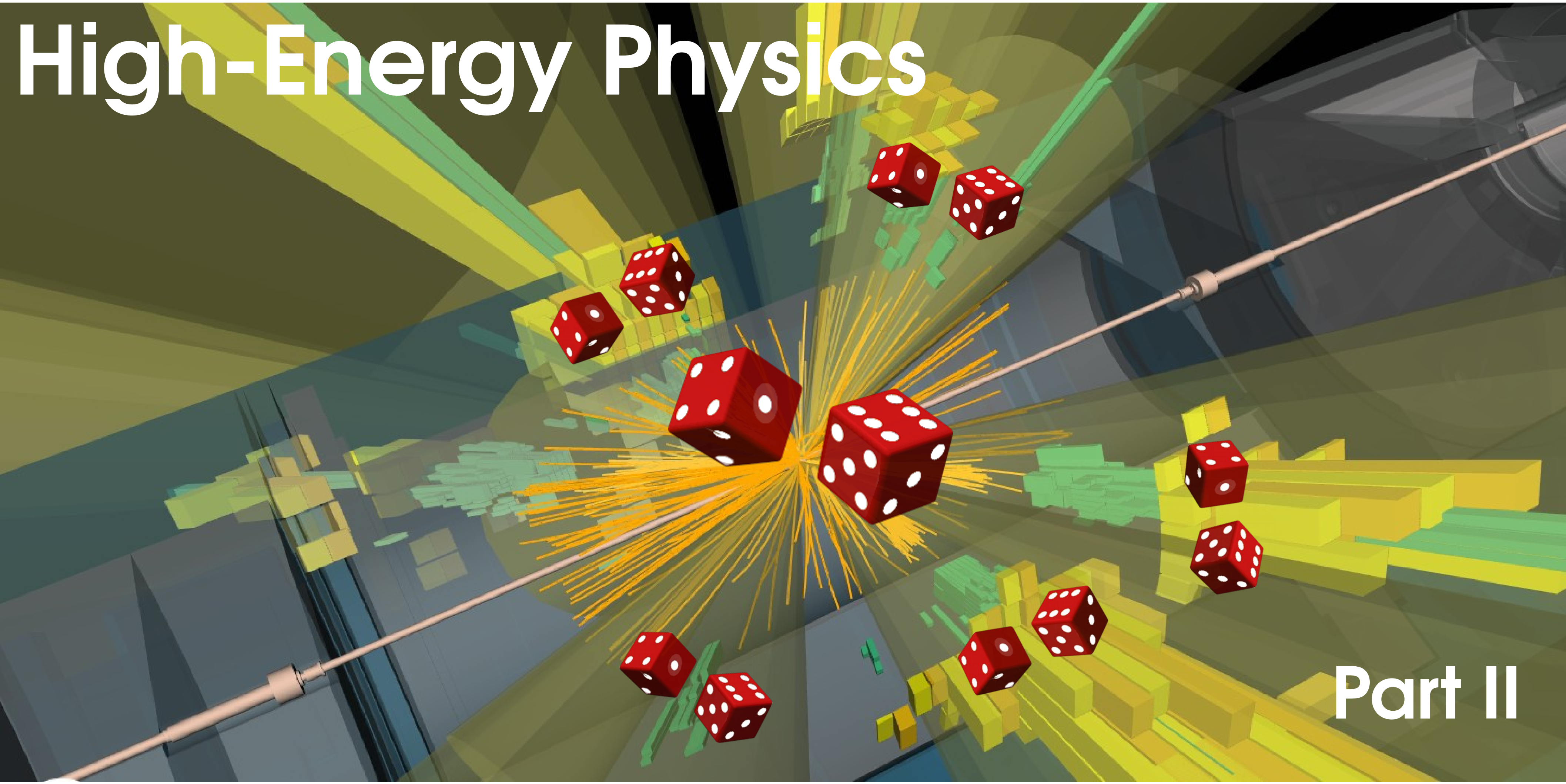
Statistical analysis methods in High-Energy Physics



Part II

Nicolas Berger (LAPP Annecy)

Computing Statistical Results
III. Discovery

(Continued from yesterday)

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

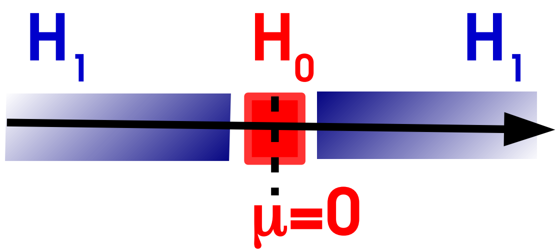
One-sided vs. Two-Sided

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

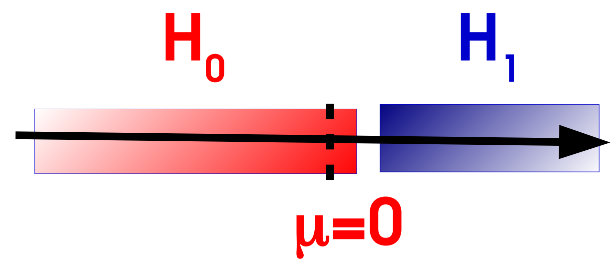
⇒ Change statistic so that $\hat{S} < 0$ ⇒ $t_0 = 0$ (perfect agreement with H_0, as for $\hat{S} = 0$)

Two-sided



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

One-sided



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

Test Statistic

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

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

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

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

⇒ Same Z in both cases for a given signal S

One-Sided Asymptotics

→ **One-sided test**:

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

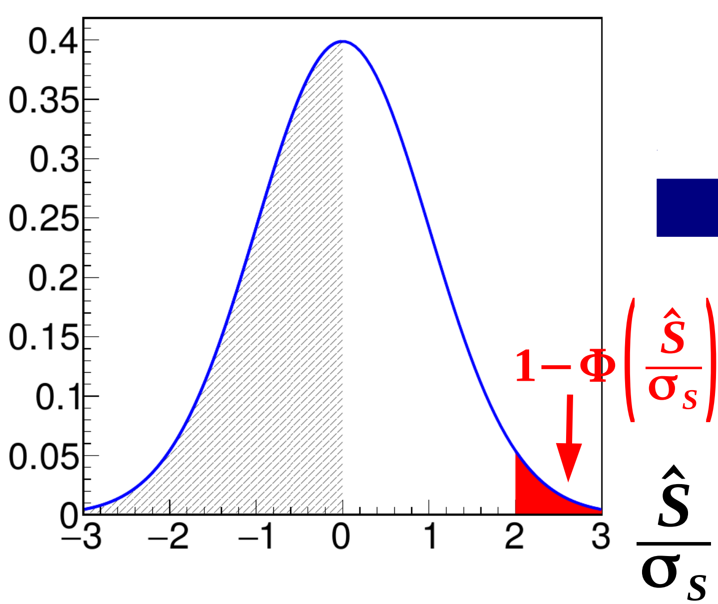
Asymptotics: "half-χ²" distribution:

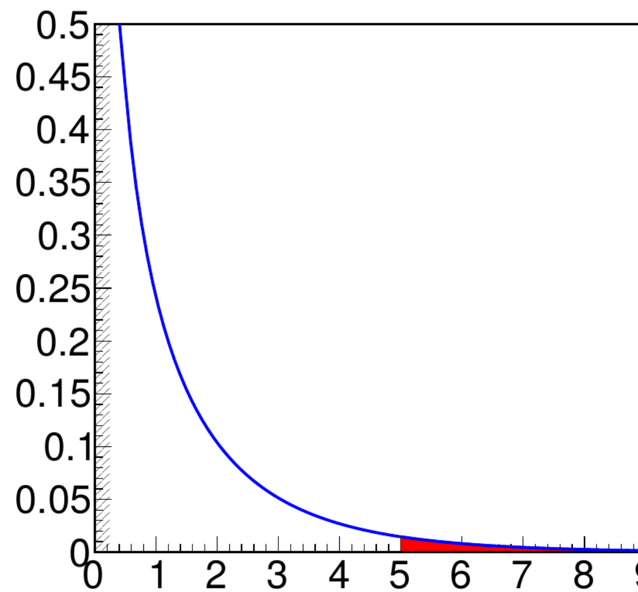
$$f(q_0 \mid S=0) = \frac{1}{2}\delta(q_0) + \frac{1}{2} f_{\chi^2(n_{dof}=1)}(q_0)$$

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

Φ : normal CDF

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





$$q_0 = \left(\frac{\hat{S}}{\sigma_S}\right)^2$$

Example: Gaussian Counting

Count number of events n in data

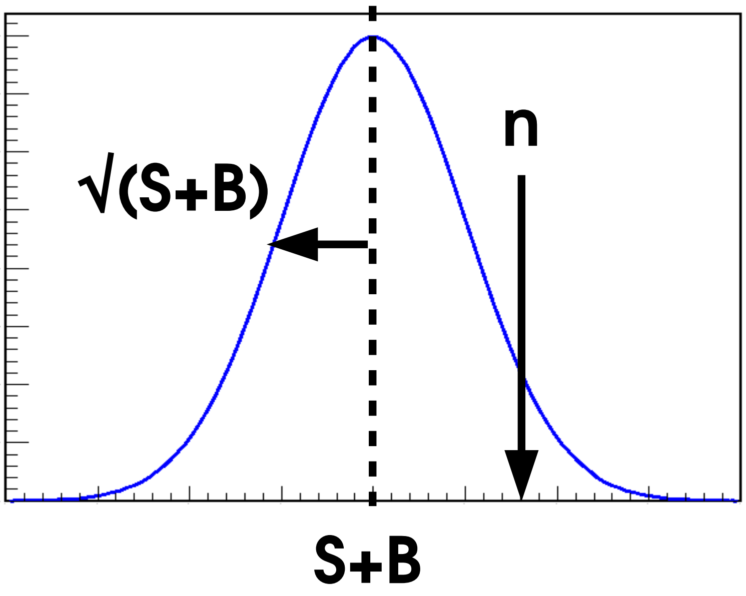
→ assume n large enough so process is Gaussian

→ assume B is known, measure S

Likelihood :

$$L(S) = e^{-\frac{1}{2}\left(\frac{n-(S+B)}{\sqrt{S+B}}\right)^2}$$

$$\lambda(S) = \left(\frac{n-(S+B)}{\sqrt{S+B}}\right)^2$$



MLE for S : **$\hat{S}$ = n – B**

Test statistic: assume $\hat{S} > 0$,

$$q_0 = -2\log\frac{L(S=0)}{L(\hat{S})} = \lambda(S=0) - \lambda(\hat{S}) = \left(\frac{n-B}{\sqrt{B}}\right)^2 = \left(\frac{\hat{S}}{\sqrt{B}}\right)^2$$

Finally:

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

Known formula!

→ Strictly speaking only valid in Gaussian regimge

Example: Poisson Counting

Same problem but now not assuming Gaussianity

$$L(S) = e^{-(S+B)}(S+B)^n \qquad \lambda(S) = 2(S+B) - 2n\log(S+B)$$

MLE: **$\hat{S} = n - B$**, same as Gaussian

Test statistic (for $\hat{S} > 0$): $$q_0 = \lambda(S=0) - \lambda(\hat{S}) = -2\hat{S} - 2(\hat{S}+B)\log\frac{B}{\hat{S}+B}$$

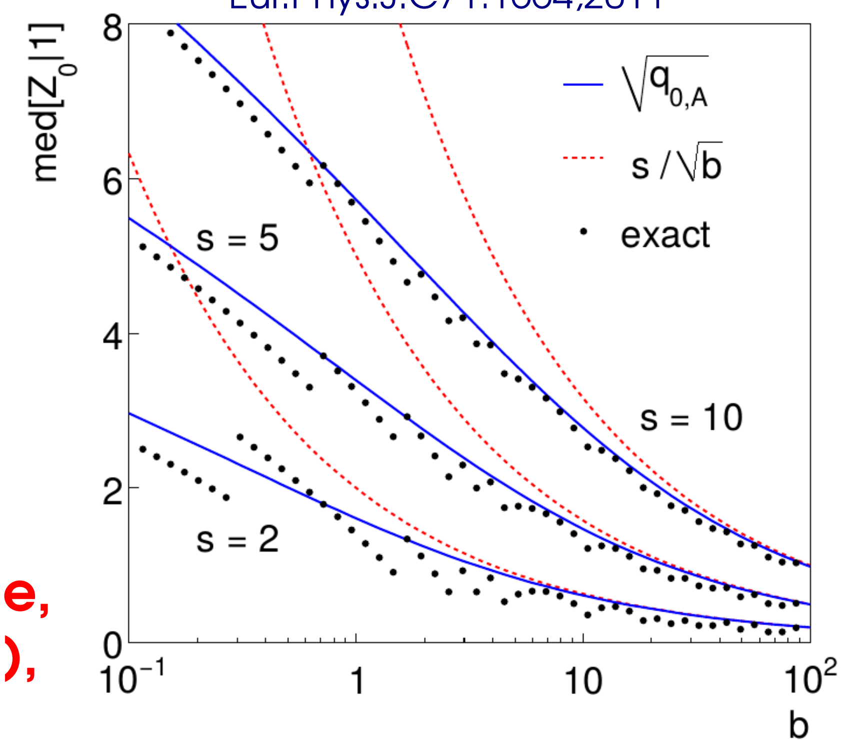
Assuming asymptotic distribution for q_0,

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

Exact result can be obtained using pseudo-experiments → close to $\sqrt{q_0}$ result

Asymptotic formulas justified by Gaussian regime, but remain valid even for small values of S+B (5!), when S itself is not Gaussian

Eur.Phys.J.C71:1554,2011



See G. Cowan's slides for case with B uncertainty

Example: Multi-bin counting

Likelihood :

$$L(S) = \prod_{i=1}^{N} \mathrm{Pois}(n_i; S f_i + B_i)$$

$$\hat{S}_i = n_i - B_i$$

Assume Gaussianity:

$$\lambda(S) = \sum_{i=1}^{N} \left(\frac{n_i - (S f_i + B_i)}{\sqrt{S f_i + B_i}} \right)^2 \qquad \hat{S} = \frac{\sum_{i=1}^{N} \frac{f_i}{B_i} \hat{S}_i}{\sum_{i=1}^{N} \frac{f_i^2}{B_i}}$$

Test statistic: assuming $\hat{S} > 0$,

$$q_0 = \lambda(S=0) - \lambda(\hat{S}) = \left(\hat{S} \sqrt{\sum_{i=1}^{N} \frac{f_i^2}{B_i}} \right)^2$$

Asymptotics:

$$Z = \sqrt{q_0} = \frac{\hat{S}}{\left(\sum_{i=1}^{N} \frac{f_i^2}{B_i} \right)^{-1/2}}$$

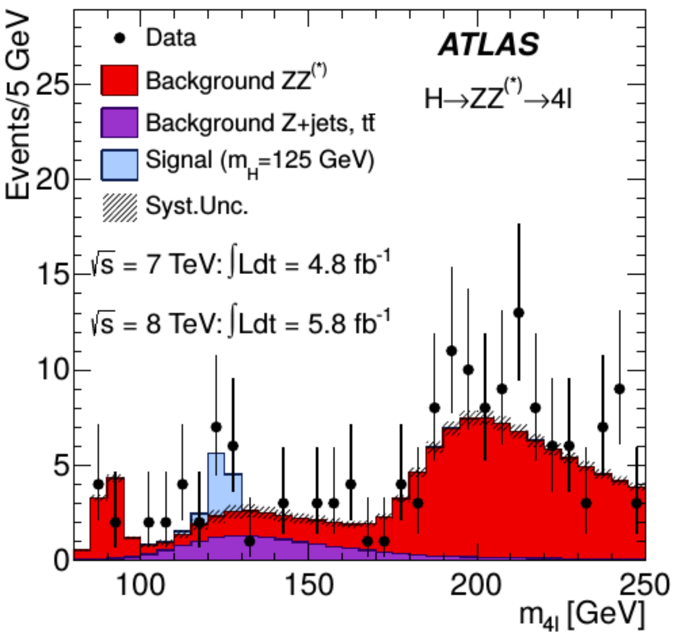
Combined uncertainty on $\hat{S}$ from all the bins

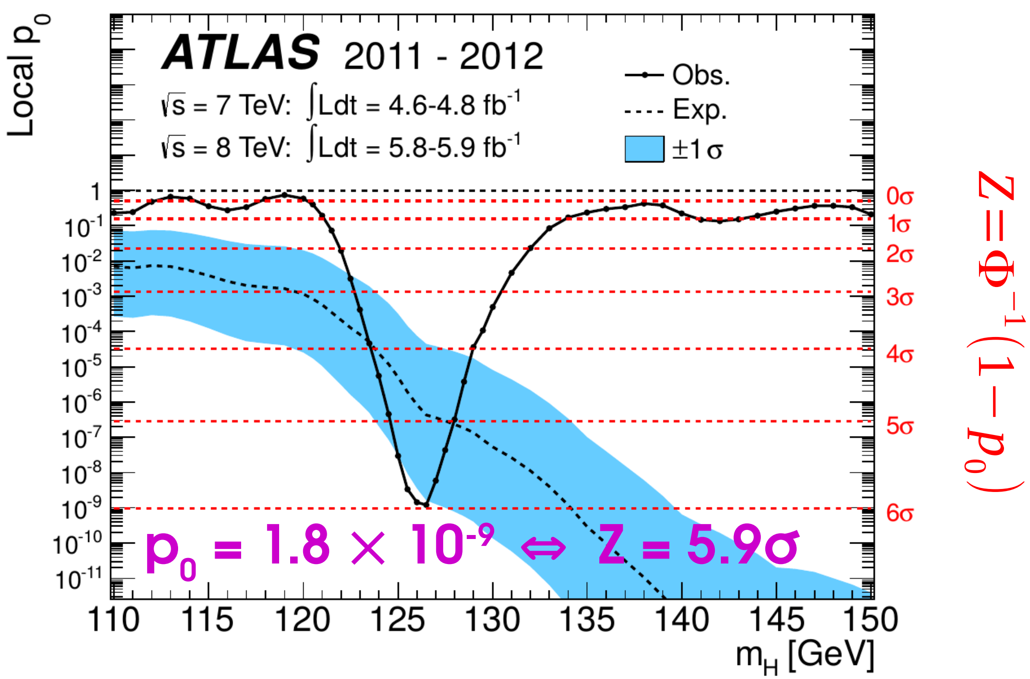
Always better than

- Any bin by itself (for same $\hat{S}$)
- All bins merged together

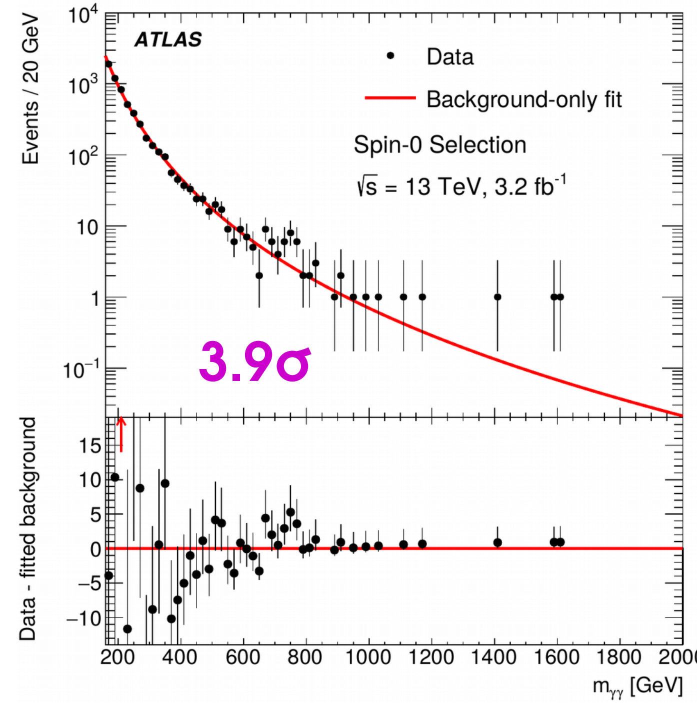
Some Examples

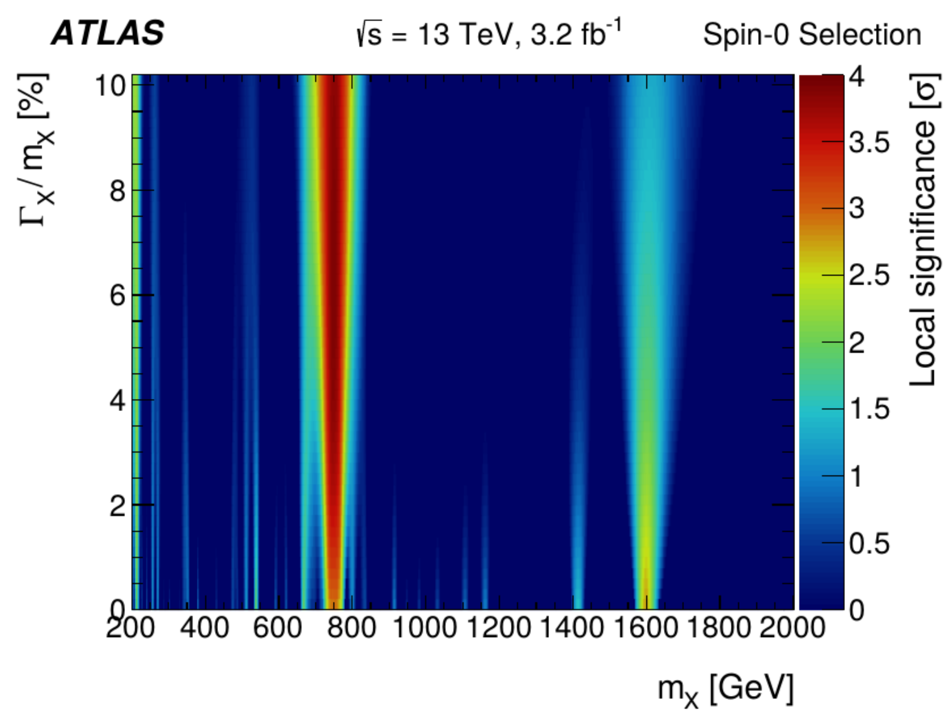
Higgs Discovery: Phys. Lett. B 716 (2012) 1-29





High-mass X→γγ Search: JHEP 09 (2016) 1





Takeaways

Given a statistical model P(data; μ), define likelihood **L(μ) = P(data; μ)**

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

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

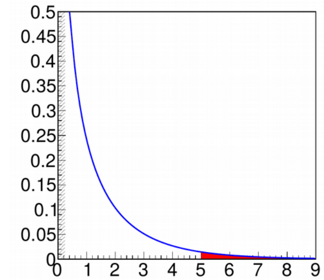
To test for **discovery**, use $q_0 = \begin{cases} -2\log\frac{L(S=0)}{L(\hat{S})} & \hat{S} \geq 0 \\ 0 & \hat{S} < 0 \end{cases}$

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

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

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

What was the question ?



Definition of the p-value:

$$\text{p-value} = \frac{\text{number of signal-like outcomes with only background present}}{\text{all outcomes with only background present}}$$

So **5σ significance (p_0~10^{-7})** ⇔ ***Occurs once in 10^7 if only background present***

However this is **NOT** "***~~One chance in 10^7 to be a fluctuation~~***"

The first statement is about **data probabilities – P(data; H_0)**

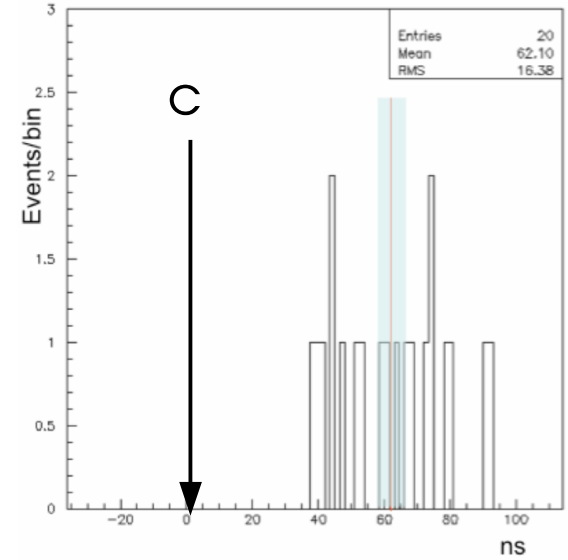
The second is on **P(H_0)** itself – not addressed in the framework described so far
→ makes sense in a ***Bayesian*** context, more on this later in these lectures.

It's also a different statement (although they sometimes get confused)
→ If a signal outcome is also very unlikely, **we may not want to reject H_0, even with p_0 ~ 10^{-7}.**

What was the question ?

e.g. Faster-than-light neutrino anomaly

$(v\text{-}c)/c = (2.37 \pm 0.32\ (\text{stat.})\ {}^{+0.34}_{-0.24}\ (\text{sys.})) \times 10^{-5}$ **6.2σ above c**



"***despite the large significance of the measurement*** *reported here and the stability of the analysis,* ***the potentially great impact of the result motivates the continuation of our studies*** *in order to investigate possible still unknown systematic effects that could explain the observed anomaly.*"

"Extraordinary claims require extraordinary evidence"

⇒ Very unlikely to be a background fluctuation, but hard to believe ***since alternative (v>c) is far-fetched***

Alternative: $$P(\text{fluctuation}) = \frac{\text{number of signal-like outcomes with only B present}}{\text{number of signal-like outcomes from any source (S or B)}}$$

$$= \frac{P(\text{deviation}|B)\,P(B)}{P(\text{deviation}|S)\,P(S) + P(\text{deviation}|B)\,P(B)}$$

→ Needs ***a priori*** **P(S) and P(B)** → Bayesian methods, discussed later

→ In frequentist context, only have $\mathbf{p_0 = P(\text{deviation}|B)}$

⇒ However usually same conclusion, assuming P(S) is not ≪ p_0...

Outline

Yesterday:

- Statistics basics for HEP
- Describing HEP measurements
- Computing statistics results:
  - Discovery

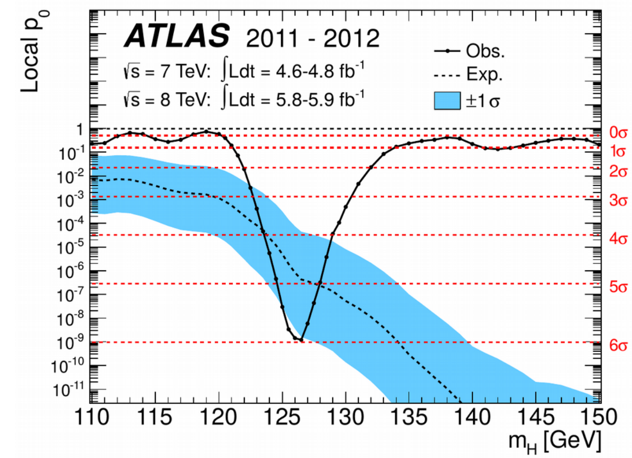
Today:

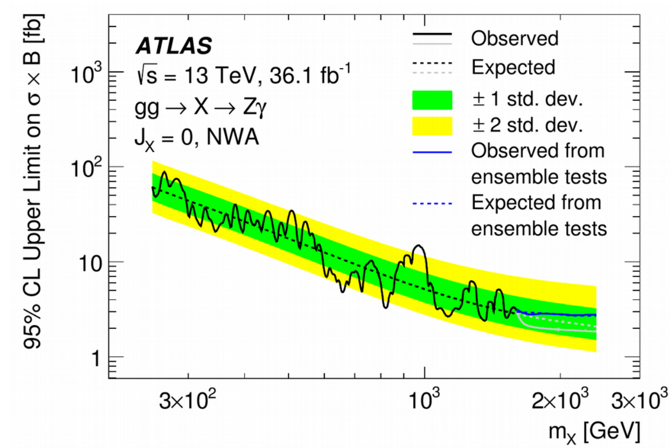
- **Computing statistics results:**
  - Limits
  - Confidence intervals
- **Profiling**
- **Look-Elsewhere Effect**
- **Bayesian methods**

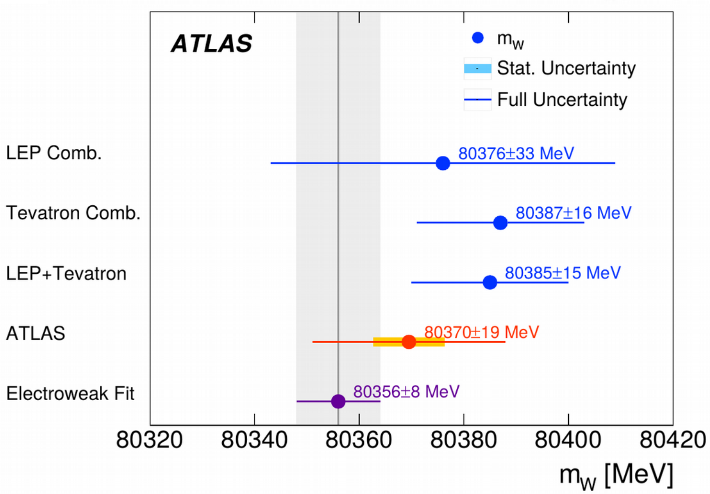
Tomorrow: Practical modeling, Unfolding

Usual Statistical Results

- **Discovery**: we see an excess – is it a (new) signal, or a background fluctuation ?
- **Upper limits**: we don't see an excess – if there is a signal present, how small must it be ?
- **Parameter measurement:** what is the allowed range ("confidence interval") for a model parameter ?





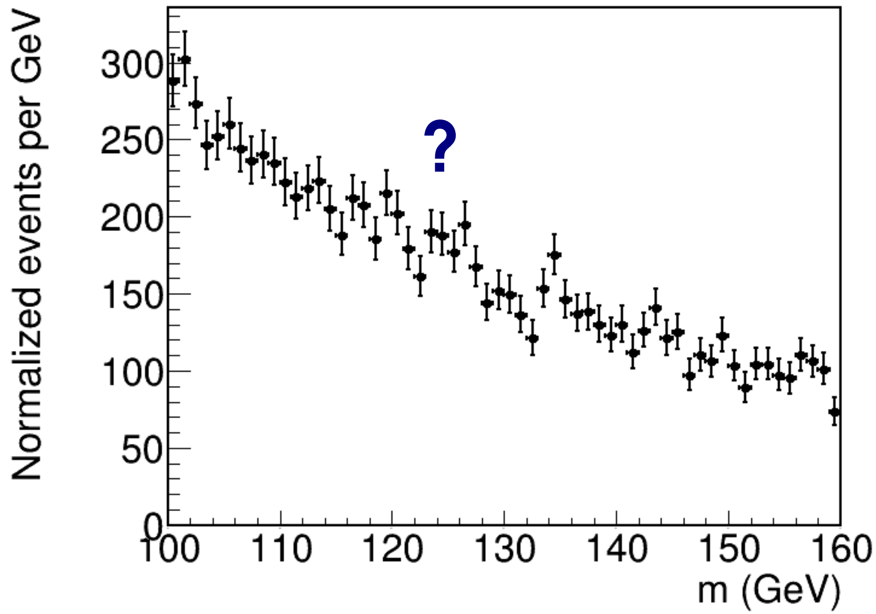


Upper Limits

Hypothesis tests for Limits

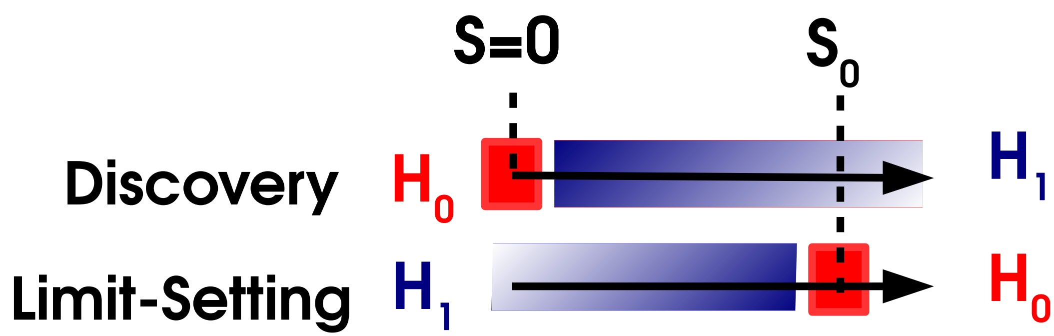
If no signal in data, testing for discovery not very relevant (report 0.2σ excess ?)

→ More interesting to **exclude large signals** → **Upper limits on signal yield**



For **discovery**

- Try to exclude **H_0 : S=0**
- Alternative : **H_1 : S > 0**
- Report p-value for the test (or Z)

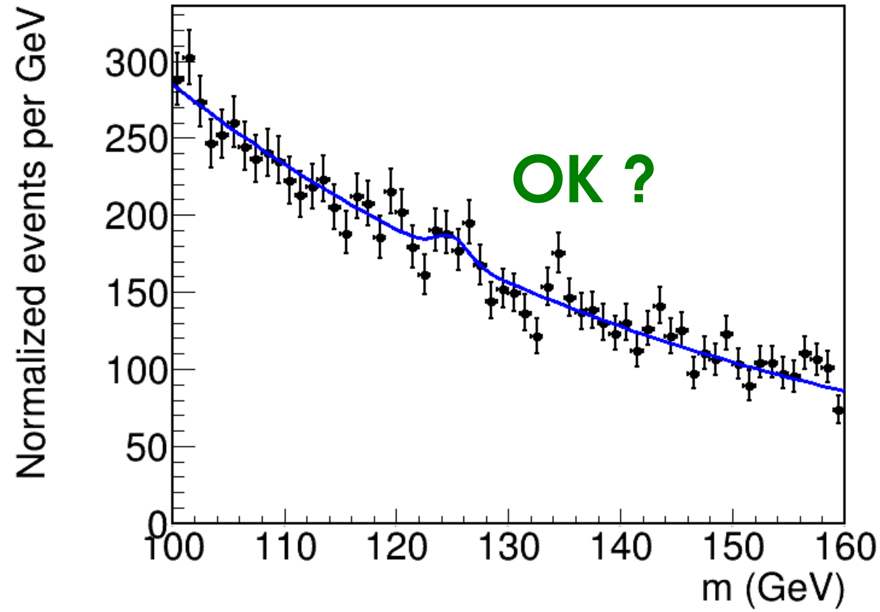


For **limit-setting**:

- Try to exclude **H_0 : $S=S_0$**
- Alternative : **H_1 : $S < S_0$**
- Usually, **adjust S_0 to get a predefined p-value** (typically **5%**)

  → ***Confidence Levels***: **CL = 1 - p** (p = 5% ⇔ 95% CL)

Hypothesis tests for Limits

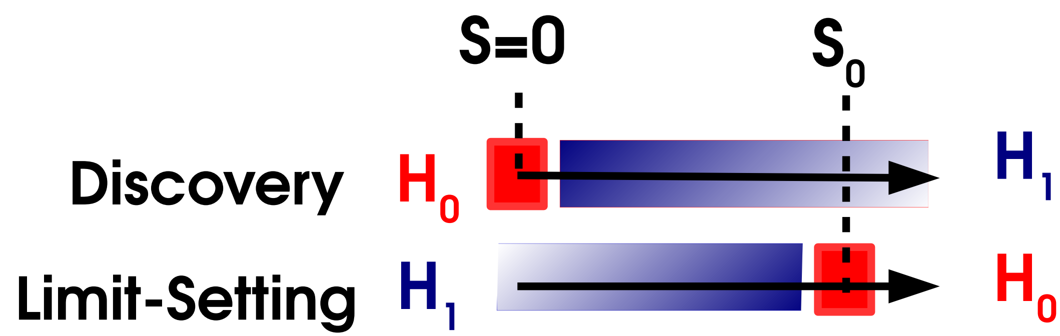


If no signal in data, testing for discovery not very relevant (report 0.2σ excess ?)

→ More interesting to **exclude large signals** → **Upper limits on signal yield**

For **discovery**

- Try to exclude **H_0 : S=0**
- Alternative : **H_1 : S > 0**
- Report p-value for the test (or Z)



For **limit-setting**:

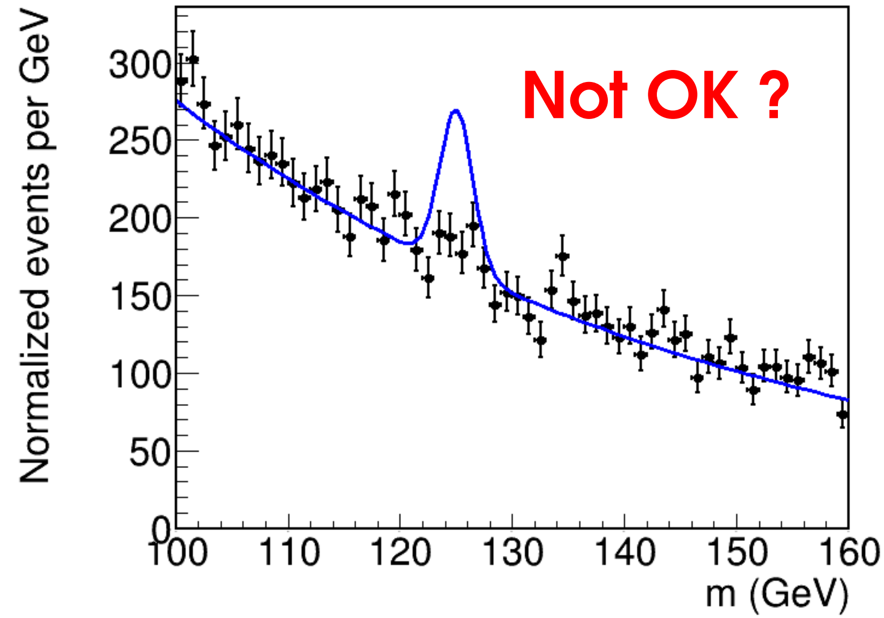
- Try to exclude **H_0 : $S=S_0$**
- Alternative : **H_1 : $S < S_0$**
- Usually, **adjust S_0 to get a predefined p-value** (typically **5%**)

  → ***Confidence Levels***: **CL = 1 - p** (p = 5% ⇔ 95% CL)

Hypothesis tests for Limits

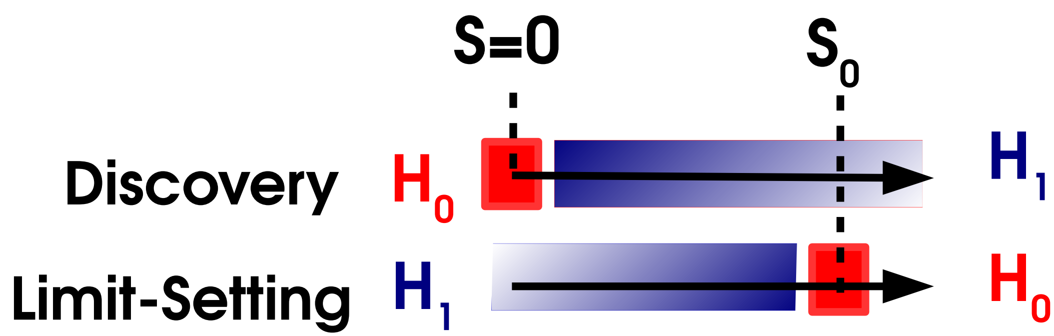
If no signal in data, testing for discovery not very relevant (report 0.2σ excess ?)

→ More interesting to **exclude large signals** → **Upper limits on signal yield**



For **discovery**

- Try to exclude **H_0 : S=0**
- Alternative : **H_1 : S > 0**
- Report p-value for the test (or Z)



For **limit-setting**:

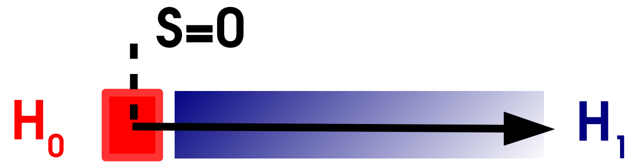
- Try to exclude **H_0 : $S=S_0$**
- Alternative : **H_1 : $S < S_0$**
- Usually, **adjust S_0 to get a predefined p-value** (typically **5%**)

  → ***Confidence Levels***: **CL = 1 - p** (p = 5% ⇔ 95% CL)

Test Statistic for Limit-Setting

Discovery :

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



Compare

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

← Likelihood of H_0

← Likelihood of H_1

Limit-setting

- $H_0 : S = \mu_0$
- $H_1 : S < \mu_0$

Compare

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

← Likelihood of H_0

← Likelihood of H_1

$\hat{S} \sim S_0$ (no exclusion) : $q_{S0} \sim 0$

$\hat{S} \ll S_0$ (good exclusion) : $q_{S0} \gg 1$

Same as q_0 : large values

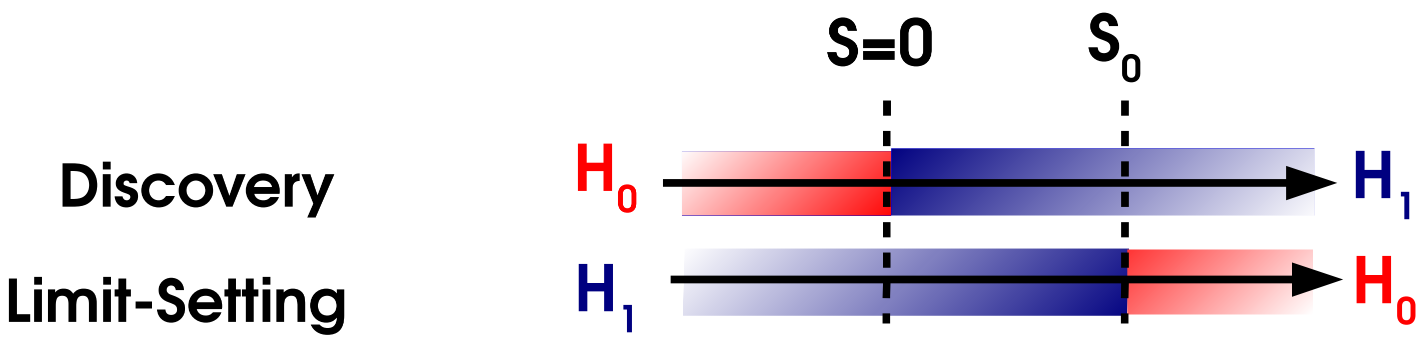
⇒ good rejection of H_0.

One-sided Test Statistic

For upper limits, alternate is **$H_1 : S < \mu_0$** :

→ If **large** signal observed ($\hat{S} \gg S_0$), does not favor H_1 over H_0

→ Only consider **$\hat{S} < S_0$** for **H_1**, and include **$\hat{S} \geq S_0$** in **H_0**.



⇒ Set **q_{S0} = 0 for $\hat{S} > S_0$** – only small signals ($\hat{S} < S_0$) help lower the limit.

→ Also treat separately the case $S < 0$ to avoid technical issues in -2logL fits.

$$\tilde{q}_{S_0} = \begin{cases} 0 & \hat{S} \geq S_0 \\ -2\log \dfrac{L(S=S_0)}{L(\hat{S})} & 0 \leq \hat{S} \leq S_0 \\ -2\log \dfrac{L(S=S_0)}{L(S=0)} & \hat{S} < 0 \end{cases}$$

Asymptotics:

q_{S0} ~ "½χ^2" under $H_0(S=S_0)$, same as q_0, except for special treatment of $\hat{S} < 0$.

$$p_0 = 1 - \Phi\left(\sqrt{q_{S_0}}\right)$$

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

Inversion : Getting the limit for a given CL

Procedure

→ Consider H_0 : $H(S=S_0)$ – alternative H_1 : $H(\hat{S} < S_0)$

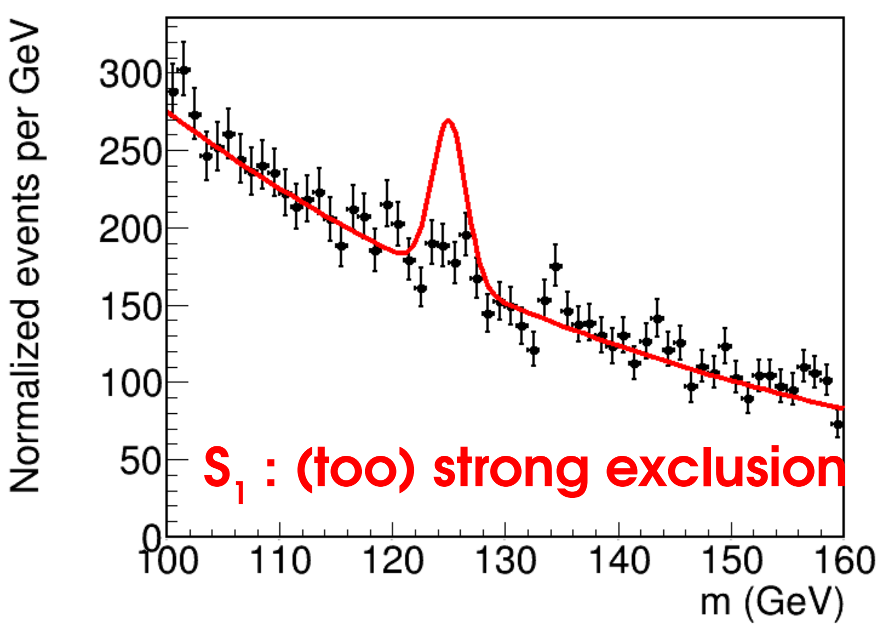
→ Compute q_{S0}, get **exclusion p-value p_{S0}.**

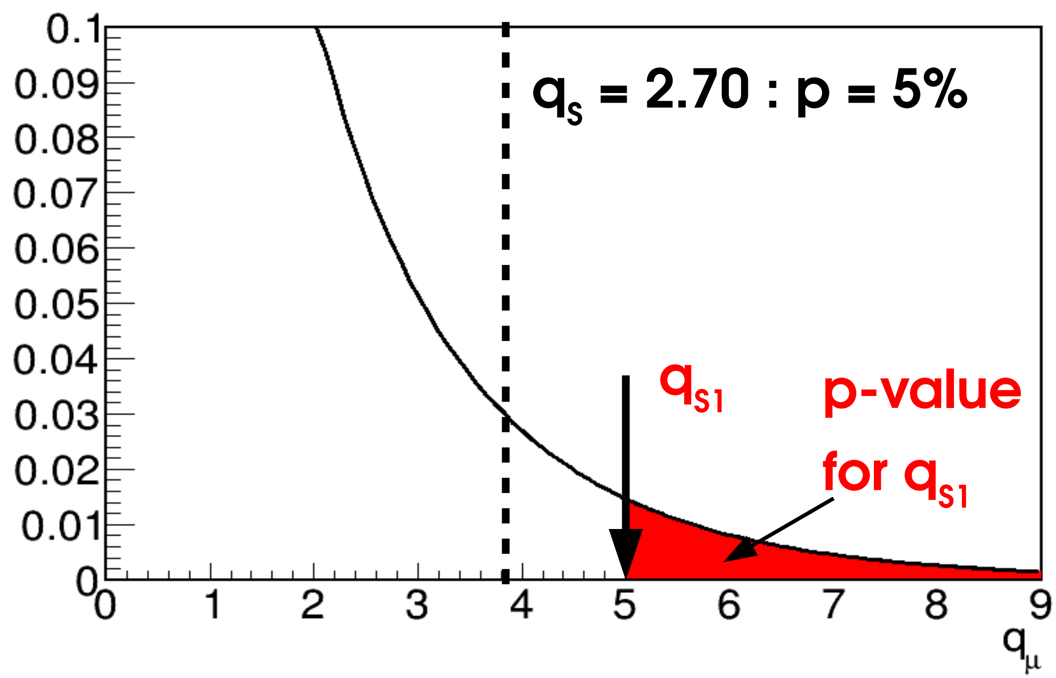
→ Adjust S_0 until 95% CL exclusion (p_{S0} = 5%) is reached

Asymptotics: set target in terms of q_{S0} : $\sqrt{q_{S_0}} = \Phi^{-1}(1 - p_0)$

Asymptotics

CL	Region
90%	$q_S > 1.64$
95%	$q_S > 2.70$
99%	$q_S > 5.41$





Inversion : Getting the limit for a given CL

Procedure

→ Consider H_0 : $H(S=S_0)$ – alternative H_1 : $H(\hat{S} < S_0)$

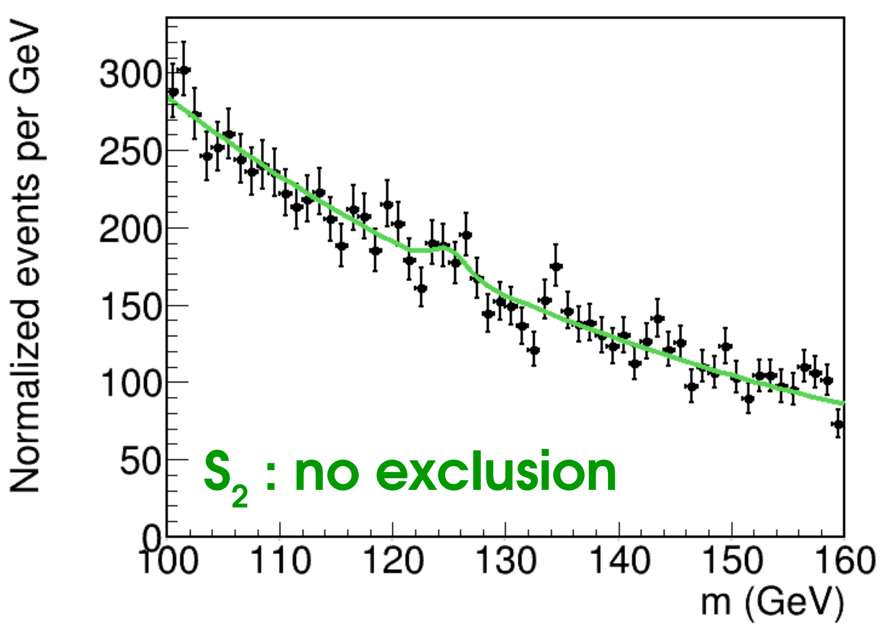
→ Compute q_{S0}, get **exclusion p-value p_{S0}.**

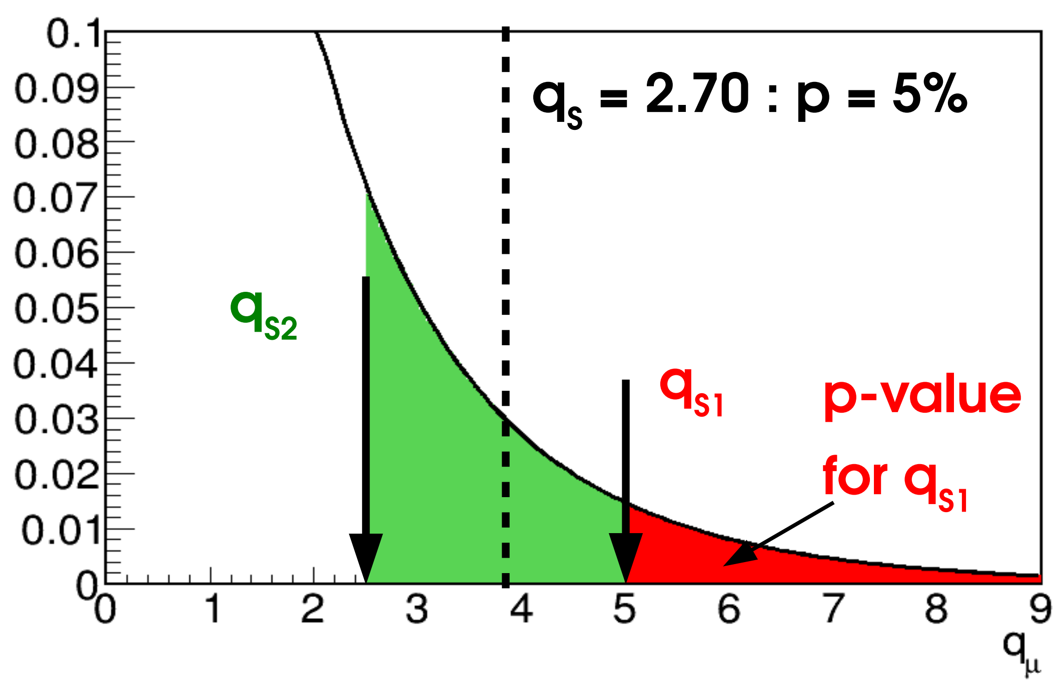
→ Adjust S_0 until 95% CL exclusion (p_{S0} = 5%) is reached

Asymptotics: set target in terms of q_{S0} : $\sqrt{q_{S_0}} = \Phi^{-1}(1-p_0)$

Asymptotics

CL	Region
90%	$q_S > 1.64$
95%	$q_S > 2.70$
99%	$q_S > 5.41$





Inversion : Getting the limit for a given CL

Procedure

→ Consider H_0 : $H(S=S_0)$ – alternative H_1 : $H(\hat{S} < S_0)$

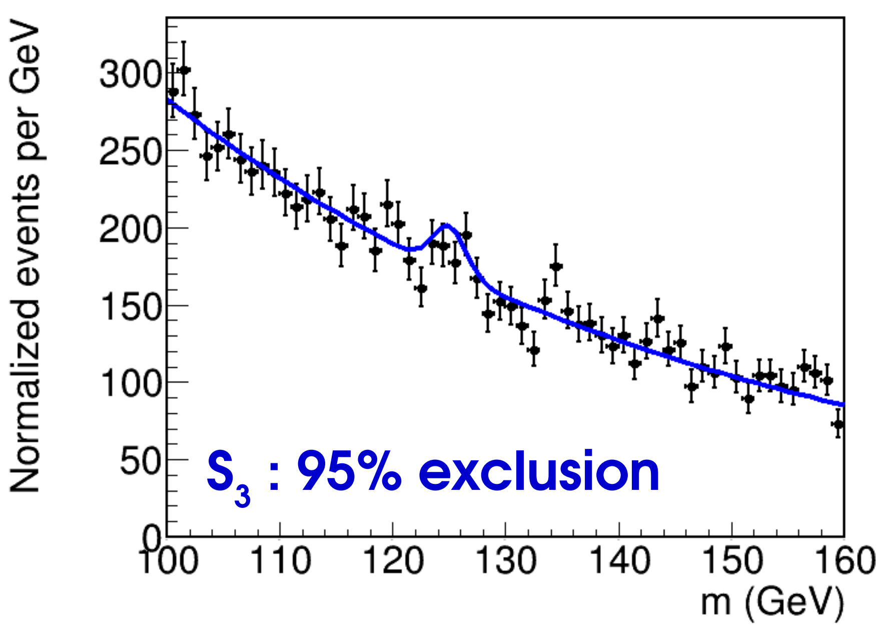
→ Compute q_{S0}, get **exclusion p-value p_{S0}.**

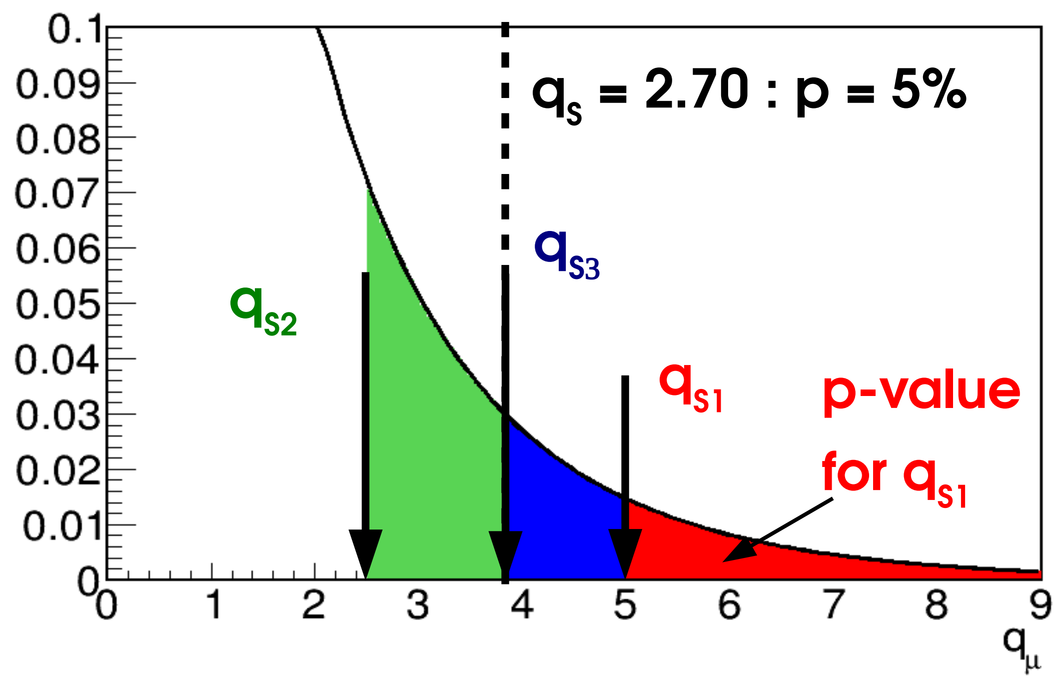
→ Adjust S_0 until 95% CL exclusion (p_{S0} = 5%) is reached

Asymptotics: set target in terms of q_{S0} : $\sqrt{q_{S_0}} = \Phi^{-1}(1 - p_0)$

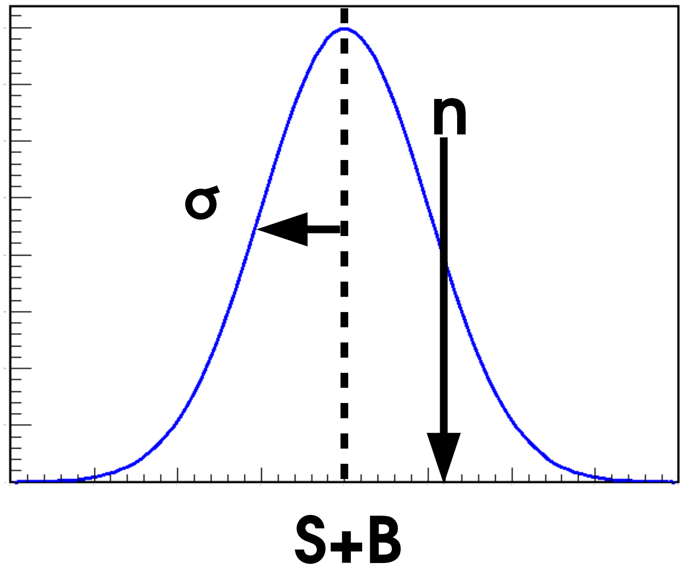
Asymptotics

CL	Region
90%	$q_S > 1.64$
95%	$q_S > 2.70$
99%	$q_S > 5.41$





Upper Limits: Gaussian Example



Usual Gaussian counting example with known B:

$$\lambda(S) = \left(\frac{n-(S+B)}{\sigma_S}\right)^2$$

Reminder:

Best fit signal : $\hat{S} = n - B$

Significance: $Z = \hat{S}/\sqrt{B}$

Compute the 95% CL upper limit on S:

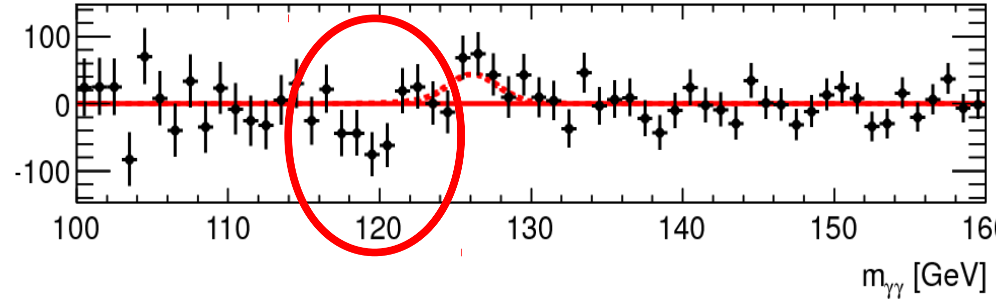
$$q_{S_0} = -2\log\frac{L(S=S_0)}{L(\hat{S})} = \lambda(S_0) - \lambda(\hat{S}) = \left(\frac{n-(S_0+B)}{\sigma_S}\right)^2 = \left(\frac{S_0-\hat{S}}{\sigma_S}\right)^2 \quad \text{for } S_0 > \hat{S}$$

so $q_{S_0} = 2.70$ for $S_0 = \hat{S} + \sqrt{2.70}\,\sigma_S$

And finally $S_{\text{up}} = \hat{S} + 1.64\,\sigma_S$ at 95 % CL

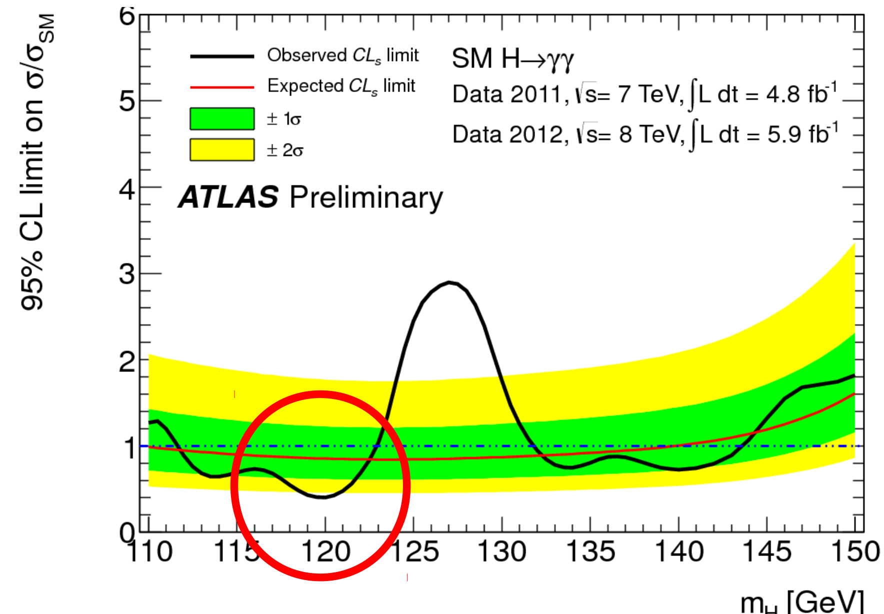
Upper Limit Pathologies

Upper limit: $S_{up} \sim \hat{S} + 1.64\,\sigma_S$.



Problem: for negative $\hat{S}$, get **very** good observed limit.

→ For $\hat{S}$ sufficiently negative, even $S_{up} < 0$!



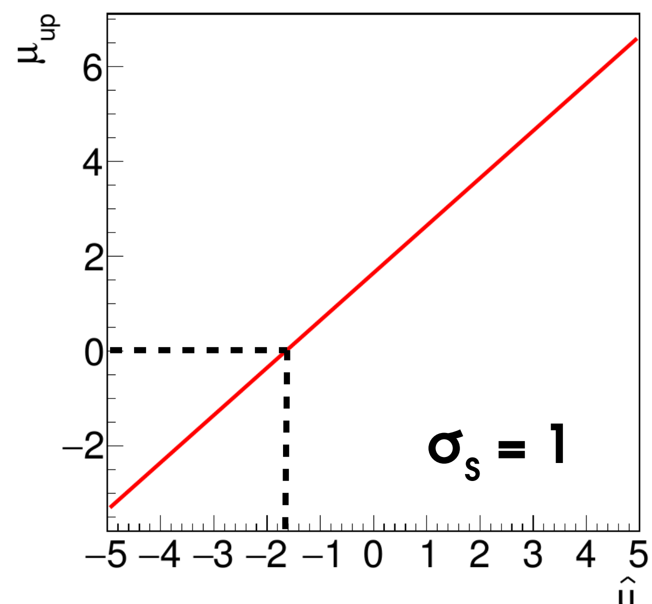
How can this be ?

→ **Background modeling issue ?…** Or:

→ This is a **95%** limit

⇒ **5% of the time, the limit wrongly excludes the true value,** *e.g.* S*=0.

But if we assume S must be >0, we know a priori this is just a fluctuation.



Options

→ **live with it**: sometimes report limit < 0

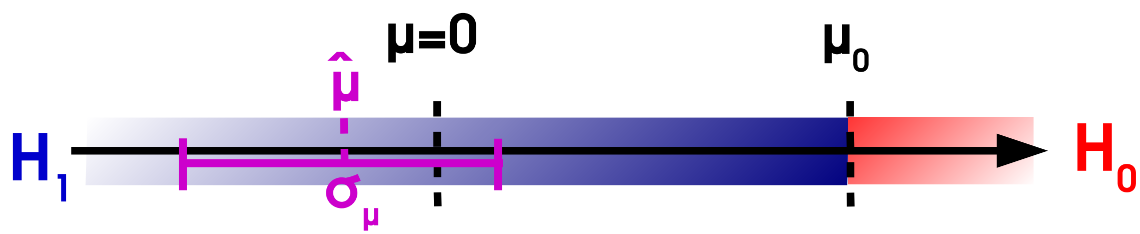
→ **Special procedure to avoid these cases**

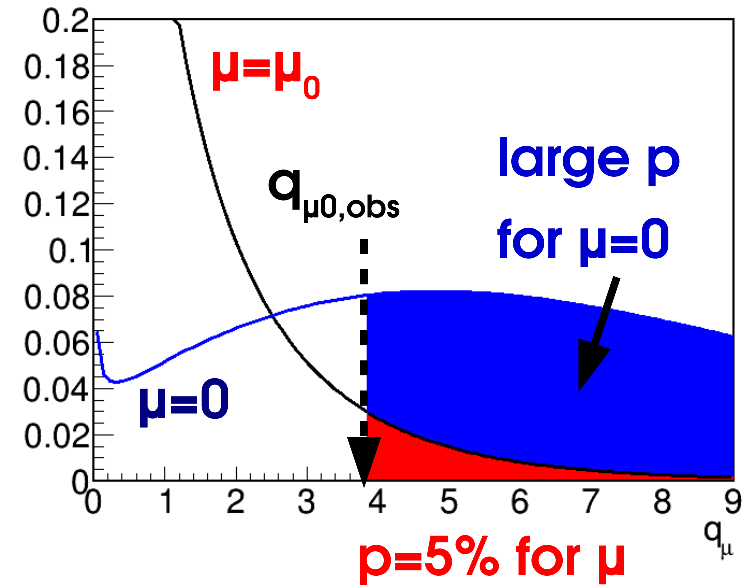
Upper Limit Pathologies

When setting limits, goal is to exclude large μ, **to indicate that μ~0.** What happens at $\mu=0$?

Normal case: $\hat{\mu}$~0, **$\mu=0$ not excluded** :

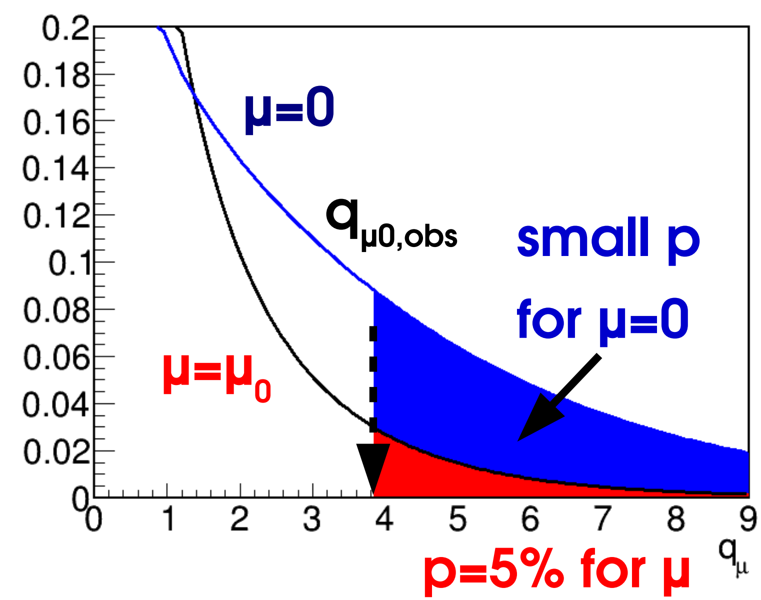
$\mu_{up} = \hat{\mu} + 1.64\,\sigma_\mu > 0$, large p-value for $\mu=0$





Pathological case, very negative $\hat{\mu}$, **$\mu=0$ also excluded** :

$\mu_{up} = \hat{\mu} + 1.64\,\sigma_\mu < 0$, p-value for $\mu=0$ also small



→ However we know a priori that $\mu \geq 0$

⇒ Inject this information into the procedure

CL_s

A. Read, J.Phys. G28 (2002) 2693-2704

Usual solution in HEP : **CL_s**.

→ Compute modified p-value $$p_{CL_s} = \frac{p_{\mu_0}}{p_0}$$

- $p_{\mu 0}$ is the usual p-value (5%)
- p_0 is the p-value computed under H(μ=0).

⇒ **Rescale** exclusion at μ_0 by exclusion at μ=0.

→ Somewhat ad-hoc, but good properties…

Good case : $p_0 \sim O(1)$

p_{CLs} **~ $p_{\mu 0}$ ~ 5%, no change**.

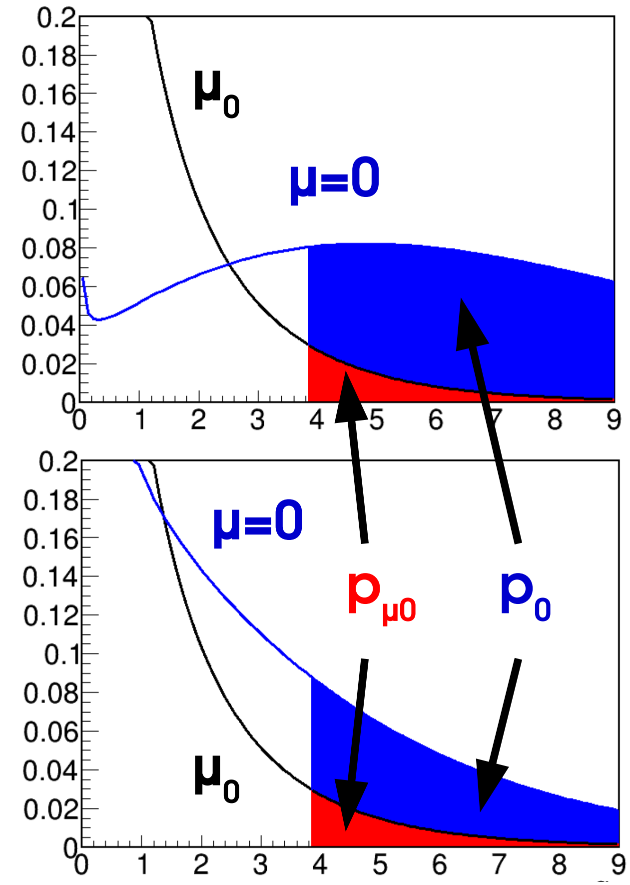
Pathological case : $p_0 \ll 1$

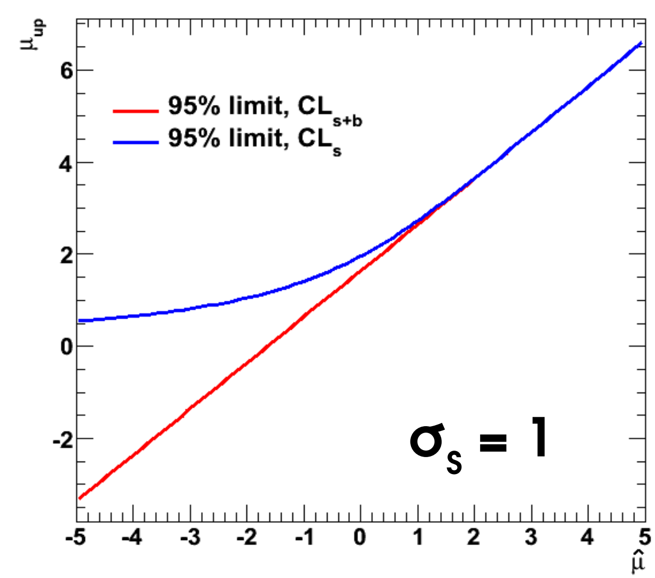
$p_{CLs} \sim p_{\mu 0}/p_0 \gg 5\%$

→ no exclusion ⇒ worse limit, usually >0 as desired

Drawback: ***overcoverage***

→ limit is actually >95% CL for small p_0.

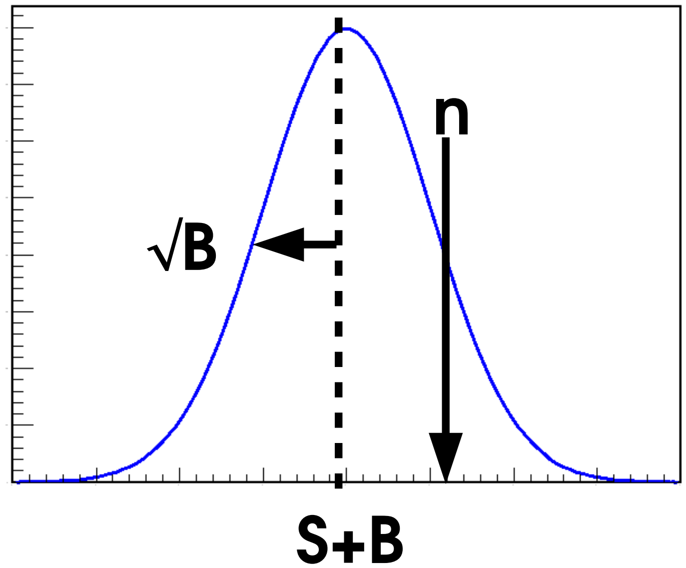




CL_s : Gaussian Example

Usual Gaussian counting example with known B:

$$\lambda(S) = \left(\frac{n-(S+B)}{\sigma_S}\right)^2$$



Reminder

Best fit signal : $\hat{S}$ **= n - B**

CL_{s+b} limit:

$$S_{up} = \hat{S} + 1.64\,\sigma_S \text{ at } 95\,\%\ \text{CL}$$

CL_s upper limit : still have

$$q_{S_0} = \left(\frac{S_0-\hat{S}}{\sigma_S}\right)^2 \quad (\text{for } S_0 > \hat{S})$$

so need to solve

$\hat{S} \sim G(S, \sigma_S)$ so

Under $H_0(S = S_0)$:

$$\sqrt{q_{S_0}} \sim G(0,1)$$

$$p_{S_0} = 1-\Phi\left(\sqrt{q_{S_0}}\right)$$

Under $H_0(S = 0)$:

$$\sqrt{q_{S_0}} \sim G(S_0/\sigma_S, 1)$$

$$p_0 = 1-\Phi\left(\sqrt{q_{S_0}} - S_0/\sigma_S\right)$$

$$p_{CL_s} = \frac{p_{S_0}}{p_0} = \frac{1-\Phi\left(\sqrt{q_{S_0}}\right)}{1-\Phi\left(\sqrt{q_{S_0}} - S_0/\sigma_S\right)} = 5\,\%$$

for $\hat{S} = 0$,

$$S_{up} = \hat{S} + \left[\Phi^{-1}\left(1 - 0.05\,\Phi\left(\hat{S}/\sigma_S\right)\right)\right]\sigma_S \text{ at } 95\,\%\ \text{CL}$$

$$\Phi(0) = 0.5 \Rightarrow \text{ at 95\% CL, } CL_s:\ S_{up} = 1.96\,\sigma_S \qquad CL_{s+b}:\ S_{up} = 1.64\,\sigma_S$$

CL_s: Poisson Rule of Thumb

Same exercise, for the Poisson case

Exact computation : sum probabilities of cases "at least as extreme as data" (n)

$$p_{S_0}(n) = \sum_0^n e^{-(S_0+B)} \frac{(S_0+B)^k}{k!}$$ and one should solve $p_{CL_s} = \frac{p_{S_{up}}(n)}{p_0(n)} = 5\%$ for S_{up}

For n = 0: $$p_{CL_s} = \frac{p_{S_{up}}(0)}{p_0(0)} = e^{-S_{up}} = 5\% \Rightarrow S_{up} = \log(20) = 2.996 \approx 3$$

⇒ **Rule of thumb: when n_{obs}=0, the CL_s 95% CL limit is 3 events (for any B)**

Asymptotics: as before, $$q_{S_0} = \lambda(S_0) - \lambda(\hat{S}) = 2(S_0 + B - n) - 2n \log \frac{S_0+B}{n}$$

For n = 0, $$q_{S_0}(n=0) = 2(S_0+B)$$

$$p_{CL_s} = \frac{p_{S_0}}{p_0} = \frac{1-\Phi\left(\sqrt{q_{S_0}(n=0)}\right)}{1-\Phi\left(\sqrt{q_{S_0}(n=0)} - \sqrt{q_{S_0}(n=B)}\right)} = 5\%$$

⇒ S_{up} ~ 2, exact value depends on B

⇒ Asymptotics not valid in this case – need to use exact results, or toys

Expected Limits: Toys

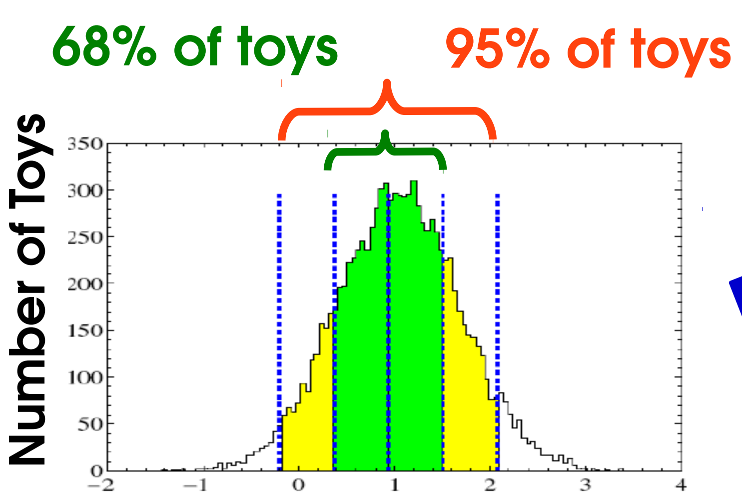
***Expected* results**: median outcome under a given hypothesis
→ usually B-only by convention, but other choices possible.

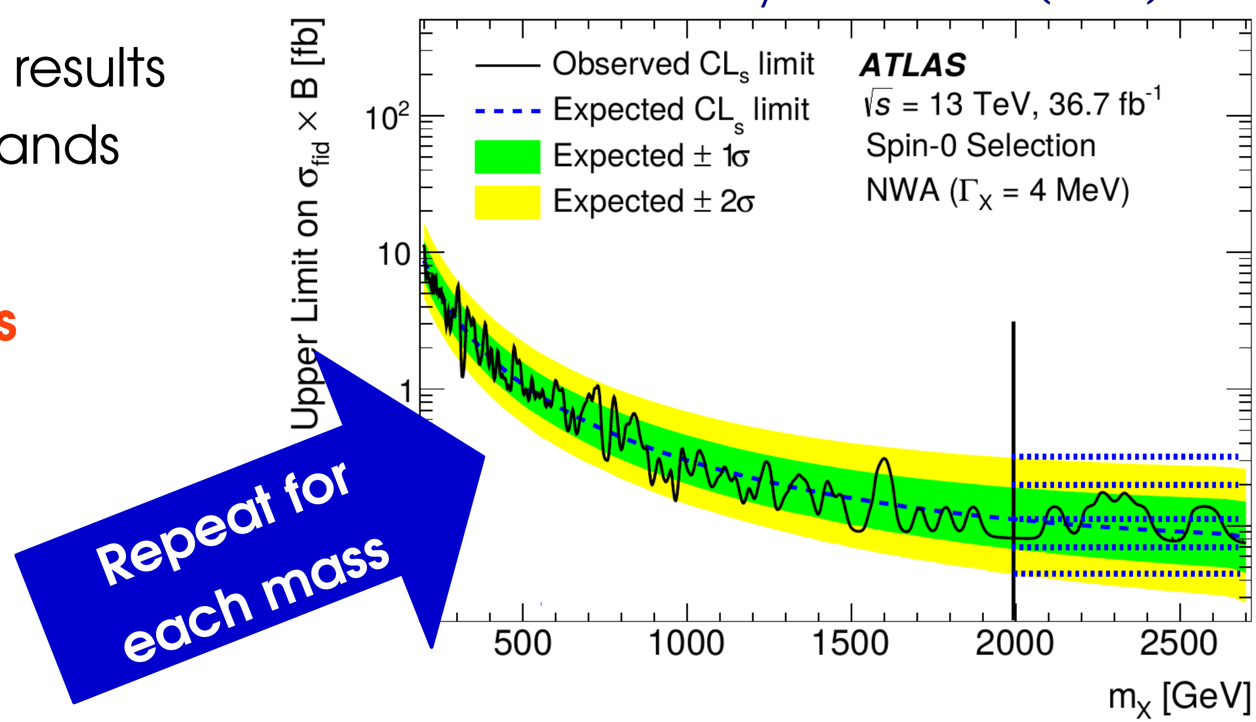
Two main ways to compute:

→ Pseudo-experiments (*toys*):

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

Phys. Lett. B 775 (2017) 105





Eur.Phys.J.C71:1554,2011 **Computed limit**

Expected Limits: Asimov

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

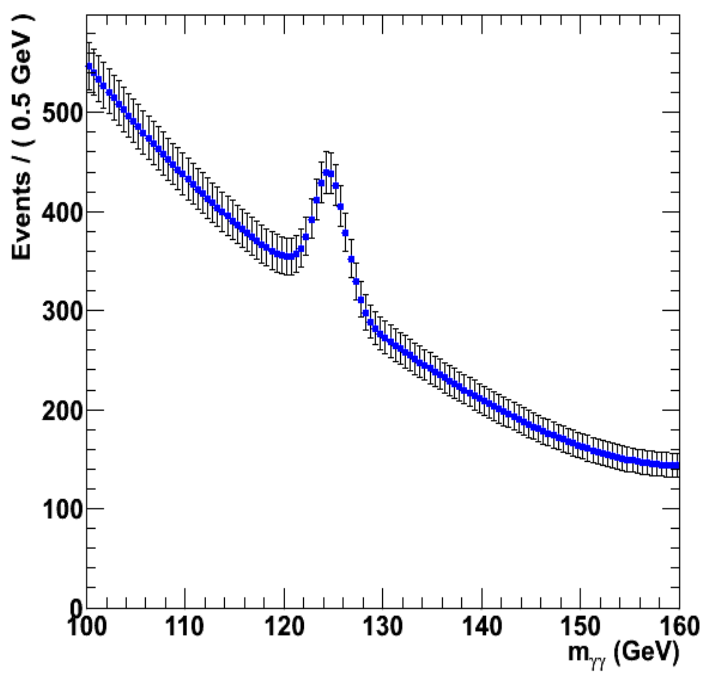
Two main ways to compute:

Strictly speaking, Asimov dataset if
$\hat{X} = X_0$ for all parameters X,
where 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

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

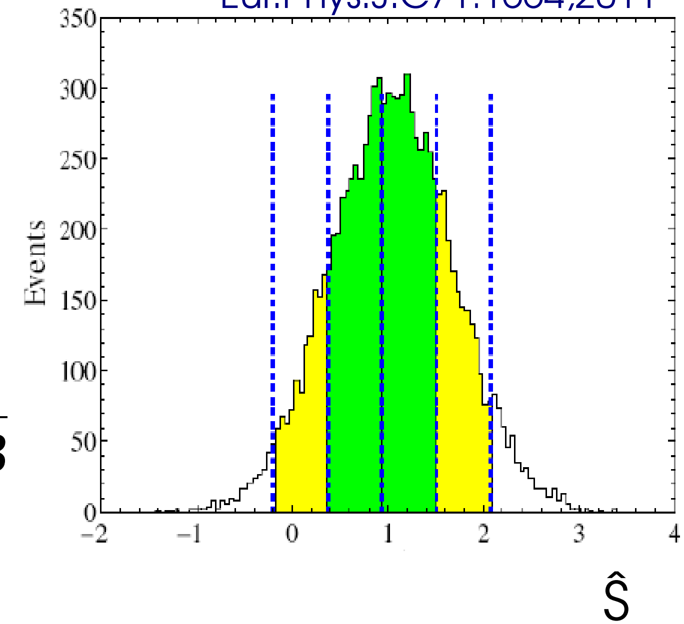


⊕ **Much faster (1 "toy")**

⊖ **Relies on Gaussian approximation**

CL_s : Gaussian Bands

Eur.Phys.J.C71:1554,2011



Usual Gaussian counting example with known B:

95% CL_s upper limit on S:

$$S_{\mathrm{up}} = \hat{S} + \left[\Phi^{-1}\left(1 - 0.05\,\Phi\left(\hat{S}/\sigma_S\right)\right)\right]\sigma_S$$

with

$$\sigma_S = \sqrt{B}$$

Compute expected bands for S=0:

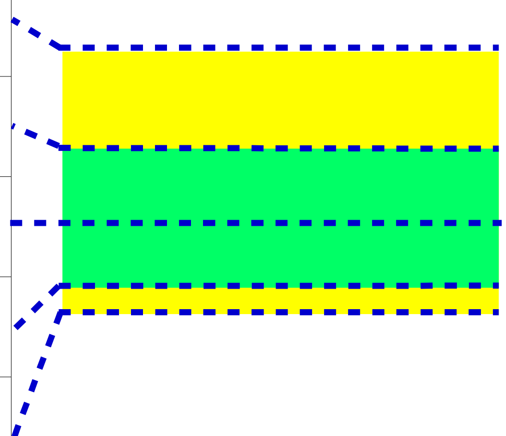
→ **Asimov dataset ⇔ $\hat{S}$ = 0** :

$$S^{0}_{\mathrm{up,exp}} = 1.96\ \sigma_S$$

→ **± nσ bands**:

$$S^{\pm n}_{\mathrm{up,exp}} = \left(\pm n + \left[1 - \Phi^{-1}\left(0.05\,\Phi\left(\mp n\right)\right)\right]\right)\sigma_S$$

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



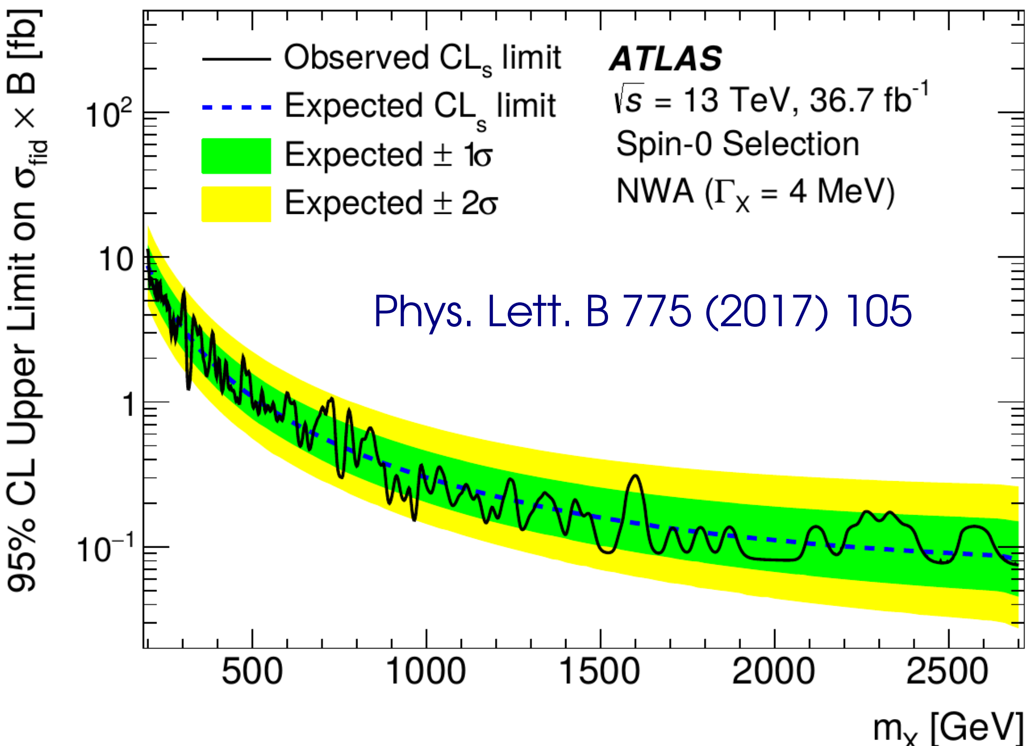
CLs :

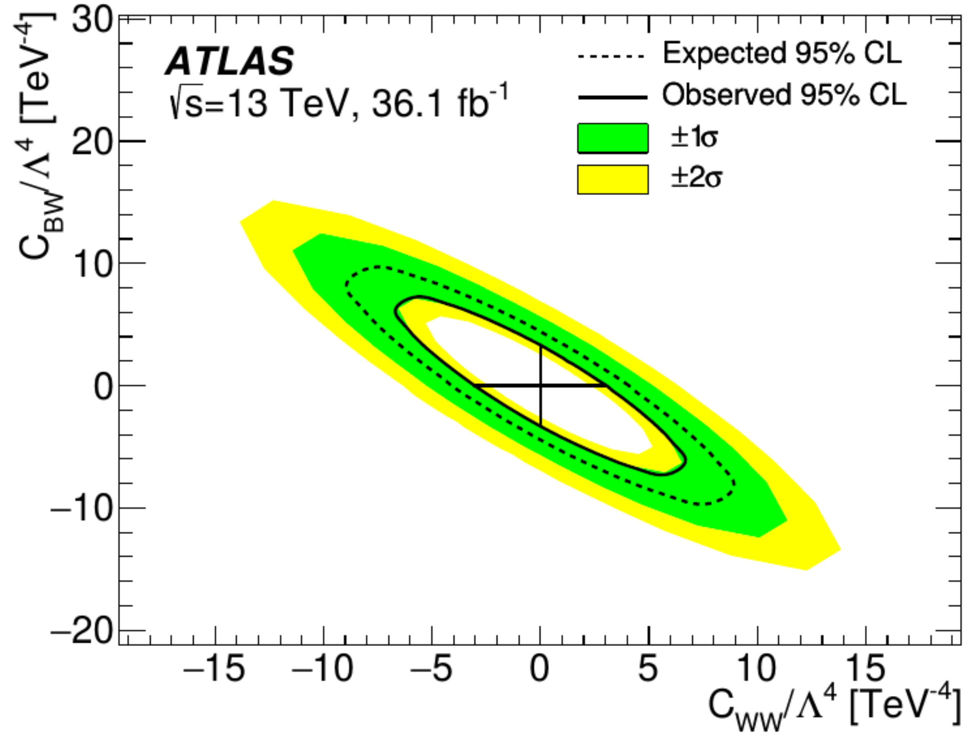
- Positive bands somewhat reduced,
- Negative ones more so

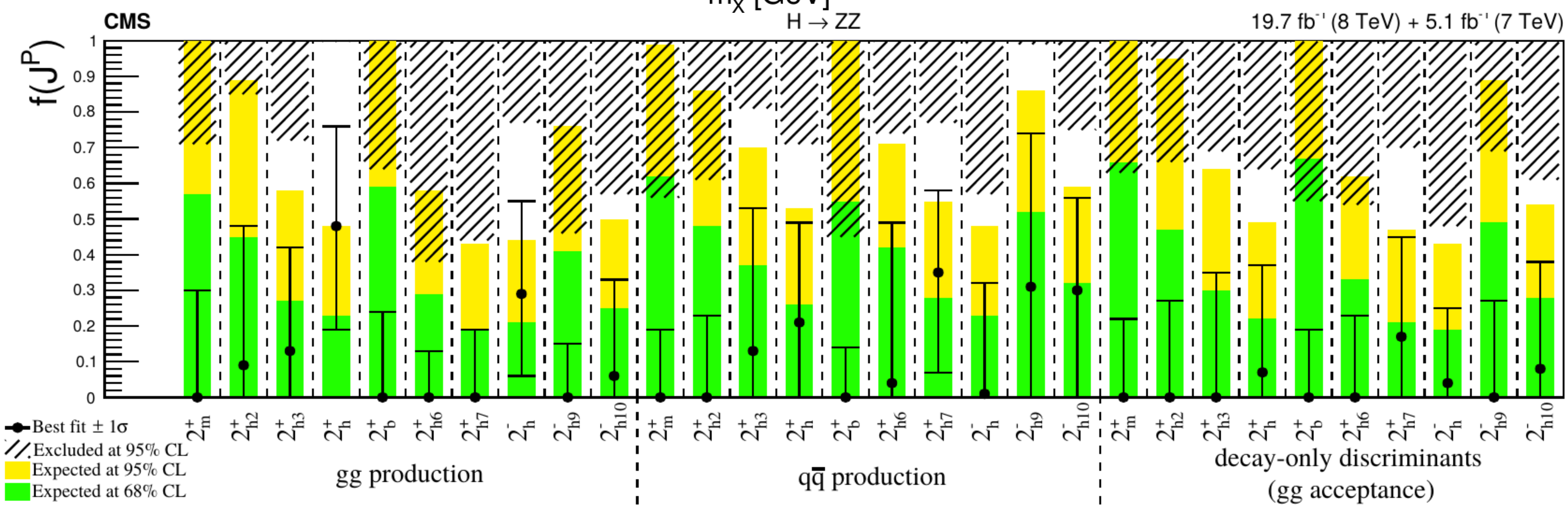
Band width from $\sigma^2_{S,A} = \dfrac{S^2}{q_S(\mathrm{Asimov})}$ depends on S, for non-Gaussian cases,different values for each band...

Upper Limit Examples

ATLAS 2015-2016 4l aTGC Search



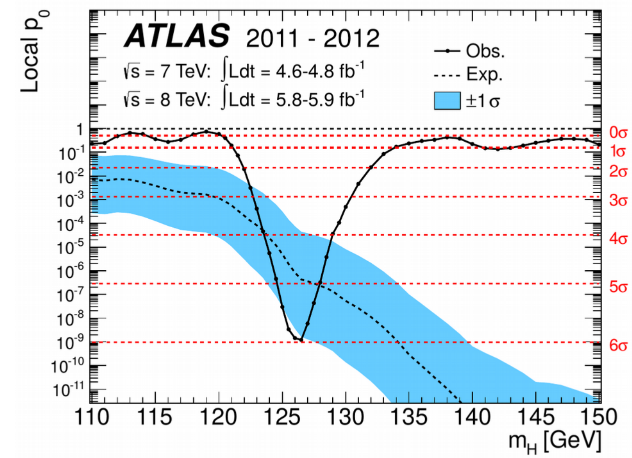




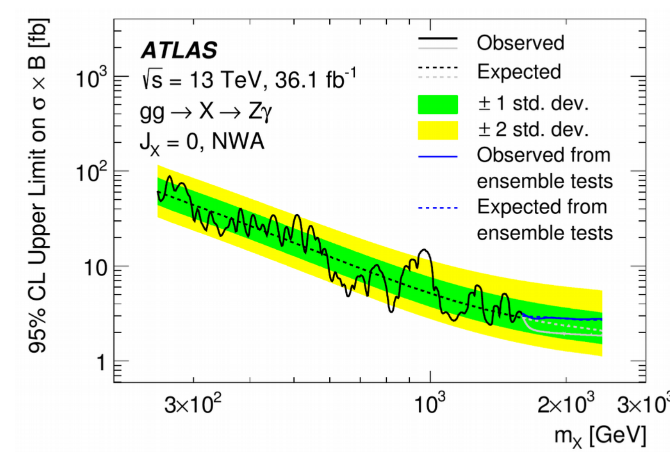
Phys. Rev. D 92 (2015) 012004

Usual Statistical Results

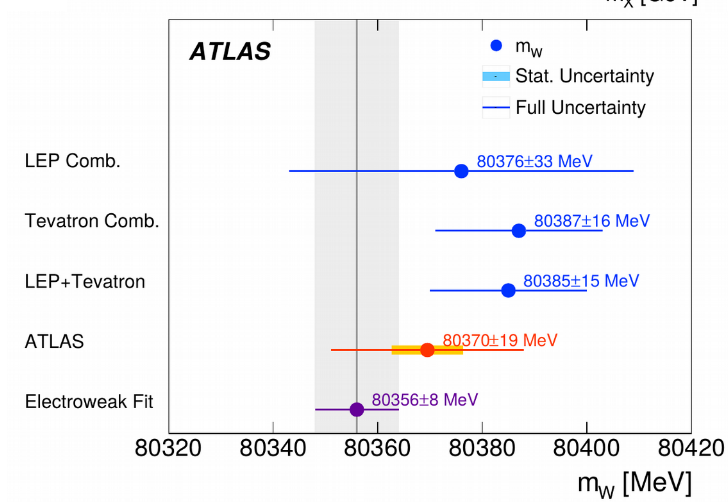
- **Discovery**: we see an excess – is it a (new) signal, or a background fluctuation ?



- **Upper limits**: we don't see an excess – if there is a signal present, how small must it be ?



- **Parameter measurement:** what is the allowed range ("confidence interval") for a model parameter ?



Outline

Computing statistics results:

- Limits
- **Confidence intervals**

Profiling

Look-Elsewhere Effect

Bayesian methods

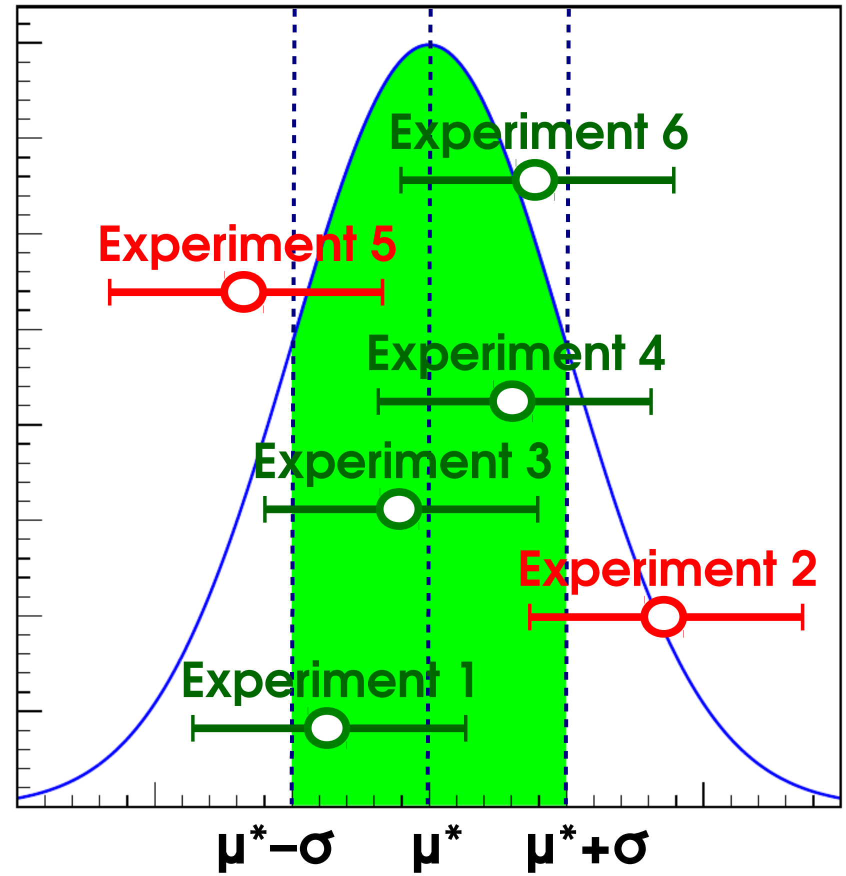
Confidence Intervals

Gaussian Inversion

If $\hat{\mu} \sim G(\mu^*, \sigma)$, known quantiles :

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

This is a probability for $\hat{\mu}$, not μ^*!
→ μ^* is a **fixed number**, **not a random variable**



But we can invert the relation:

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

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

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

→ This gives the desired statement on μ^* : *if we repeat the experiment many times, **[$\hat{\mu}$ - σ, $\hat{\mu}$ + σ]** will **contain the true value** 68% of the time*

This is a statement ***on the interval** [$\hat{\mu}$ - σ, $\hat{\mu}$ + σ]* obtained for each experiment

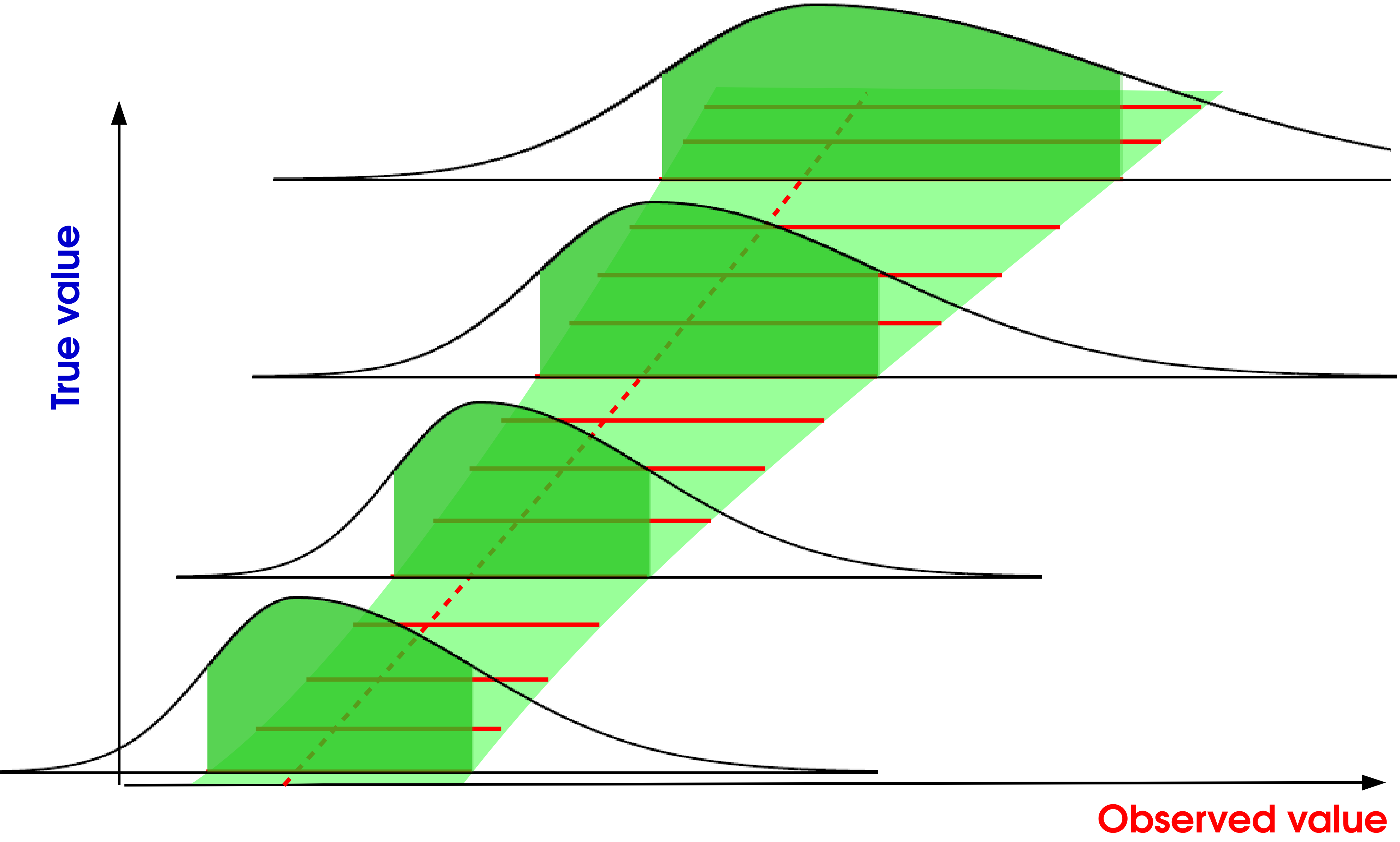
Works in the same way for other interval sizes: **[$\hat{\mu}$ - Zσ, $\hat{\mu}$ + Zσ]** with

Z	1	1.96	2
CL	0.68	0.95	0.955

Neyman Construction

General case: Build 1σ intervals of observed values for each true value

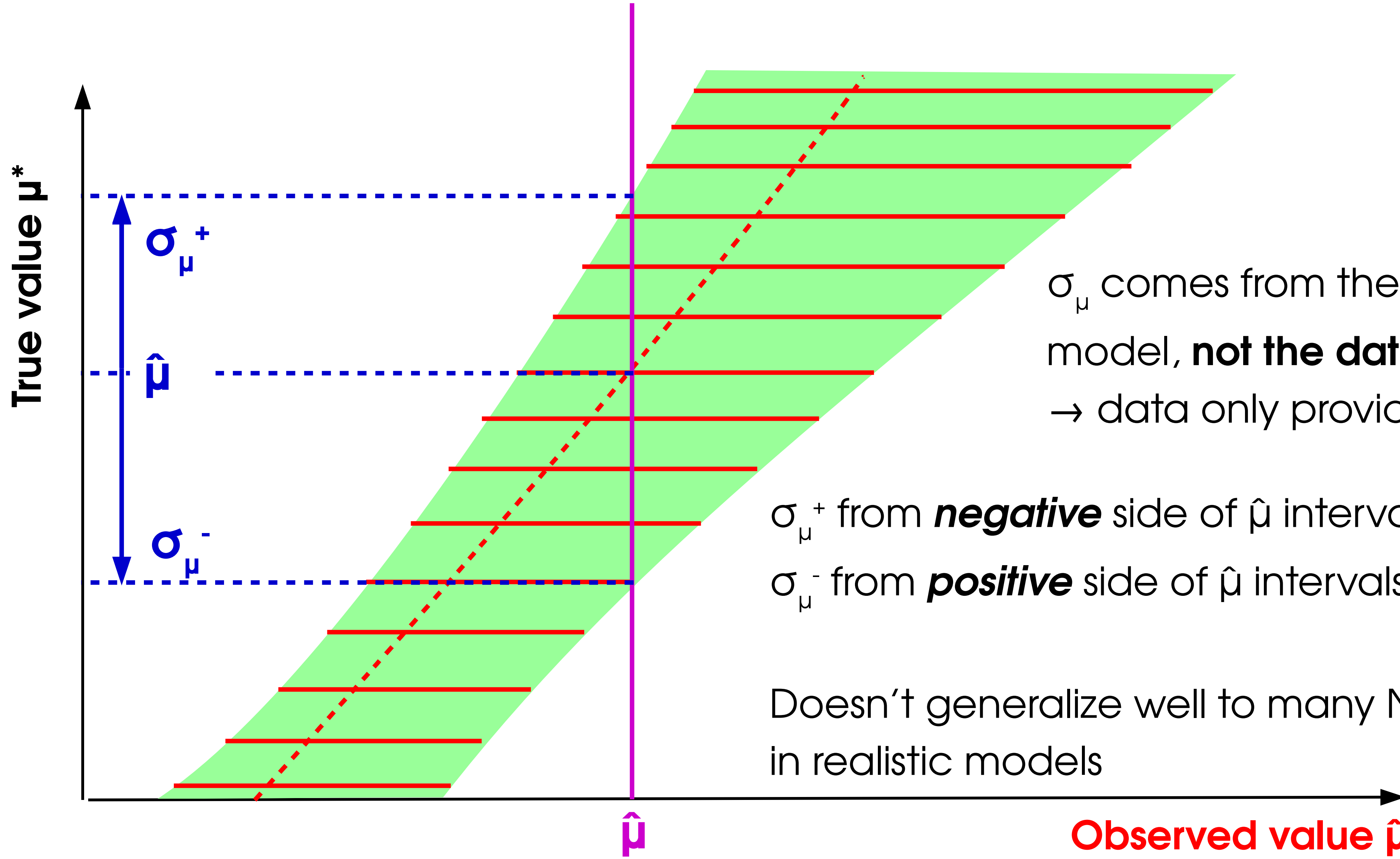
⇒ ***Confidence belt***



Inversion using the Confidence Belt

General case: Intersect belt with given $\hat{\mu}$, get $P\left(\hat{\mu} - \sigma_\mu^- < \mu^* < \hat{\mu} + \sigma_\mu^+\right) = 68\,\%$

→ Same as before for Gaussian, works also when $P(\mu^{obs} \mid \mu)$ varies with μ.



σ_μ comes from the model, **not the data**

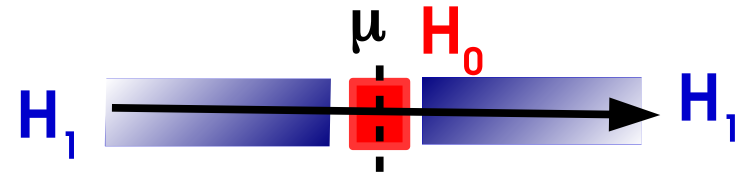
→ data only provides $\hat{\mu}$.

σ_μ^+ from ***negative*** side of $\hat{\mu}$ intervals

σ_μ^- from ***positive*** side of $\hat{\mu}$ intervals

Doesn't generalize well to many NPs in realistic models

Likelihood Intervals



Confidence intervals from L:

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

μ can be several POI!

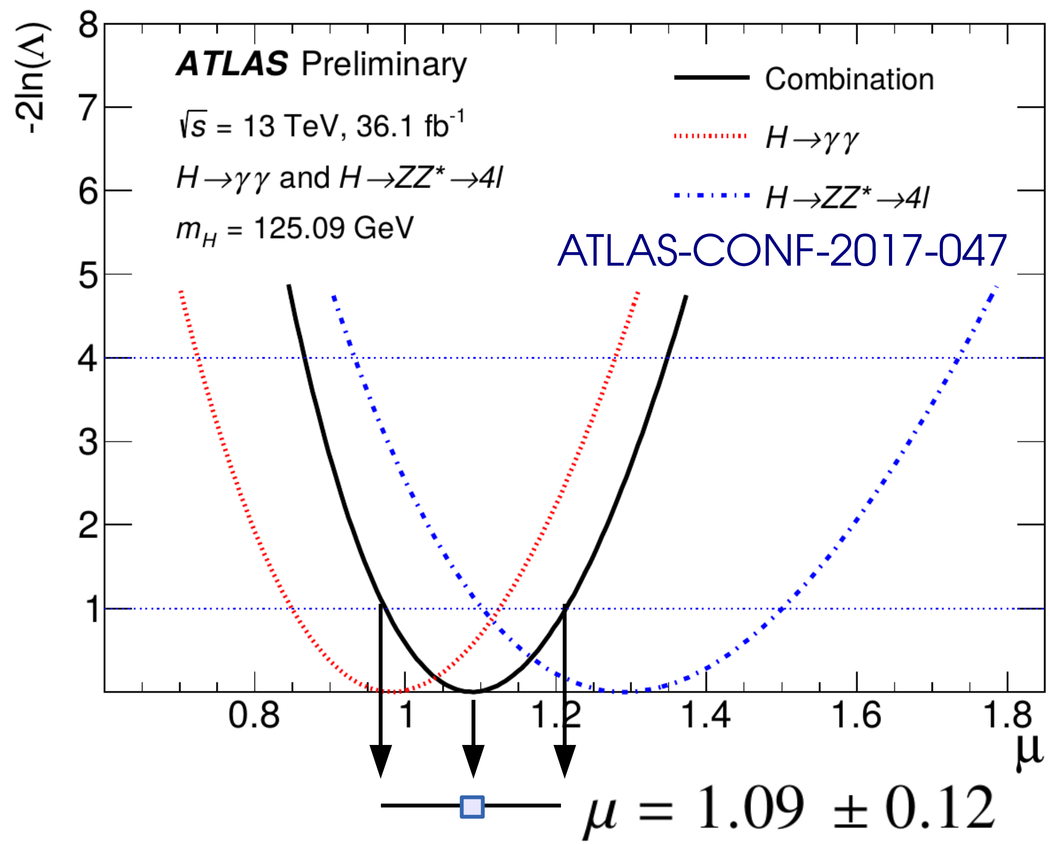
- Test $H(\mu_0)$ against alternative using
- Two-sided test since true value can be higher or lower than observed

Asymptotics:

- **$t_\mu \sim \chi^2(N_{POI})$ under $H(\mu_0)$**
- **$\sqrt{t_\mu} \sim G(0,1)$** (Gaussian with $d=N_{POI}$)

In practice:

- Plot t_μ vs. μ
- The minimum occurs at **$\mu = \hat{\mu}$**
- Crossings with **$t_\mu = Z^2$** give the **$\pm Z\sigma$ uncertainties** (for $N_{POI}=1$)



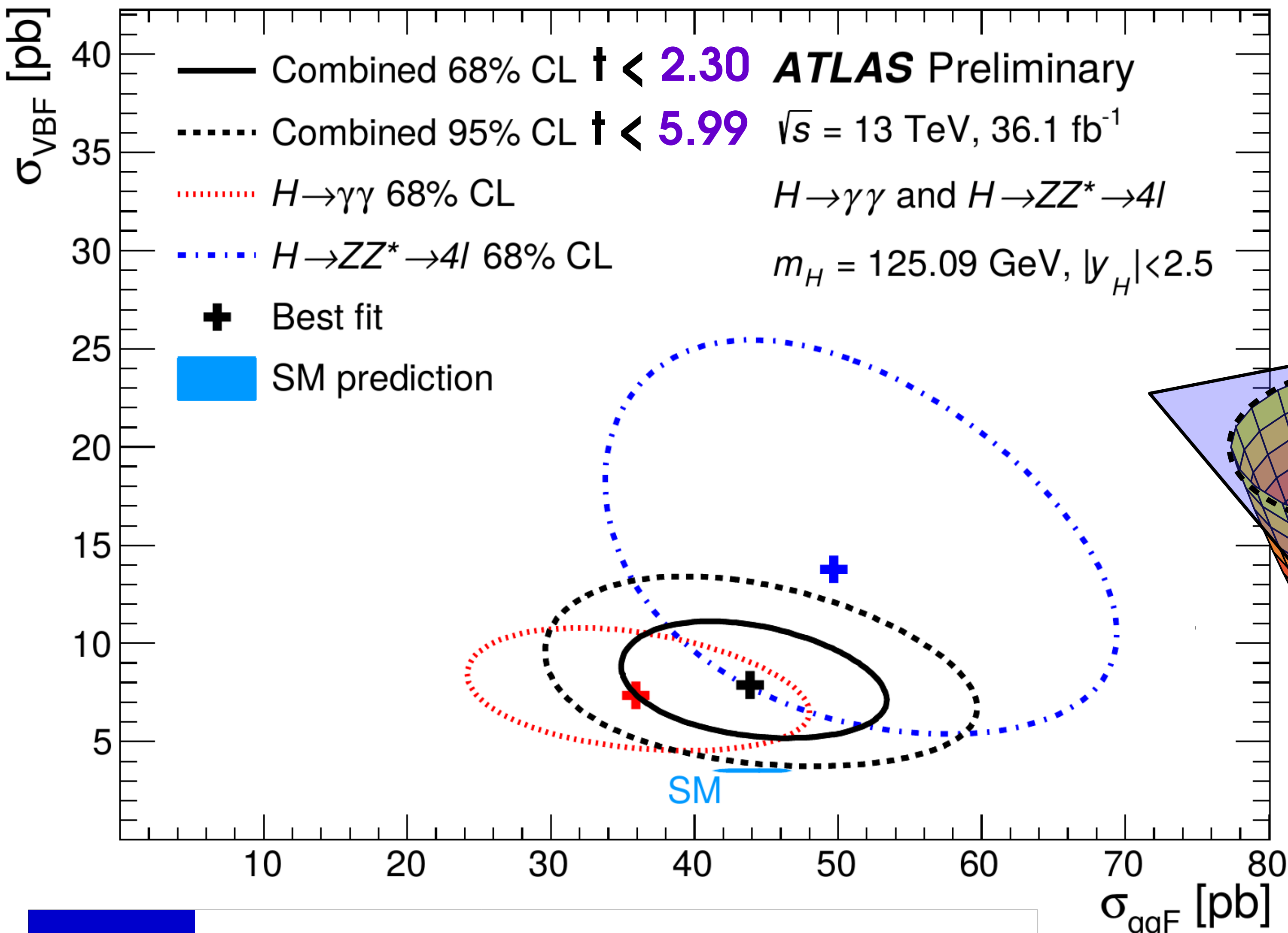
→ **Gaussian case**: parabolic profile, $t_\mu = \left(\frac{\mu - \hat{\mu}}{\sigma}\right)^2 \Rightarrow \mu_\pm = \hat{\mu} \pm \sigma$ **at** $t_\mu = 1$
same result as Neyman construction, also robust against non-Gaussian effects.

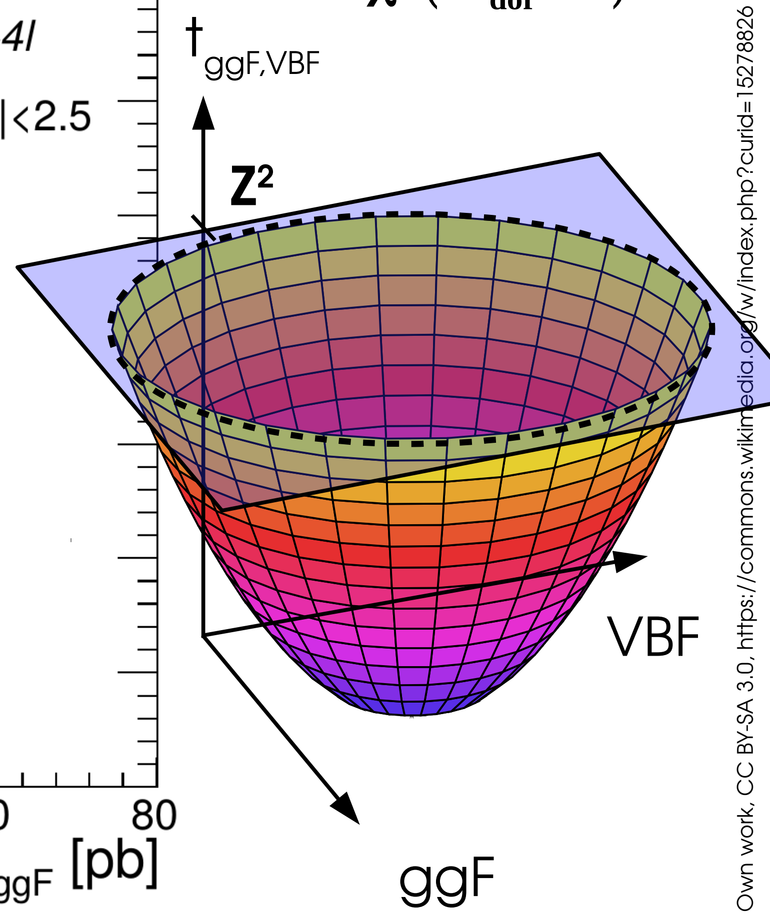
2D Example: Higgs σ_{VBF} vs. σ_{ggF}

ATLAS-CONF-2017-047

$$t = -2\log\frac{L(X_0, Y_0)}{L(\hat{X}, \hat{Y})}$$

$$\sim \chi^2(N_{dof}=2)$$





CL	68% (1σ)	95%	95.5% (2σ)
1D Z^2	1	3.84	4
2D Z^2	2.30	5.99	6.18

Gaussian case: elliptic paraboloid surface

By Krishnavedala - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15278826

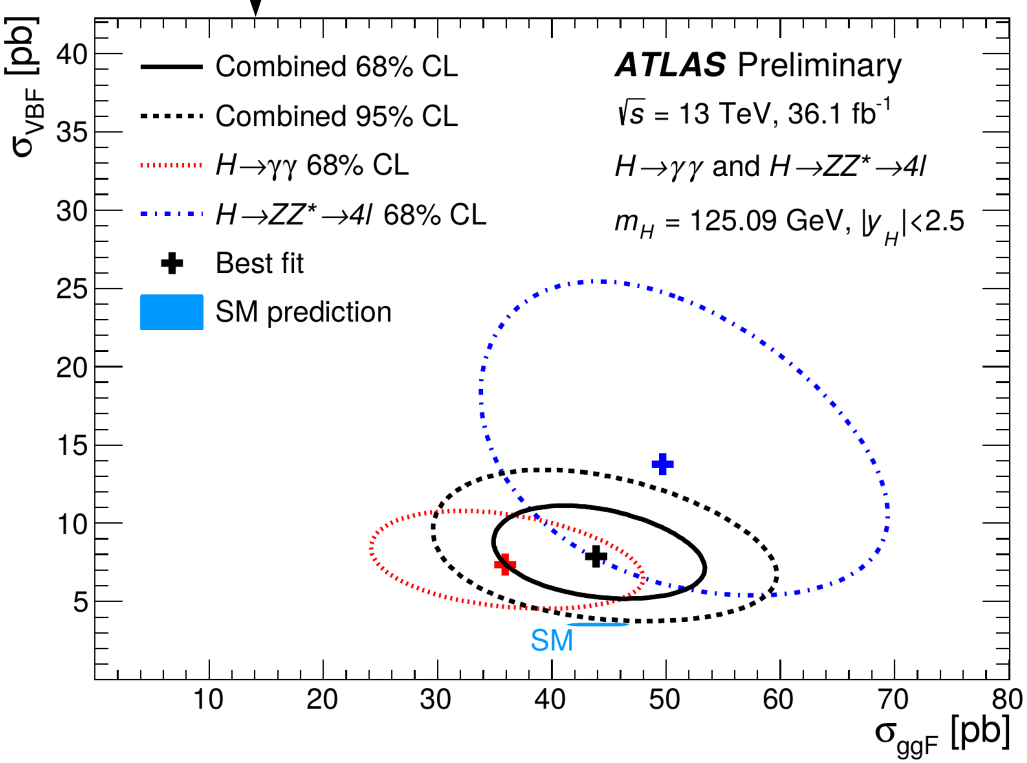
Reparameterization

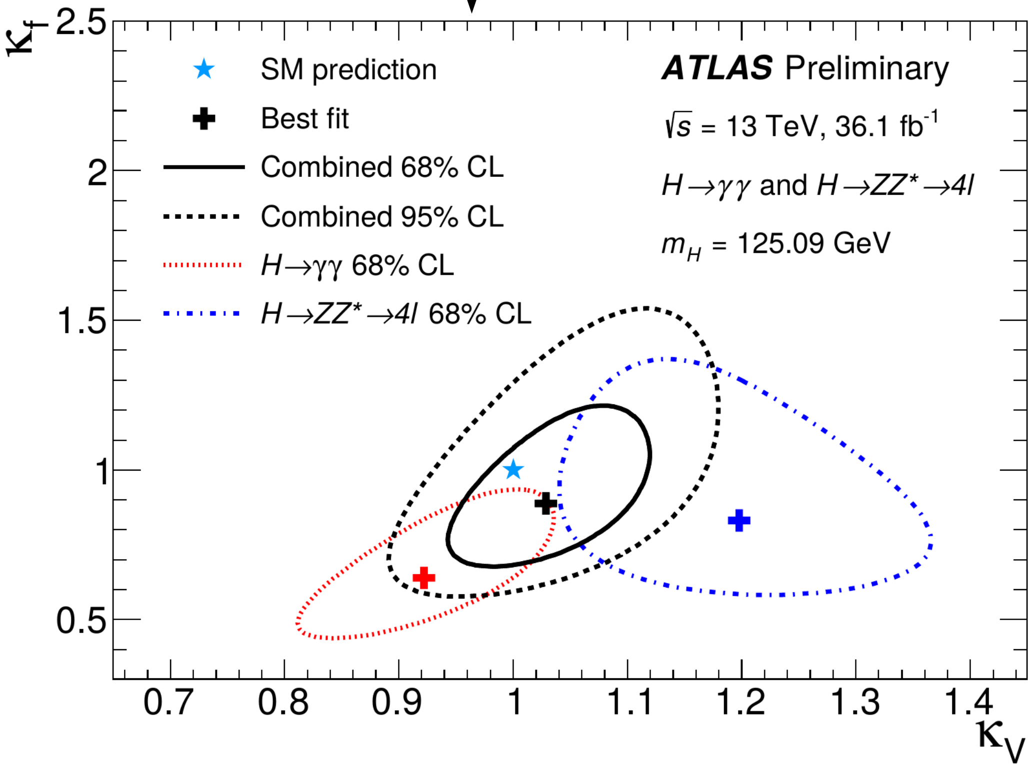
Start with basic measurement in terms of e.g. **σ×B**

→ 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.

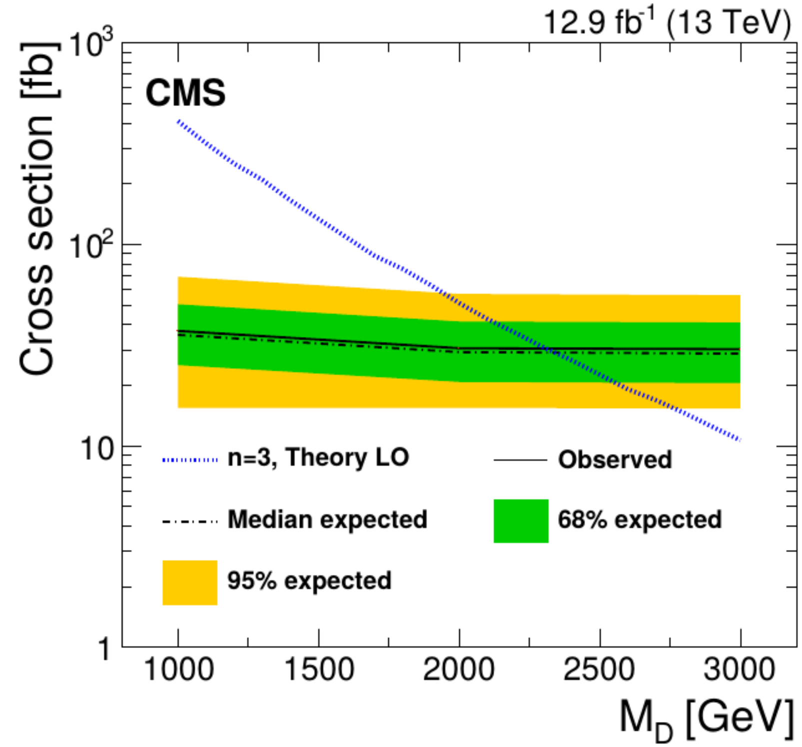
$$L(\sigma_{ggF}, \sigma_{VBF}) \xrightarrow[\sigma_{VBF} \rightarrow \sigma_{VBF}(\kappa_V, \kappa_F)]{\sigma_{ggF} \rightarrow \sigma_{ggF}(\kappa_V, \kappa_F)} L(\sigma_{ggF}(\kappa_V, \kappa_F), \sigma_{VBF}(\kappa_V, \kappa_F)) \equiv L'(\kappa_V, \kappa_F)$$

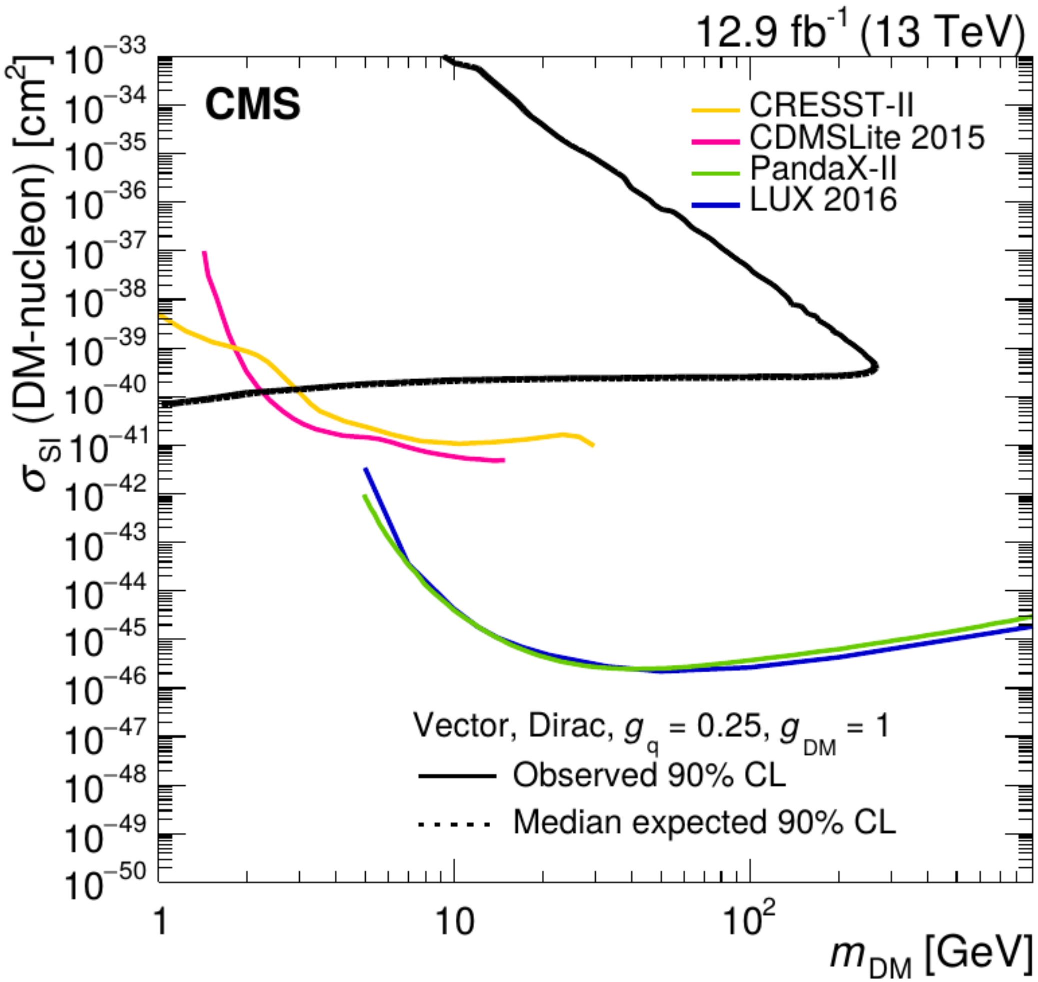


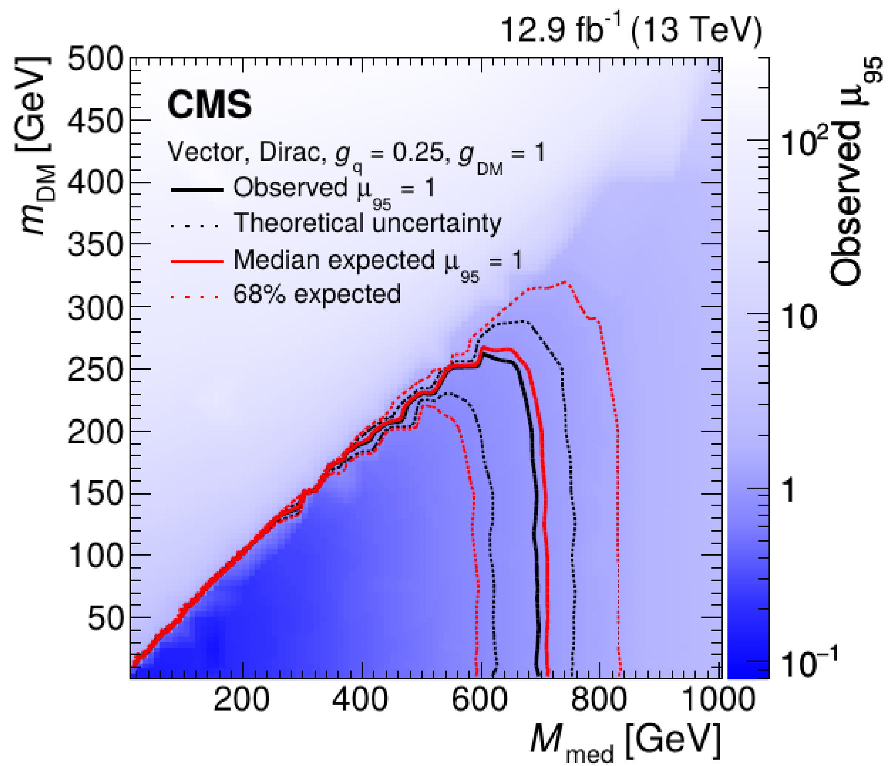


Reparameterization: Limits

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







Takeaways

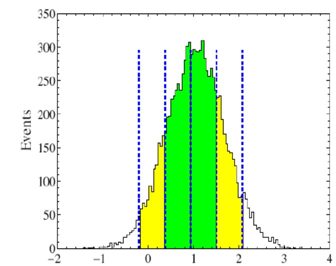
Limits : use LR-based test statistic:

$$\tilde{q}_{\mu_0} = \begin{cases} 0 & \hat{\mu} \geq \mu_0 \\ -2\log \dfrac{L(\mu=\mu_0)}{L(\hat{\mu})} & 0 \leq \hat{\mu} \leq \mu_0 \\ -2\log \dfrac{L(\mu=\mu_0)}{L(\mu=0)} & \hat{\mu} < 0 \end{cases}$$

→ Use **$\mathbf{CL_s}$ procedure** to avoid negative limits

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

Gaussian regime, n=0 : **$\mathbf{S_{up} = 1.96\ \sigma_{Gauss}}$**



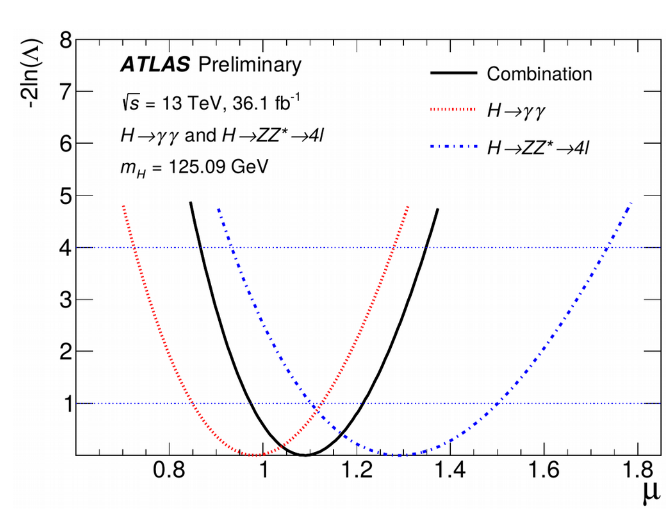
Uncertainty bands: obtain from toys or from Asimov

$$\sigma^2_{S,A} = \frac{S^2}{q_S(\text{Asimov})}$$

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

→ 1D: crossings with $t_{\mu 0} = Z^2$ for $\pm Z\sigma$ intervals

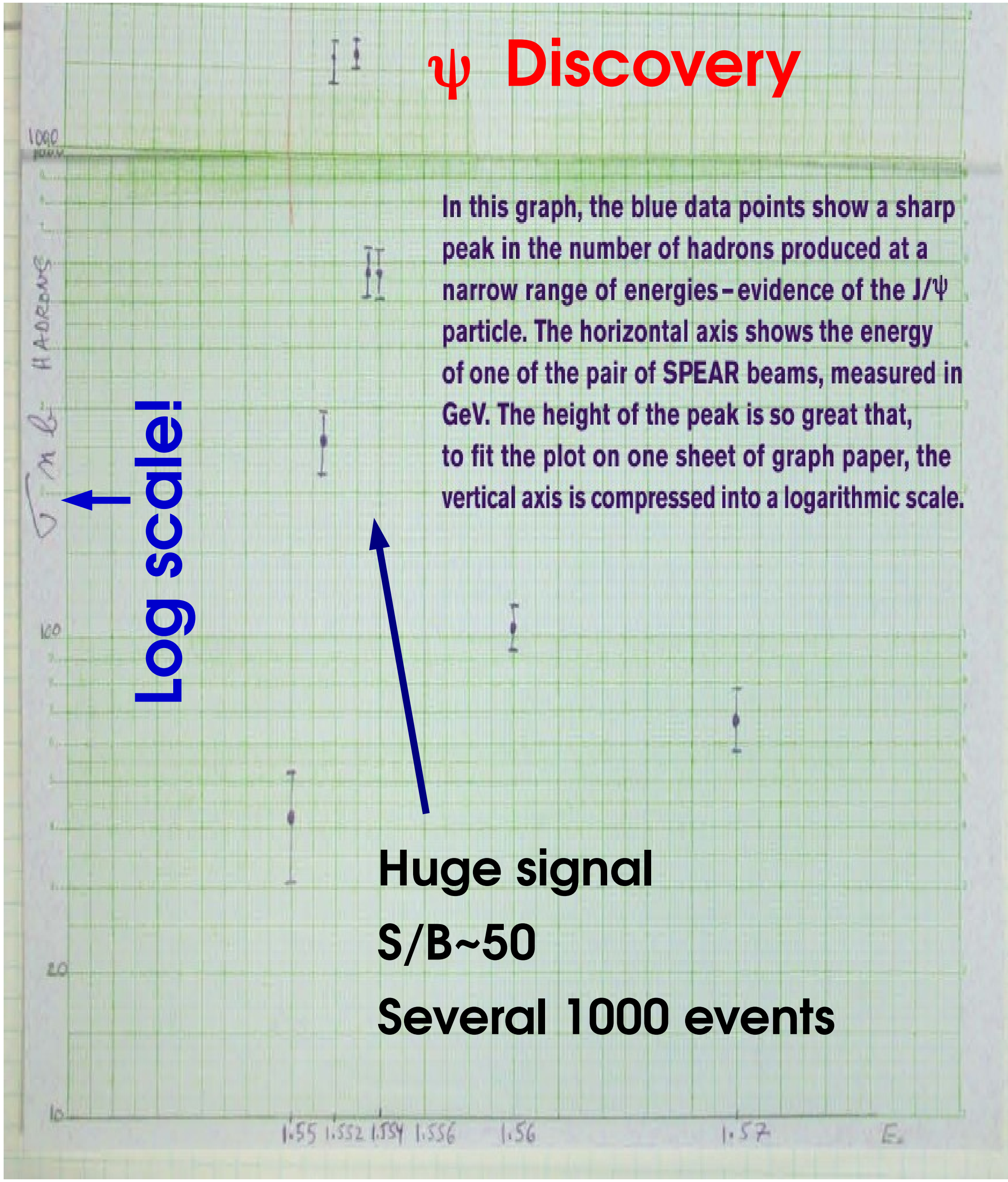
Gaussian regime: $\mu = \hat{\mu} \pm \sigma_{Gauss}$ (1σ interval)



Historical Aside

Classic Discoveries (1)

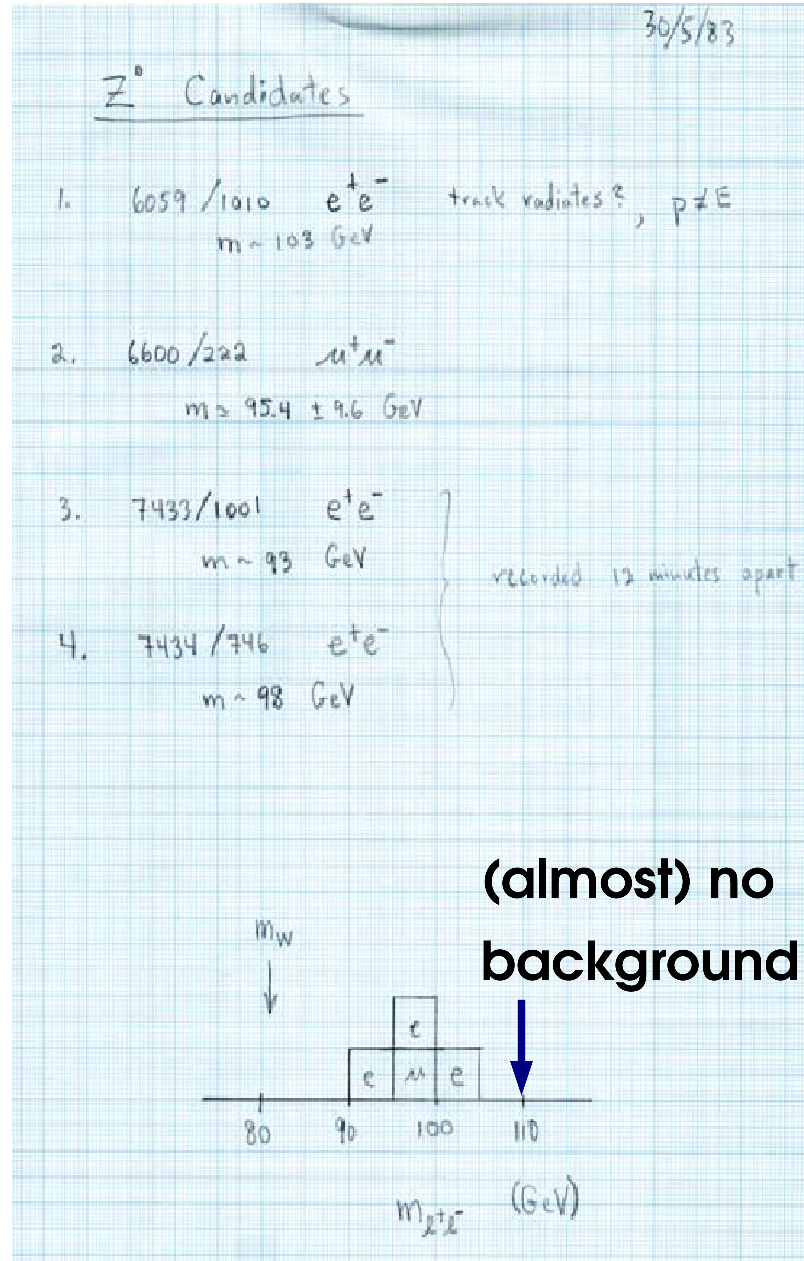
Z^0 Discovery



Huge signal

S/B~50

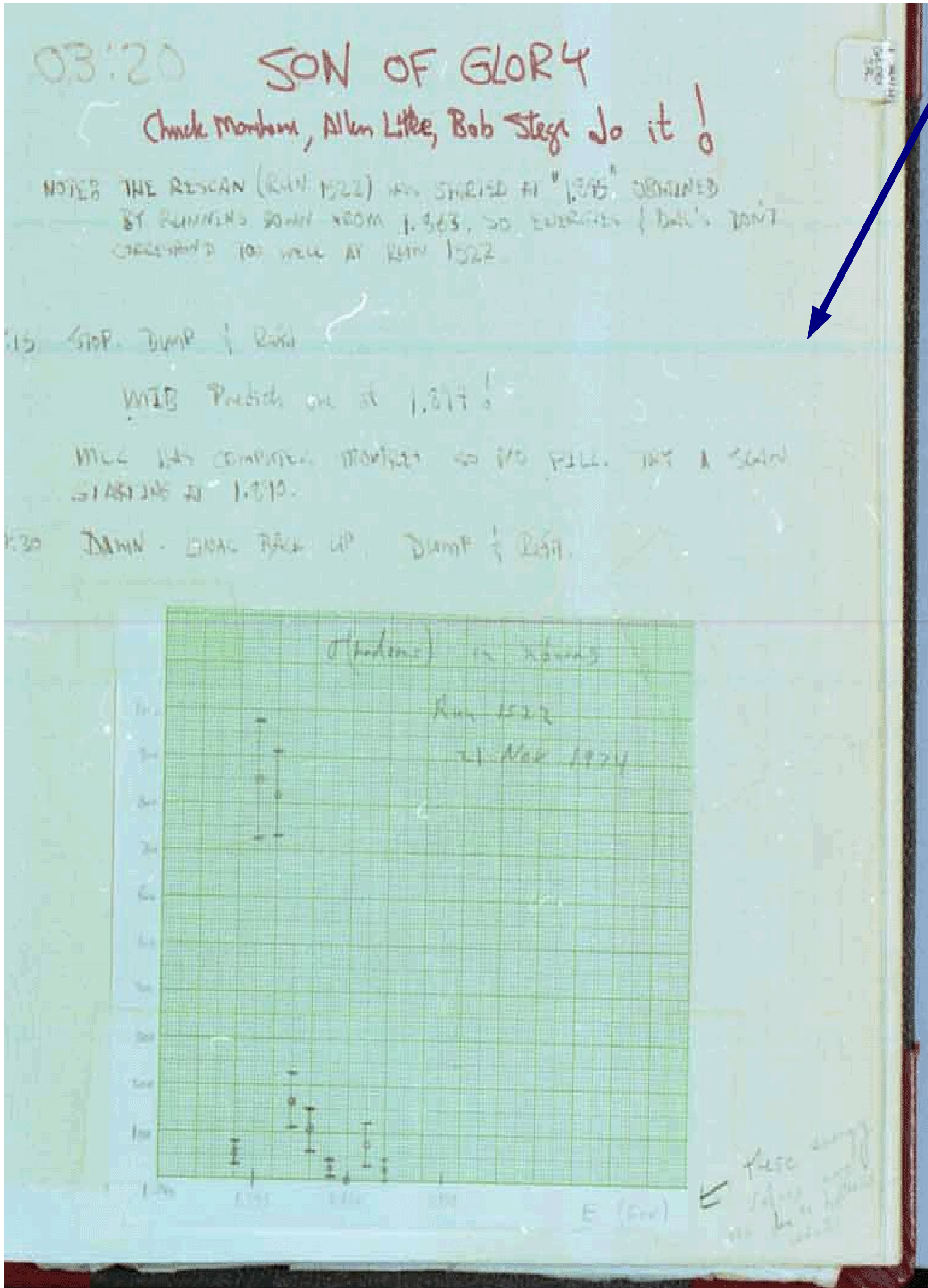
Several 1000 events



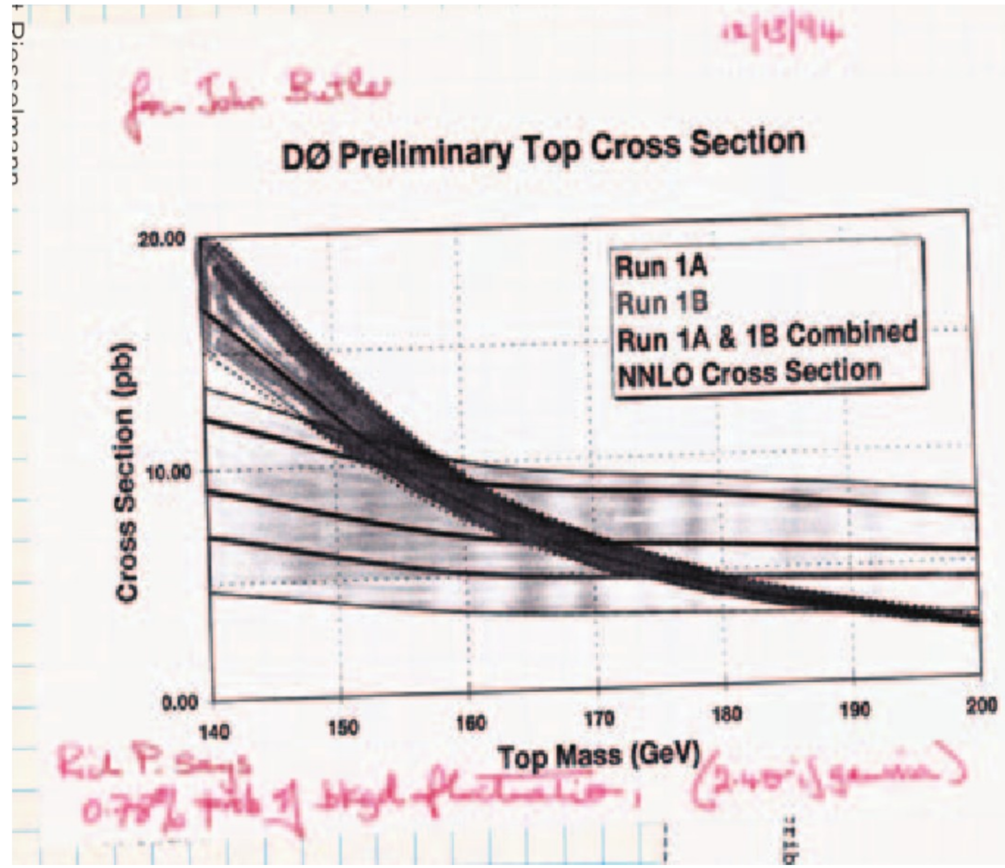
(almost) no background

Logbook of J. Rohlf, 1983-05-30

Classic Discoveries (2)



ψ' : discovered online by the (lucky) shifters



First hints of top at D0:

O(10) signal events,

a few bkg events, 2.4σ

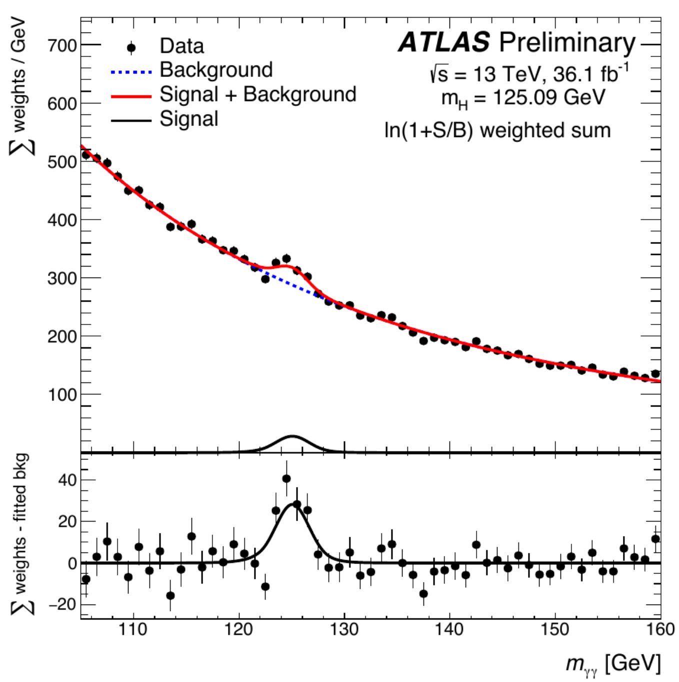
And now ?

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

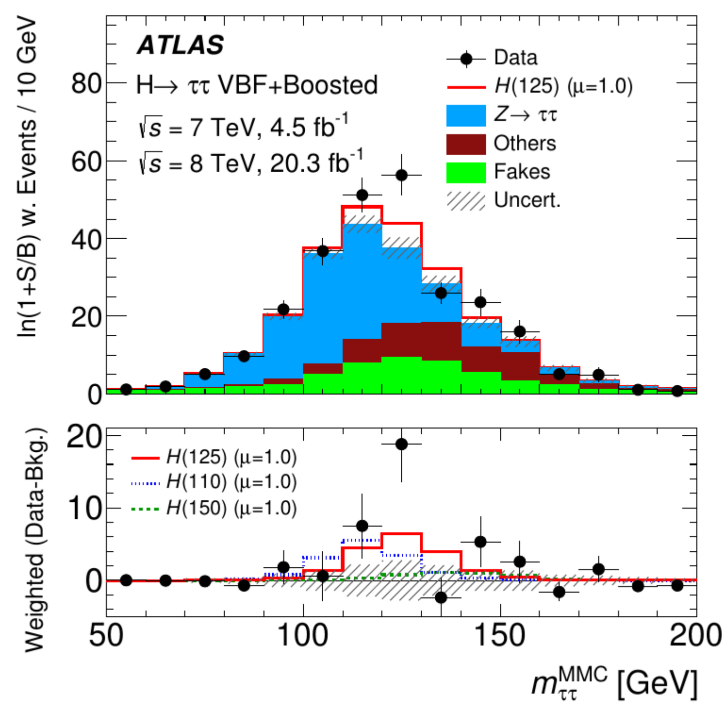
e.g. at LHC:

- **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



Discoveries that weren't

UA1 Monojets (1984)

Volume 139B, number 1,2 PHYSICS LETTERS 3 May 1984

EXPERIMENTAL OBSERVATION OF EVENTS WITH LARGE MISSING TRANSVERSE ENERGY ACCOMPANIED BY A JET OR A PHOTON (S) IN $p\bar{p}$ COLLISIONS AT $\sqrt{s}$ = 540 GeV

UA1 Collaboration, CERN, Geneva, Switzerland

At the present time we can only speculate about the origin of this new effect. The missing transverse energy can be due either to:

(i) One or more prompt neutrinos.

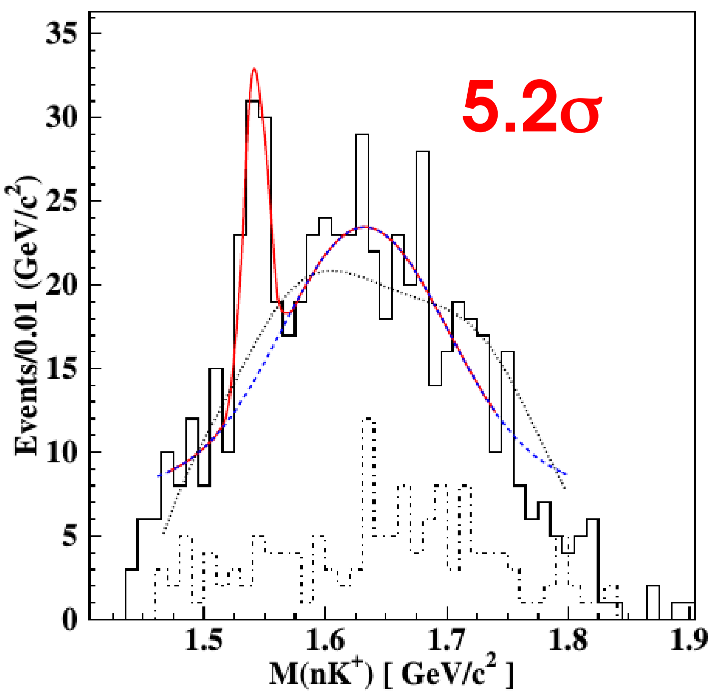
(ii) Any invisible Z^0, such as $Z^0 \to \nu\bar{\nu}$ decay, which is expected to have a large (18%) branching ratio. Note that the corresponding decays into charged lepton pairs $Z^0 \to e^+e^-$, $Z^0 \to \mu^+\mu^-$ have lower branching ratios (~3%) and may not have yet been produced within the present statistics.

(iii) New, non-interacting neutral particles.

The jets appear somewhat narrower and with lower multiplicities than the corresponding QCD jets, although it might be premature to draw conclusions on such limited statistics.

A number of theoretical speculations [9] may be relevant to these results. We mention briefly the possibilities of excited quarks or leptons and of composite or coloured or supersymmetric W's and Higgs. A recent calculation [10] ‡8 has been made in the context of the present collider experiment, on the rate of events with large missing transverse energy from gluino pair production with each gluino decaying into a quark, antiquark, and photino. The non-interacting photinos may produce large apparent missing energy. For instance, the calculation gives an expectation of about 100 single-jet events with $\Delta E_M > 20$ GeV for a gluino mass of 20 GeV/c^2. Taking our excess of 5 events above background as an upper limit for such a process, we deduce that the gluino mass must be greater than about 40 GeV/c^2.

Pentaquarks (2003)



Phys. Rev. Lett. 91, 252001 (2003)

BICEP2 B-mode Polarization (2014)

Selected for a Viewpoint in *Physics*

PRL **112,** 241101 (2014) PHYSICAL REVIEW LETTERS week ending 20 JUNE 2014

Detection of *B*-Mode Polarization at Degree Angular Scales by BICEP2

$$r = 0.20^{+0.07}_{-0.05}, \text{ with } r = 0 \text{ disfavored at } 7.0\sigma.$$

Avoid spurious discoveries!

→ Treatment of modeling uncertainties, systematics in general

Outline

Computing statistics results:

- Limits
- Confidence intervals

Profiling

Look-Elsewhere Effect

Bayesian methods

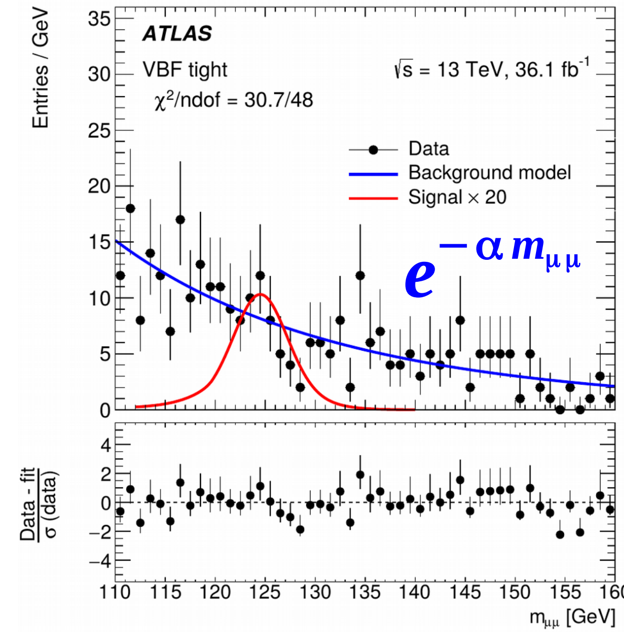
Profiling

Nuisances and Systematics

Phys. Rev. Lett. 119 (2017) 051802

Likelihood typically includes

- **Parameters of interest** (POIs) : **S**, **σ×B**, **m_W**, …
- **Nuisance parameters** (NPs) : other parameters needed to define the model
  → Ideally, **constrained by data** like the POI
  e.g. shape of H→μμ continuum bkg



What about systematics ?

= what we don't know about the random processs

⇒ Parameterize using additional NPs

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

⇒ Cannot be free, or would spoil the measurement (lumi free ⇒ no σ×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 ?***

$$L(\mu, \theta; \mathrm{data}) = L_{\mathrm{measurement}}(\mu, \theta; \mathrm{data})\ C(\theta)$$

POI

Systematics NP

Measurement Likelihood

NP Constraint term

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

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: $L_{\text{aux}}(\theta; \text{aux. data}) = G(\theta^{\text{obs}}; \theta, \sigma_{\text{syst}})$

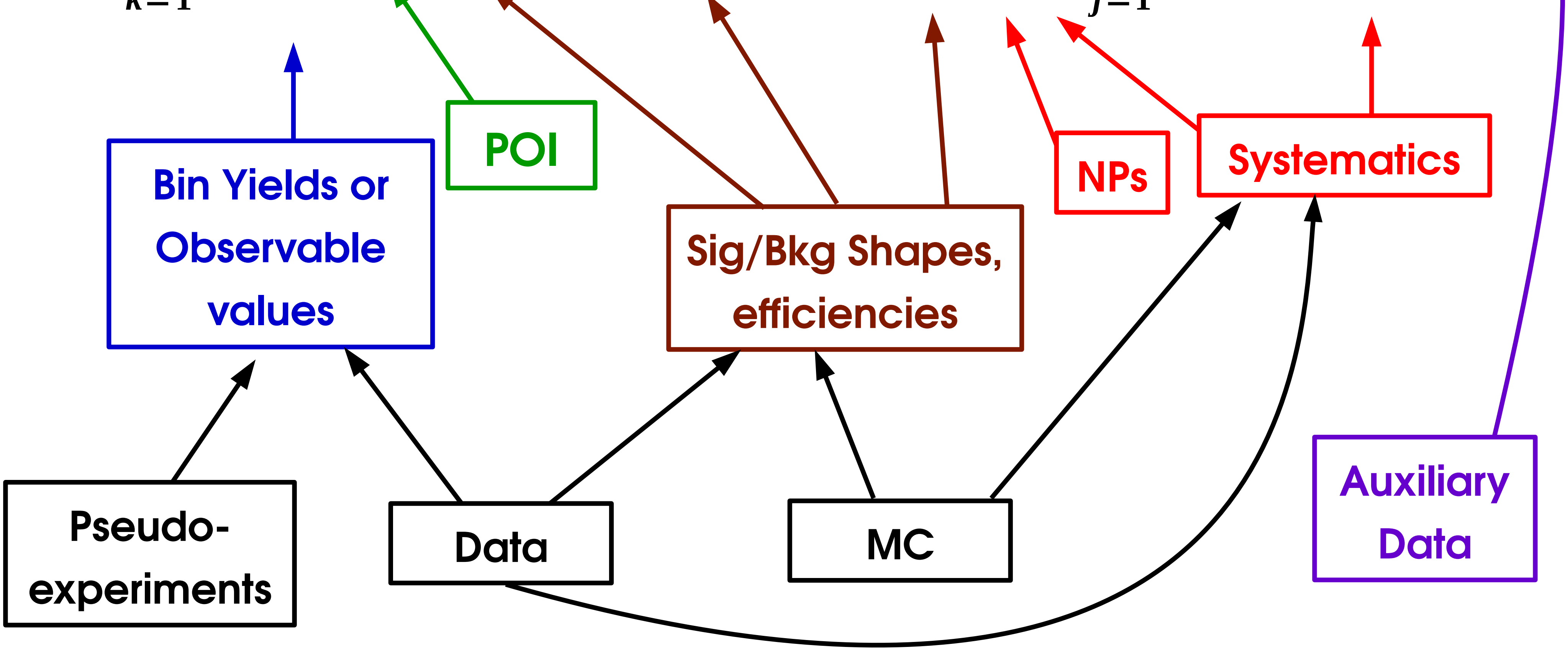
In the combined likelihood, **systematic NPs are constrained**
→ now same as other NPs: **all uncertainties statistical in nature**

→ 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")

Likelihood, the full version (binned case)

$$L\left(\mu,\{\theta_j\}_{j=1\ldots n_{NP}};\{n_i^{(k)}\}_{i=1\ldots n_{data}^{(k)}}^{k=1\ldots n_{cat}},\{\theta_j^{obs}\}_{j=1..n_{NP}}\right)=$$

$$\prod_{k=1}^{n_{cat}} P\left[n_i;\overbrace{\mu\,\epsilon_{i,k}(\vec{\theta})\,N_{S,i,k}(\vec{\theta})+B_{i,k}(\vec{\theta})}^{\text{Expected bin yield}}\right]\prod_{j=1}^{n_{syst}} G\left(\theta_j^{obs};\theta_j;1\right)$$



× number of categories!

Wilks' Theorem

The likelihood usually has NPs:

- **Systematics**
- Parameters fitted in data

→ **What values to use when defining the hypotheses ?** → H(μ=0, θ=?)

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 ! $f\left(t_{\mu_0} \mid \mu=\mu_0\right) = f_{\chi^2(n_{dof}=1)}\left(t_{\mu_0}\right)$

also with NPs present

→ Profiling "builds in" the effect of the NPs

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

Effect of Profiling

Systematics still affect the result even after profiling their NPs!

e.g. **Simple counting experiment: $N(S,\theta) = S + \theta$**, measure N_{obs}, constraint on θ.

1. No NP: $N(S) = S$

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

→ **$\hat{S}$ fit**: adjust S to **$N(\hat{S}) = \hat{S} = n_{obs}$**

→ **$S=S_0$ fit**: $S=S_0$ fixed ⇒ $N(S_0) = S_0$, **cannot adjust**

⇒ **tension** between $N(S_0)=S_0$ and S_{obs} ⇒ large t_{S0} ⇒ **strong exclusion of $H(S_0)$**

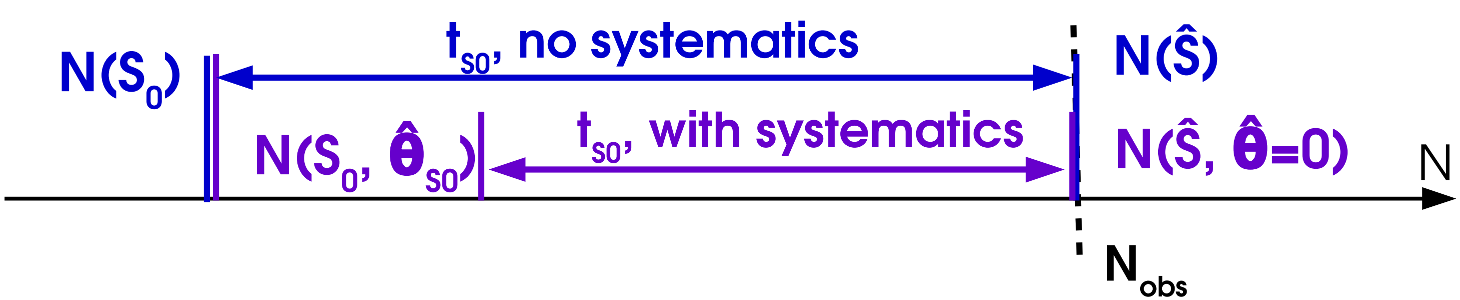
$$t_{S_0} = -2\log\frac{L(S=S_0, \hat{\hat{\theta}}_{S_0}; n_{obs})}{L(\hat{S}, \hat{\theta}; n_{obs})}$$

2. With NP: $N(\mu,\theta) = S + \theta$

→ **$\hat{S}$ fit**: adjust **$N(\hat{S}, \hat{\theta}) = N(\hat{S}, \hat{\theta}=0) = n_{obs}$** using S only (avoid penalty on θ)

→ **$S=S_0$ fit**: $S=S_0$ fixed, but $\hat{\theta}_{S0}$ **can still pull $N(S_0,\hat{\theta}_{\mu0})$ towards N_{obs}**

⇒ smaller t_{S0} ⇒ **reduced exclusion of $H(S_0)$**



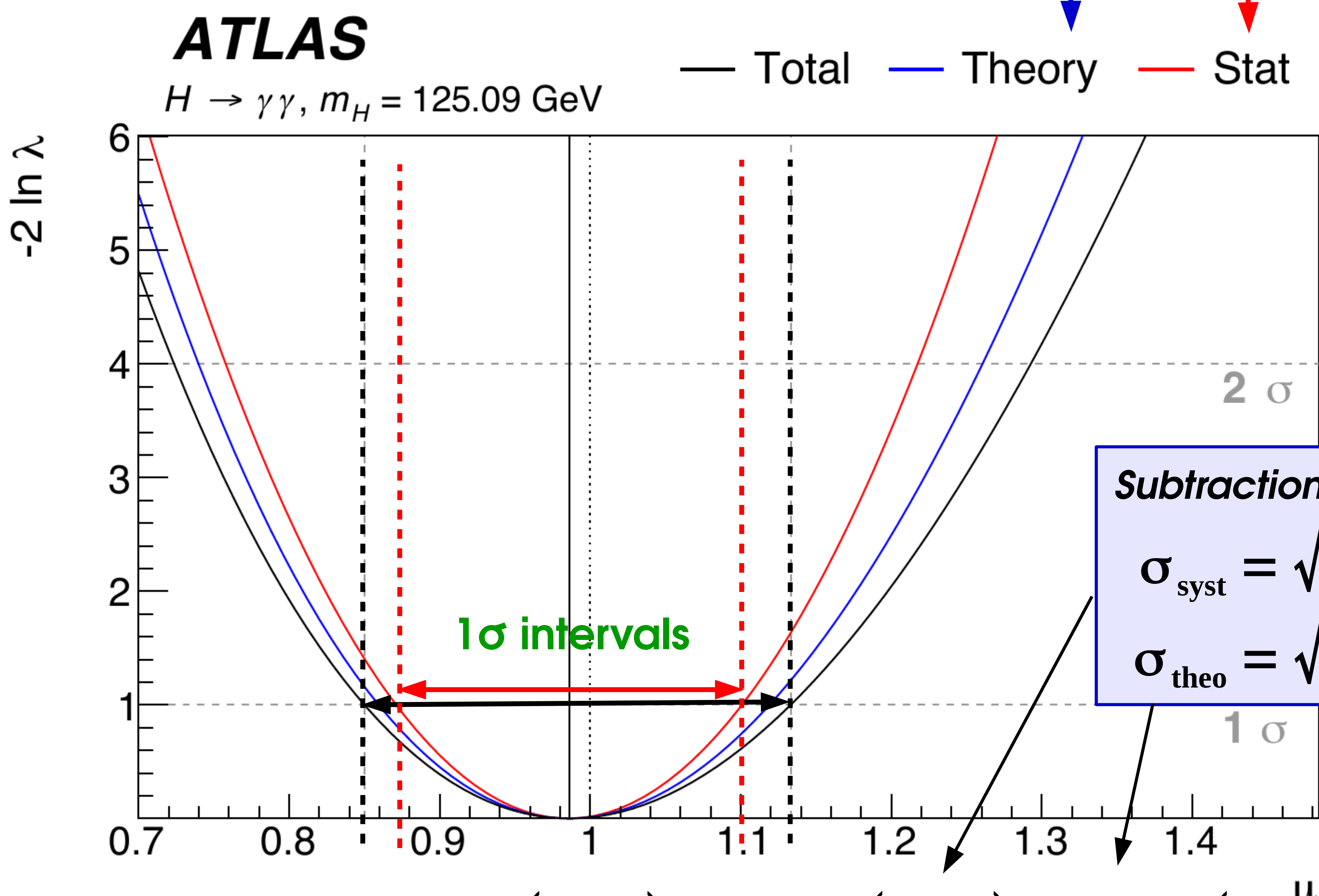
More freedom

⇒ Weaker exclusion

Uncertainty decomposition

All systematics NPs fixed to 0 : statistical uncertainty only

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



Subtraction in quadrature

$$\sigma_{syst} = \sqrt{\sigma^2_{total} - \sigma^2_{stat}}$$

$$\sigma_{theo} = \sqrt{\sigma^2_{stat+theo} - \sigma^2_{stat}}$$

$$\mu = 0.99 \pm 0.12\,(\text{stat}) \pm 0.06\,(\text{syst}) \pm 0.06\,(\text{theo})$$

Gaussian Profiling

Gaussian measurement with 1 POI μ and 1 NP θ:

$$L(\mu,\theta;\hat{\mu},\hat{\theta}) = \exp\left[-\frac{1}{2}\begin{pmatrix}\mu-\hat{\mu}\\ \theta-\hat{\theta}\end{pmatrix}^T C^{-1}\begin{pmatrix}\mu-\hat{\mu}\\ \theta-\hat{\theta}\end{pmatrix}\right] \qquad C = \begin{bmatrix}\sigma_\mu^2 & \gamma\sigma_\mu\sigma_\theta\\ \gamma\sigma_\mu\sigma_\theta & \sigma_\theta^2\end{bmatrix}$$

"data"

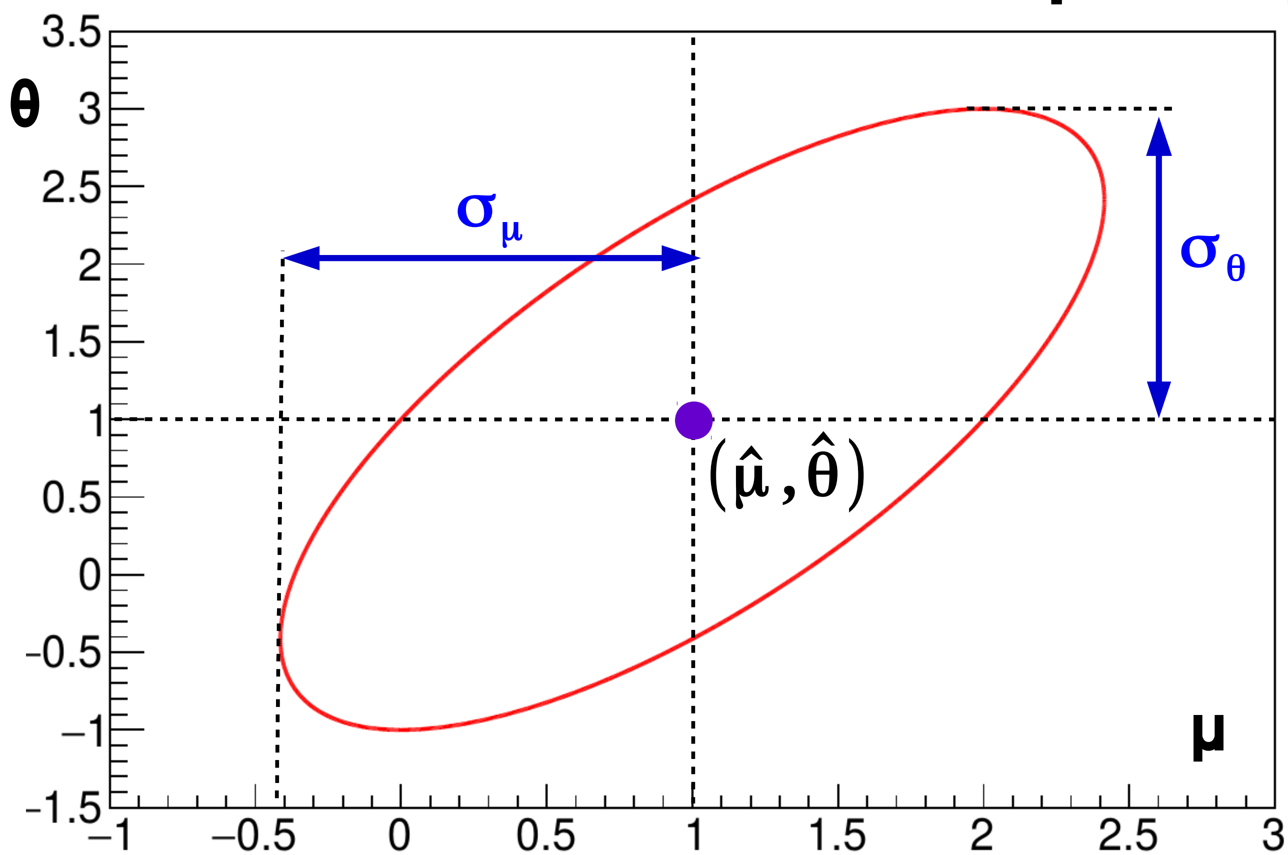
→ λ(μ, θ) defines an **ellipse:**

$$\lambda(\mu,\theta;\hat{\mu},\hat{\theta}) = F_{\mu\mu}(\mu-\hat{\mu})^2 + 2F_{\mu\theta}(\mu-\hat{\mu})(\theta-\hat{\theta}) + F_{\theta\theta}(\theta-\hat{\theta})^2$$

$$F \equiv C^{-1} = \begin{bmatrix}F_{\mu\mu} & F_{\mu\theta}\\ F_{\mu\theta} & F_{\theta\theta}\end{bmatrix}$$

Uncertainty on μ:

- **From C, with θ included:** σ_μ



Gaussian Profiling

$$C = \begin{bmatrix} \sigma_\mu^2 & \gamma\sigma_\mu\sigma_\theta \\ \gamma\sigma_\mu\sigma_\theta & \sigma_\theta^2 \end{bmatrix}$$

$$F = \begin{bmatrix} F_{\mu\mu} & F_{\mu\theta} \\ F_{\mu\theta} & F_{\theta\theta} \end{bmatrix}$$

$$\lambda(\mu,\theta;\hat{\mu},\hat{\theta}) = F_{\mu\mu}(\mu-\hat{\mu})^2 + 2F_{\mu\theta}(\mu-\hat{\mu})(\theta-\hat{\theta}) + F_{\theta\theta}(\theta-\hat{\theta})^2$$

Profiled θ (minimize λ at fixed μ) :

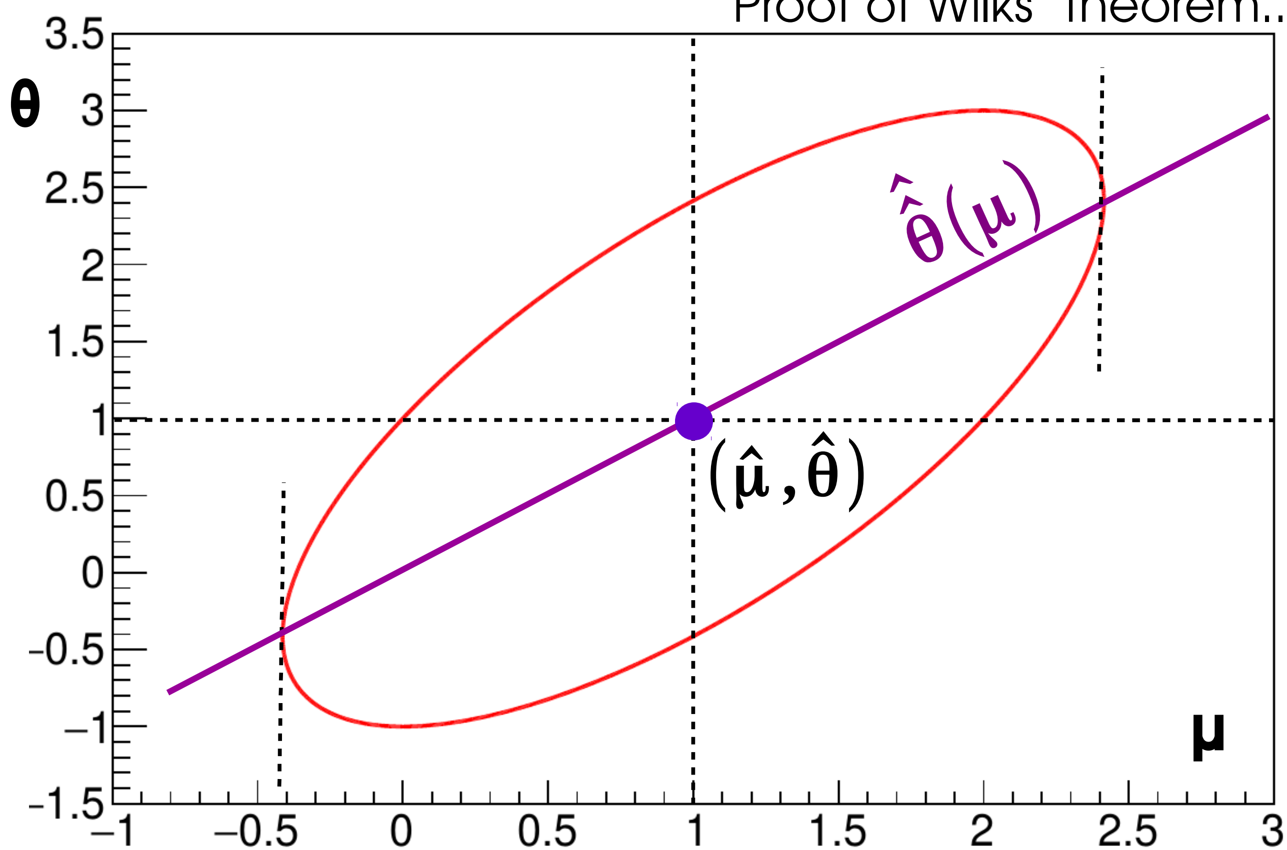
$$\hat{\hat{\theta}}(\mu) = \hat{\theta} - F_{\theta\theta}^{-1}F_{\theta\mu}(\mu-\hat{\mu})$$

Profile likelihood ratio:

$$\lambda(\mu,\hat{\hat{\theta}}(\mu);\hat{\mu},\hat{\theta}) = \left(F_{\mu\mu} - F_{\mu\theta}F_{\theta\theta}^{-1}F_{\theta\mu}\right)(\mu-\hat{\mu})^2 = C_{\mu\mu}^{-1}(\mu-\hat{\mu})^2 = \left(\frac{\mu-\hat{\mu}}{\sigma_\mu}\right)^2$$

$F_{\mu\mu} \neq C_{\mu\mu}^{-1}$!!

Proof of Wilks' theorem...



Uncertainty on μ:

- **From C:** σ_μ
- **From PLR:** σ_μ

Profiled θ **crosses ellipse at vertical tangents** by definition (L is lower at other points on the tangent)

Gaussian Profiling

$$\lambda(\mu,\theta;\hat{\mu},\hat{\theta}) = F_{\mu\mu}(\mu-\hat{\mu})^2 + 2F_{\mu\theta}(\mu-\hat{\mu})(\theta-\hat{\theta}) + F_{\theta\theta}(\theta-\hat{\theta})^2$$

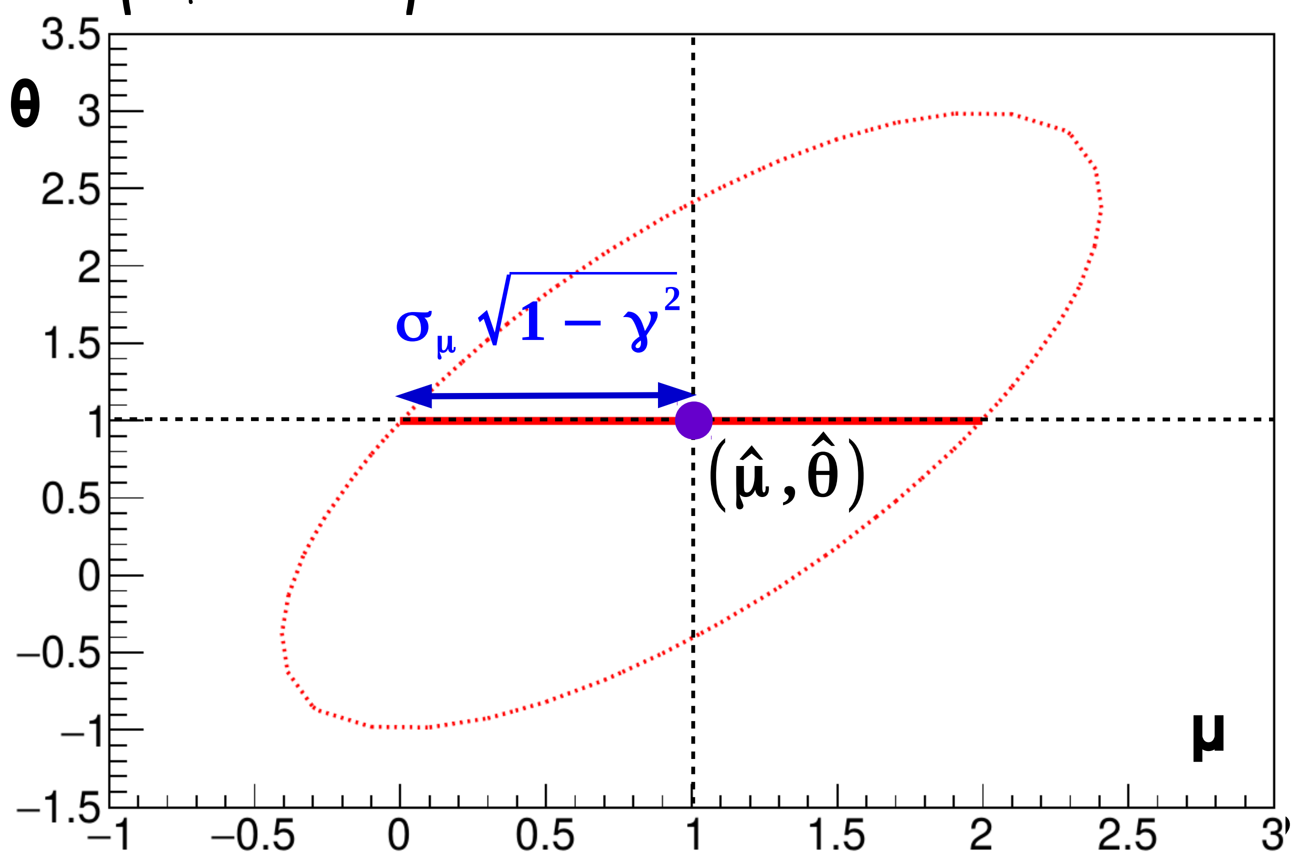
$$F \equiv C^{-1} = \frac{1}{1-\gamma^2}\begin{bmatrix} \frac{1}{\sigma_\mu^2} & \frac{\gamma}{\sigma_\mu\sigma_\theta} \\ \frac{\gamma}{\sigma_\mu\sigma_\theta} & \frac{1}{\sigma_\theta^2} \end{bmatrix}$$

→ For fixed $\theta = \hat{\theta}$, $\lambda(\mu)$ defines an interval:

$$\lambda(\mu,\theta=\hat{\theta};\hat{\mu},\hat{\theta}) = F_{\mu\mu}(\mu-\hat{\mu})^2 = \left(\frac{\mu-\hat{\mu}}{\sigma_\mu\sqrt{1-\gamma^2}}\right)^2$$

Uncertainty on μ:

- **From C:** σ_μ
- **From PLR:** σ_μ
- **From λ(μ):** $\sigma_\mu\sqrt{1-\gamma^2}$



Gaussian Profiling

$$\lambda(\mu,\theta;\hat{\mu},\hat{\theta}) = F_{\mu\mu}(\mu-\hat{\mu})^2 + 2F_{\mu\theta}(\mu-\hat{\mu})(\theta-\hat{\theta}) + F_{\theta\theta}(\theta-\hat{\theta})^2$$

$$F \equiv C^{-1} = \frac{1}{1-\gamma^2}\begin{bmatrix} \frac{1}{\sigma_\mu^2} & \frac{\gamma}{\sigma_\mu\sigma_\theta} \\ \frac{\gamma}{\sigma_\mu\sigma_\theta} & \frac{1}{\sigma_\theta^2} \end{bmatrix}$$

→ For fixed $\theta = \hat{\theta}$, $\lambda(\mu)$ defines an interval:

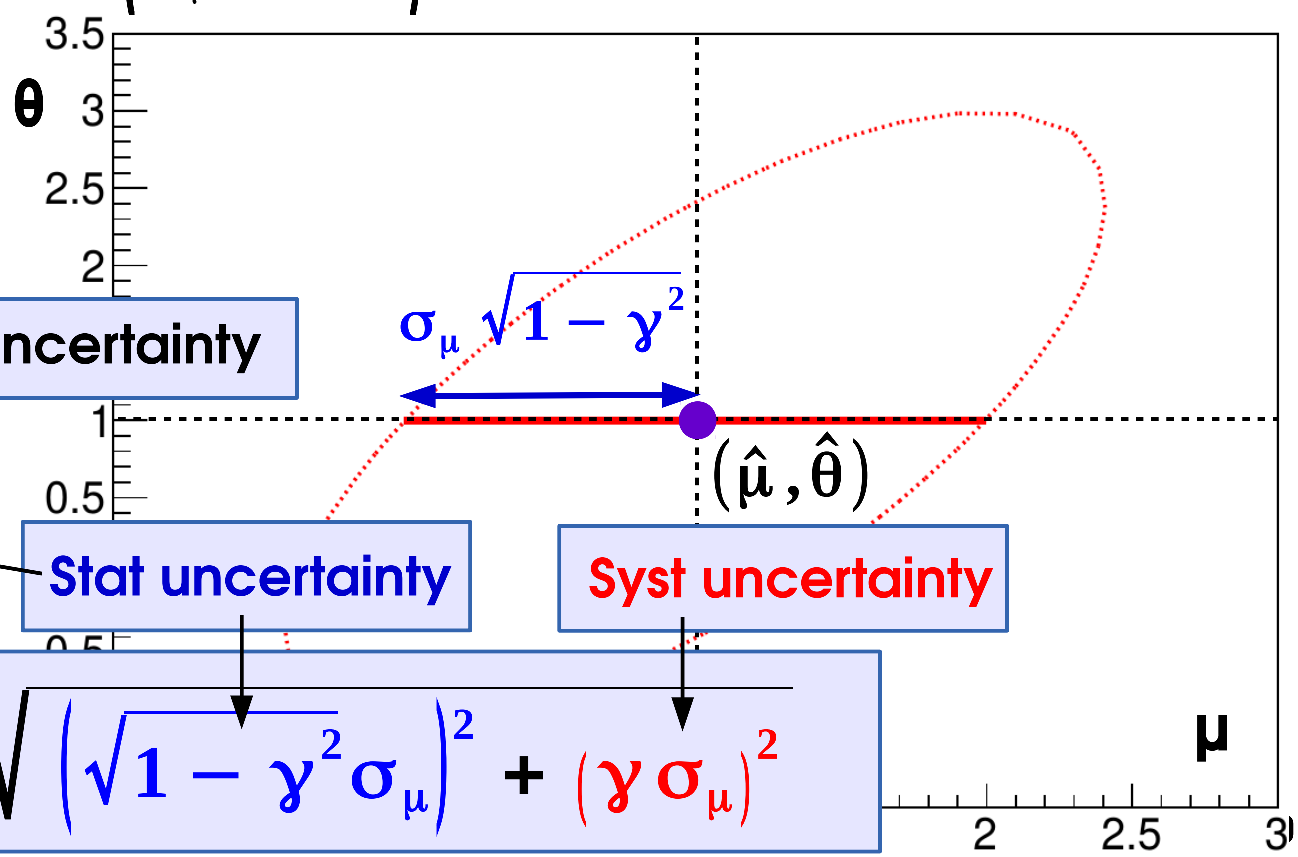
$$\lambda(\mu,\theta=\hat{\theta};\hat{\mu},\hat{\theta}) = F_{\mu\mu}(\mu-\hat{\mu})^2 = \left(\frac{\mu-\hat{\mu}}{\sigma_\mu\sqrt{1-\gamma^2}}\right)^2$$

Uncertainty on μ:

- **From C:** σ_μ
- **From PLR:** σ_μ

Total uncertainty

- **From λ(μ):** $\sigma_\mu\sqrt{1-\gamma^2}$



Stat uncertainty

Syst uncertainty

$$\sigma_\mu = \sqrt{\left(\sqrt{1-\gamma^2}\,\sigma_\mu\right)^2 + (\gamma\,\sigma_\mu)^2}$$

Gaussian Profiling

Back to **$N(S,\theta) = S + \theta$** :

→ Measure **$N_{obs} \sim G(N^*, \sigma_N)$**

→ constraint **$G(\theta, \sigma_\theta)$** on θ

→ everything still Gaussian:

Then:

$$\left.\begin{array}{ll} \sqrt{1-\gamma^2}\,\sigma_\mu = \sigma_N & \textbf{Stat. uncertainty} \\ \gamma\,\sigma_\mu = \sigma_\theta & \textbf{Syst. uncertainty} \end{array}\right\} \quad \sigma_\mu = \sqrt{\sigma_{stat}^2 + \sigma_{syst}^2}$$

⇒ Stat uncertainty (on N) and syst (on θ) add in quadrature as expected

Executive summary:

→ **Systematic = NP with an external constraint** (auxiliary measurement)

→ Profiling systematics includes their effect into the total uncertainty, as desired

→ 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.

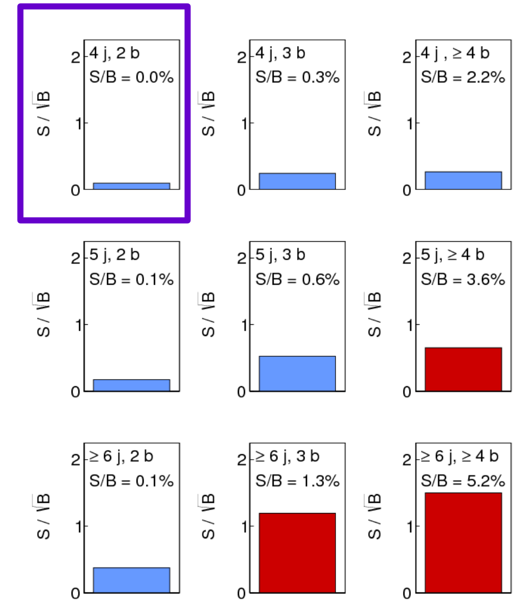
Profiling Example: ttH→bb

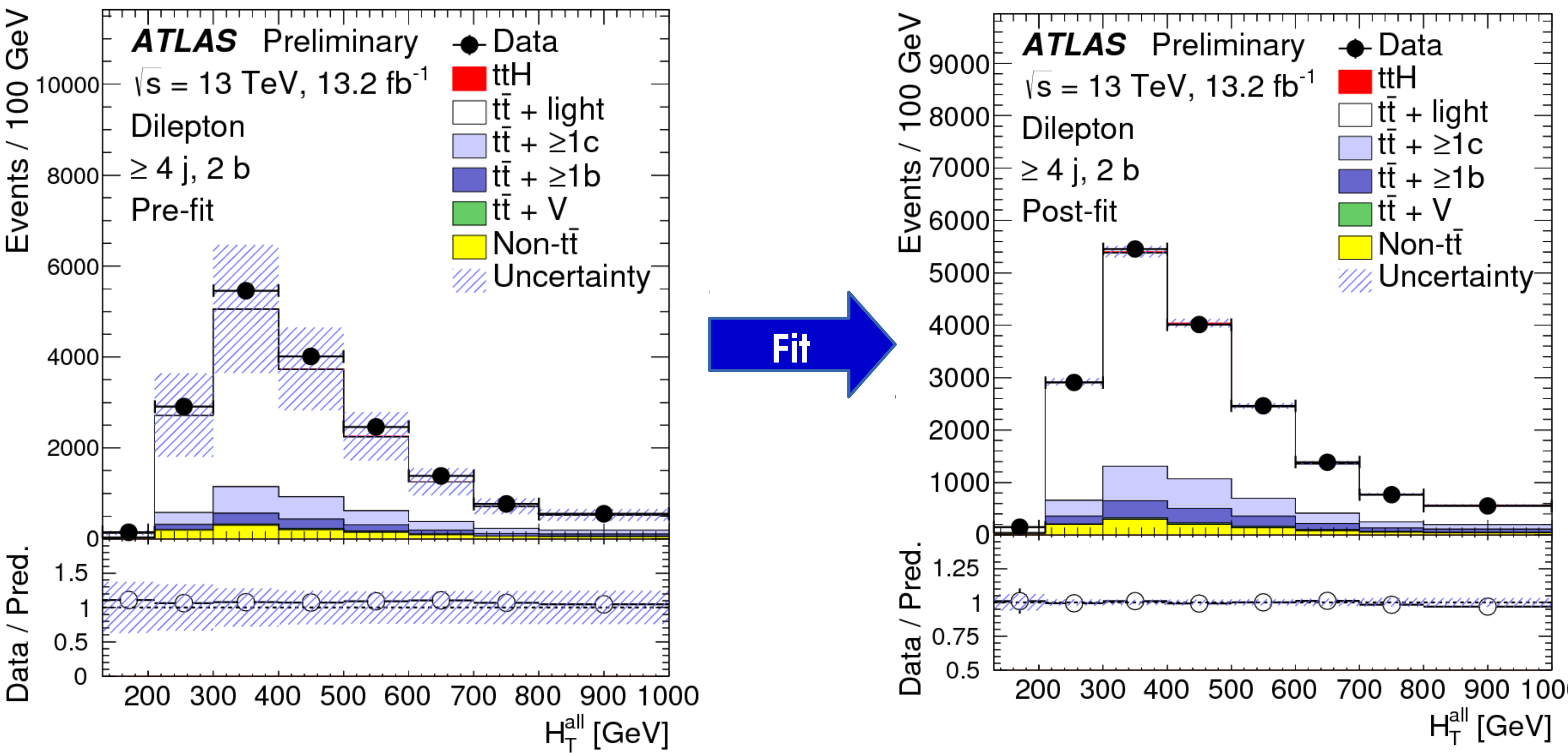
Analysis uses low-S/B categories to constrain backgrounds.

→ **Reduction in large uncertainties on tt bkg**

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

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



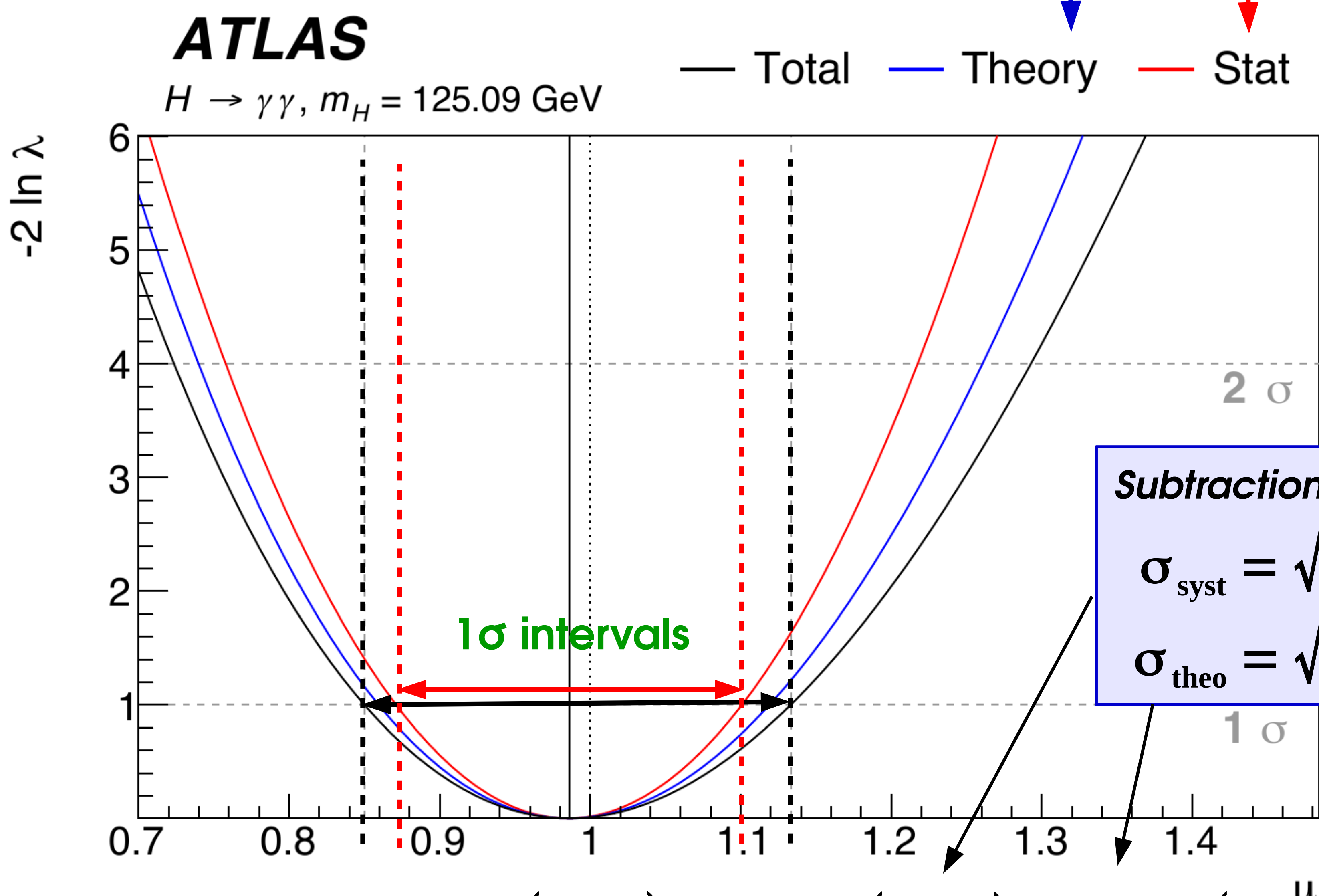


ATLAS-CONF-2016-080

Uncertainty decomposition

All systematics NPs fixed to 0 : statistical uncertainty only

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



Subtraction in quadrature

$$\sigma_{syst} = \sqrt{\sigma^2_{total} - \sigma^2_{stat}}$$

$$\sigma_{theo} = \sqrt{\sigma^2_{stat+theo} - \sigma^2_{stat}}$$

$$\mu = 0.99 \pm 0.12\,(\text{stat}) \pm 0.06\,(\text{syst}) \pm 0.06\,(\text{theo})$$

Pull/Impact plots

ATLAS-CONF-2016-058

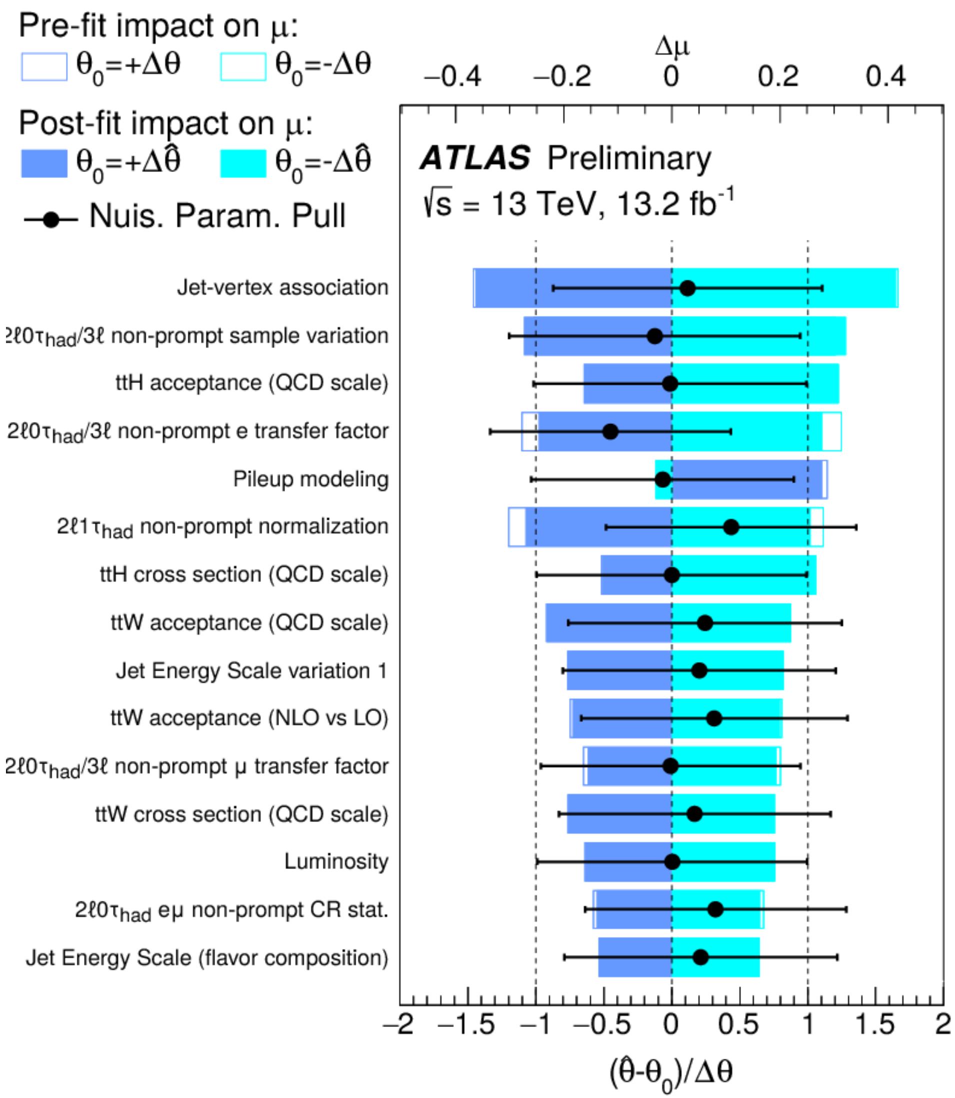
Systematics are described by NPs included in the fit. Nominally:

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

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

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

- **If central value ≠ 0**: some data feature absorbed by nonzero value ⇒ Need investigation if large pull
- **If uncertainty < 1** : systematic is constrained by the data ⇒ Needs checking if this legitimate or a modeling issue
- **Impact on result** of ±1σ shift of NP



Pull/Impact plots

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

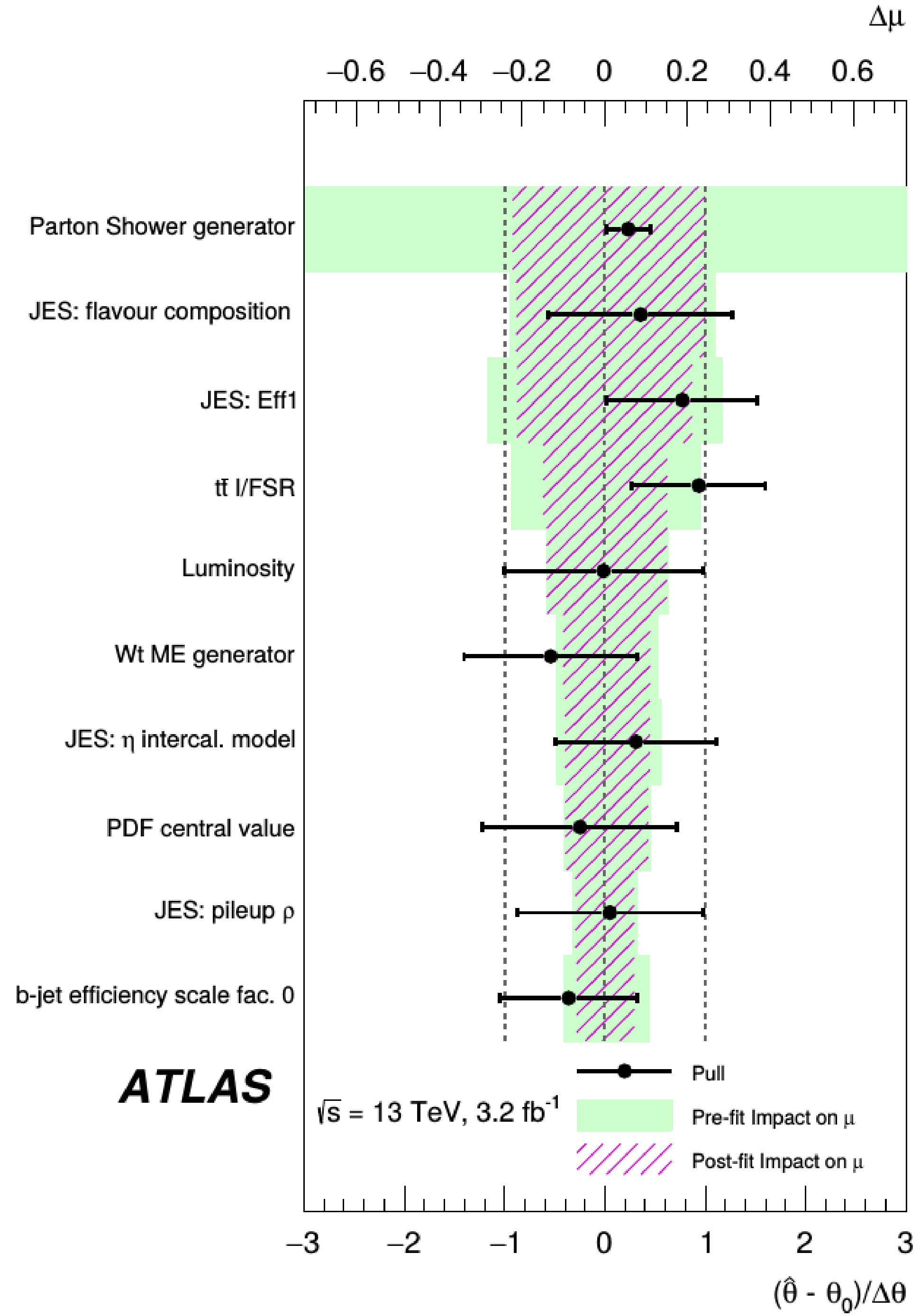
- **NP central value = 0** : corresponds to the pre-fit expectation (usually MC)
- **NP uncertainty = 1** : since NPs normalized to the value of the syst. :

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

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

- **If central value ≠ 0**: some data feature absorbed by nonzero value ⇒ Need investigation if large pull
- **If uncertainty < 1** : systematic is constrained by the data ⇒ Needs checking if this legitimate or a modeling issue
- **Impact on result** of ±1σ shift of NP

13 TeV single-t XS (arXiv: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.

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

⇒ **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***.

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

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:

$$\sigma_{\mathbf{total}} = \sqrt{\sigma^2_{\mathbf{stat}} + \sigma^2_{\mathbf{syst}}}$$

Profiling can have unintended effects – need to carefully check behavior

Summary of 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

Comparison with LEP/TeVatron definitions

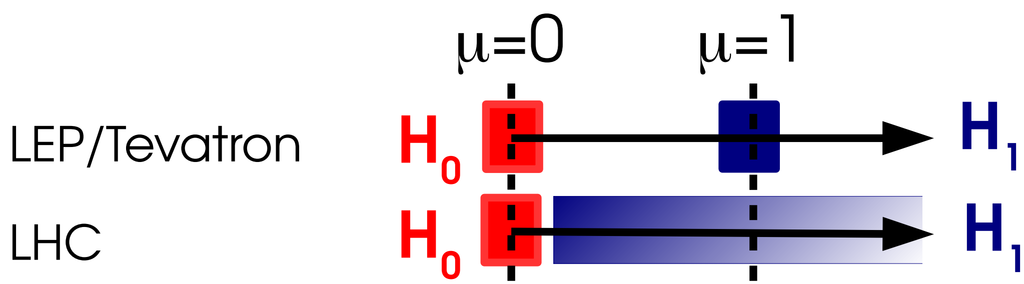
Likelihood ratios are not a new idea:

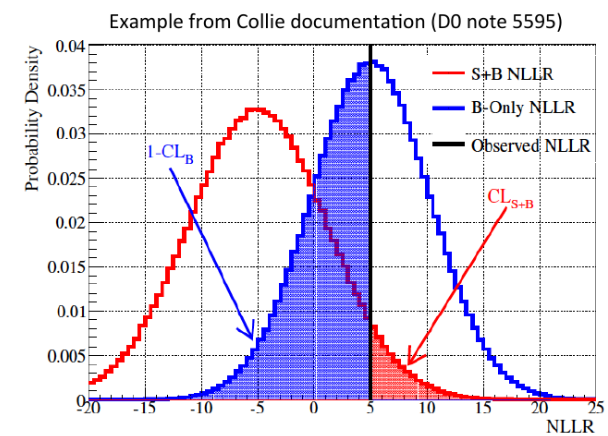
- **LEP**: Simple LR with NPs from MC
  - Compare μ=0 and μ=1
- **Tevatron**: PLR with profiled NPs

$$q_{LEP} = -2\log\frac{L(\mu=0,\tilde{\theta})}{L(\mu=1,\tilde{\theta})}$$

$$q_{Tevatron} = -2\log\frac{L(\mu=0,\hat{\hat{\theta}}_0)}{L(\mu=1,\hat{\hat{\theta}}_1)}$$

Both compare to **μ=1** instead of best-fit **$\hat{\mu}$**

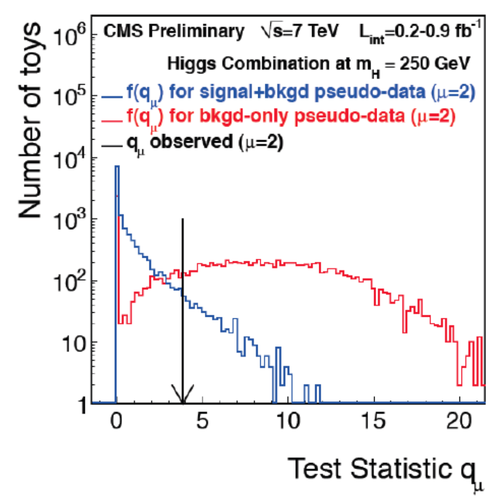




→ Asymptotically:

- **LEP/Tevaton**: q linear in μ ⇒ **~Gaussian**
- **LHC**: q quadratic in μ ⇒ **~χ2**

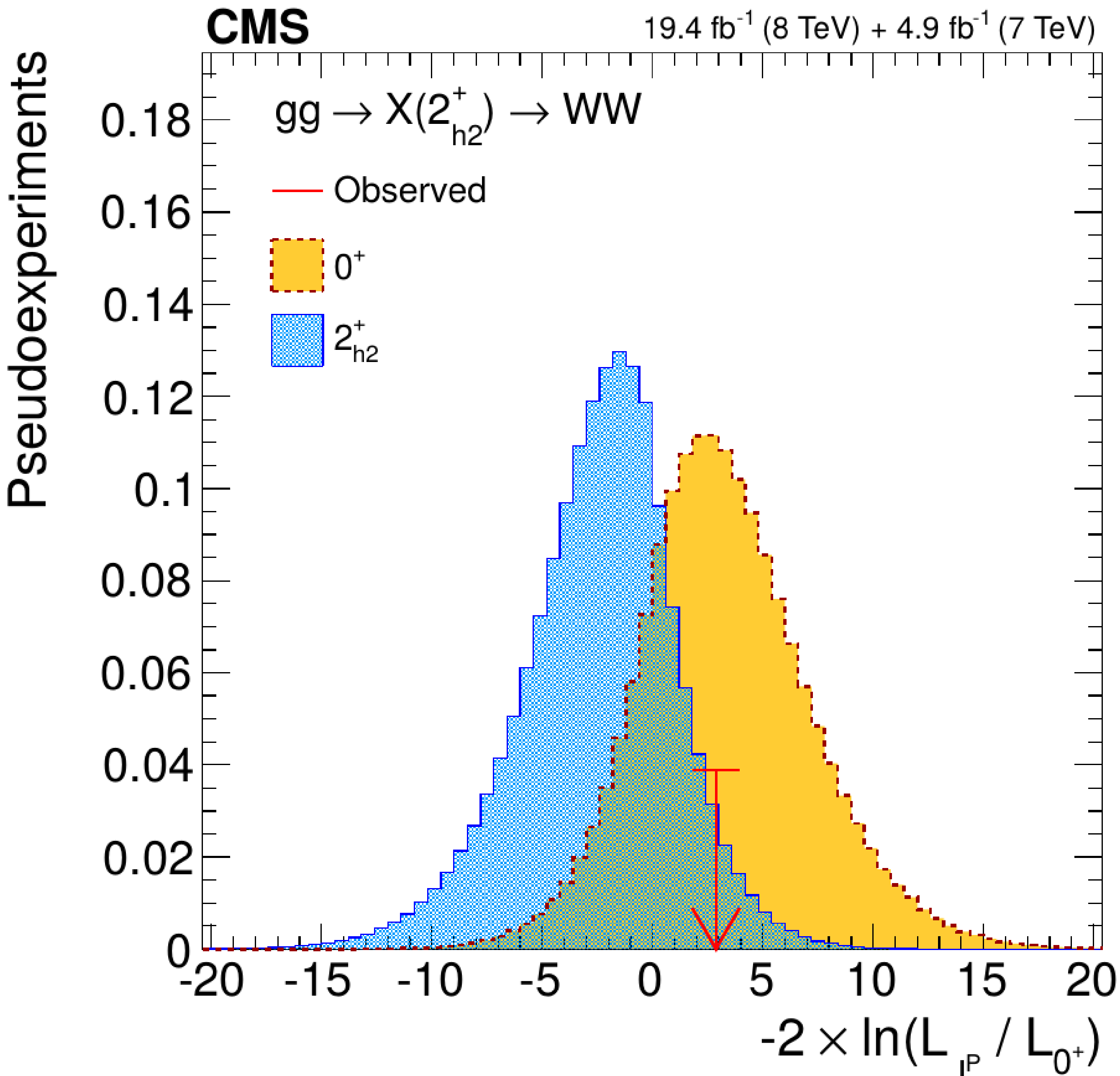
→ Still use TeVatron-style for discrete cases



Andrey Korytov, EPS 2011

Spin/Parity Measurements

Phys. Rev. D 92 (2015) 012004



Beyond Asymptotics: Toys

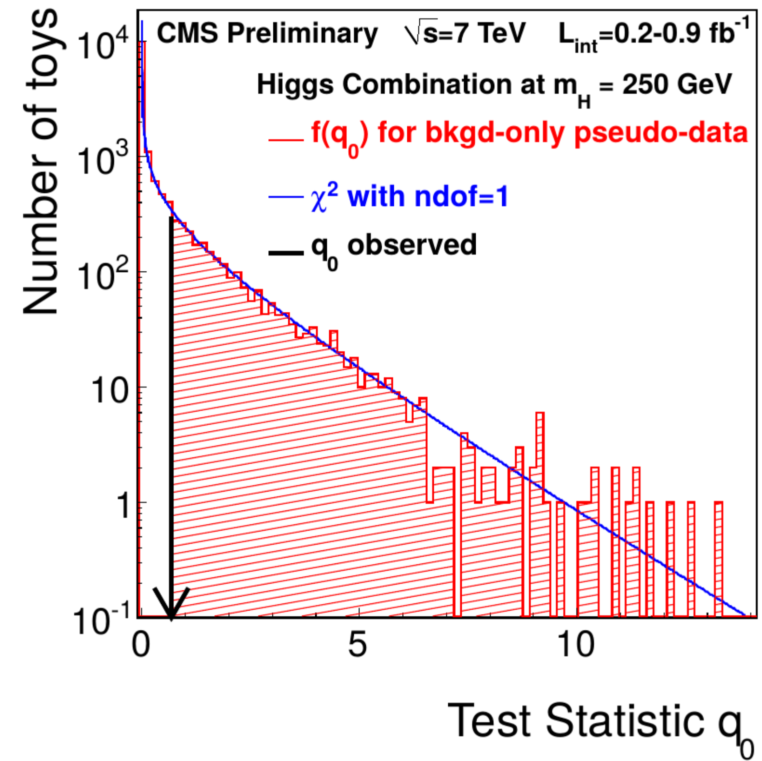
CMS-PAS-HIG-11-022

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

→ Samples the true distribution of the PLR

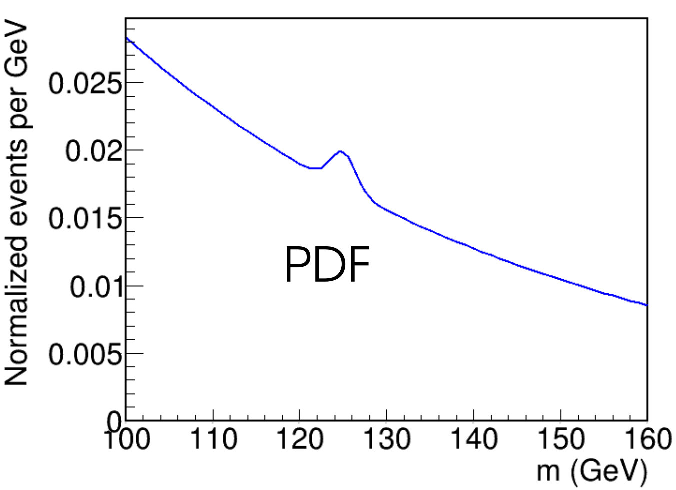


⇒ Integrate above observed PLR to get the p-value

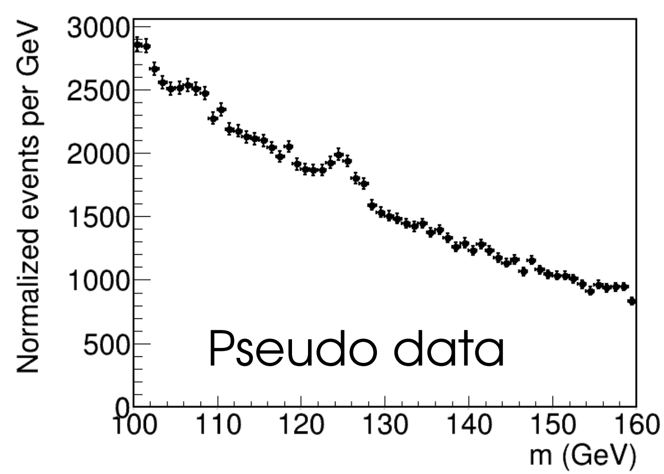
→ Precision limited by number of generated toys,

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

Repeat N_{toys} times



p(data | x) →



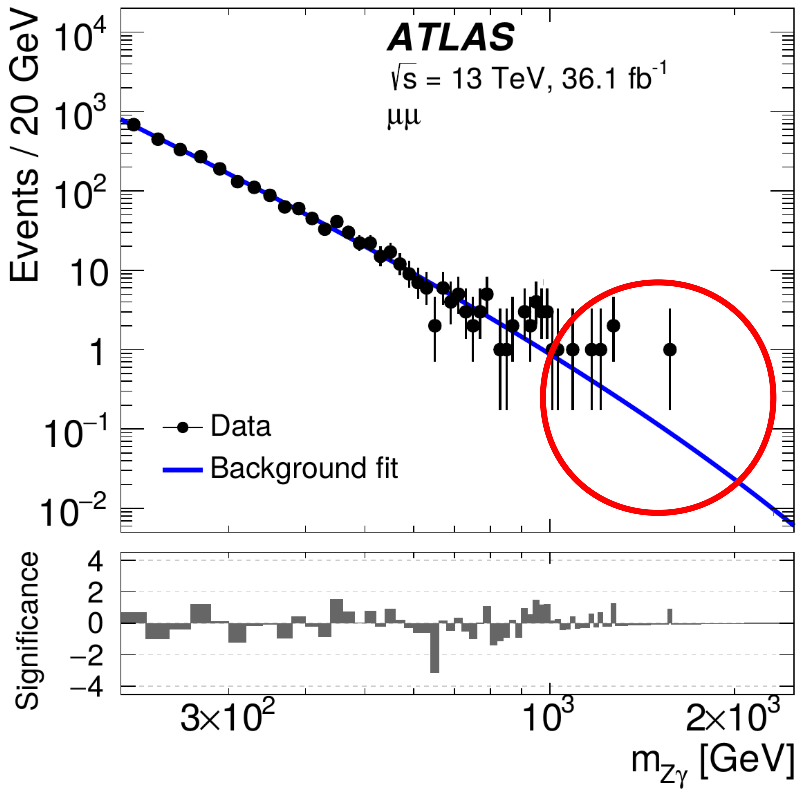
→ q_0

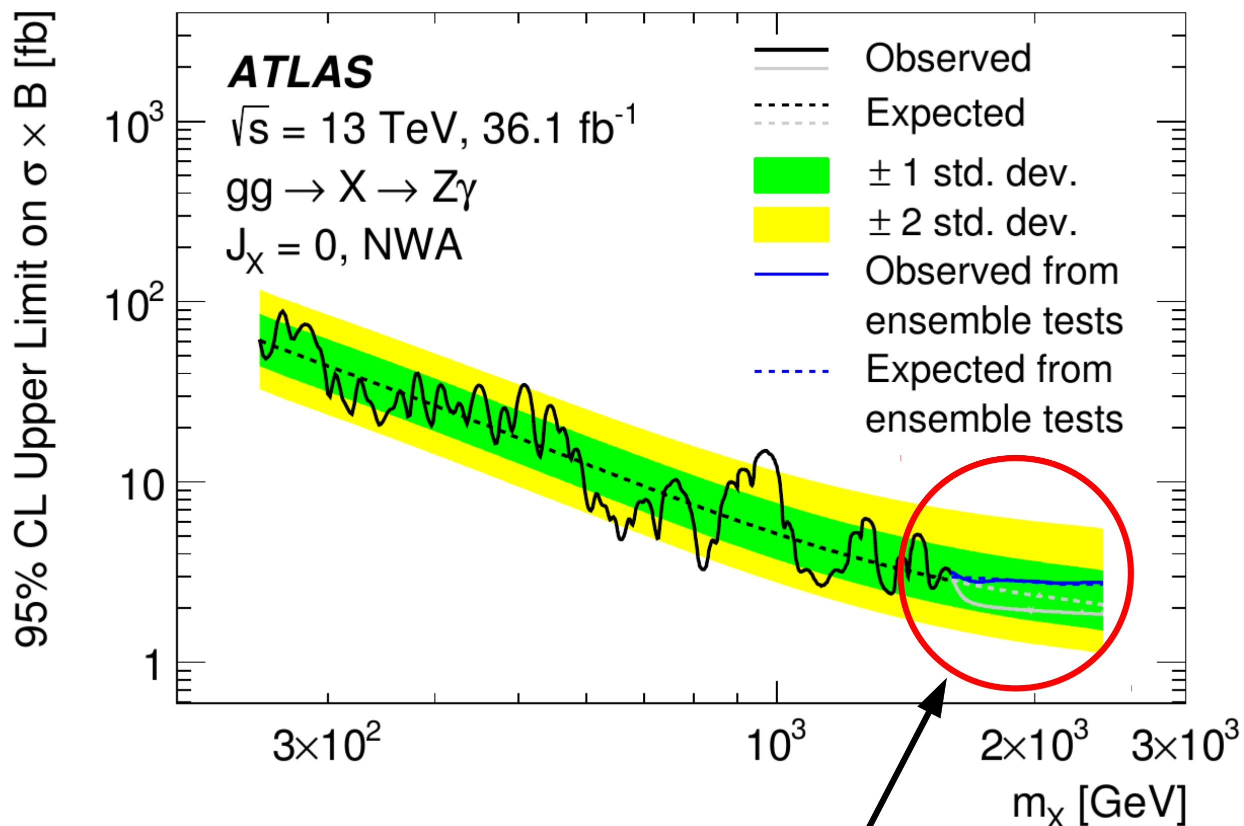
Toys: Example

arXiv:1708.00212

ATLAS $X \to Z\gamma$ Search: covers 200 GeV < m_X < 2.5 TeV

→ for m_X > 1.6 TeV, low event counts ⇒ derive results from toys





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

Outline

Computing statistics results:

- Limits
- Confidence intervals

Profiling

Look-Elsewhere Effect

Bayesian methods

Look-Elsewhere Effect

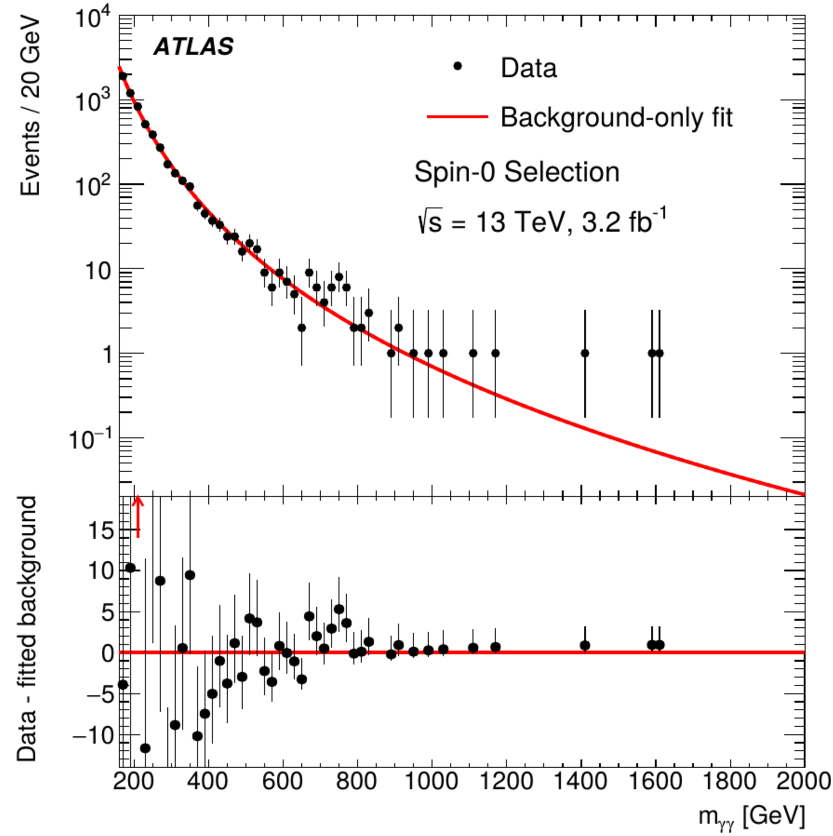
Look-Elsewhere effect

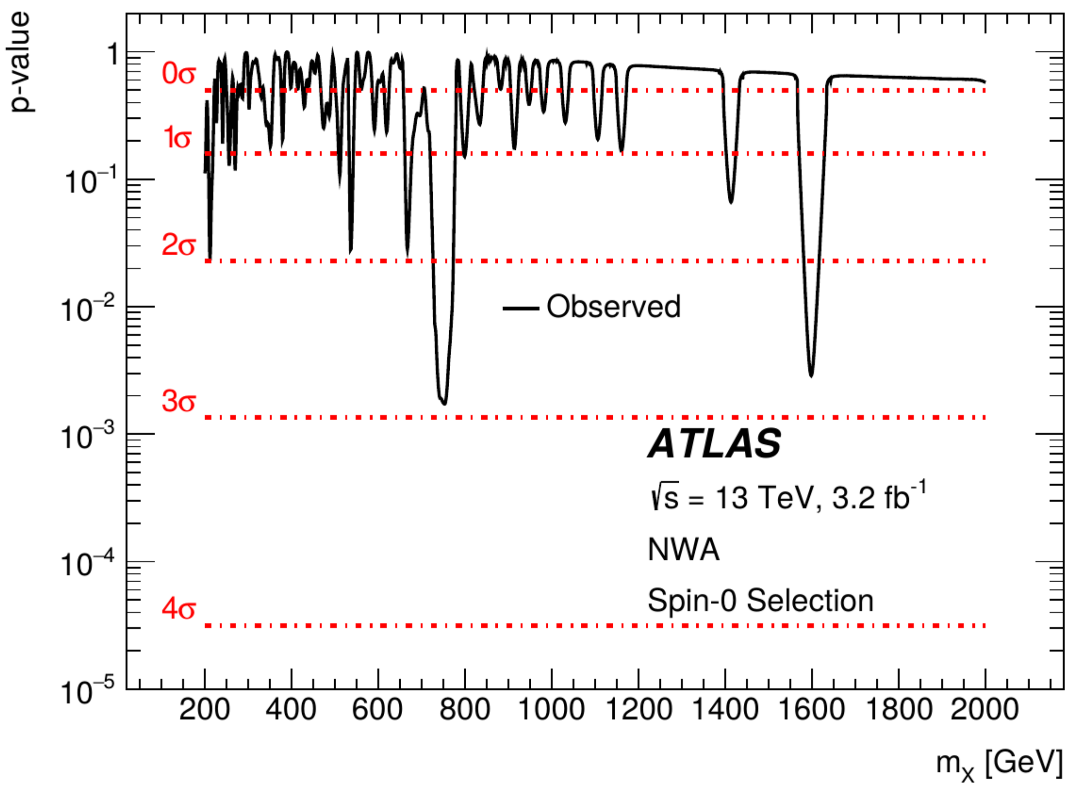
Sometimes, unknown parameters in signal model

e.g. p-values as a function of m_X

⇒ Effectively performing **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)

Testing the same H_0, but against different alternatives

⇒ different p-values

Global Significance

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

Global p-value → $p_{global} = 1 - (1 - p_{local})^N \approx N \; p_{local}$ ← **Local p-value**

Trials factor → N

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

⚠

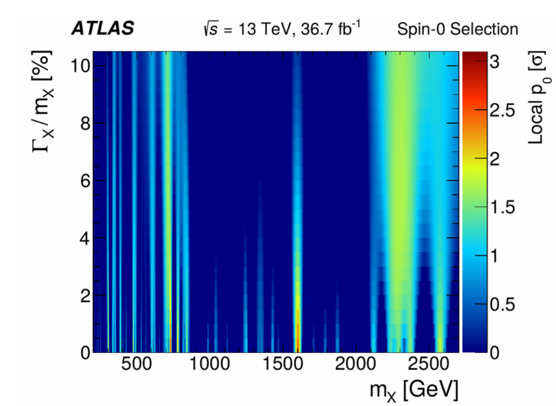
Trials factor : **naively** = # of independent intervals: $N_{\text{trials}} \overset{??}{=} N_{\text{indep}} = \dfrac{\text{scan range}}{\text{peak width}}$
However this is usually **wrong** – more on this later

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

→ Depends on the scanned parameter ranges

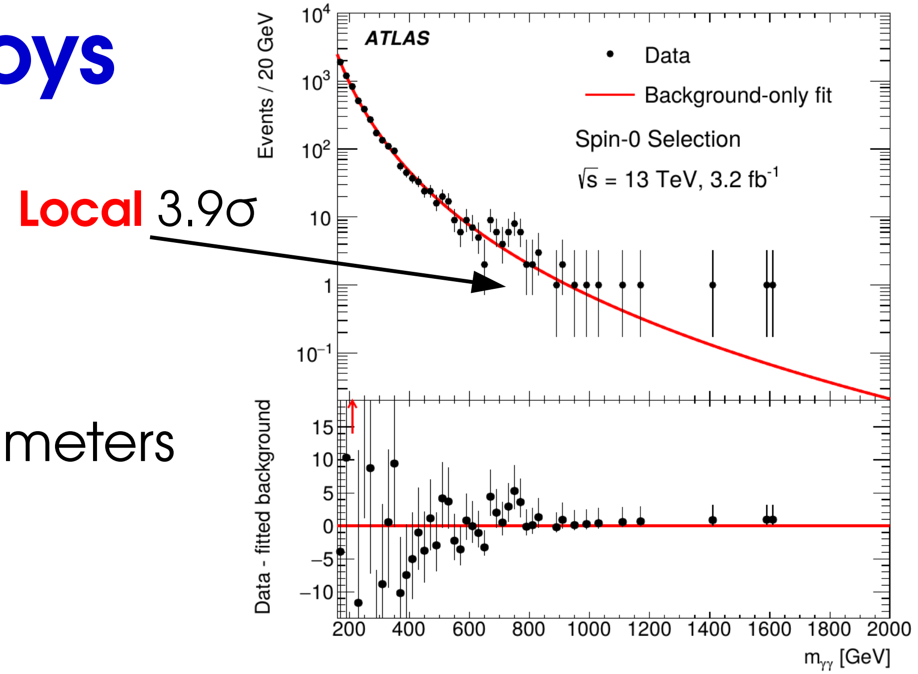
e.g. X→γγ : 200 < m_X< 2000 GeV, 0 < Γ_X < 10% m_X.

→ 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:

→ a 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→γγ Search**: Z_{local} **= 3.9σ** (⇒ $p_{local} \sim 5\ 10^{-5}$),
scanning $200 < m_X < 2000$ GeV and $0 < \Gamma_X < 10\%\ m_X$

→ In toys, find such an excess 2% of the time
⇒ $p_{global} \sim 2\ 10^{-2}$, Z_{global} **= 2.1σ** Less exciting...

⊕ **Exact treatment**
⊖ **CPU-intensive** especially for large Z (need ~$O(100)/p_{global}$ 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} !

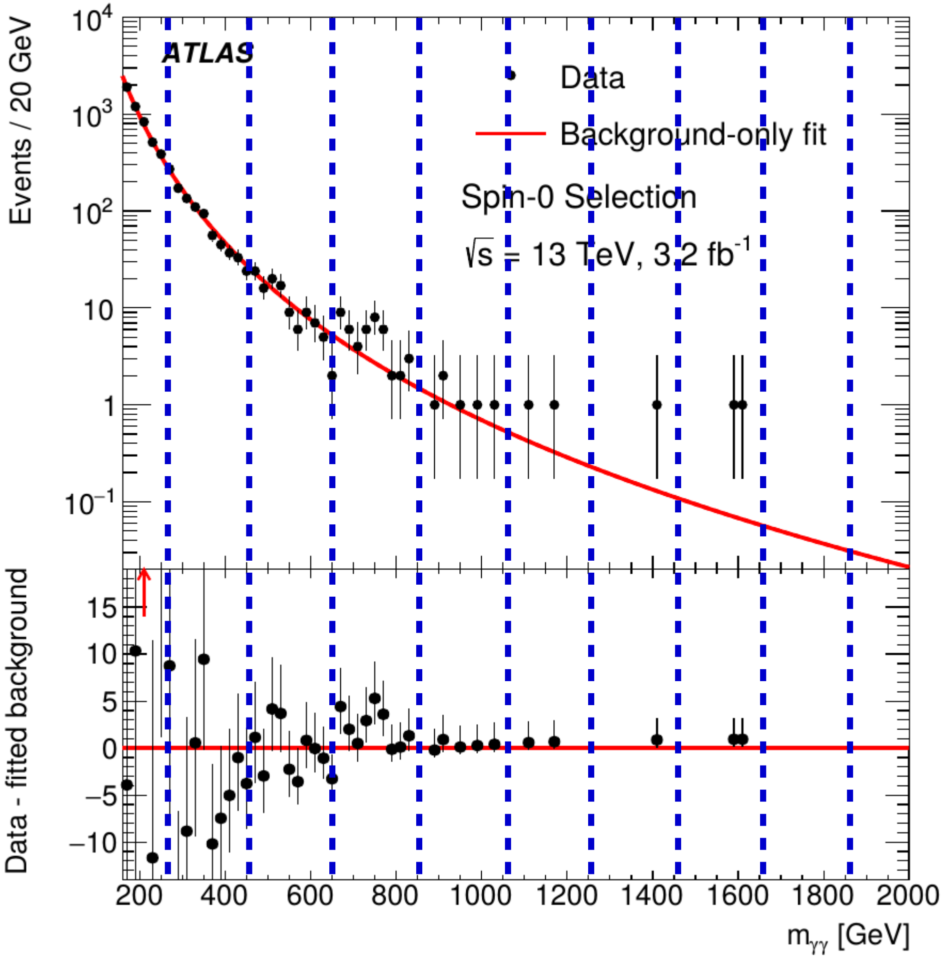
Why ?

→ slice scan range into N_{indep} regions
of size ~ 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 ?

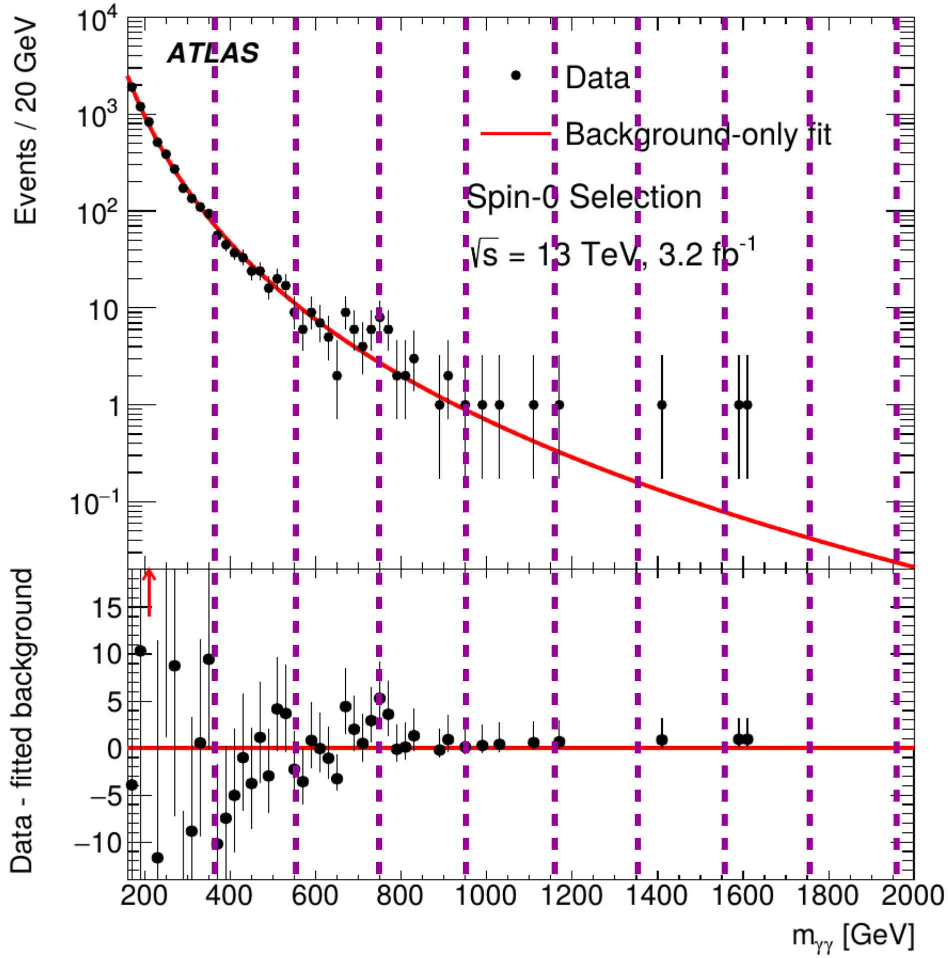
→ slice scan range into N_{indep} regions of size ~ peak width

→ 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

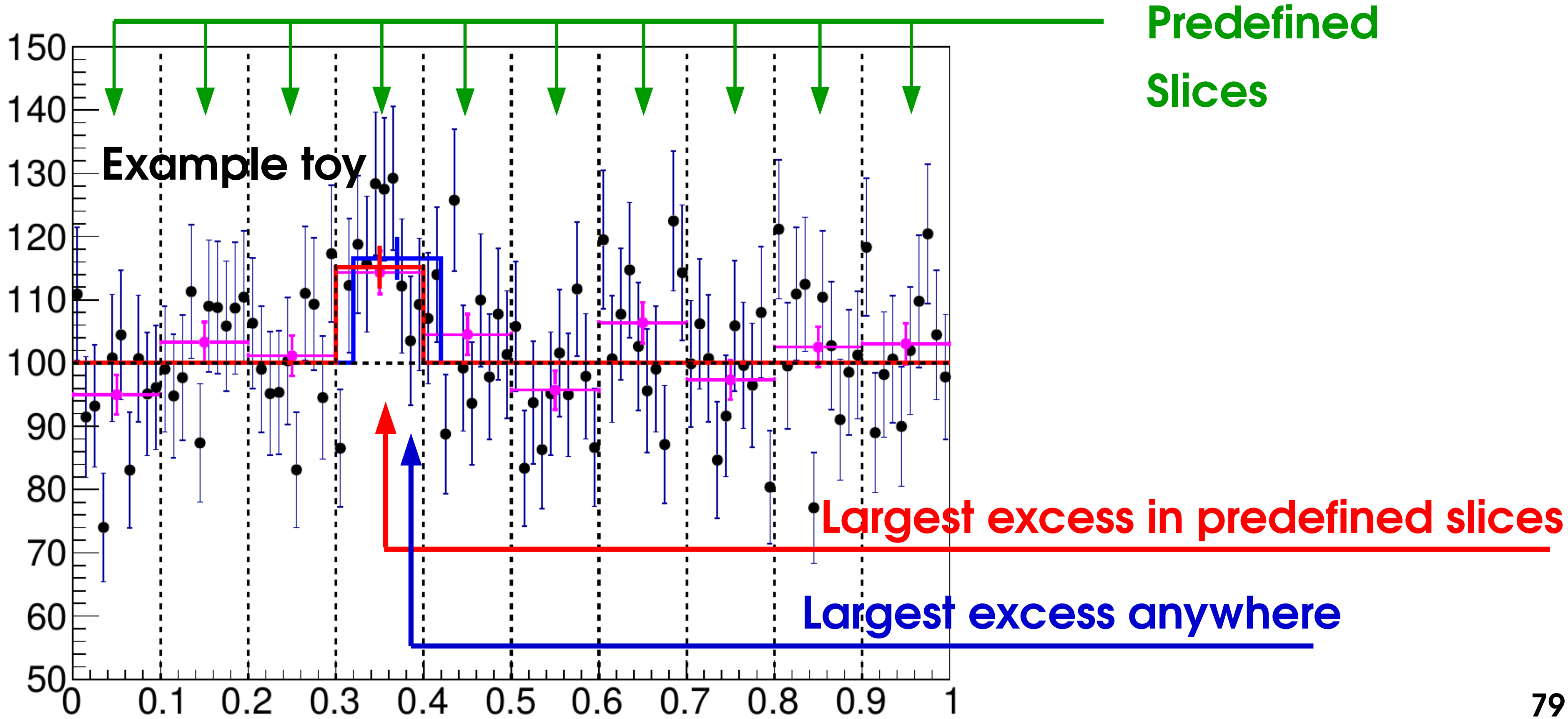
Test on a simple example: generate toys with

→ flat background (100 events/bin)

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

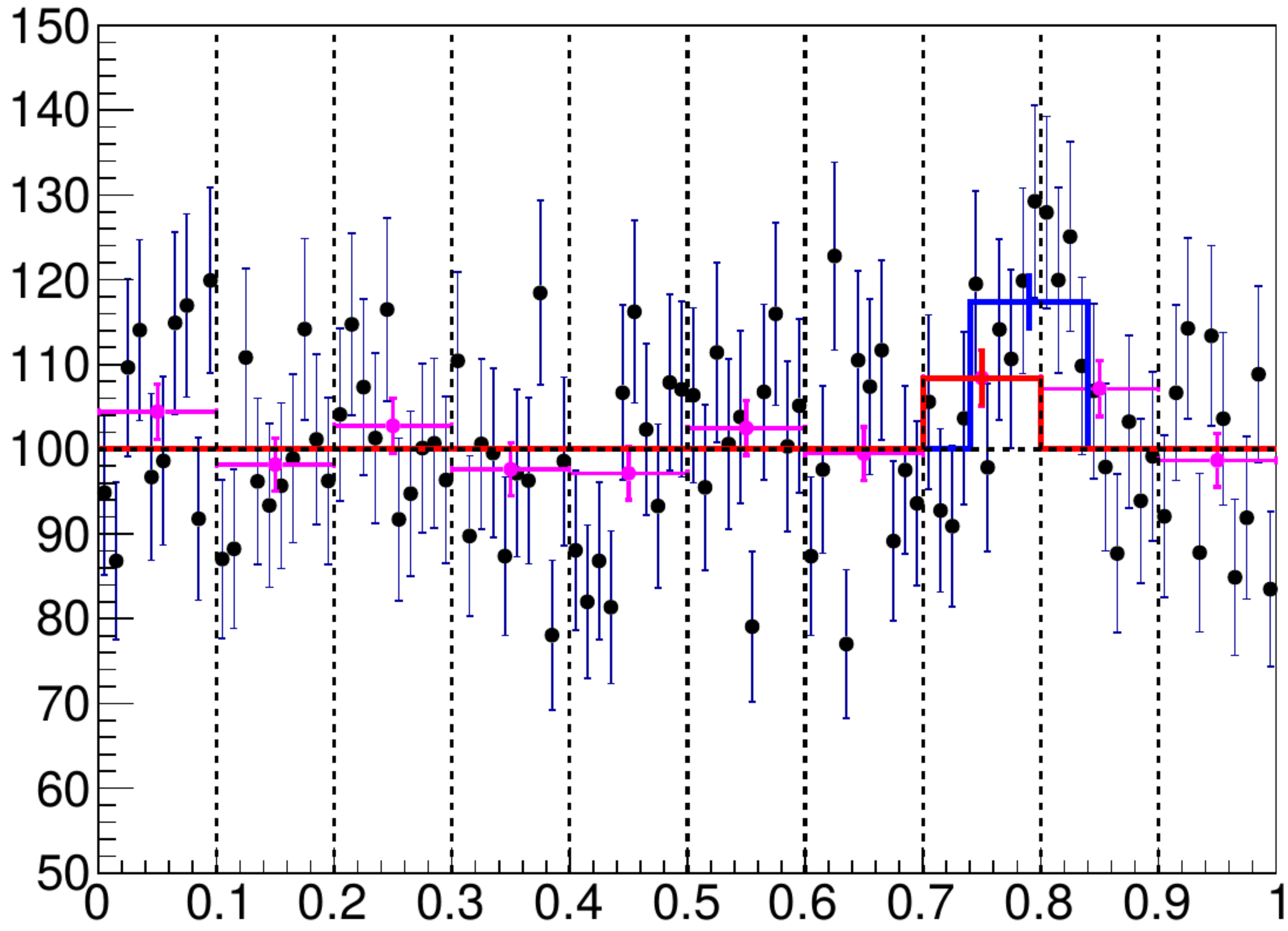
Two configurations:

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



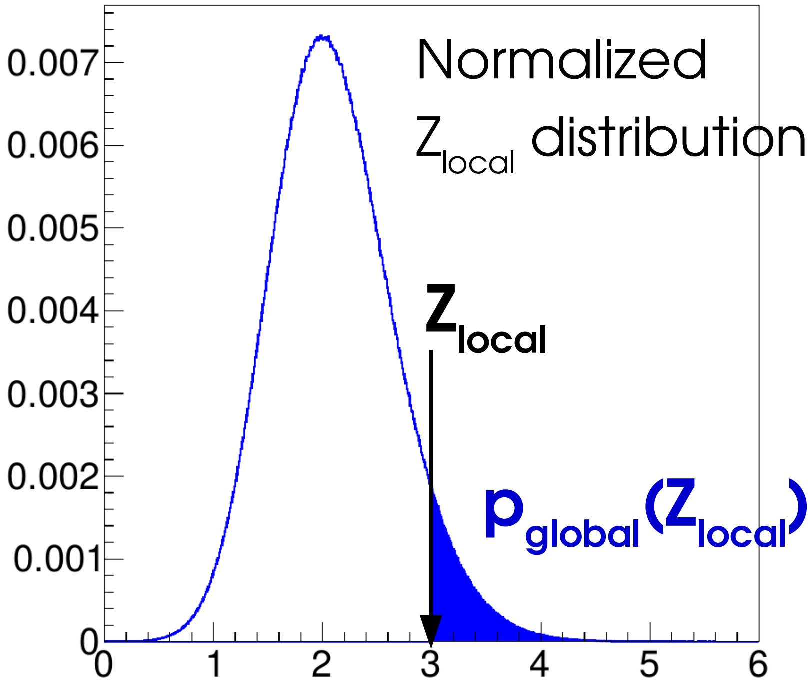
Illustrative Example (2)

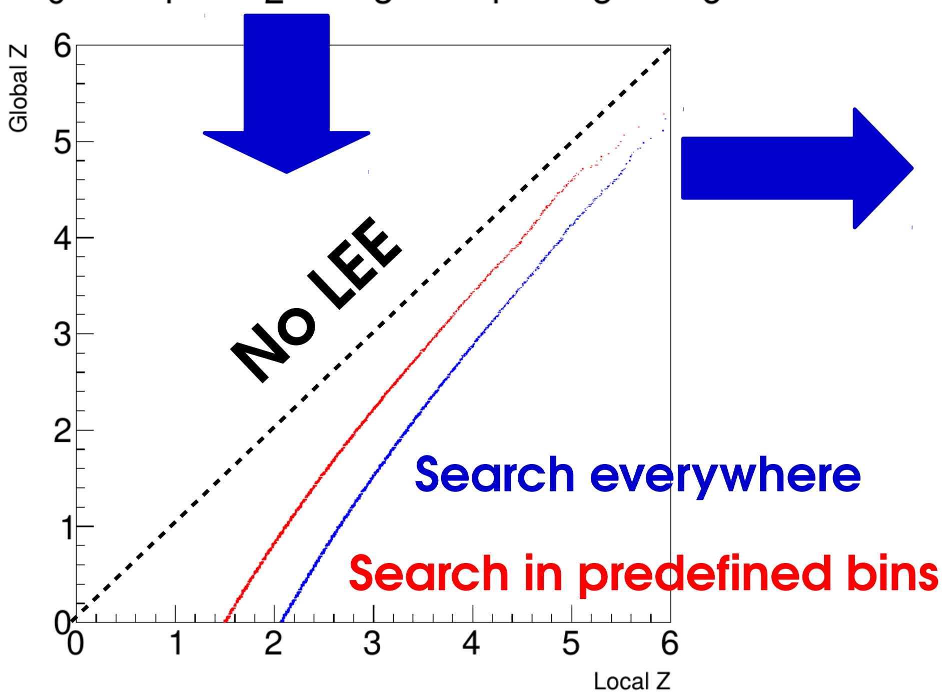
Very different results if the excess is **near a boundary :**

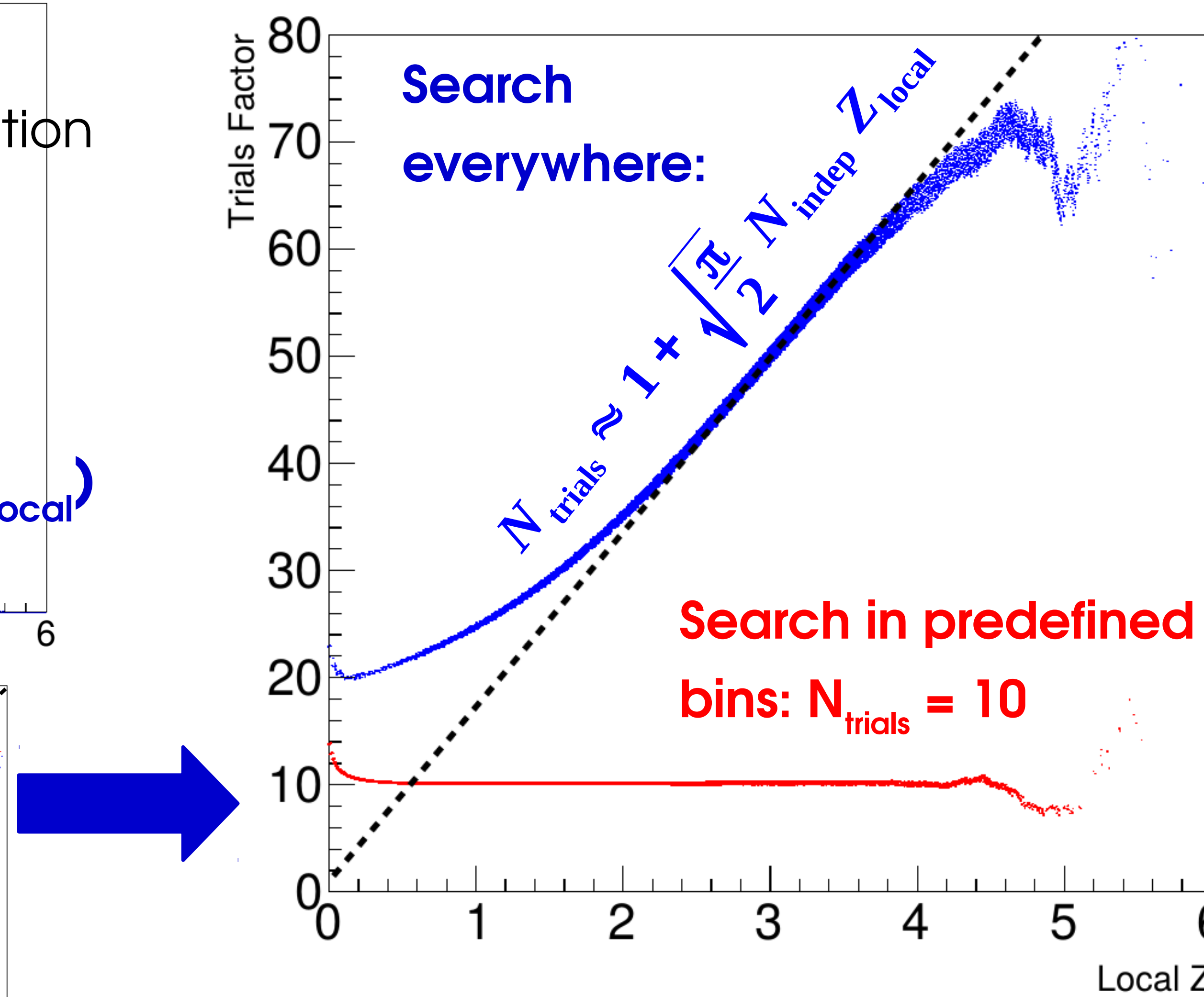


1. Look only in 10 slices of the full spectrum
2. Look in any window of same size as above, anywhere in the spectrum

Illustrative Example (3)







Searching everywhere gives the extra Z_{local} dependence

Z_{Global} Asymptotics Extrapolation

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

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

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

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

→ **Get N_{up} From toys ? but high Z_{local} ⇒ many toys needed**

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

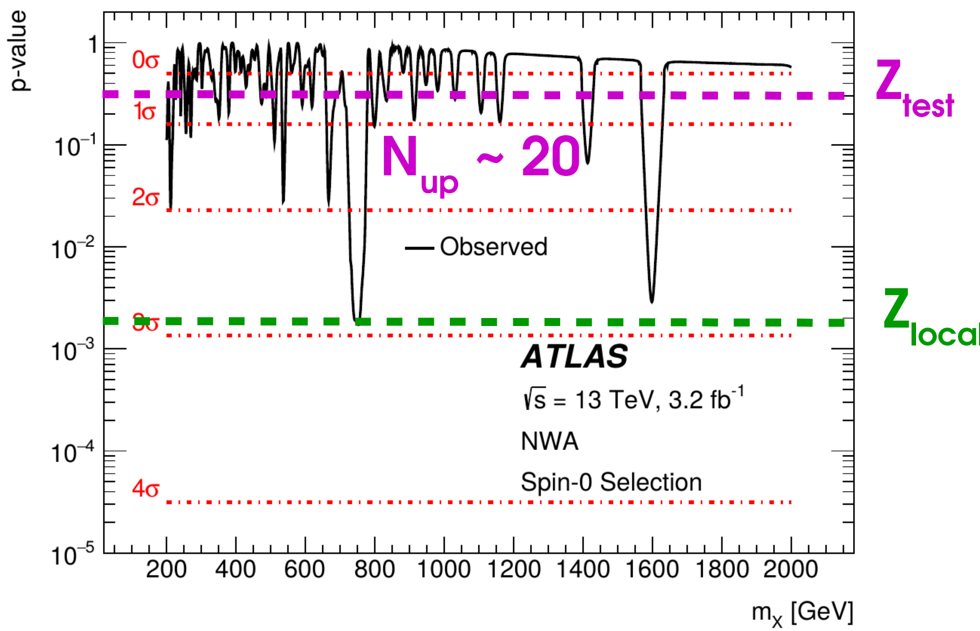
Can choose arbitrarily small Z_{test}

⇒ many excesses

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

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

if large stat uncertainty from
too few excesses

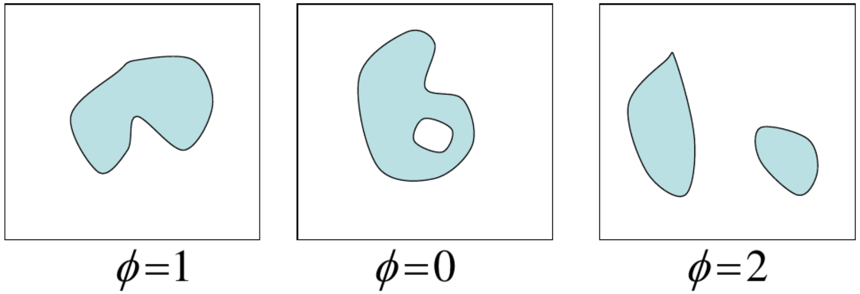


In 2D

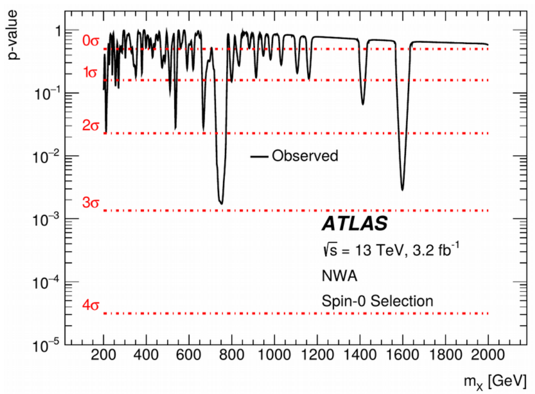
O. Vitells and E. Gross, Astropart. Phys. 35 (2011) 230

Generalization to 2D scans: consider sections at a fixed Z_{test}, compute its ***Euler characteristic*** φ, and use

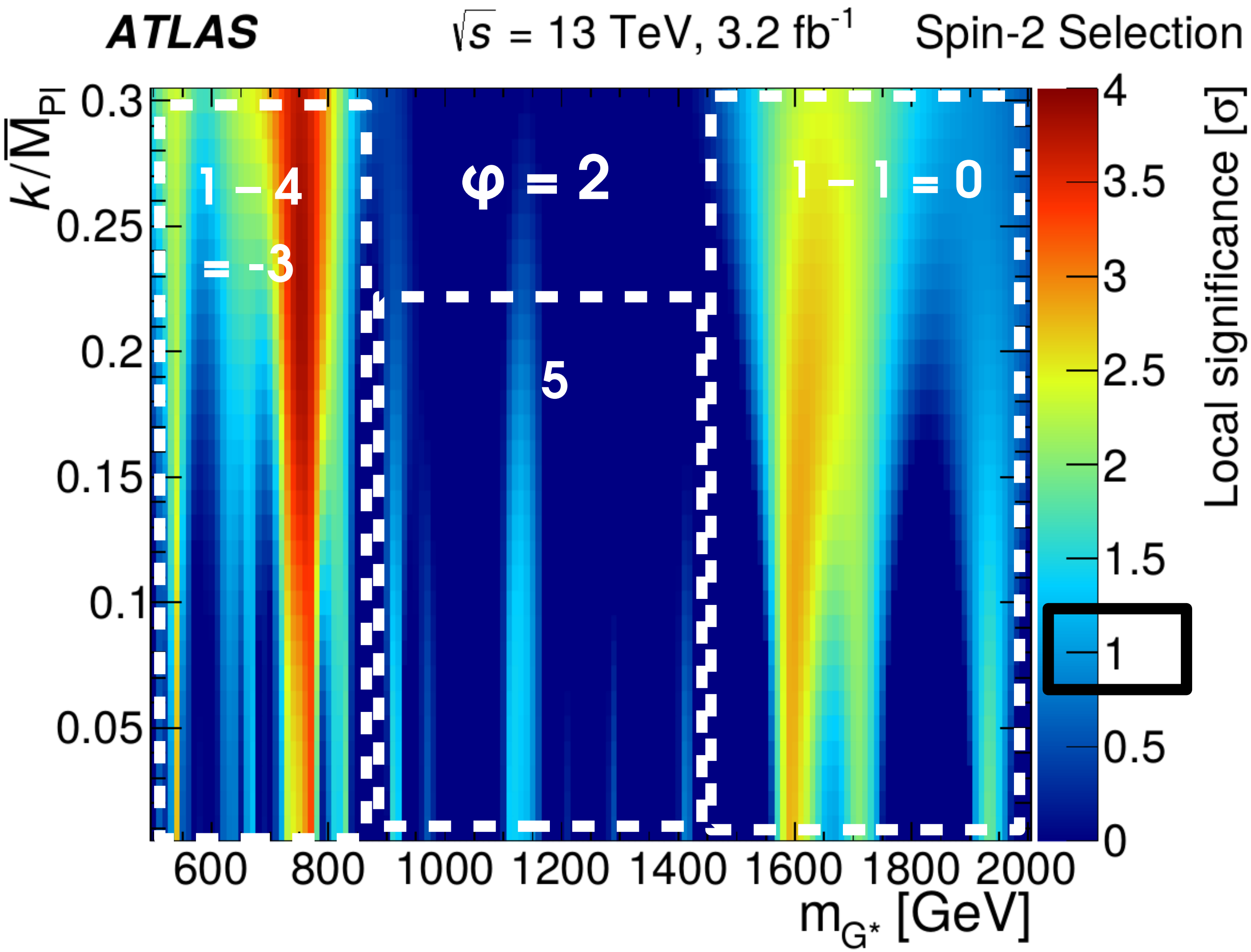
$$p_{\text{global}} \approx E[\phi(A_u)] = p_{\text{local}} + e^{-u/2}(N_1 + \sqrt{u}N_2)$$



→ Generalizes 1D bump counting



Now need to determine 2 constants N_1 and N_2, from Euler φ measurements at 2 different Z_{test} values.



Outline

Computing statistics results:

- Limits
- Confidence intervals

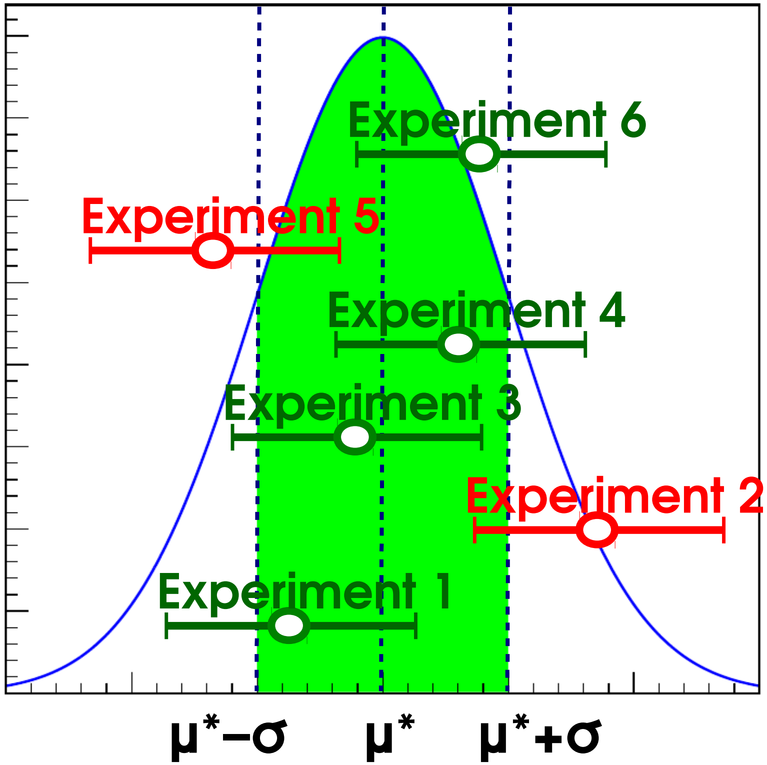
Profiling

Look-Elsewhere Effect

Bayesian methods

Bayesian Methods

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 ± 0.24 GeV"** :

→ i.e. $[125.09 - 0.24\ ;\ 125.09 + 0.24\ ]$ 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.

→ **"5σ 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(μ | data), **if we provide P(μ)**

→ 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(μ) ***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(θ)
not auxiliary measurements $P(\theta^{mes}; \theta)$

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

→ Use probability distribution P(μ) 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 κ…

→ 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(data | μ) in a frequentist way

  → "Bayesian NPs, frequentist POIs"
- Some use in Run 1, now phased out in favor of frequentist PLR.

Bayesian methods and CL_s: CL_s computation

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

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

$$\left.\begin{array}{l} \textbf{MLE: } \hat{S} = n - B \\[2em] \textbf{Conditional MLE: } \hat{\hat{\theta}}(\mu) = \dfrac{\sigma_{syst}}{\sigma_{stat}^2+\sigma_{syst}^2}(n-S-B) \end{array}\right\} \quad \textbf{PLR : } \lambda(\mu) = \left(\frac{S+B-n}{\sqrt{\sigma_{stat}^2+\sigma_{syst}^2}}\right)^2$$

Gaussian ⇒ from previous studies, CL_s limit is

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

Bayesian methods and CL$_s$: Bayesian case

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

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

Bayesian: G(θ) is actually a ***prior*** on θ ⇒ perform integral (***marginalization***)

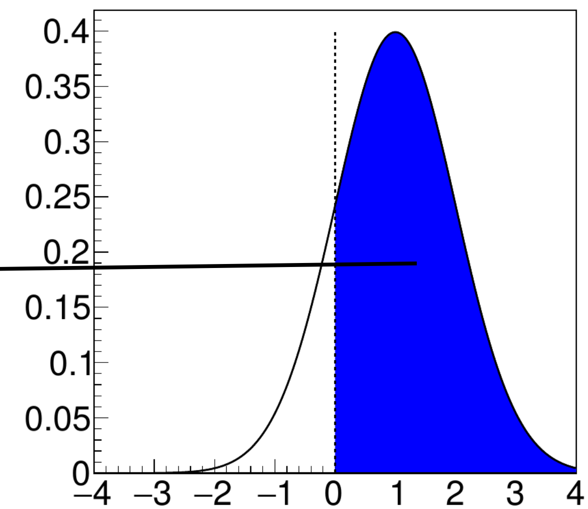
$$P(n\,|\,S) = G\left(S;\, n-B,\, \sqrt{\sigma_{stat}^2+\sigma_{syst}^2}\right)$$

same effect as profiling!

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

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

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



Bayesian Limit:

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

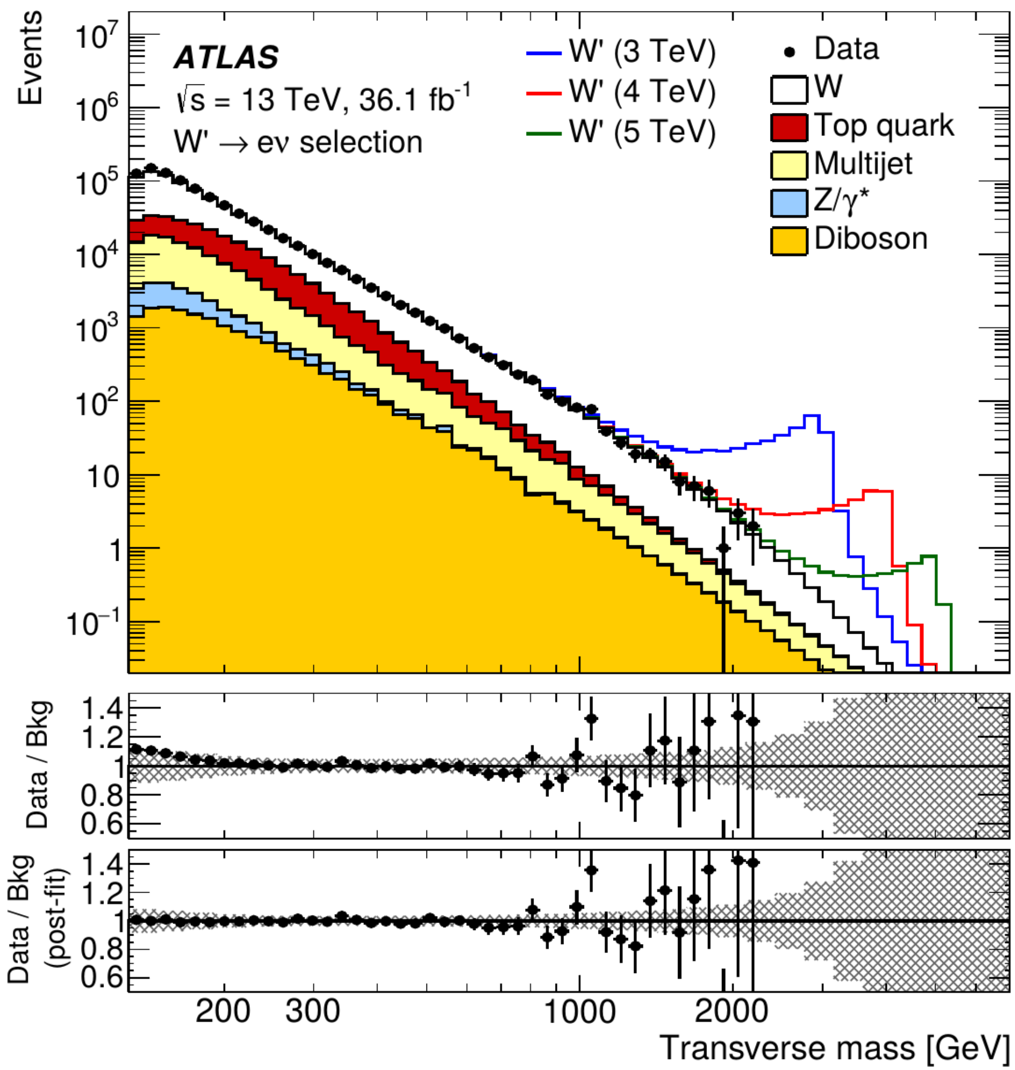
$$S_{up}^{Bayes} = n-B+\left[\Phi^{-1}\left(1-0.05\,\Phi\left(\frac{n-B}{\sqrt{\sigma_{stat}^2+\sigma_{syst}^2}}\right)\right)\right]\sqrt{\sigma_{stat}^2+\sigma_{syst}^2}$$

same result as CLs!

Example: W'→lν Search

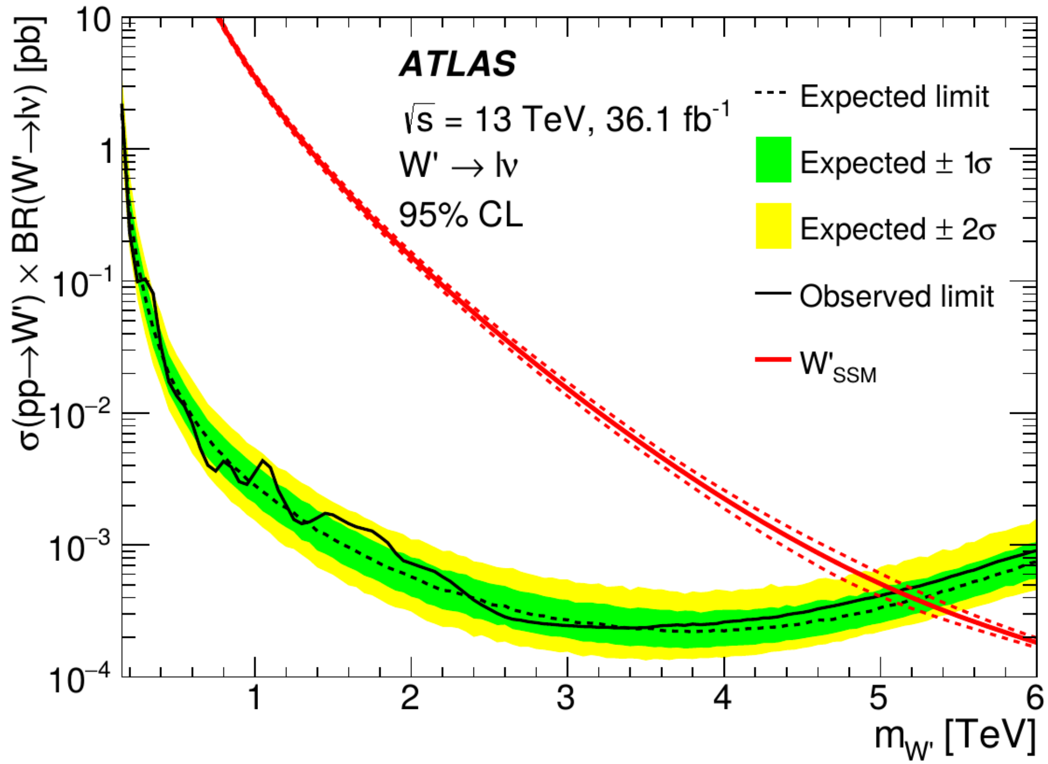
arXiv:1706.04786

- **POI**: **W' σ×B** → use **flat prior over** [0, +∞[.
- **NPs**: syst on **signal ε** (6 NPs), **bkg** (6), **lumi** (1) → integrate over Gaussian priors



Trigger
Lepton reconstruction and identification
Lepton momentum scale and resolution
$E_{\mathrm{T}}^{\mathrm{miss}}$ resolution and scale
Jet energy resolution
Pile-up

Multijet background
Top extrapolation
Diboson extrapolation
PDF choice for DY
PDF variation for DY
EW corrections for DY
Luminosity



Why 5σ ?

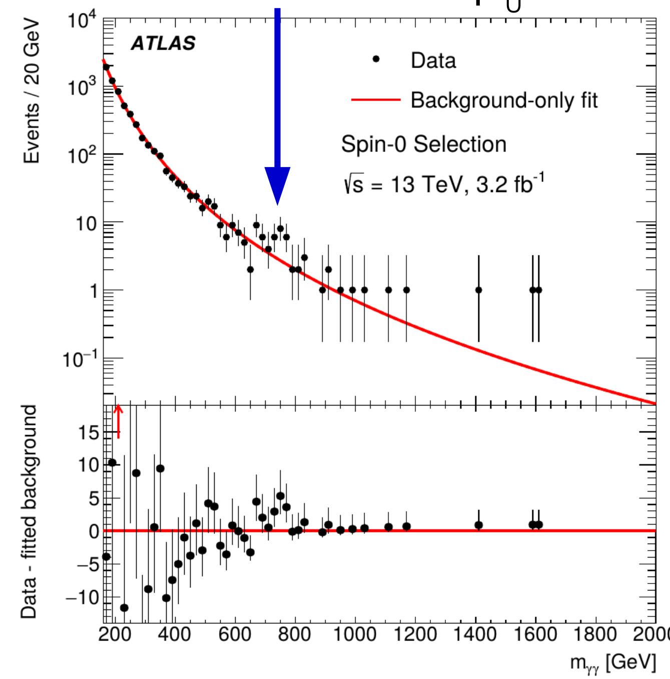
One-sided discovery: **5σ ⇔ p_0 = 3 10-7 ⇔ 1 chance in 3.5M**

→ Overly conservative ?

→ Do we even know the sampling distributions so far out ?

Local 3.9σ, p_0 = 5E-5

Global 2.1σ, p_0 = 2E-2



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

- **LEE** : searches typically cover multiple independent regions
  ⇒ Global p-value is the relevant one
  N_{trials} ~ 1000 : local 5σ ⇔ 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):

$$P(\mathrm{fluct}) = \frac{P(\mathrm{fluct}|B)\,P(B)}{P(\mathrm{fluct}|S)\,P(S) + P(\mathrm{fluct}|B)\,P(B)}$$

Extraordinary claims require extraodinary evidence

⇒ Stay with 5σ...