework, as well as other tools (BAT)
- Still many open questions and areas that could use improvement
→ e.g. how to present results with all available information