Statistics for (LHC Higgs) Physics

Nicolas Berger (LAPP Annecy)

Disclaimer

- This will not be a general statistics course:
  - I will deal mostly with topics relevant to Higgs searches at LHC (already a big task)
  - Most of the concepts will be introduced in the context of these searches, rather than in full generality
  - I won't be talking about Bayesian methods. The focus will be entirely on likelihood-based frequentist techniques.
- Focus on $H\rightarrow\gamma\gamma$ to introduce concepts, then generalize.

Outline

What are the goals ?

Setting up the problem : Maximum likelihood and Likelihood ratios

Discovery

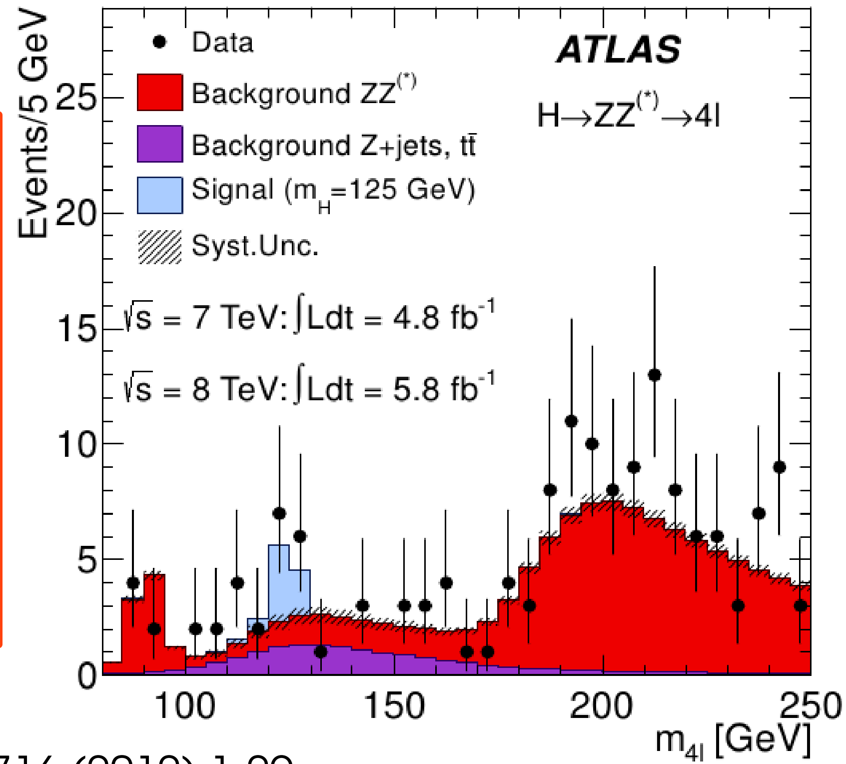
Additional wrinkles (NPs, categories)

Limit setting

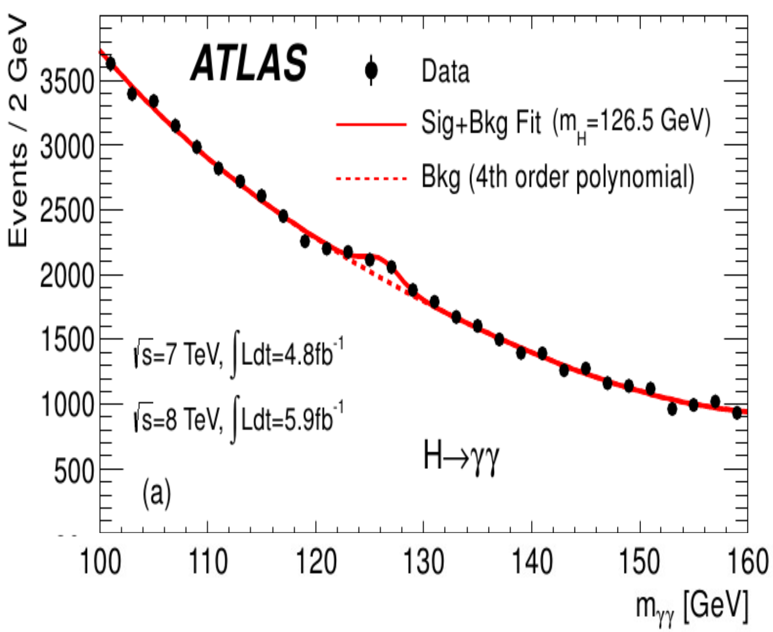
Further topics

The starting point

Statistical treatment starts when the analysis is already 99% done:
→ We have identified variables which are useful for our search : for Higgs analysis: mass (or m_T) spectra
→ We have already taken the data
→ We have already processed the data and reconstructed the quantities of interest

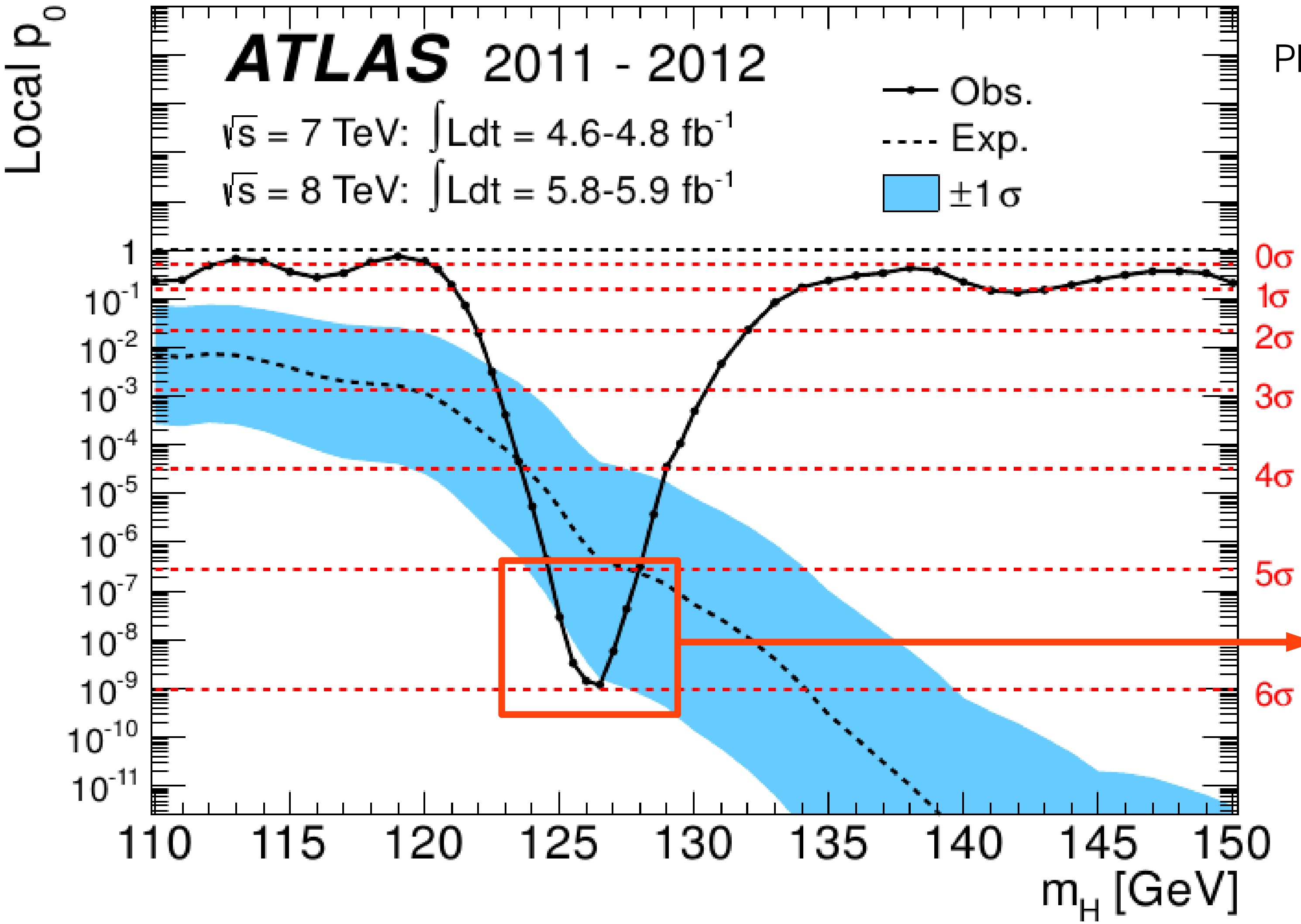


PLB 716 (2012) 1-29



However still need to quantify observations:
→ maybe we an see peaks by eye (or not)
→ need to understand chances that this comes from a real signal.

The challenge: what we want



PLB 716 (2012) 1-29

How do we get there ?

p_0 (p-value) : if there is no Higgs, probability to still get a fluctuation at least as large as this one.

Outline

What are the goals ?

Setting up the problem : Maximum likelihood and Likelihood ratios

Discovery

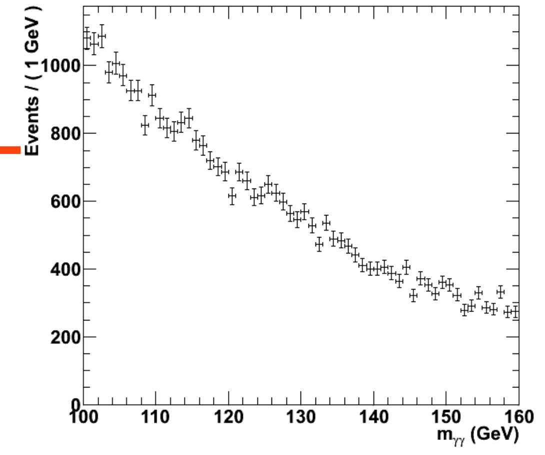
Additional wrinkles (categories, LEE)

Limit setting

Further topics

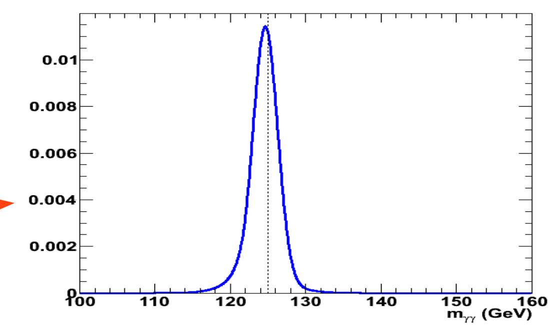
What do we need ?

→ Measurements! (observables)



→ A theoretical model to test (the SM or some extension)

→ An experimental model that describes how the measurements are obtained from the theory

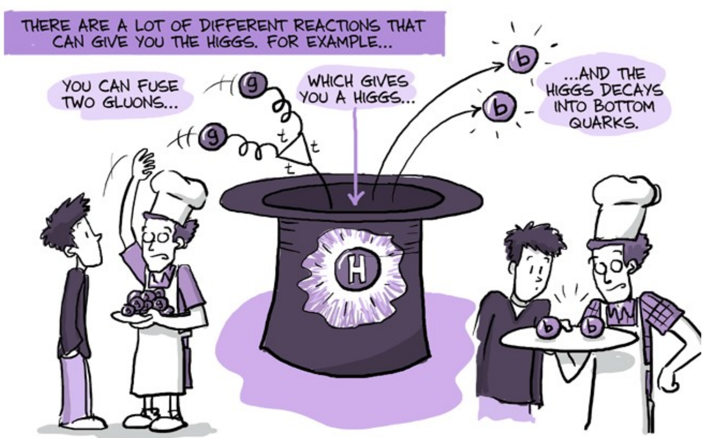


→Lots of randomness involved!

→ Quantum uncertainty

→ Measurement errors

=> Need a **Statistical** Model



http://www.phdcomics.com/comics/archive.php?comicid=1489

How to describe it

In general use **$P(m; \theta)$**

→ **m** = measurements (observables): random variables

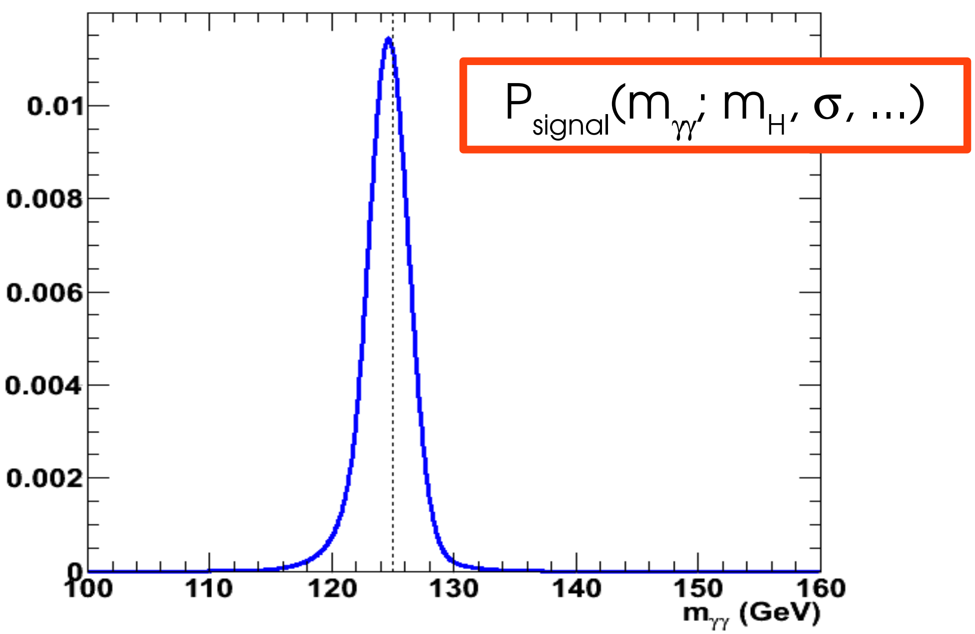
→ **θ** = parameters, with fixed (but often unknown!) values

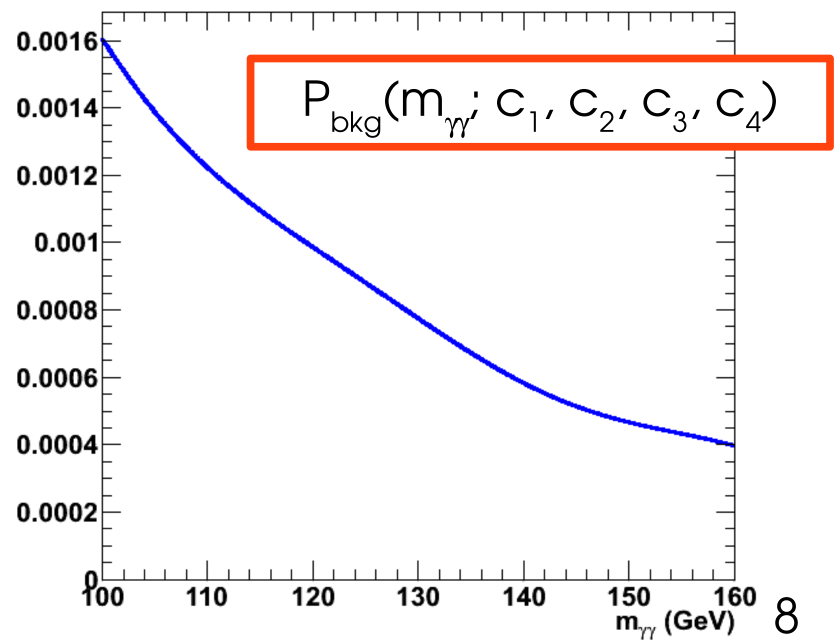
Measurements can be

→ Discrete observables : e.g. Event counts $\Sigma_i P(m_i; \theta) = 1$

→ Continous observables : (e.g. $m_{\gamma\gamma}$)

=> probability **density** function, $\int P(m; \theta)\, dm = 1$





Likelihood

Defined simply as

$$\mathbf{L(\theta;\, m_{obs}) = P(m_{obs};\, \theta)}$$

Where L is now a function of θ with the **measured** m_{obs} as parameter

The meaning is different:

→ P : **probability** to observe m **for a given θ**
(useful e.g. For MC generation)

→ L : **likelihood** of θ **given that m_{obs} has been observed**
sets up the problem of determining θ.

...But the information content is exactly the same.

Defining the correct likelihood is the hard part! The rest is just turning the crank.

Common likelihood definitions

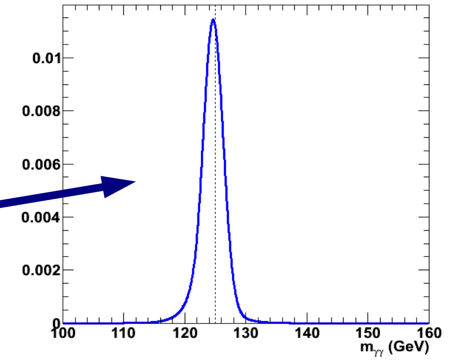
Method	Observable	Likelihood
Cut-and-count	**n** : measured number of events	Poisson $L(n;s,b)=e^{-(s+b)}\frac{(s+b)^n}{n!}$ b : expected background
Binned shape analysis	**n_i**, i=1..N_{bins} : measured events in each bin.	Multi-Dimensional Poisson $$L(n;s,f_i,b_i)=\prod_{i=1}^{N_{bins}} e^{-(sf_i+b_i)}\frac{(sf_i+b_i)^{n_i}}{n_i!}$$ f_i : fraction of signal in each bin b_i : expected background in each bin
Unbinned shape analysis	**m_i**, i=1..N_{events} : observable value for each event	Extended Likelihood $$L(m_i;s,b)=e^{-(s+b)}\prod_{i=1}^{N_{events}} sP_S(m_i)+bP_B(m_i)$$ P_S, P_B : PDFs for x in signal and background

The (unbinned) likelihood for H→γγ

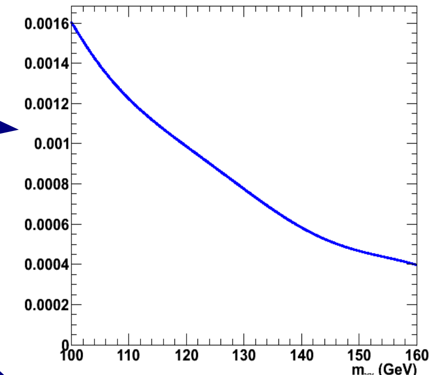
One observable: $m_{\gamma\gamma}$

How is it distributed ?

For signal $P_s(m_{\gamma\gamma})$

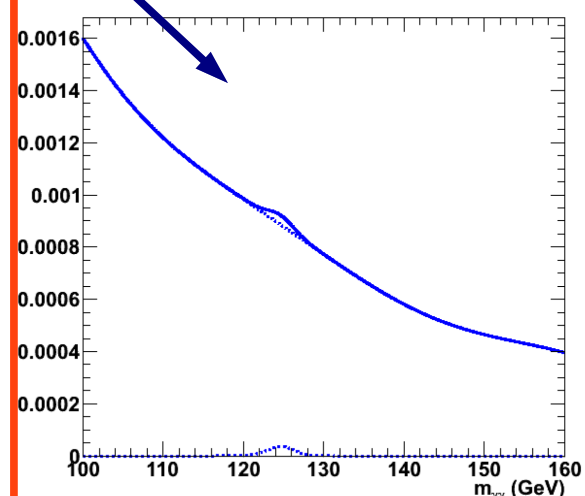


For background $P_b(m_{\gamma\gamma})$



So in total (1 event)

$$P(m_{\gamma\gamma}) = \frac{N_s}{N_s+N_b} P_s(m_{\gamma\gamma}) + \frac{N_b}{N_s+N_b} P_b(m_{\gamma\gamma})$$



For N_{obs} events:

N_{obs} can fluctuate, so include a Poisson ("extended") term :

$$e^{-(N_s+N_b)} \frac{(N_s+N_b)^{N_{obs}}}{N_{obs}!}$$

Finally:

$$L(m_{\gamma\gamma,1} .. m_{\gamma\gamma,N_{obs}}) = e^{-(N_s+N_b)} \prod_{i=1}^{N_{obs}} N_s P_s(m_{\gamma\gamma,i}) + N_b P_b(m_{\gamma\gamma,i})$$

"Unbinned Extended Likelihood"

Maximum likelihood

Idea: estimate θ by picking the **most likely** value, where L is maximal

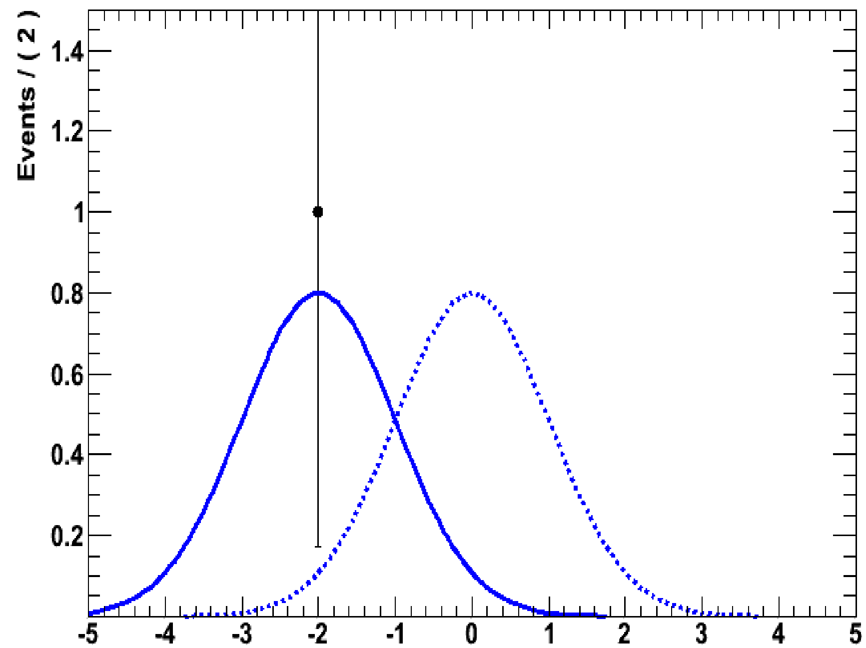
Maximum likelihood (ML) estimates denoted by "hat" : $\hat{\theta}$

Good properties:

→ **Asymptotically Efficient**: Maximum information (=> smallest error) for large N

→ **Asymptotically Gaussian** for large N

→ **Unbiased** : correct on average even for small N.

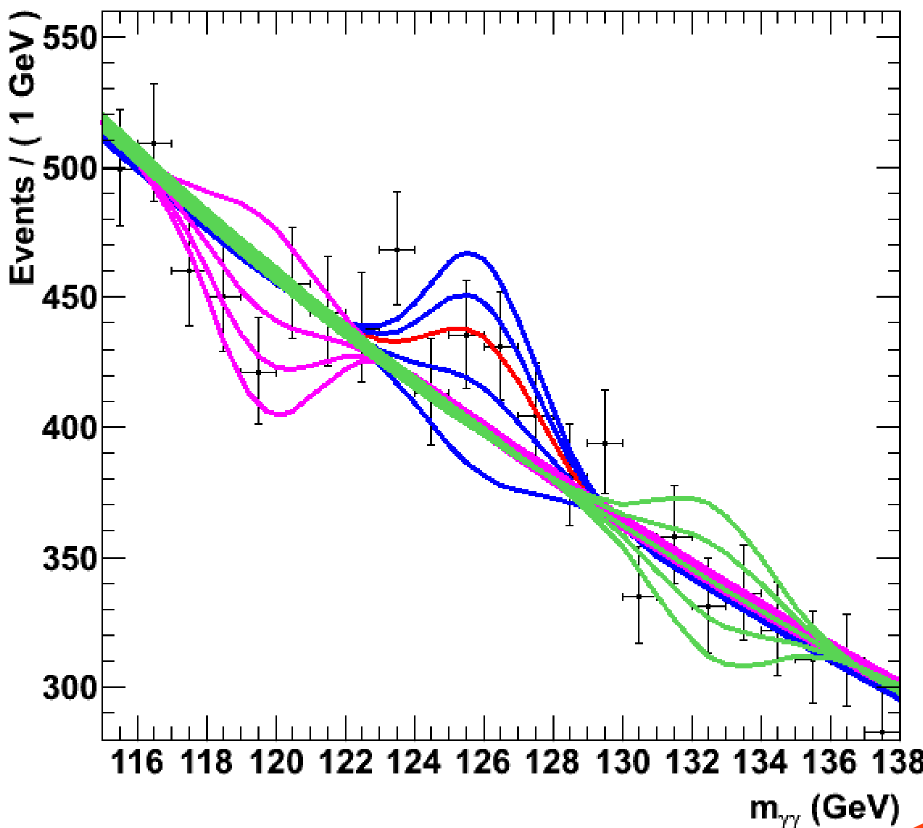


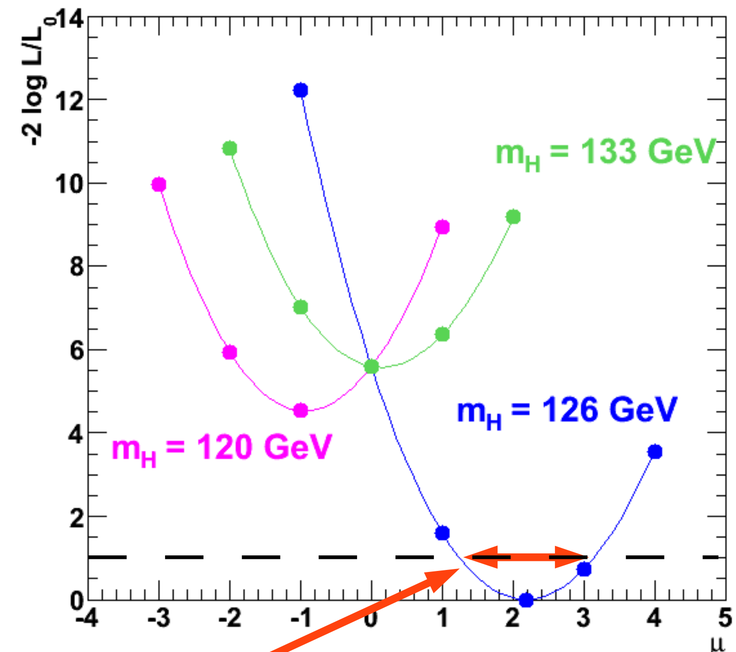
Single-event Gaussian example

Note that errors on data points don't influence the fit – uncertainties come from the model.

H→γγ-inspired example

Simple template fit using fixed shapes for signal and background
Free parameters: N_{bkg} and $\boldsymbol{\mu = N_{signal}/N_{signal}^{SM}}$





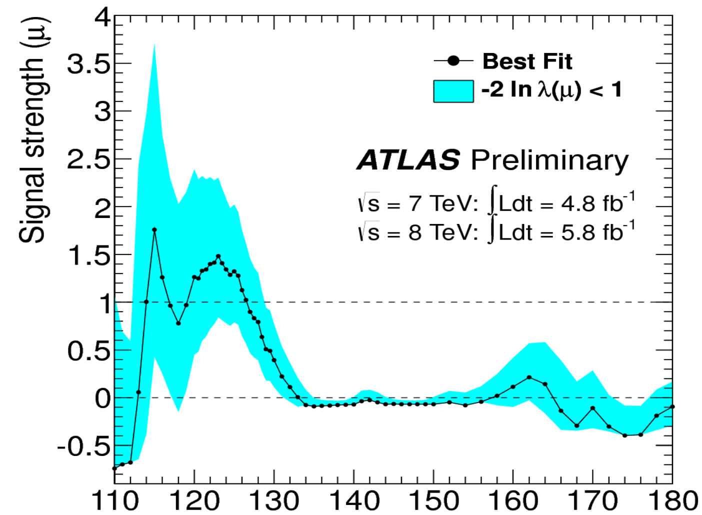
Size of -2ΔlogL=1 contour gives +/-1 sigma (68%) error

Check 3 mass points, scan over μ for each one

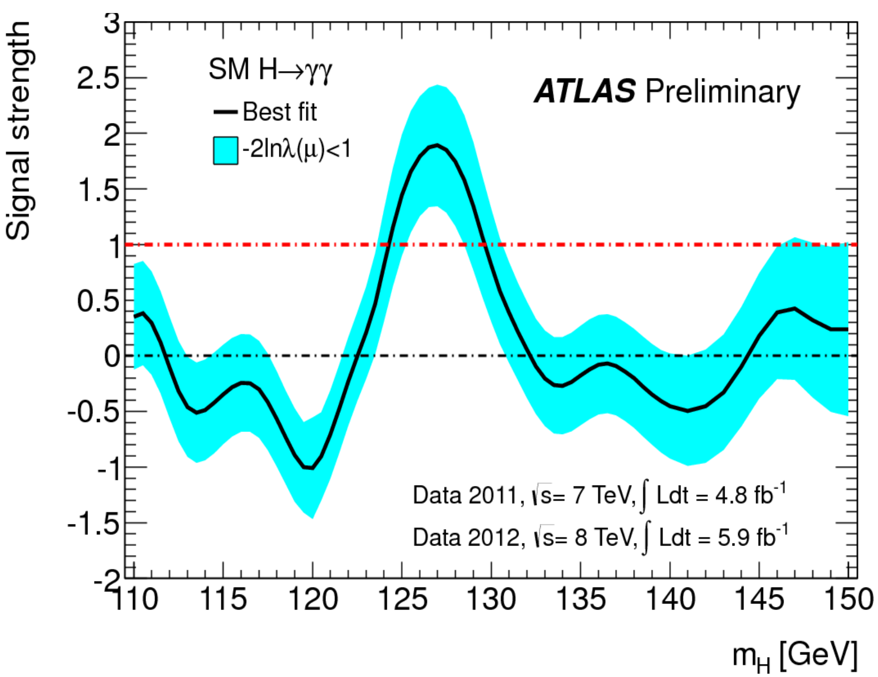
Compute -log L for each, find minimum (=max of L) => $\hat{\mu}$

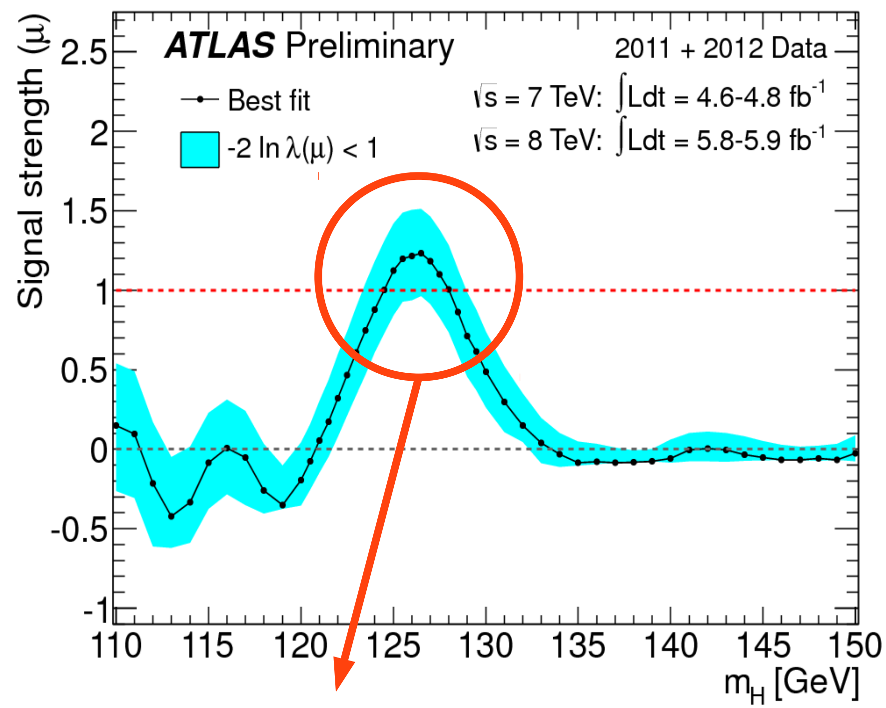
“Blue band” plots

Same principle for the “Blue band” plots:
→ Scan over m_H values
→ For each m_H, find $\hat{\mu}$ and its error
Done for each channel and combination (details later)



ATLAS-CONF-2012-093





So, do we have a discovery ?

Outline

What are the goals ?

Setting up the problem : Maximum likelihood and Likelihood ratios

Discovery

Additional wrinkles (categories, LEE)

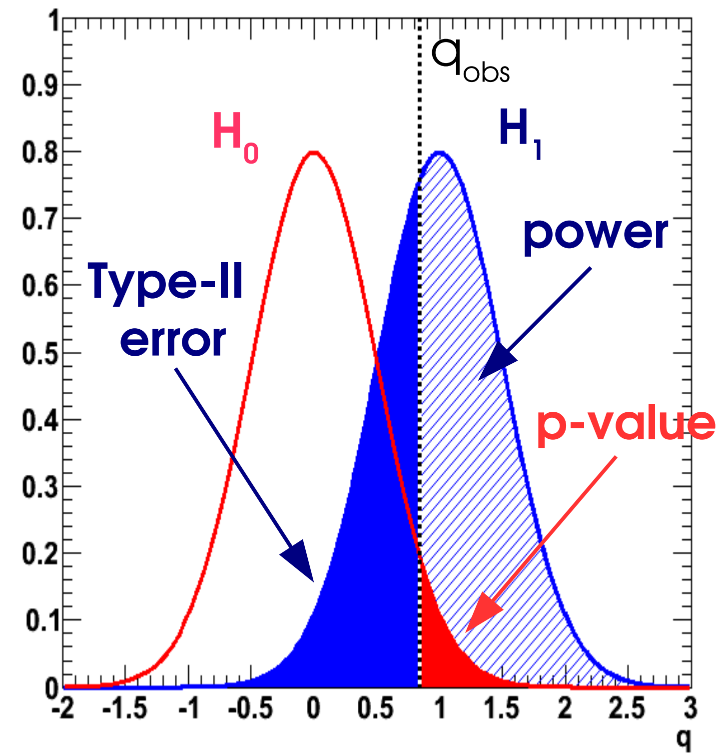
Limit setting

Further topics

Hypothesis testing

Hypothesis = a region of parameter space.
We will use: "SM without Higgs" : $\mu = 0$
Define:
→ A "null" hypothesis H_0 to reject (here μ=0)
→ An alternate hypothesis H_1 (∃ Higgs)

Strategy:
→ Define some function q of the observables
→ Find the distributions for H_0 and H_1
→ See where is the value q_{obs} from the data



Two ways to make a mistake:
Type I :
→ q_{obs} is H_1-like (Higgs?), but actually H_0 was true (no Higgs)
=> wrongly claim a discovery (bad!).
→ Probability is the **p-value**. For a discovery, need <2.9E-7
Type II:
→q_{obs} is H_0-like but H_1 was true
→ Leads to missed discovery: less bad, but still to be avoided! Probability is 1-power.

Goal: find q with **max power** for a given p-value (=> max separation)

Neyman-Pearson lemma

Define q from likelihoods:

$\rightarrow$ Compute $L(\text{data}; H_1) = L(\text{data}; \theta(H_1))$ for H_1

$\rightarrow$ Compute $L(\text{data}; H_0) = L(\text{data}; \theta(H_0))$ for H_0

Then $\boldsymbol{\lambda = L(\text{data}; H_1)/L(\text{data}; H_0)}$ obviously carries information on the hypothesis test:

$\rightarrow$ If data is H_0-like, $L(\text{data}; H_0)$ is large, $L(\text{data}; H_1)$ small => small λ

$\rightarrow$ If data is H_1-like, $L(\text{data}; H_1)$ is large, $L(\text{data}; H_0)$ small => large λ

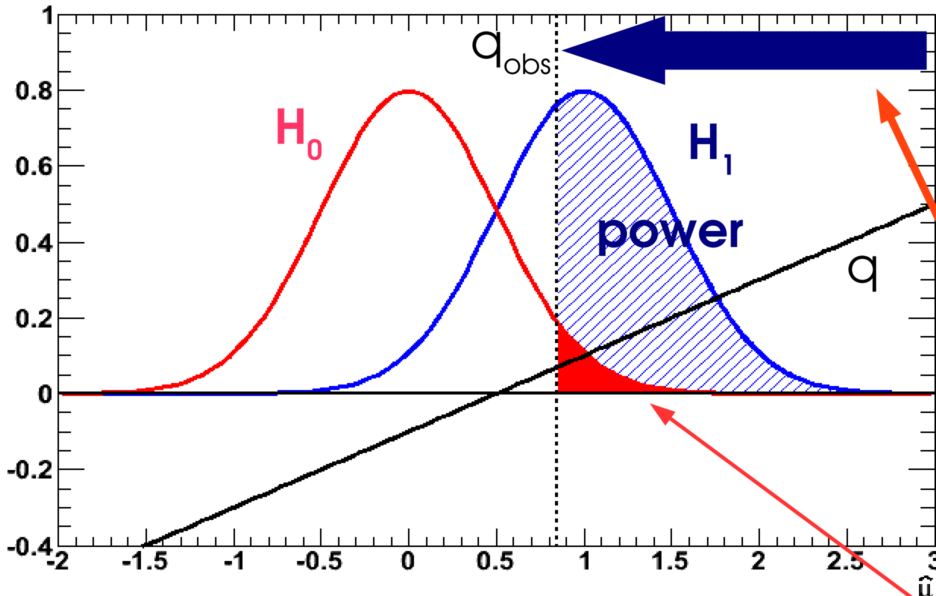
Neyman-Pearson lemma:

Use $\lambda > A$ as the test. This is actually **optimal** (carries the maximum available information)

In practice use:

$$q = -2 \log(L(\text{data}; H_0)/L(\text{data}; H_1))$$

Simple Gaussian Example



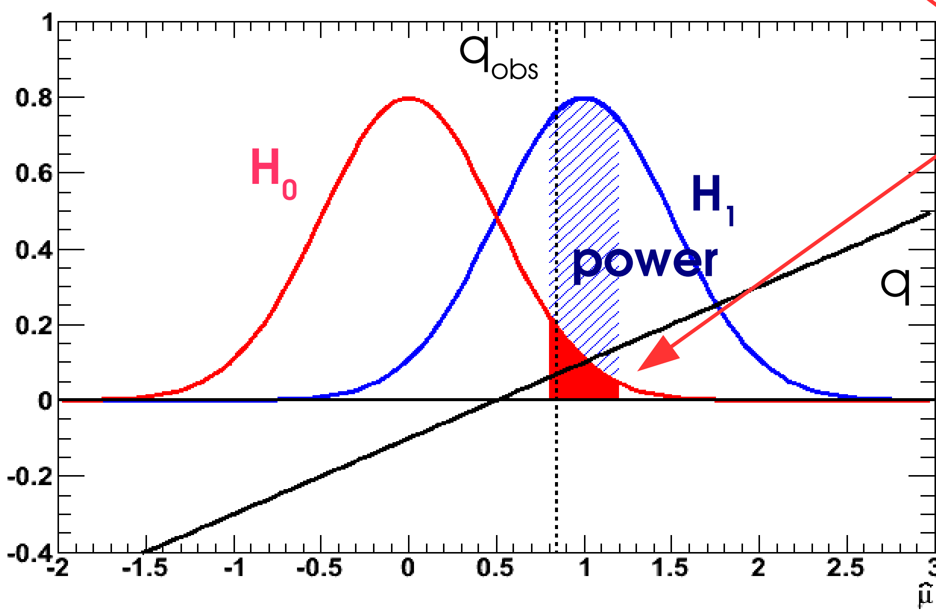
Assume $\hat{\mu}$ is Gaussian for both **H_0 (μ=0)** and **H_1 (μ=1)**.

Then **$q = (1-2\hat{\mu})/\sigma$ ~$\hat{\mu}$**

Neyman-Pearson lemma:

use q>A as selection, i.e. $\hat{\mu}$>X

Same p-values



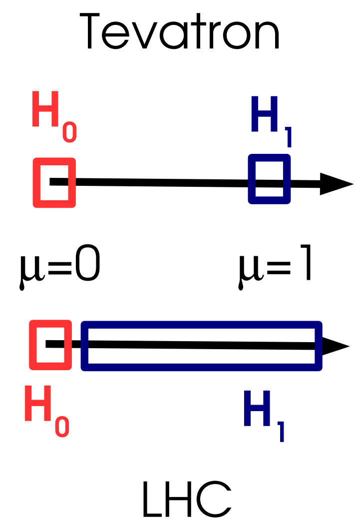
What if we were to use another test ?

For the same p-value, we would have less power

Profile-likelihood Statistic

Setup of previous example sometimes called **"Tevatron-style"** : 2 "simple" (single μ value) hypos.

At LHC usually use a different definition: **H_0 :μ=0**, **H_1 : μ>0.**
Why ?more general definition of discovery: clearly μ=2 still counts.



Now **H_1** is **composite** (range of μ values). What μ to use for L(data; H_1) ?
=> The one that maximizes the likelihood ("give H_1 its best shot")

Use: **$q_0 = -2 \log L(\text{data}; \mu=0)/L(\text{data}; \hat{\mu})$**

Closely related to $\hat{\mu}$:
→ Small $\hat{\mu}$ (no signal seen) => $L(\hat{\mu}) \sim L(\mu=0)$ => small q_0
→ Large $\hat{\mu}$ (signal!) => $L(\hat{\mu}) >> L(\mu=0)$ => large q_0

→ **q_0>0** since best-fit L always larger than fixed L(μ=0).
→ For simple Gaussian case, $q_0 = (\hat{\mu}/s)^2$.
→ For a one-sided test (μ>0), still optimal although H_1 is composite

Distributions for q_0

Asymptotically, q_0 is distributed as:

→ $\mu = 0$: a **$\chi_2(n_{dof}=1)$** distribution

→ $\mu \neq 0$: non-central **$\chi_2(n_{dof}=1, \lambda)$,** $\lambda = \mu/\sigma$

This is **Wilks' theorem**

=> Can easily convert a q_0 value to a p-value:

$$p_0 = \int_{q_0}^{+\infty} \chi^2(q, n_{dof}=1)\, dq$$

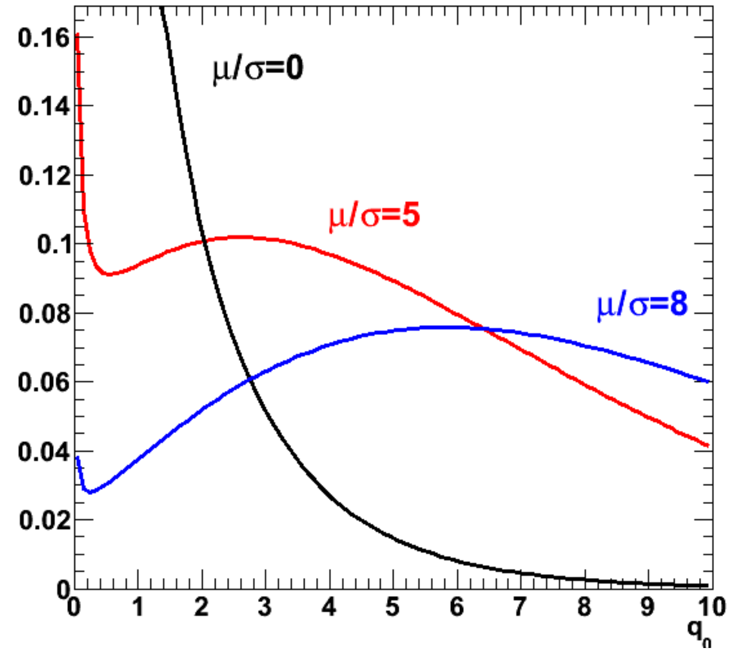
The key property for this is that $\hat{\mu}$ is Gaussian-distributed (σ = Gaussian width)

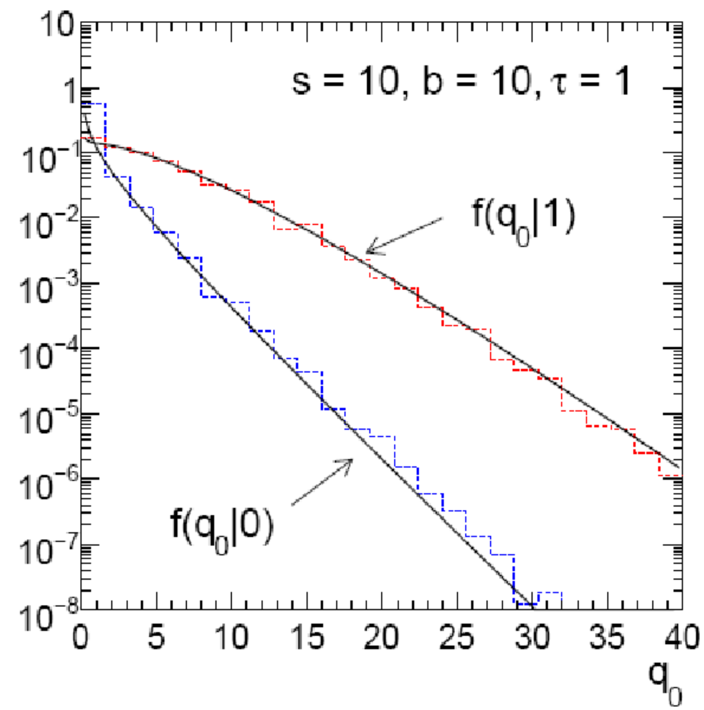
If this is not true (**small stats, LEE issues**), need to determine distribution "by hand":

→ Generate toys (pseudo-data) for some μ.

→ For each pseudo-dataset, compute q_0 and histogram the results

→ May need many toys to populate the tails!
  ($5\sigma \to 2.9\ 10^{-7}$!)





One-sided or Two-sided ?

$$q_0 = -2\log\frac{L(\mu=0\,;data)}{L(\hat{\mu}\,;data)}$$

As defined, we have

→ $\hat{\mu} \sim 0$ => small q_0

→ Large $\hat{\mu} > 0$ => large q_0

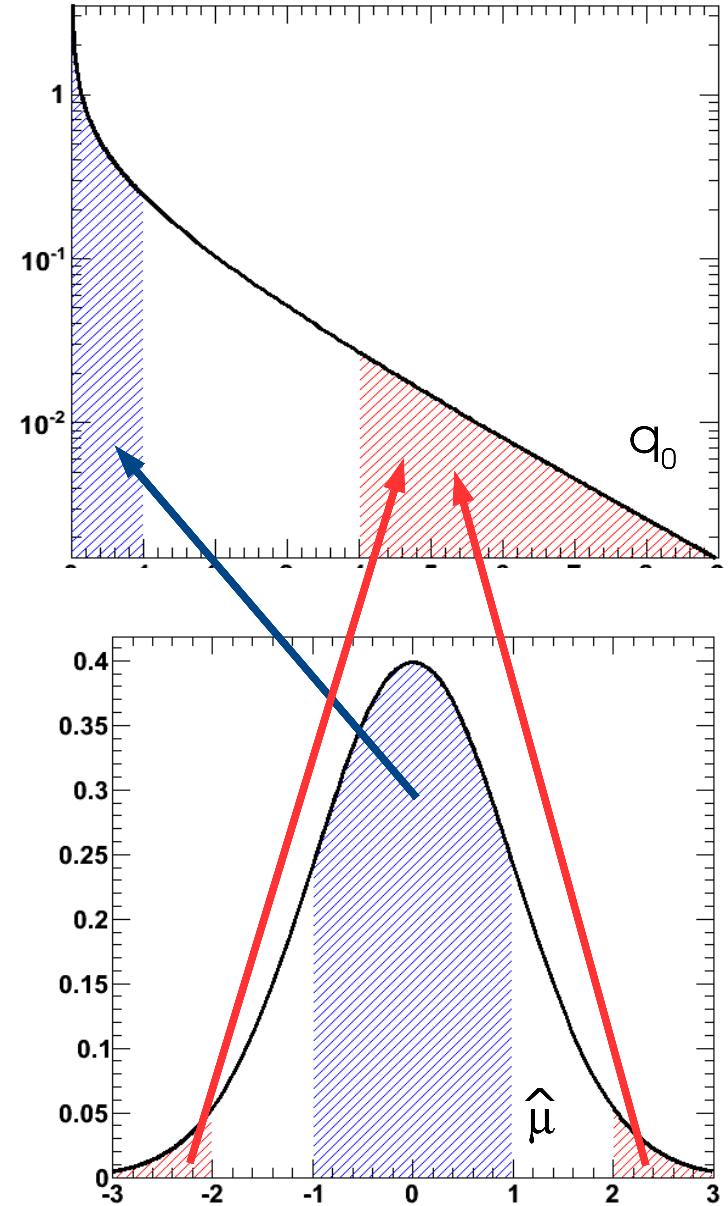
But also

→ "Very negative" $\hat{\mu} < 0$

=> also large q_0

However we know these cases are not evidence for signal!

Since we also compute $\hat{\mu}$, use this extra information to improve the procedure



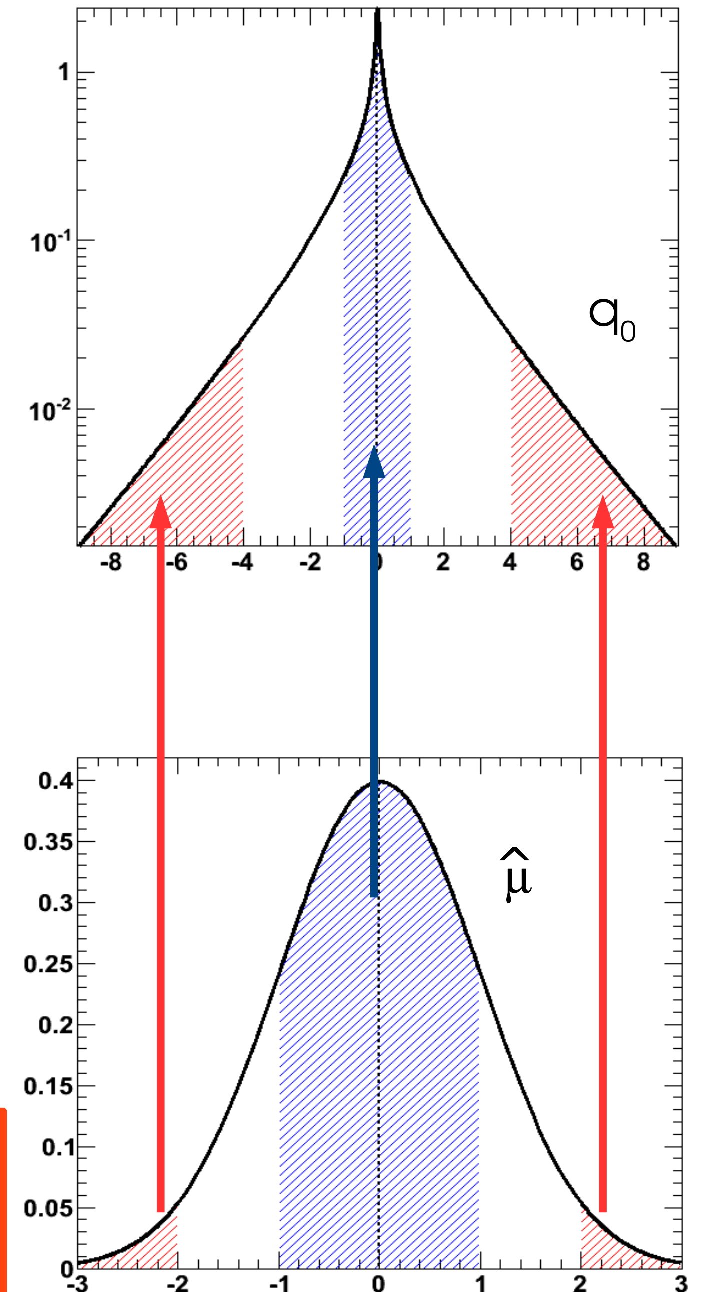
Uncapped q_0

Uncapped p0:
If $\hat{\mu}<0$, give q_0 a negative sign:

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

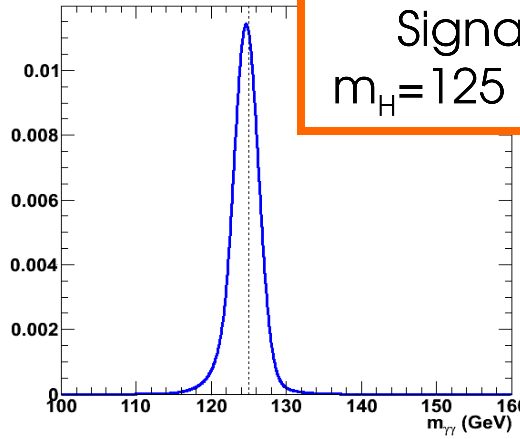
Distribution: "double half-$\chi 2$"
→ For $\hat{\mu}>0$, p-values are half those of the two-sided case: (adding more information gives a better result).
→ $p_0 < 0.5$ For $\hat{\mu}>0$,
→ $0.5 < p_0 < 1$ for $\hat{\mu} < 0$

Capped p_0: same but set $q_0=0$ for all $\hat{\mu}<0$
→ Simpler, but negative fluctuations not shown
→ Used in Higgs results before Summer 2012

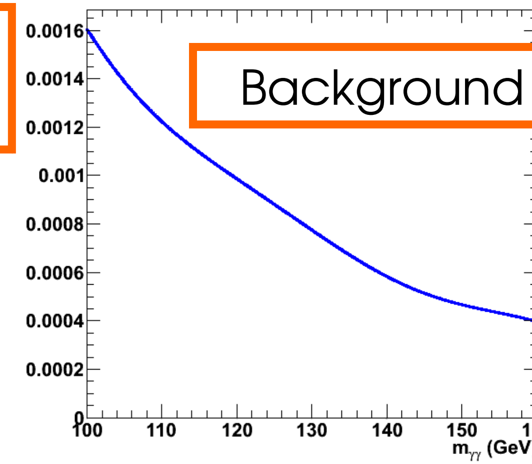


In practice

An almost-real-life example: Hgg with fixed templates.



Signal, m_H=125 GeV

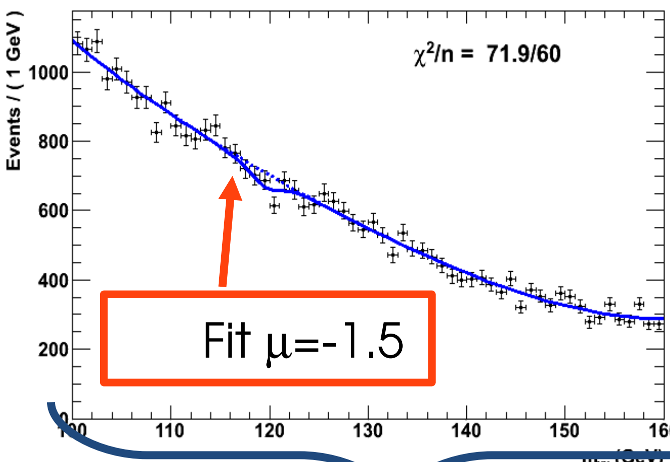


Background

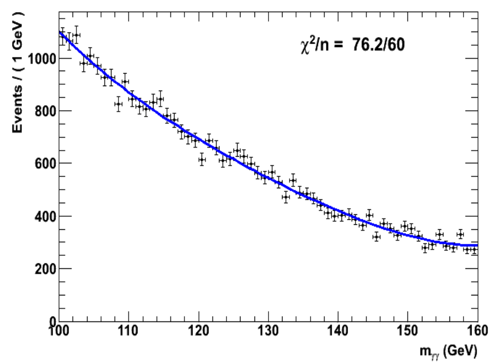
Free μ at m_H=118 GeV

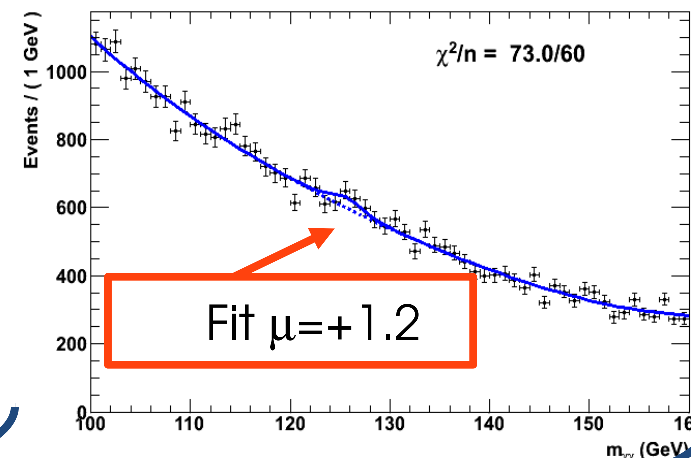
Background only

Free μ at m_H=126 GeV



Fit μ=-1.5





Fit μ=+1.2

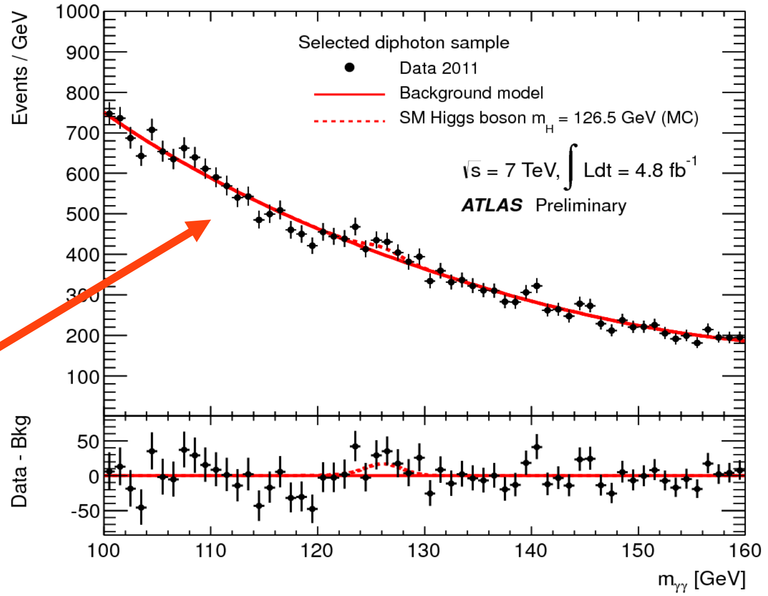
q_0 = -4.66 for m_H=120 GeV
p_0 = 98% or **-2.2σ**

q_0 = 3.13 for m_H=126 GeV
p_0 = 4% or **+1.8σ**

Nuisance parameters

Usually Likelihood involves more parameters than just the ones of interest:
→ **nuisance** parameters: PDF parameters, backgrounds, efficiencies…

e.g background slope and N_{bkg}.



Good cases: parameter can be reliably estimated from the data: "profiling"
Compute q_0 using the ML estimates of θ within the hypothesis:

$$q_0 = -2\log \frac{L(data\,;\mu=0,\hat{\hat{\theta}})}{L(data\,;\hat{\mu},\hat{\theta})}$$

Best-fit of θ in H_0 (μ=0 fixed)

Best-fit of θ in H_1 (μ floating)

Wilks' theorem: this q_0 still asymptotically distributed as a $\chi^2(n_{dof}=1)$!

Note: this isn't exactly new…

Now consider the likelihood ratio

$$l = \frac{L(x \mid \theta_{r0}, \hat{\hat{\theta}}_s)}{L(x \mid \hat{\theta}_r, \hat{\theta}_s)}. \tag{24.4}$$

Intuitively, l is a reasonable test statistic for H_0: it is the maximum likelihood under H_0 as a fraction of its largest possible value, and large values of l signify that H_0 is reasonably acceptable. The critical region for the test statistic is therefore

$$l \leqslant c_\alpha, \tag{24.6}$$

where c_α is determined from the distribution $g(l)$ of l to give a size-α test, i.e.

$$\int_0^{c_\alpha} g(l)\, dl = \alpha. \tag{24.7}$$

Kendall and Stuart, *The Advanced Theory of Statistics, vol. 2* (**1961**)

Inclusive $H\rightarrow\gamma\gamma$

Scan over **$110 < m_H < 150$ GeV**

For each value compute q_0 using

→ The signal template for this m_H

→ A free 4th-order polynomial shape for the background

→ also add systematics, but little effect here

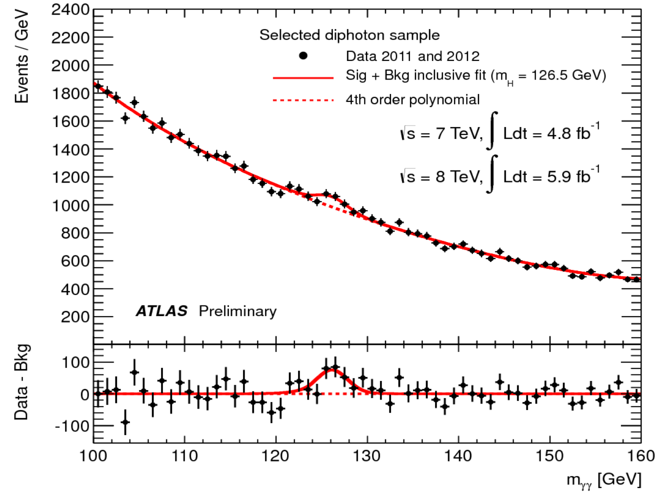
Convert q_0 to p-value using the asymptotic distribution

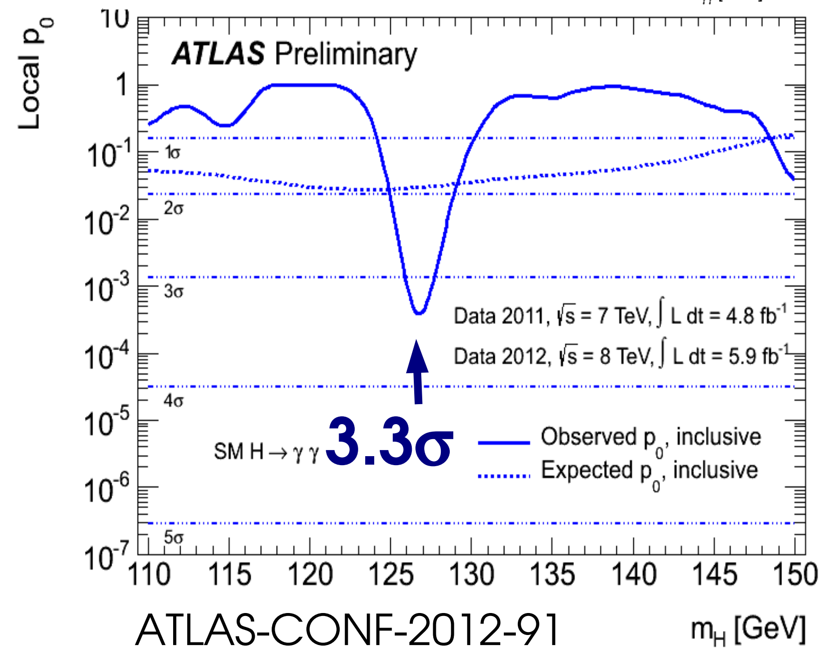
Expected p_0

→ Generate toys (usuallly for μ=0)

→ Compute p_0, histogram results

→ Report median of distribution





ATLAS-CONF-2012-91

Convert p-value to significance :

For uncapped: **$Z = \Phi^{-1}(1 - p_0)$**

(Φ = Gaussian cumulative distribution)

Outline

What are the goals ?

Setting up the problem : Maximum likelihood and Likelihood ratios

Discovery

Additional wrinkles (categories, LEE)

Limit setting

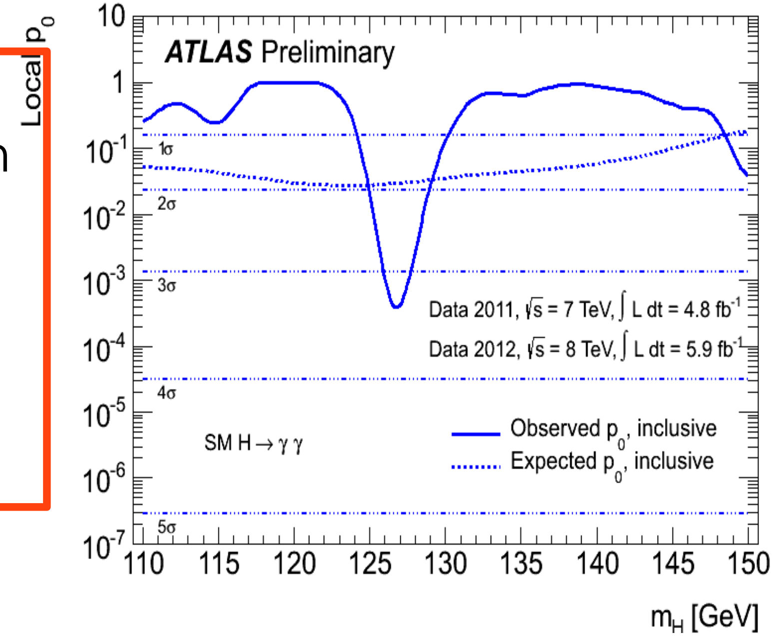
Further topics

Look-elsewhere effect

Search for a particle with unknown mass:
→Scan p_0 as a function of mass, find minimum
→ Better: include mass in the statistic:

$$q_0 = -2\log \frac{L(\mu=0, \hat{\hat{m}}_H; data)}{L(\hat{\mu}, \hat{m}_H; data)}$$

Wilks' theorem: should be $\chi^2(n_{dof}=2)$?



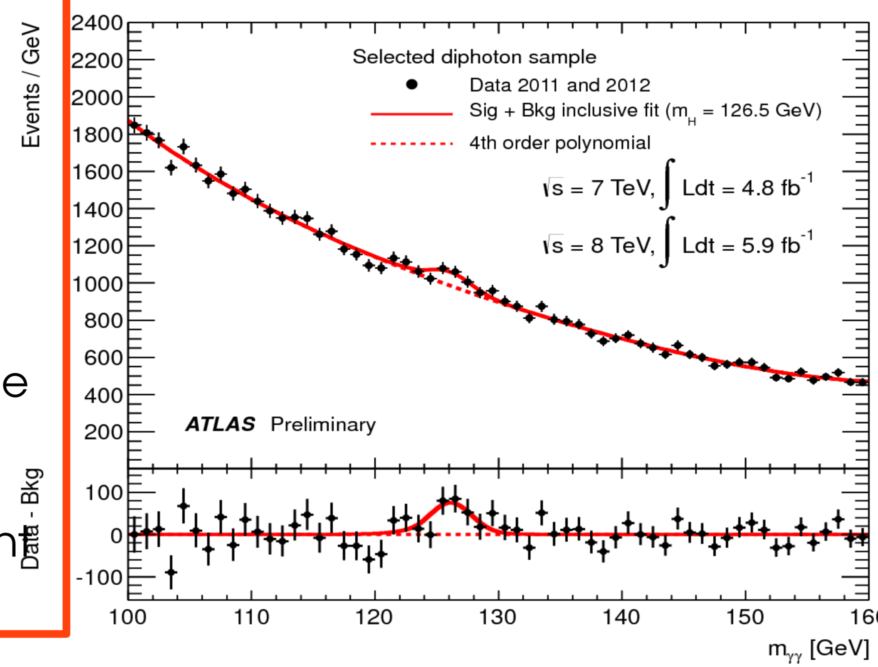
No: finding a fluctuation at **any** mass is much more likely than finding one at a **given** mass.

$$p_0^{float} = p_0^{fix} \times N$$

N the "**trials factor**" = number of independent regions in mass range, ~ $(m_{H,max}-m_{H,min})/(2\sigma_{peak})$

As the search interval increases, probability to find fluctuations of arbitrary size becomes large

Technically, the problem is that μ plays a role only in the $\mu>0$ hypothesis; for $\mu=0$, μ is irrelevant
=> **Wilks' theorem not valid.**



Look-elsewhere effect (2)

Solution: get distribution of floating-mass q_0 from toys – expensive in CPU for 5σ..
$\rightarrow$ Another approach: note that

$$p(q_0>X \text{ in } (m_{H,min}, m_{H,max})) = p(q0>X \text{ @ } m_{H,min}) + p(q0<X \text{ @ } m_{H,min})\ p(q_0 \text{ crosses} > X)$$

(To be >X somewhere, you either start >X or you cross into it at some point)

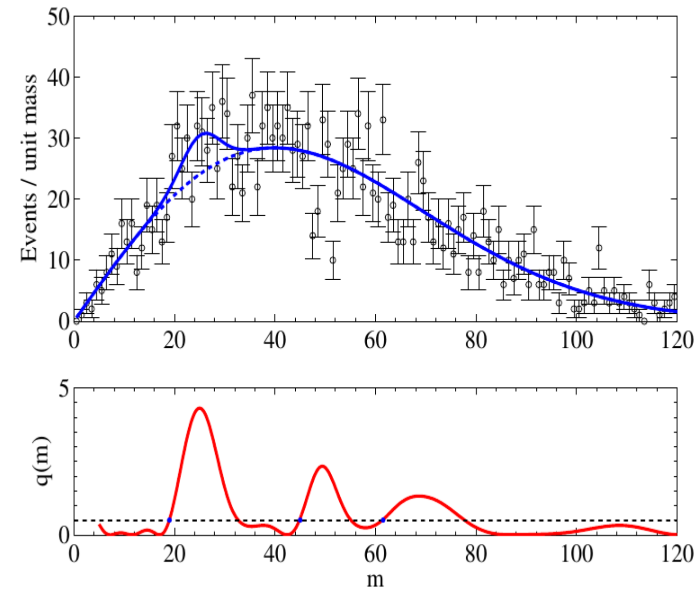
$\rightarrow$ For large X, p(upcrossing above X) ~ $\langle N_c(X)\rangle$, the average # of upcrossings,

$$p(q_0>X \text{ with floating mass}) = p(q_0>X \text{ at fixed mass}) + \langle N_c(X)\rangle$$

Interesting since $\langle N_c(X)\rangle$ has known dependence on X (for large enough X) :

$$\langle N_c(X)\rangle \sim e^{-X/2}.$$

So we can use toys to compute $\langle N_c(X)\rangle$ for small values of X (which is cheap), then extrapolate to 5σ



E. Gross and O. Vitells, Eur. Phys. J. C70 (2010) 525–530
R. B. Davies, Biometrika 74 no. 1, (1987) 33–43

Categories

→ Dataset contains "good" regions: Higher S/B, better resolution, etc.

There is a tradeoff:

→**Select good regions only**: gain on performance, lose on statistics (fewer events)

→ **Select everything**: more events, but good regions get diluted.

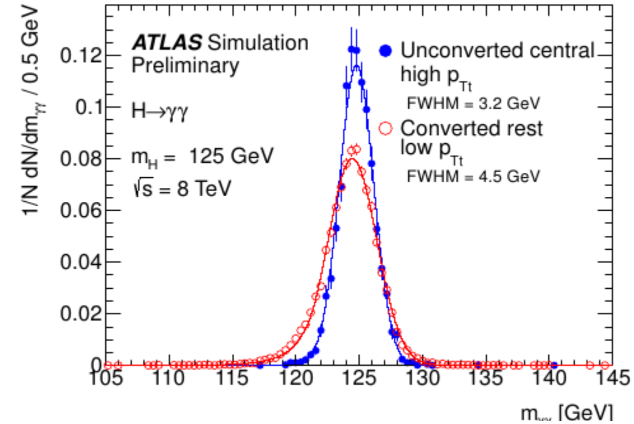
Categories : split dataset into subsets. For instance "good region" and "the rest"

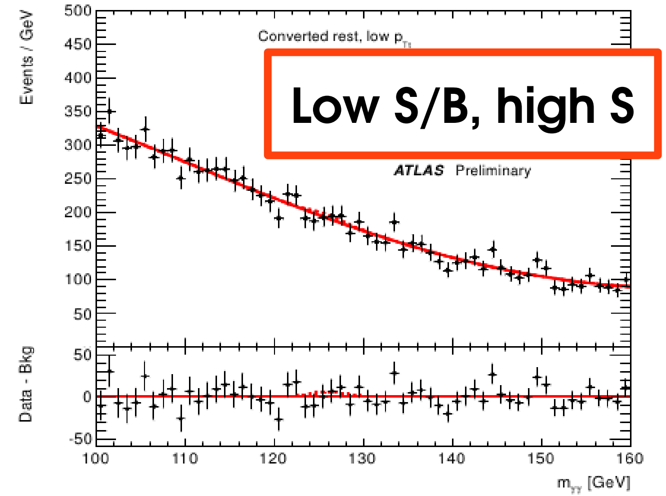
→ **Each subset modeled separately**, so can take advantage of better regions

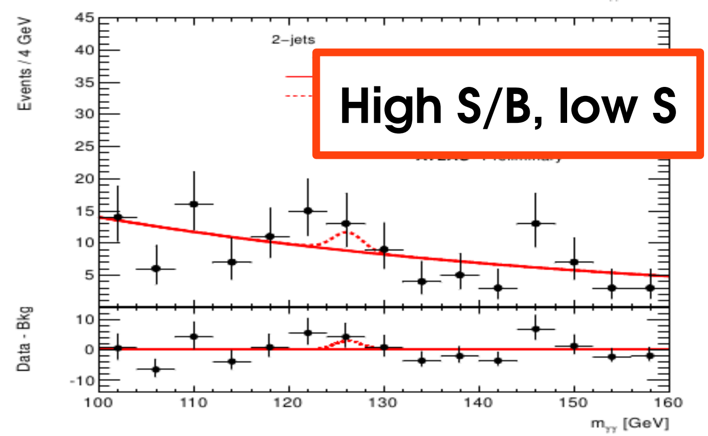
→ Fits are done simultaneously, so **some parameters can be common (μ, m_H)**

→ Fitted values are **automatically "combined"** across categories.

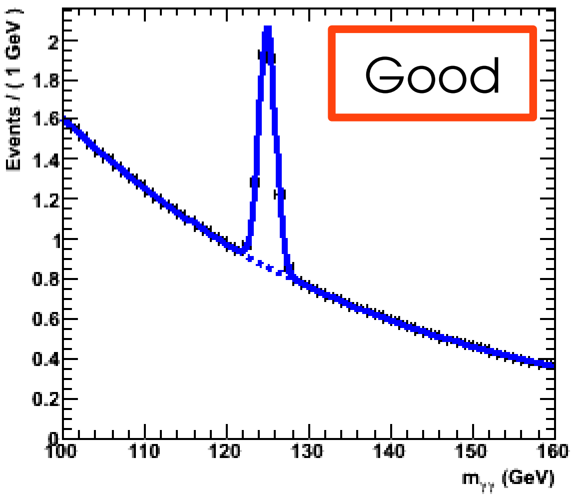
Technically $$L = \prod_{i=1}^{N_{cat}} L_i(\mu, \theta; data_i)$$

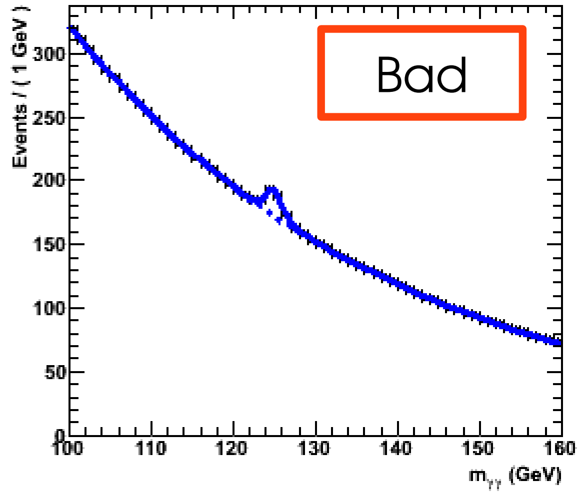


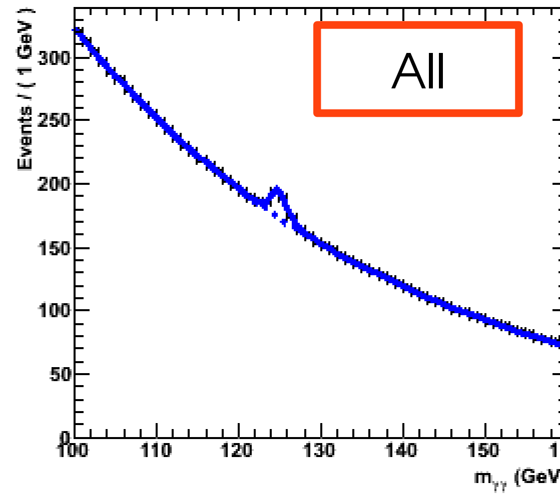




Categories: purity example







Good category

$N_S = 3$

$N_B = 50$

$\sigma = 1$ GeV

$\mathbf{Z_1 = 1.4\sigma}$

Bad category

$N_S = 50$

$N_B = 10000$

$\sigma = 1$ GeV

$\mathbf{Z_2 = 1.9\sigma}$

Inclusive

$N_S = 53$

$N_B = 10050$

$\sigma_1, \sigma_2 = 1$ GeV

$\mathbf{Z_I = 2.0\sigma}$

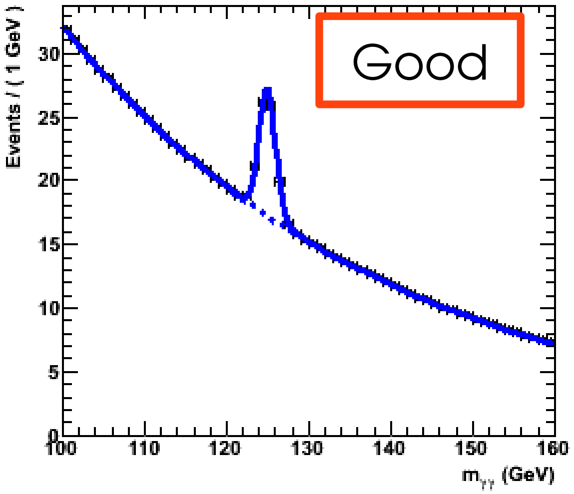
2-category result

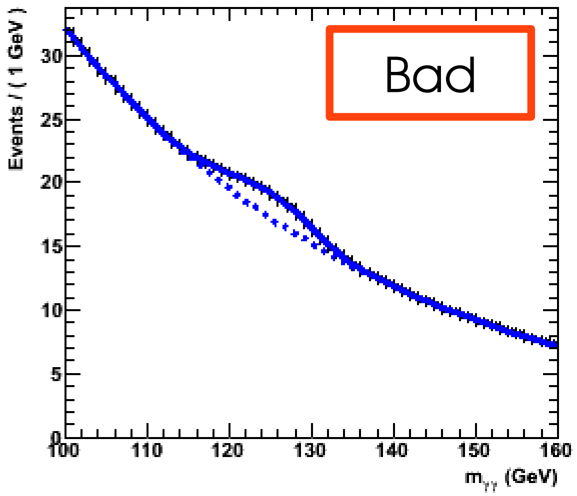
$\mathbf{Z_C = 2.4\sigma}$

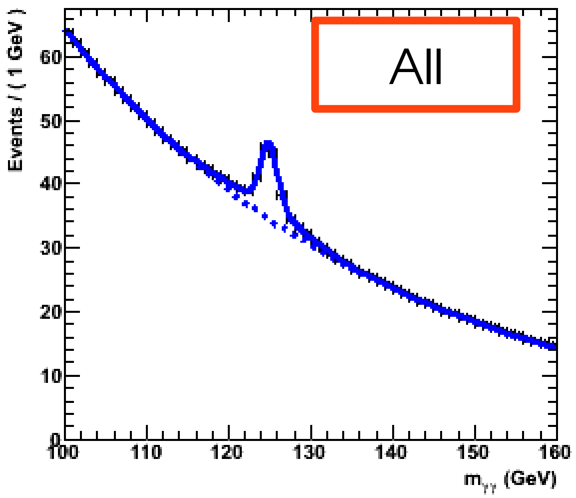
Dilution : cat1 adds little to inclusive result

Categories give the expected $Z_C \sim Z_1 \oplus Z_2$

Categories: resolution example







Good category

$N_S = 25$

$N_B = 1000$

$\sigma = 1$ GeV

$\mathbf{Z_1 = 2.9\sigma}$

Bad category

$N_S = 25$

$N_B = 1000$

$\sigma = 5$ GeV

$\mathbf{Z_2 = 1.2\sigma}$

Inclusive

$N_S = 50$

$N_B = 2000$

$\sigma_1, \sigma_2 = 1,5$ GeV

$\mathbf{Z_I = 2.6\sigma}$

2-category result

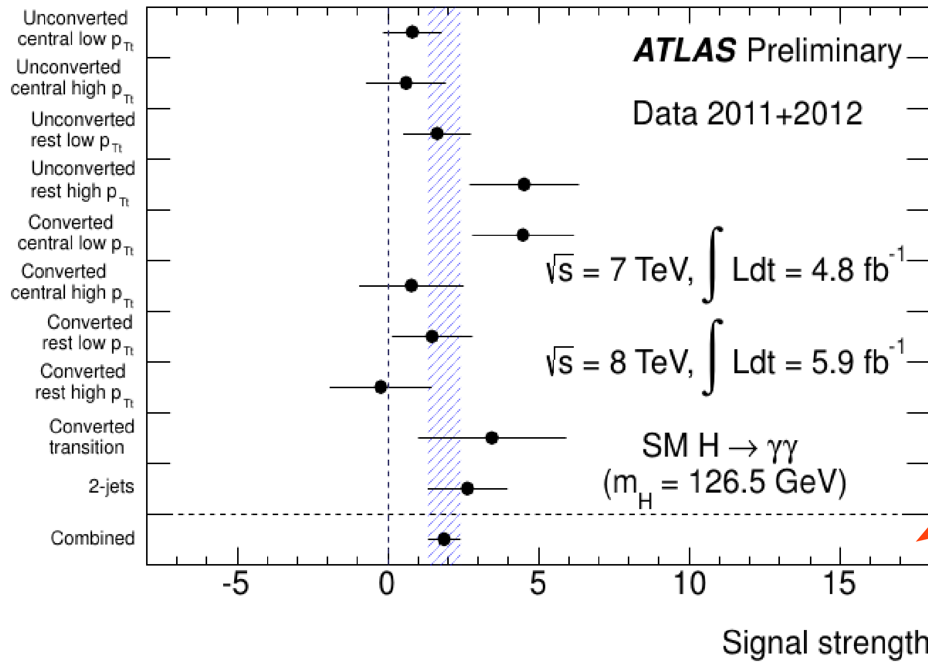
$\mathbf{Z_C = 3.1\sigma}$

Dilution : Inclusive result worse than 1 alone

Categories give the expected $Z_C \sim Z_1 \oplus Z_2$

H→γγ Category results

ATLAS-CONF-2012-091



Per-category $\hat{\mu}$ from fits to individual categories

Global $\hat{\mu}$ from simultaneous fit with single μ for all categories

H->γγ: separate "good" regions:

→ **Central γs** (better S/B, resolution)

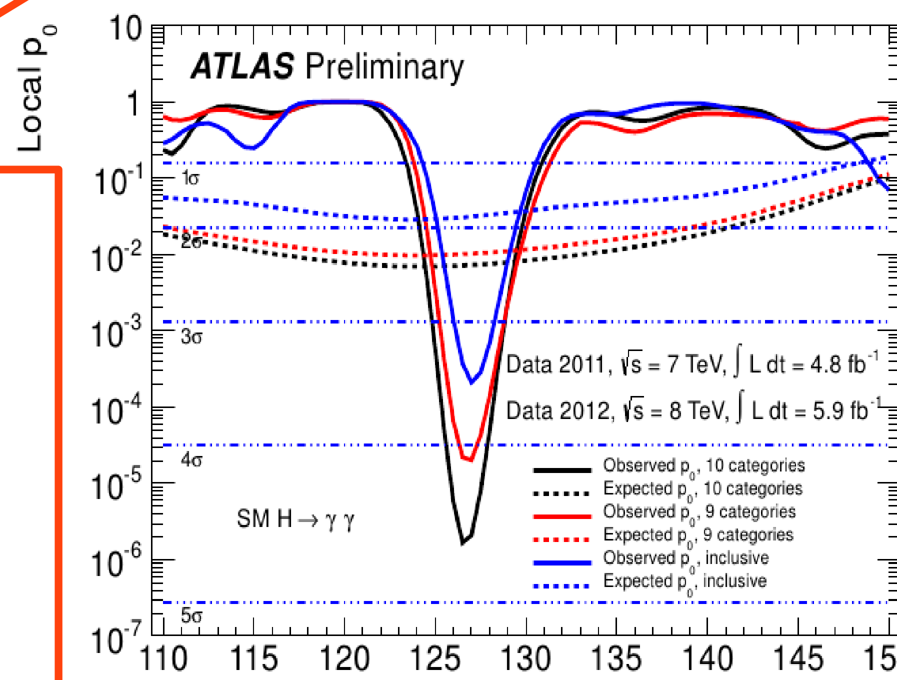
→ **Unconverted γs** (resolution)

→ **High p_T(t)** : higher S/B

→ **2-jet "VBF" topology** : higher S/B

=> 10 categories

Significant improvement in overall performance



ATLAS-CONF-2012-91

Binned ML

So far we have discussed the unbinned case, with parametrized PDFs, e.g.

$$L(m_{\gamma\gamma,1}..m_{\gamma\gamma,N_{obs}};\mu)=e^{-(\mu N_s^{SM}+N_b)}\prod_{i=1}^{N_{obs}}\mu N_s^{SM}P_s(m_{\gamma\gamma,i})+N_bP_b(m_{\gamma\gamma,i})$$

Another common approach is the binned likelihood, based on histograms:

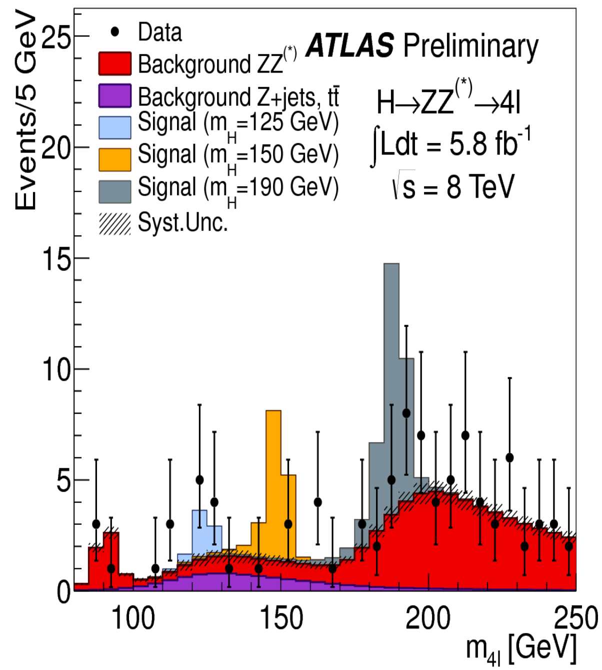
→ Define a binning in the variable(s) of interest, say m_i, i=1..N_{bins}

→The model gives the bin contents, for instance:

$$\mathbf{N_{model,i} = \mu N_{s,i}(\theta) + N_{b,i}(\theta)}$$

→ The per-bin likelihood just describes Poisson fluctuations around these values

$$P(N_{data,i};\mu,\theta)=e^{-N_{model,i}(\theta)}\frac{N_{model,i}(\theta)^{N_{data,i}}}{N_{data,i}!}$$



And the full likelihood is

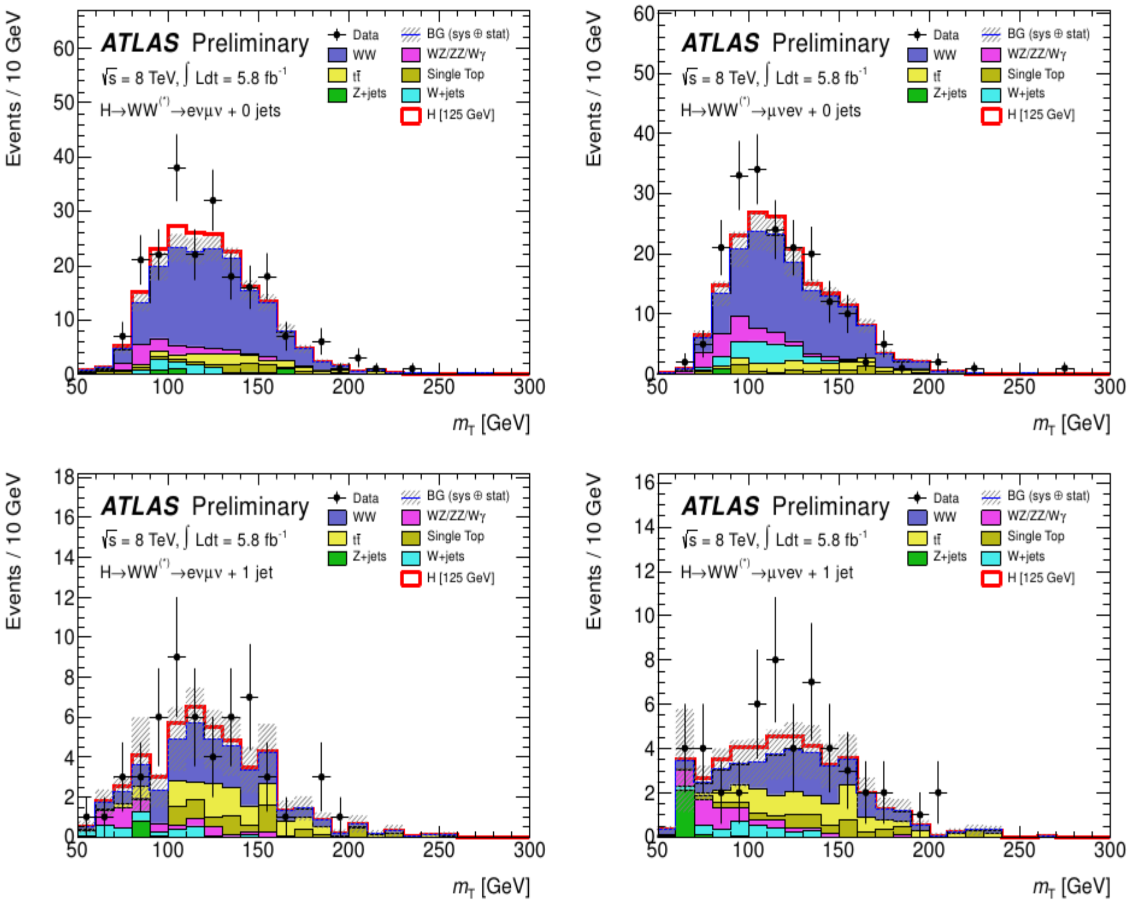
$$L(N_{data,1}..N_{data,N_{bins}};\mu,\theta)=\prod_{i=1}^{N_{bins}}P(N_{data,i};\mu,\theta)\propto e^{-(\mu N_s^{SM}+N_b)}\prod_{i=1}^{N_{bins}}N_{model,i}(\theta)^{N_{data,i}}$$

Binned vs. Unbinned

Binned	Unbinned
	Not dependent on binning (!)
Can use histogram templates directly	Need to fit templates to an analytic shape, include modeling error
Usually faster ($N_{bins} < N_{events}$)	
Sensitive to statistical fluctuations of templates	Fits to analytic shape usually removes effect of fluctuations

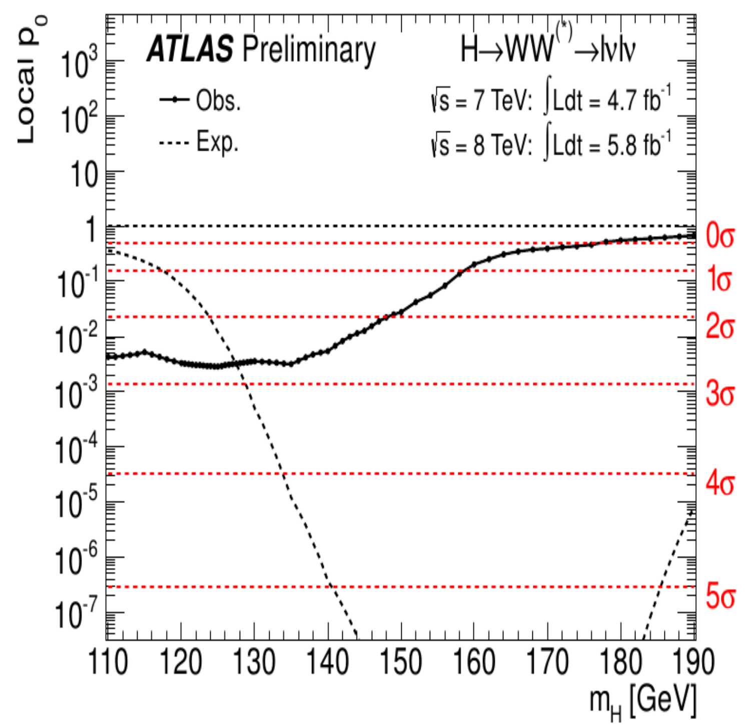
Which one to use: it depends!
→ Do we have high-statistics templates ?
→ Is there a convenient/simple analytic shape to use ?

Results : 2012 H→WW→lνlν



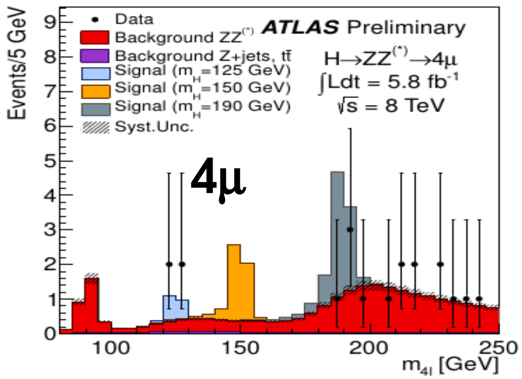
ATLAS-CONF-2012-098

Category splittings:
→ Association with 0,1,2 jets
→ ee,eμ,μμ

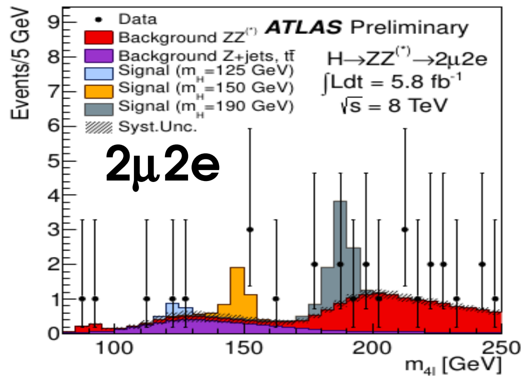


Analysis techniques:
→ Binned likelihood in 0, 1-jet categories
→ Counting analysis in 2-jet catgories

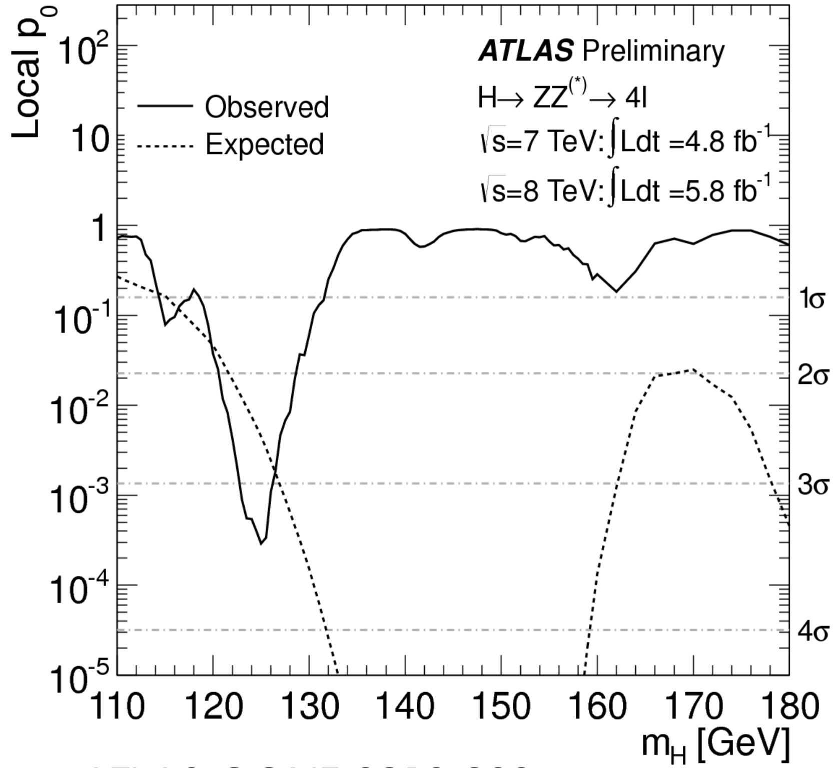
Results : H→ZZ→4l



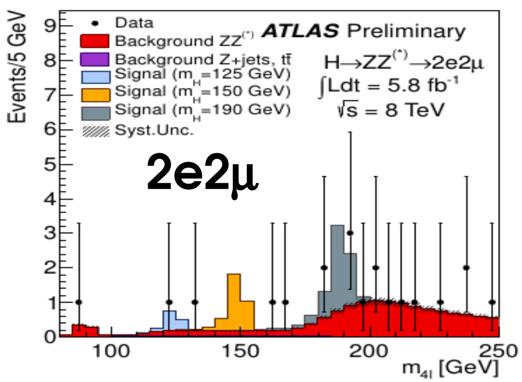
(a)

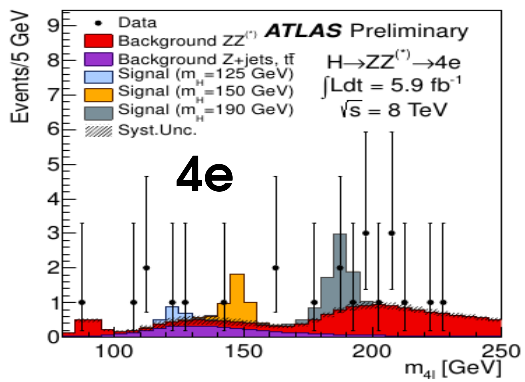


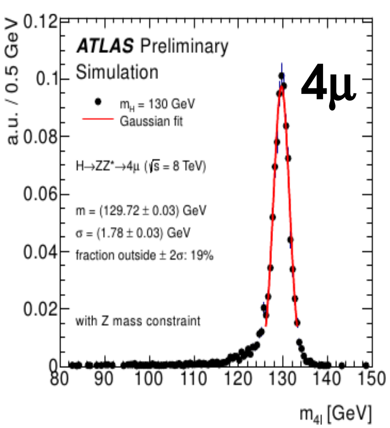
(b)

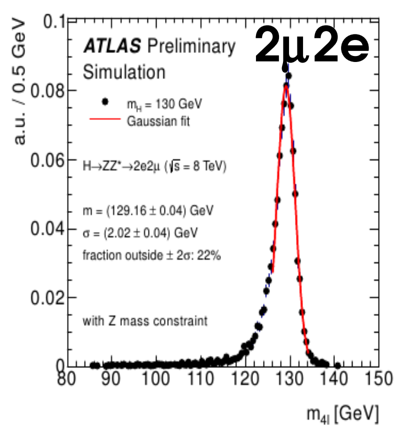


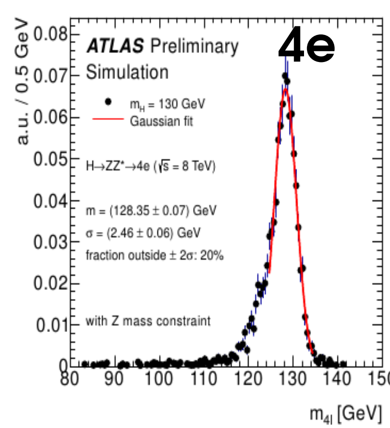
ATLAS-CONF-2012-092











Categories:
→ 4e, 2e2μ, 2μ2e,4μ
→ 2011, 2012

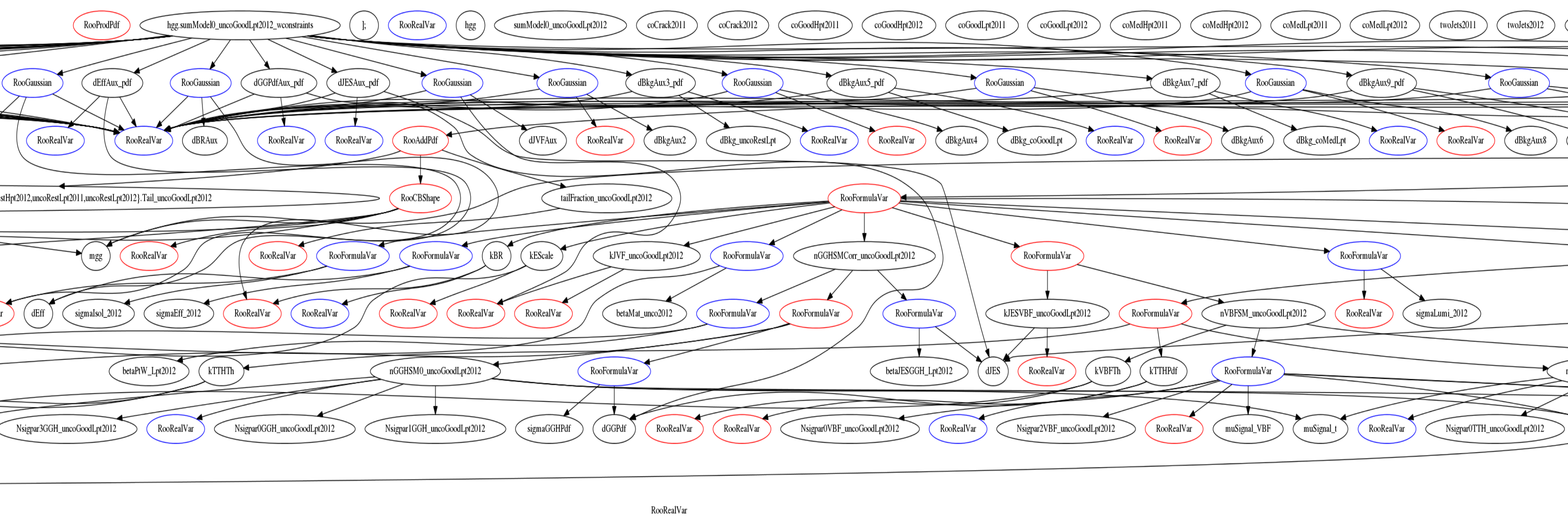
Use binned ML model everywhere

Combination

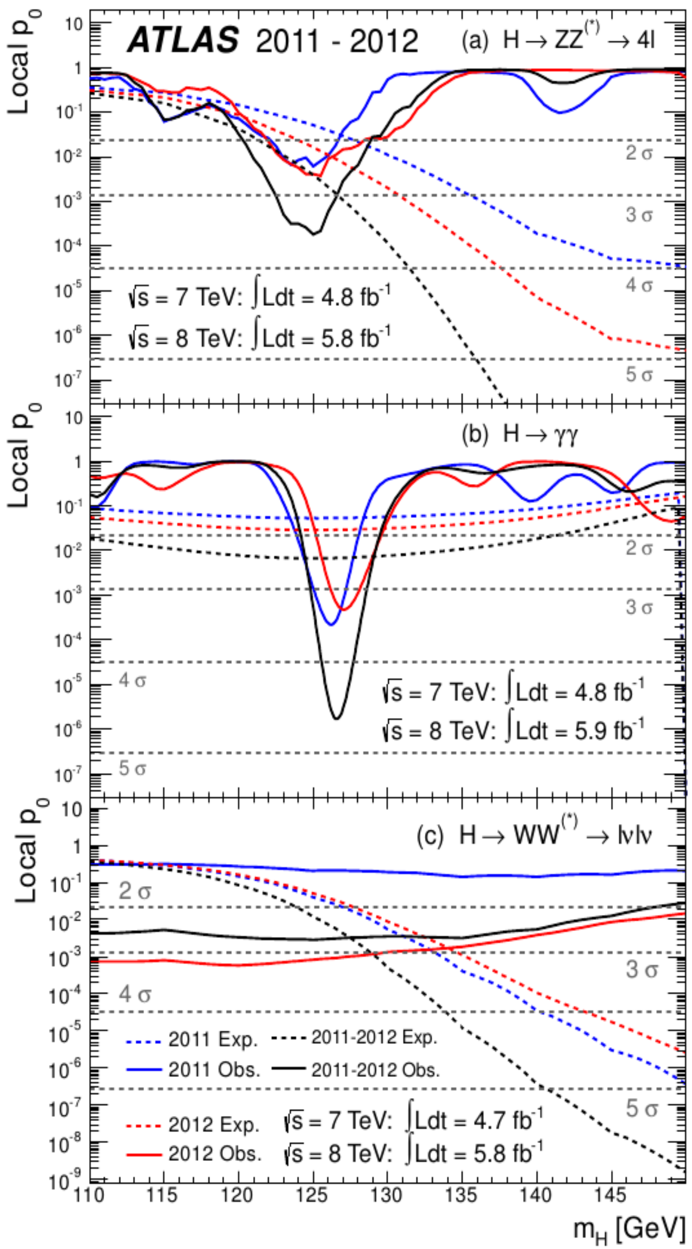
To combine channels together: just use categories
→ Each channel is one category (or several)
→ Share parameters:
→ Common physics parameters (μ, m_H...)
→ Common systematics

Global Higgs model:
78 categories in all, each separately parametrized

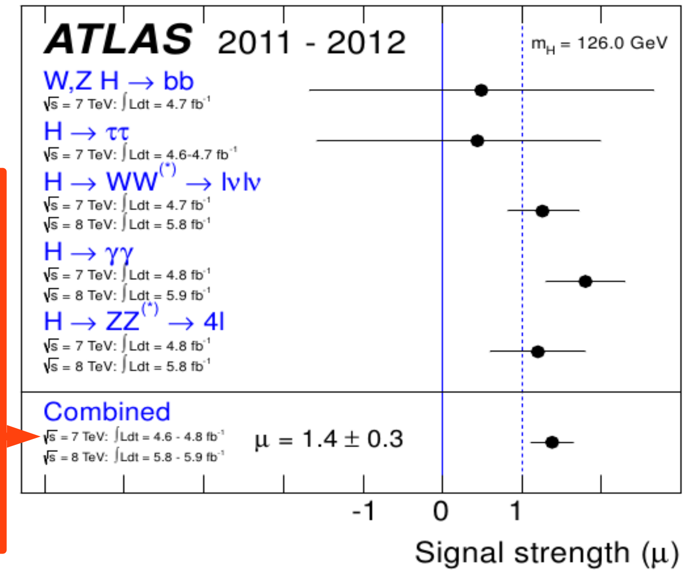
Below: part of 1 category (of 20) of the $H\to\gamma\gamma$ component of the combined model (each node is a parameter or a PDF)

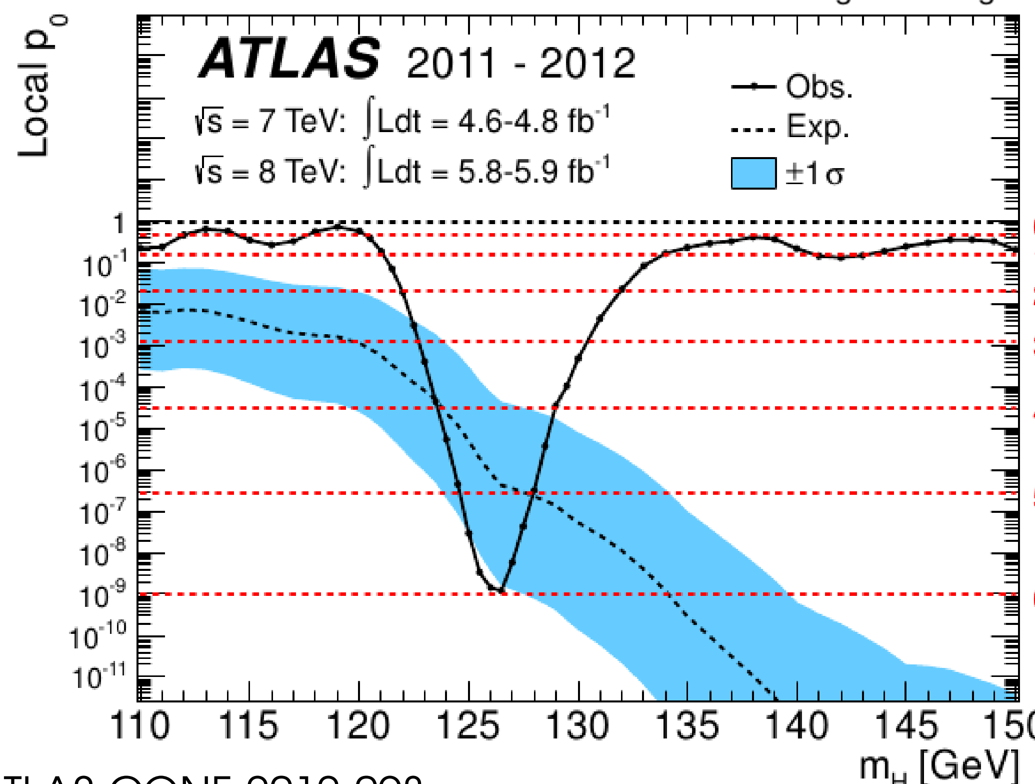


Combination results



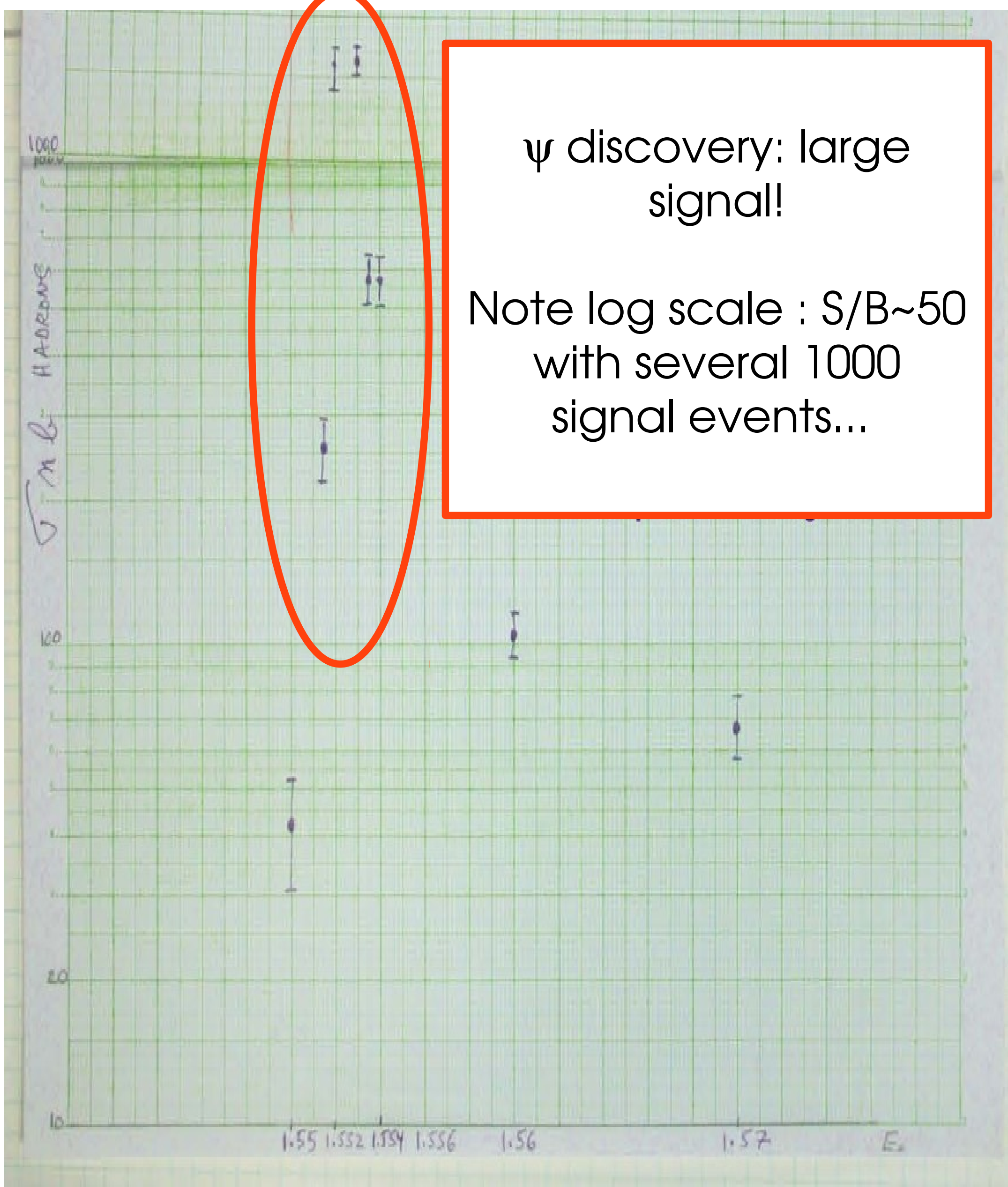
Global $\hat{\mu}$ value from simultaneous fit to all channels (with common μ for all)





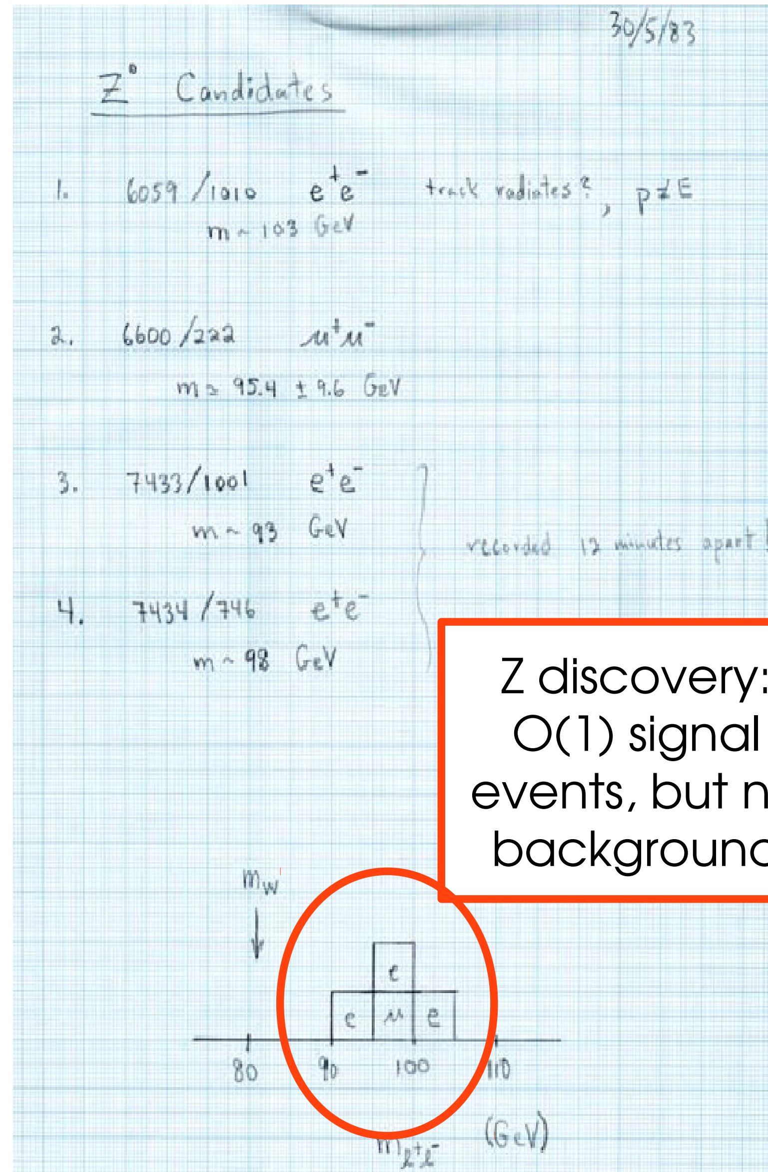
ATLAS-CONF-2012-093

Interlude: why didn't we need this before ?



ψ discovery: large signal!

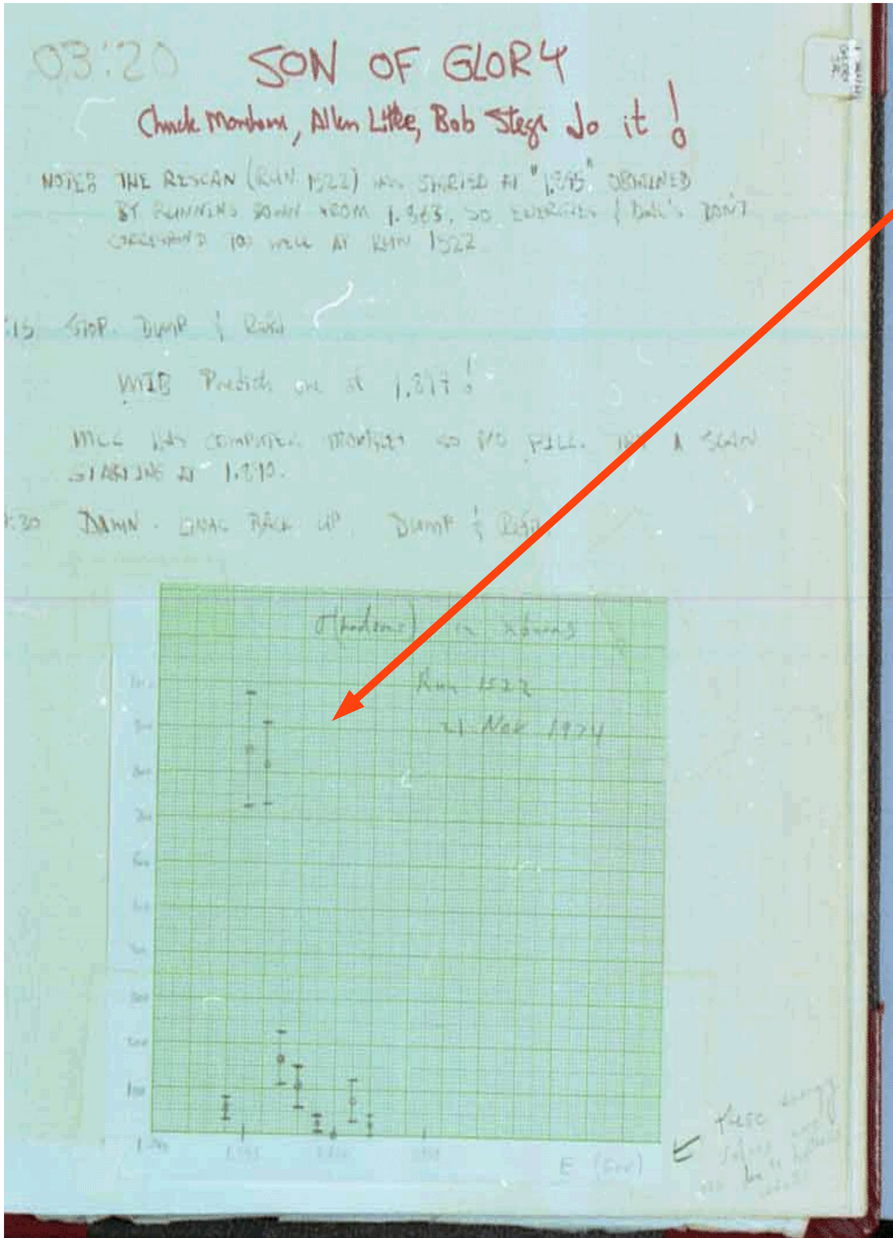
Note log scale : S/B~50 with several 1000 signal events...



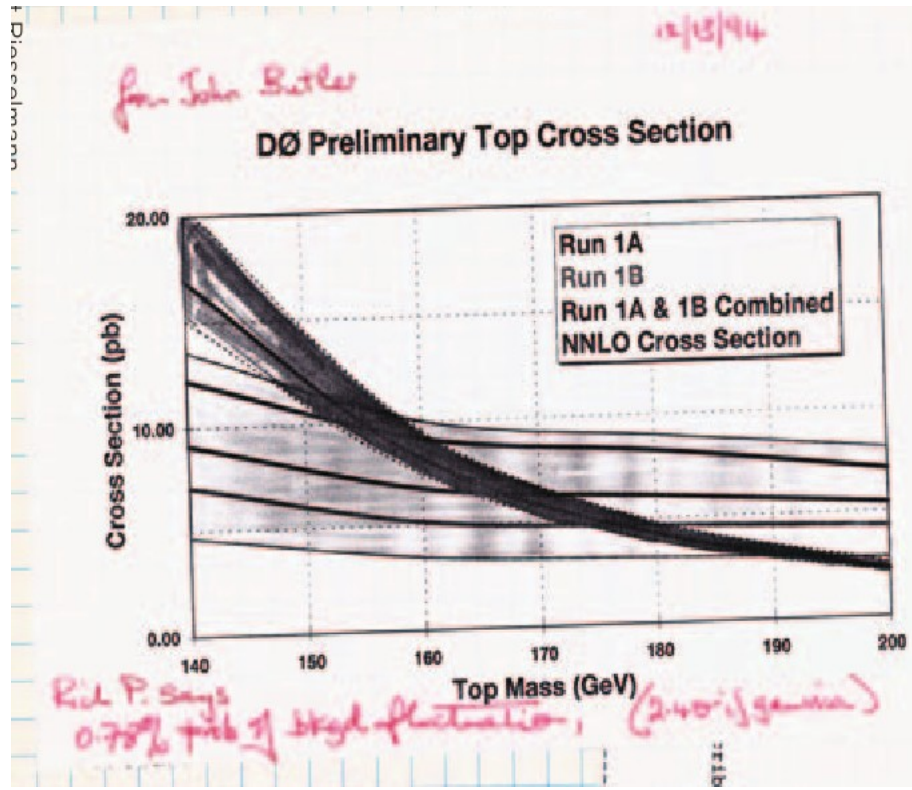
Z discovery: O(1) signal events, but no background

Logbook of J. Rohlf, 1983-05-30

More examples



ψ' : discovered online by the (lucky) shifters, similar story to ψ…



First hints of top at D0: O(10) signal events, a few background events, 0.78% p-value

Why we need this now

→ The high-signal, low-background experiments have been done already (but a surprise at 5 TeV would be welcome...)

At LHC:

→ High background levels, need precise modeling

→ Large systematics, need to be treated correctly

→ Small signals: need optimal use of available information :

- → shape analyses instead of counting
- → Isolation of signal-enriched regions (categories)

Outline

What are the goals ?

Setting up the problem : Maximum likelihood and Likelihood ratios

Discovery

Additional wrinkles (categories, LEE)

Limit setting

Further topics

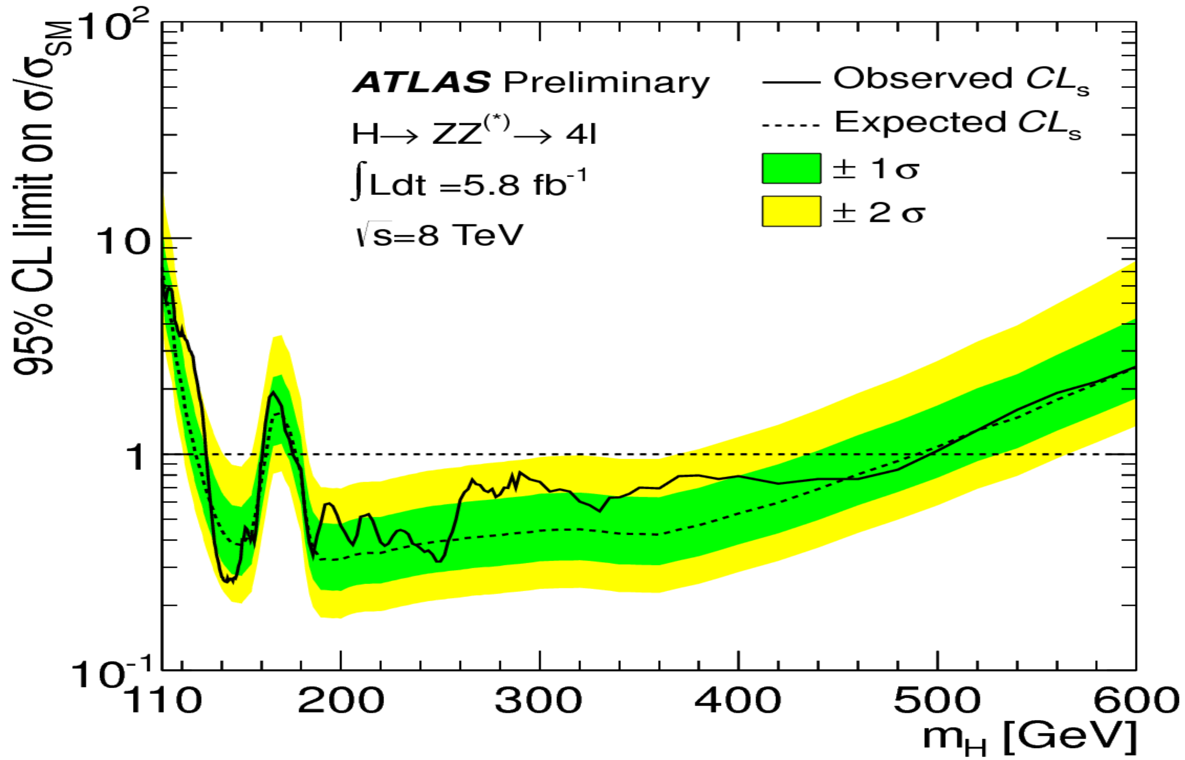
Limits ?

Do we still need them ?

Yes, need to find out if there are other bosons out there!

The goal:

Actually for other bosons need limits on σ, not σ/σ_{SM}. But not available yet ...



"**Set upper limit on μ**" = Try to exclude the $\mu S+B$ hypothesis for μ above some value. Similar situation to discovery, can reuse the same tools.

Statistic for limit-setting

Following our usual procedure, use q_μ = -2 log L(data; μ)/L(data; $\hat{\mu}$) to exclude the μS+B hypothesis.
→ If μ<<$\hat{\mu}$, this is large (bad agreement, good exclusion)
→ If μ~$\hat{\mu}$, this is small (good agreement, bad exclusion)

Problem: if $\hat{\mu}$>>μ, large as well. But too-large $\hat{\mu}$ shouldn't give good upper limit! => again, use a one sided version

$$\tilde{q}_\mu = \begin{cases} -2\log\frac{L(\mu;data)}{L(\hat{\mu};data)} & 0<\hat{\mu}<\mu \\ 0 & \hat{\mu}\geq\mu \\ -2\log\frac{L(\mu;data)}{L(\mu=0;data)} & \hat{\mu}<0 \end{cases}$$

Also separate $\hat{\mu}$<0 for "technical" reasons: fits can be unstable. In this case, use the value of q for μ=0

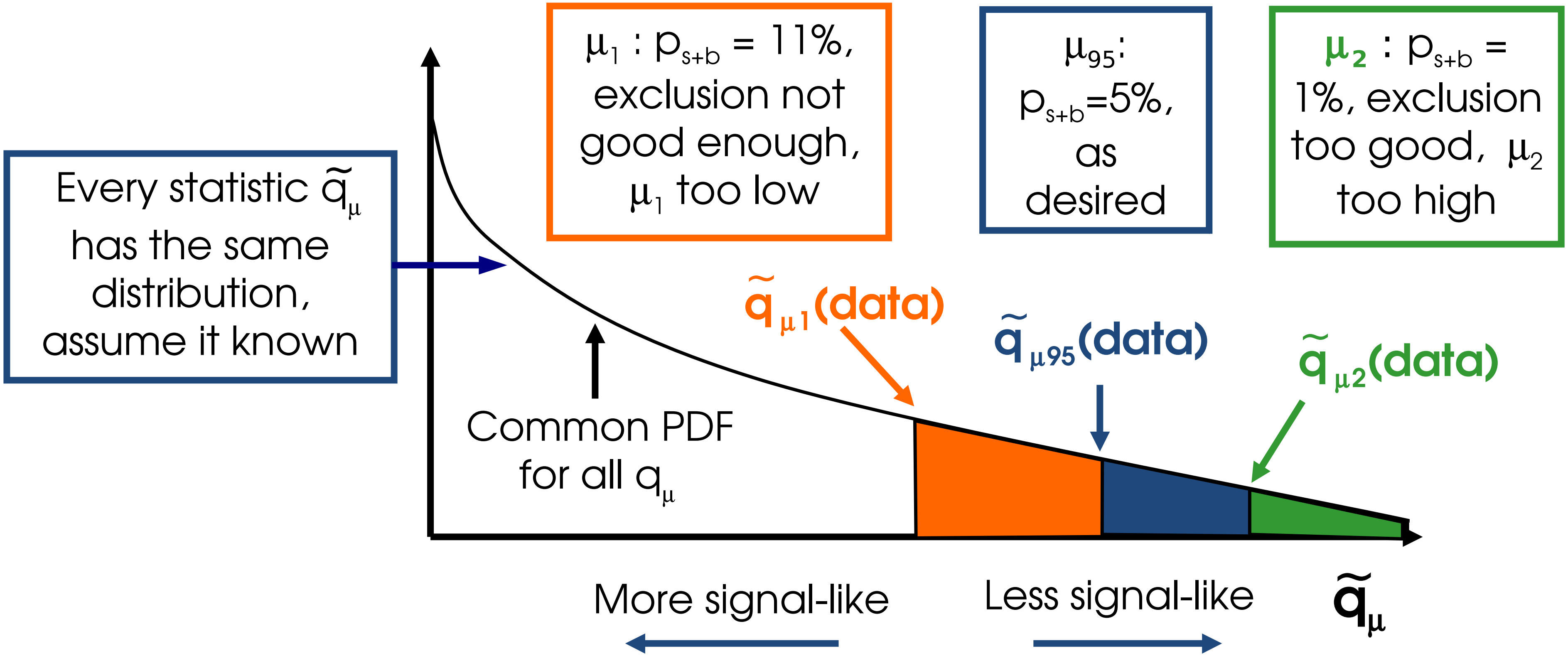
Again, Wilks' theorem gives the distribution
(need to measure one parameter (σ) separately...)

$$f(\tilde{q}_\mu|\mu) = \frac{1}{2}\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}e^{-\tilde{q}_\mu/2} & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2\,, \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu+\mu^2/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2\,. \end{cases}$$

The inversion problem

For each μ, we can compute the $q_{\mu,obs}$ of our data and the p-value.

However what is usually needed is instead the value of μ which yields a given p-value, usually p=0.05 (95% exclusion)
=> need to solve for μ



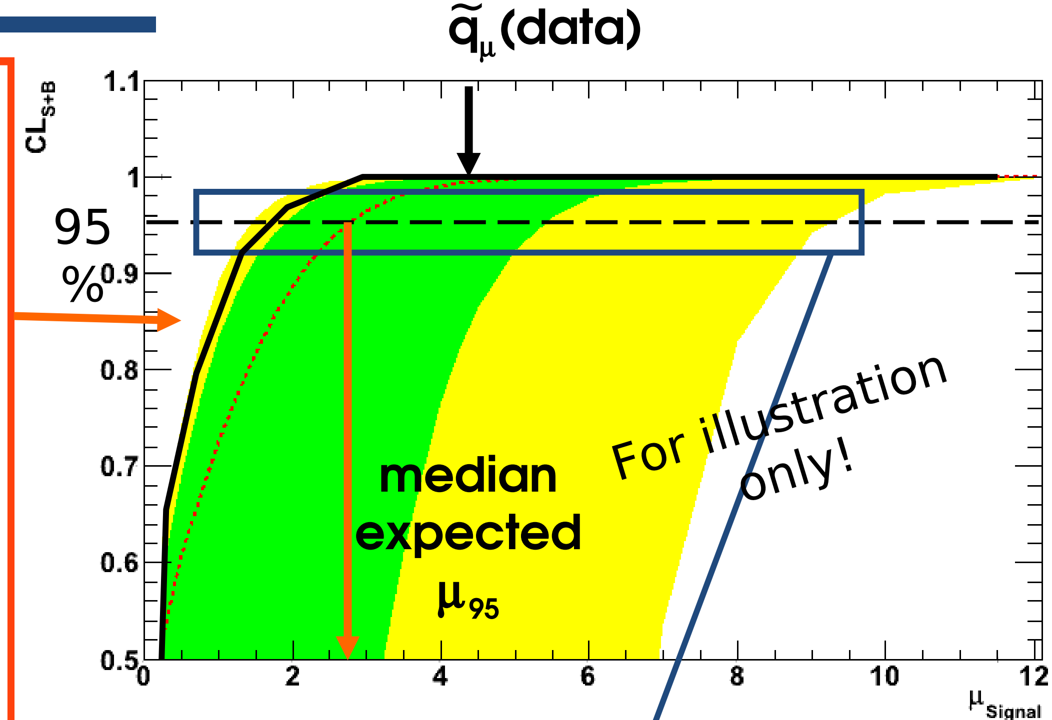
Inversion in practice

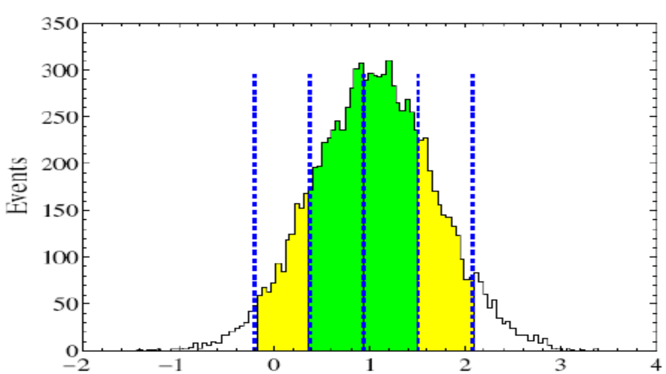
In practice, inversion procedure done as follows:

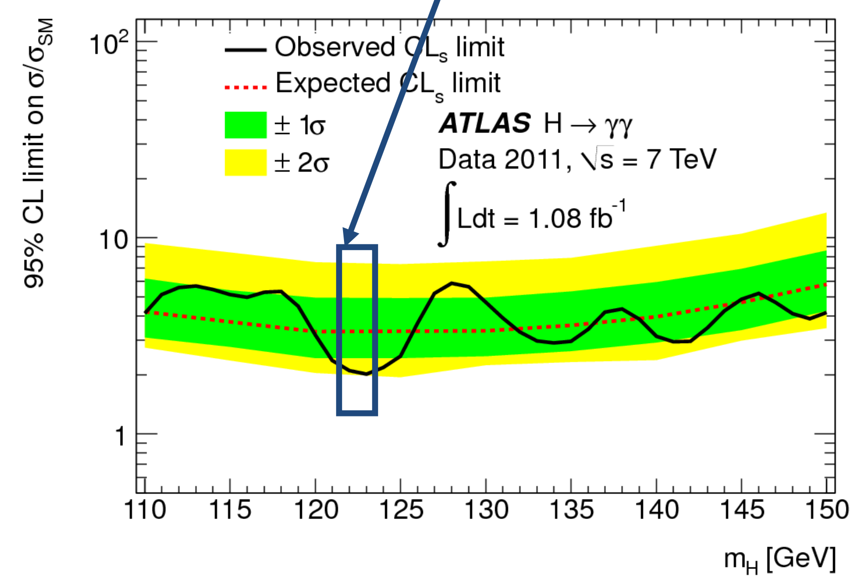
→ Define a set of values to scan (here 0-12 with varying step sizes)

→ Compute p_{s+b} for each value, find crossing with 95%

→ Expected: Generate toys (usually for μ=1) and histogram values of μ_{95}. Report median and +/- 1,2 σ quantiles.







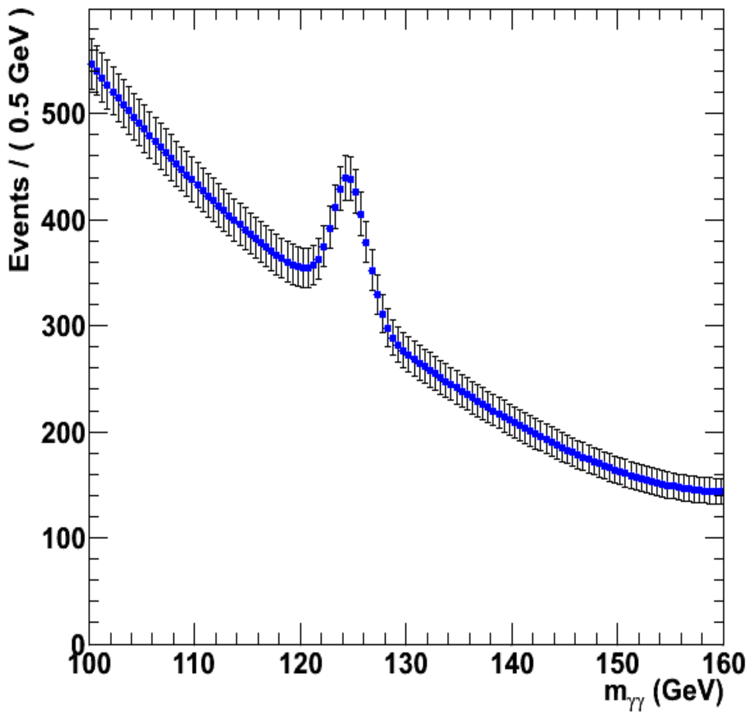
Asimov datasets

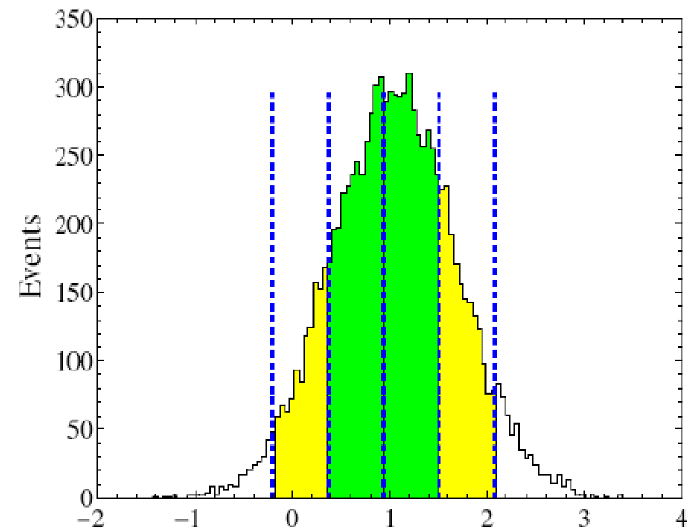
Cases when toys are needed:
→ Compute expected p_0, upper limits
→ Compute σ parameter of q_μ asymptotic distribution

$$f(\tilde{q}_\mu|\mu) = \frac{1}{2}\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}e^{-\tilde{q}_\mu/2} & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2 , \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)} \exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu + \mu^2/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2 . \end{cases}$$

In both cases, goal is to determine a quantity in a given μ scenario. Need to run toys to average over statistical fluctuations.

Another approach: **Asimov dataset** = "perfect" dataset with no statistical fluctuations. (technically such that ML estimate of all parameters are equal to predefined values)
=> Get quantities from a single determination
For limit quantiles, get bands from value of σ





Systematics

Good case: parameter constrained by data: use **profiling.**

Bad case: parameter not constrained by data.
e.g. signal efficiency, energy scale…
"**Systematic error**"

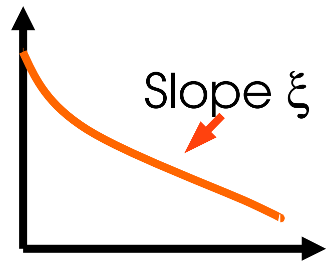
Assume an **auxiliary measurement** constrains it (e.g. Z→ee)
In the combined experiment, ε is constrained and we can fit its value.

Dragging Z data into fit is not practical! What we care about is the measured value $\varepsilon_{mes} \pm \delta\varepsilon$. so parameterize the auxiliary measurement as
L(Z data; ε) = L(ε_{mes}; ε; σ_ε)

And usually: $L(\varepsilon_{mes};\varepsilon,\sigma_\varepsilon)=\exp\left[\frac{-(\varepsilon_{mes}-\varepsilon)^2}{2\sigma_\varepsilon^2}\right]$

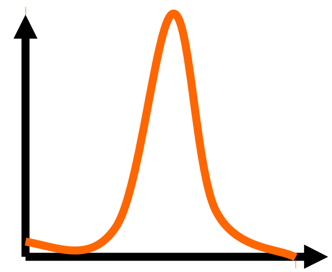
Then profile ε like ξ.

H→γγ



L(γγ data; μ, ξ, ε)

Z→ee



L(Z data; ε)

L(γγ data; μ, ξ, ε) L(Z data; ε)
=
L(γγ+Z data; μ, ξ, ε)

Constrained by H→γγ

Constrained by Z→ee

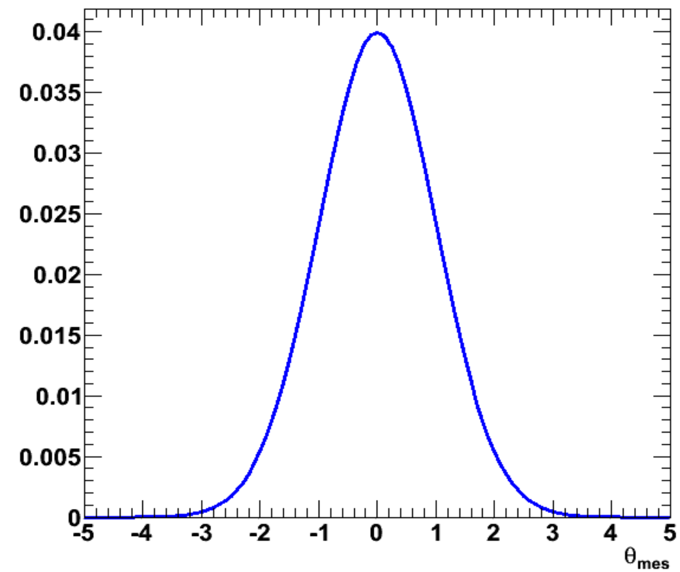
L(γγ data; μ, ξ, ε) L(Z data; ε)

L(γγ data; μ, ξ, ε) G(e_{mes}; ε, σ_ε)

Choice of constraints

Ideally, choice driven by properties of the auxiliary measurement.

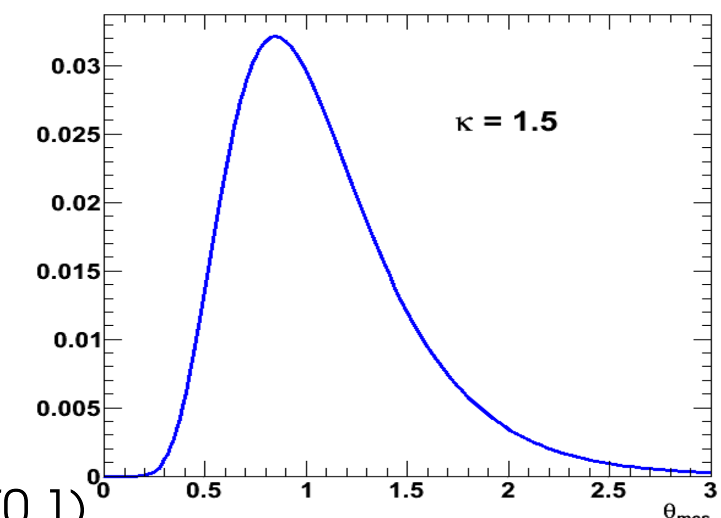
In practice, often use **Gaussians**:
→ Implement systematic effects as
$X \rightarrow X(1 + \sigma\theta)$ where $\theta \sim G(0,1)$
→ Reasonable approximation to most cases
→ Computationally efficient



Other choices
→ **Bifurcated Gaussian**: for asymmetric errors

→ **Log-normal**: for corrections on positive numbers (normalizations).

$$f(\theta;\theta_{0,}\kappa) = \frac{1}{\theta\sqrt{2\pi}\log\kappa}\exp\left[\frac{-1}{2}\left(\frac{\log(\theta/\theta_0)}{\log\kappa}\right)^2\right]$$



Represents a multiplicative uncertainty.
e.g. κ=1.50 represents an errors by x/÷ 1.50
Can implement as **$X \rightarrow X \exp(\sigma\theta)$** with $\theta \sim G(0,1)$

Systematics example

Use again the toy $H\to\gamma\gamma$ setup with fixed templates, just μ as free parameter

Look at m_H=120 GeV, μ=4 hypothesis
Best-fit is $\hat{\mu}$=0.85 (<<4), q_4 = 3.14 => p_{s+b}= 4%
μ=4 excluded at 95% CL

$$\tilde{q}_4 = -2\log\frac{L(\mu=4;data)}{L(\hat{\mu};data)}$$

Now add a systematic on efficiency, say
$\varepsilon = \varepsilon_0(1+\sigma_{\delta\varepsilon}\,\delta\varepsilon)$ and Gaussian constraint on $\delta\varepsilon$
For dramatic effect, use $\sigma_{\delta\varepsilon}$ = **30%**

$$L_S(\mu,\delta\varepsilon;data) = L(\mu\exp(\sigma_{\delta\epsilon}\delta\epsilon);data)\exp\left[-\frac{\delta\varepsilon^2}{2}\right]$$

Again m_H=116 GeV, μ=4 hypothesis
Best-fit $\hat{\mu}$ = 0.85 (<< 4) still, now q_4 = 2.17, **p_{s+b}=7%**
μ=4 **not excluded** at 95% CL

$$\tilde{q}_4 = -2\log\frac{L(\mu=4,\delta\hat{\hat{\varepsilon}};data)}{L(\hat{\mu},\delta\hat{\varepsilon};data)}$$

In fit with fixed μ=4, can now drag $\delta\varepsilon$ down => fit **$\delta\varepsilon$ = -24.6%.**
Mitigates tension between fixed μ=4 and best-fit $\hat{\mu}$=0.85 => μ=4 not excluded

Systematic parameter gives more freedom for the fixed hypothesis, makes it easier to reconcile hypo with data =>decreases exclusion potential.

Sensitivity issues

So far, use $\mathbf{CL_{s+b}}$ limits

asymptotically, $\mu_{95} \sim \hat{\mu} + 1.64\sigma$

Problem:

for negative $\hat{\mu}$, get very good (too good) limits.

For $\hat{\mu}$ sufficiently negative, can have limit < 0!

What is happening ?

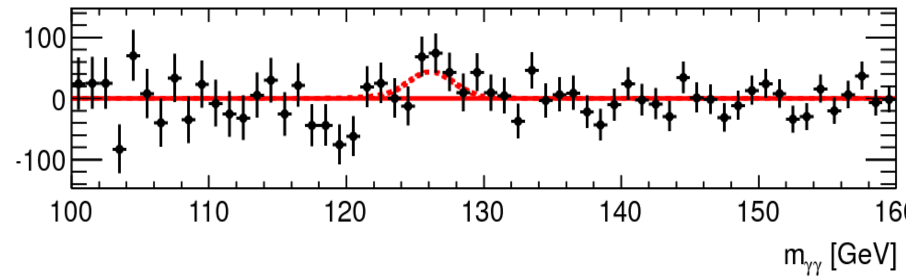
Remember this is a **95%** limit.

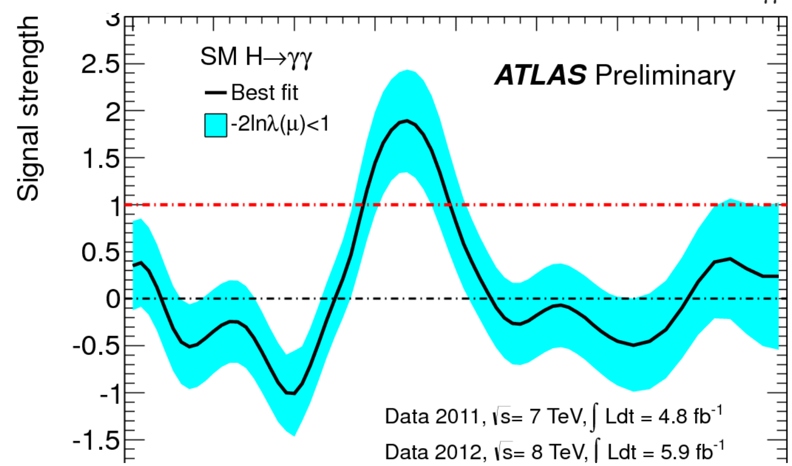
In other words, **5% of the time, the limit wrongly excludes the true value.**

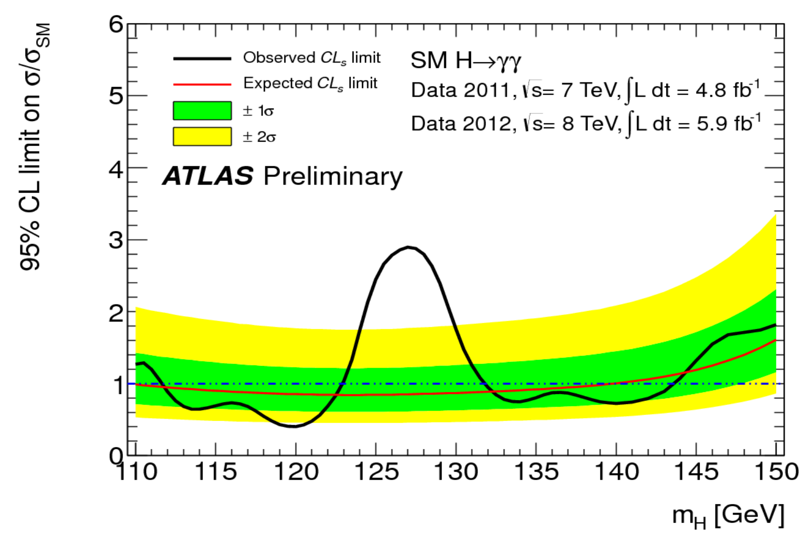
What can we do ?

Live with it ? Move to 99% ?

Understand what happens and fix it







Sensitivity issues

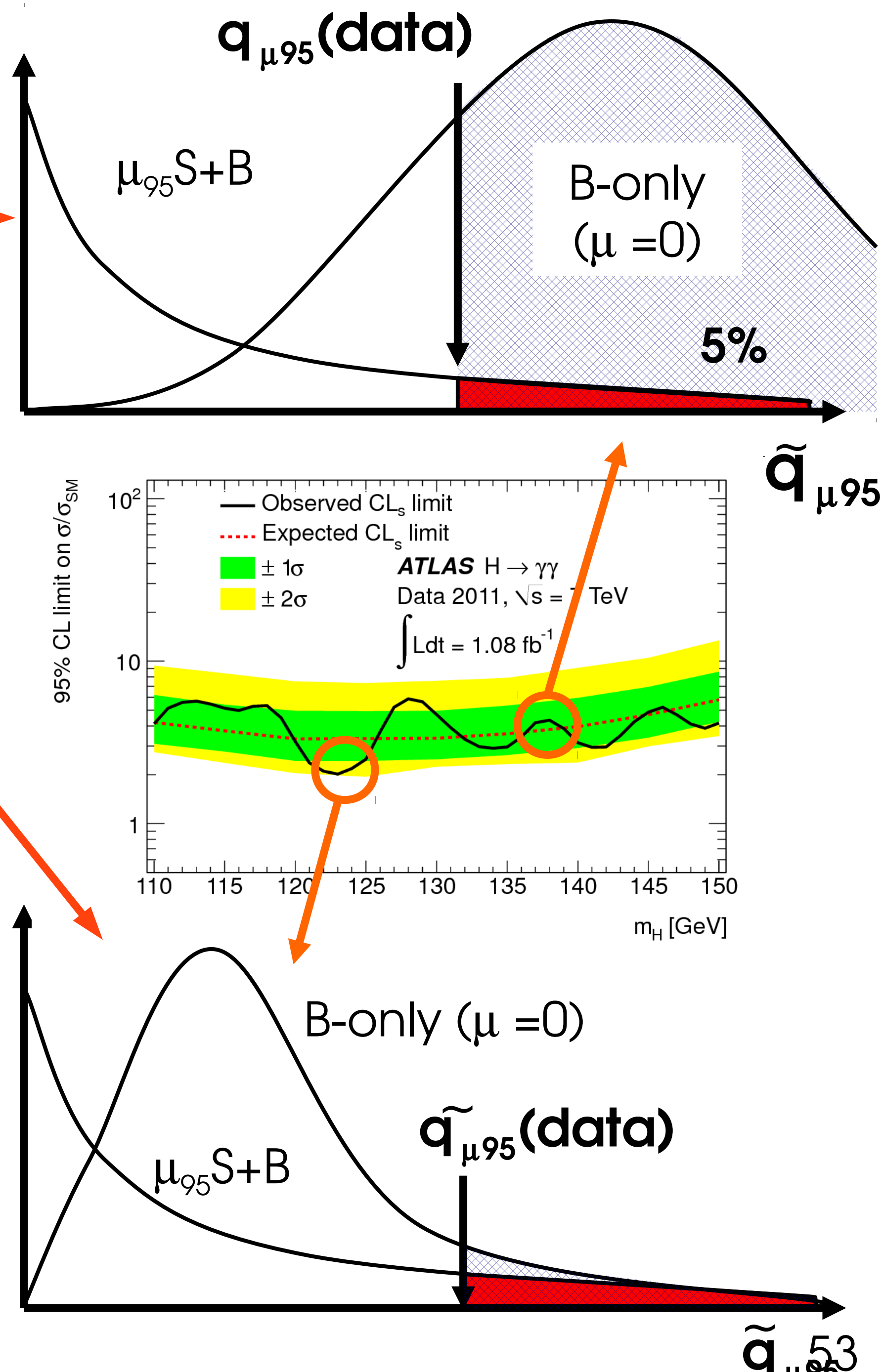
Usual situation: data is consistent with background-only hypothesis

Pathological case: we end up in the tails of B as well as most S+B, e.g. because data fluctuated below expected background

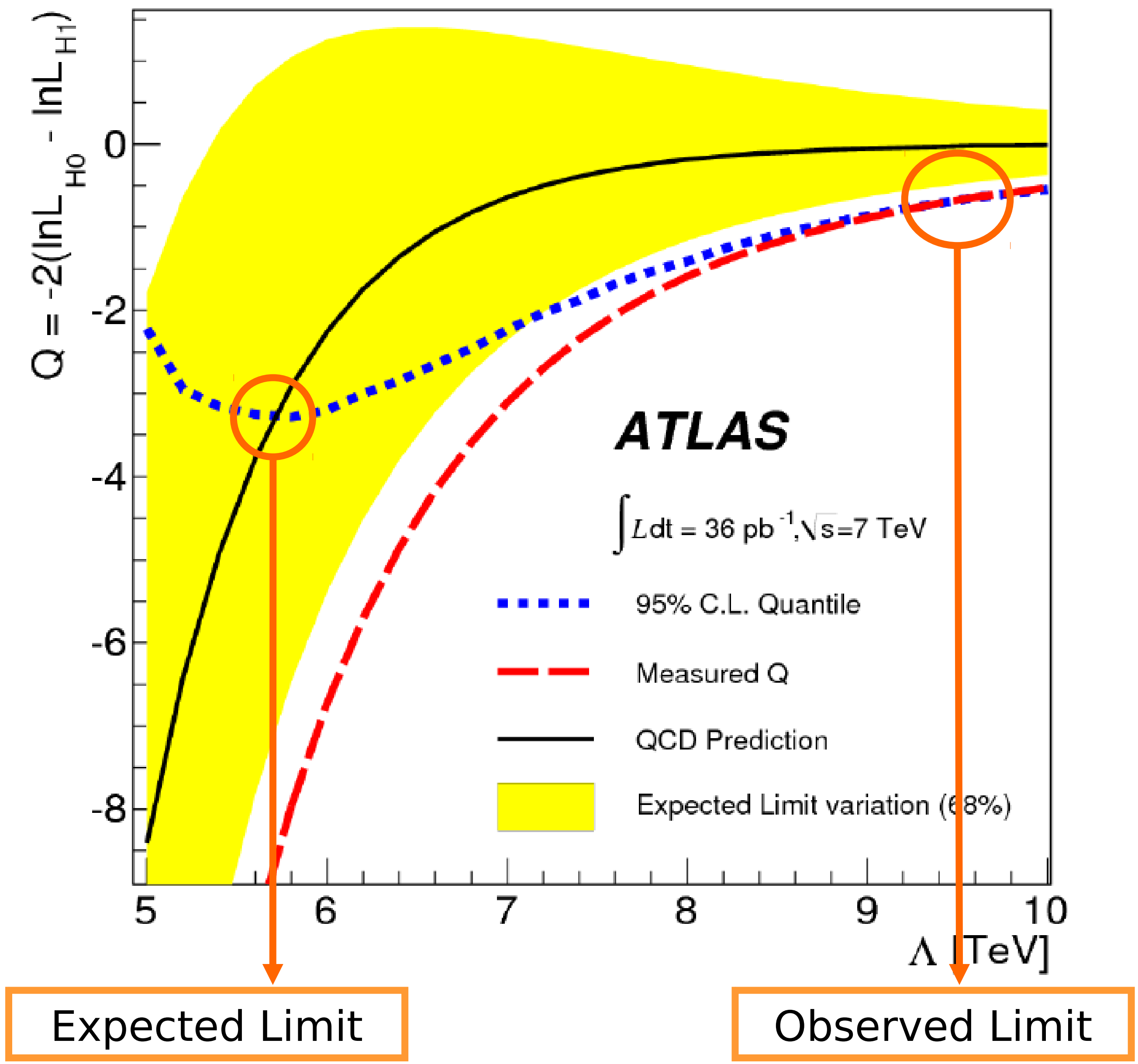
Intuitive conclusion: we have **no sensitivity** on B vs. S+B

But what the method tells us: we can set an **excellent** upper limit!

Symptom of sensitivity problems: **Power** (a.k.a. CL_b) becomes small



A real-life example



Solution: CLs

Since we can identify these cases, try to correct for them to avoid spurious exclusion claims.

CL$_s$

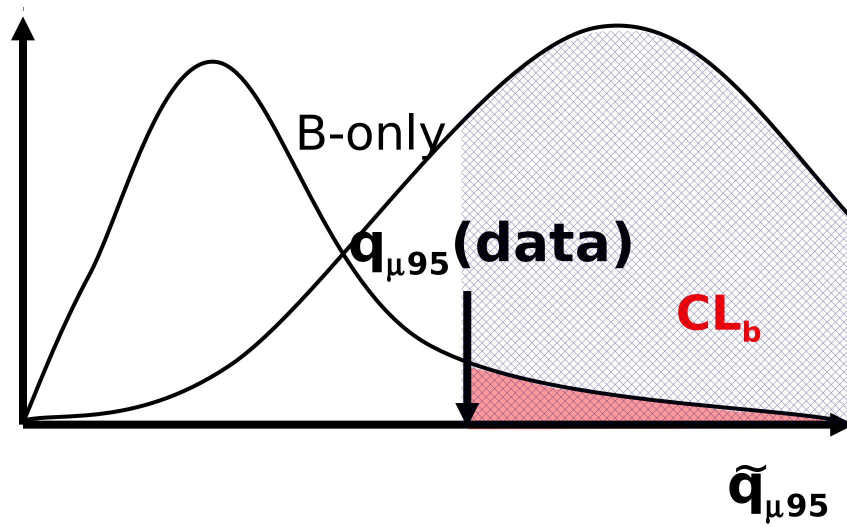
use $\mathbf{CL_s = \underbrace{CL_{s+b}}_{=p_{s+b}} / \underbrace{CL_b}_{=power}}$ to set the limits.

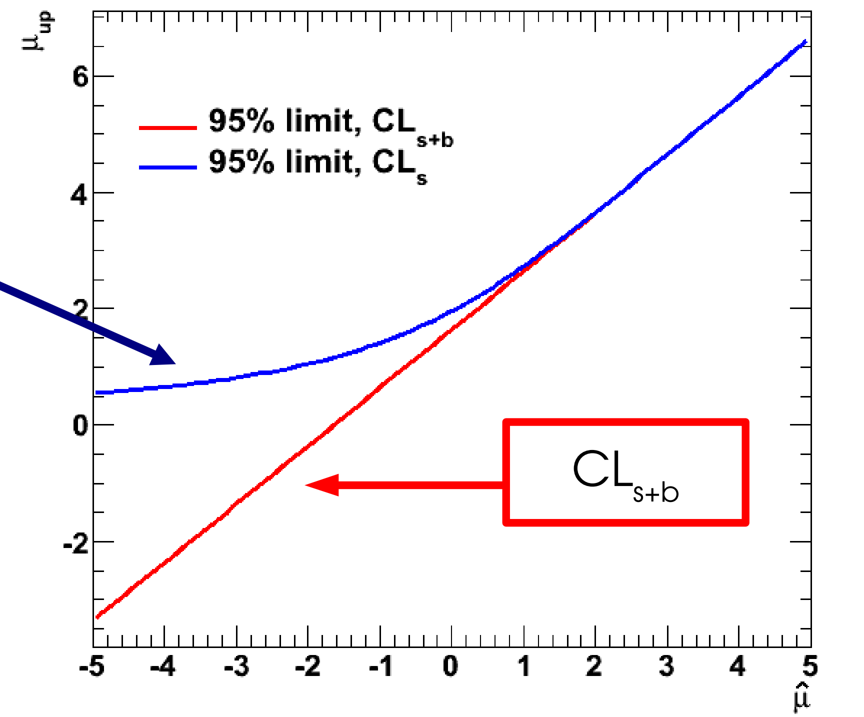
For data compatible with bkg hypo, CL$_b$ ~1 and nothing changes

if CL$_b$ << 1, then CL$_s$>>CL$_{s+b}$ and prevents too-good limits.

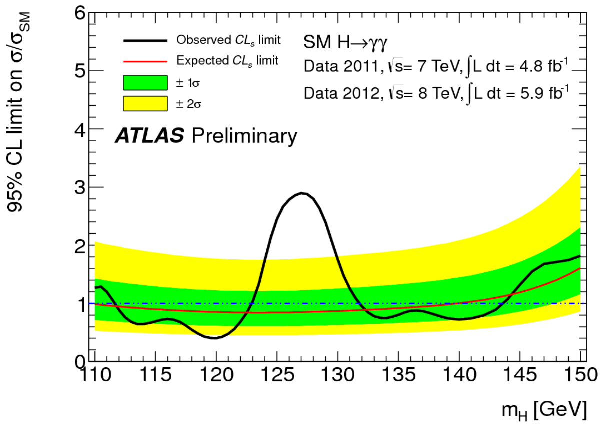
CLs is frowned upon by some statisticians:
Not well-motivated in theory
A side effect is **overcoverage** (e.g. 95% CL is in fact 98%) but can't be avoided.

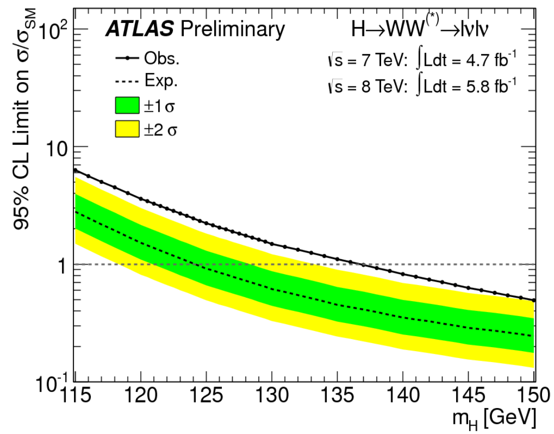
In HEP it is the de facto standard

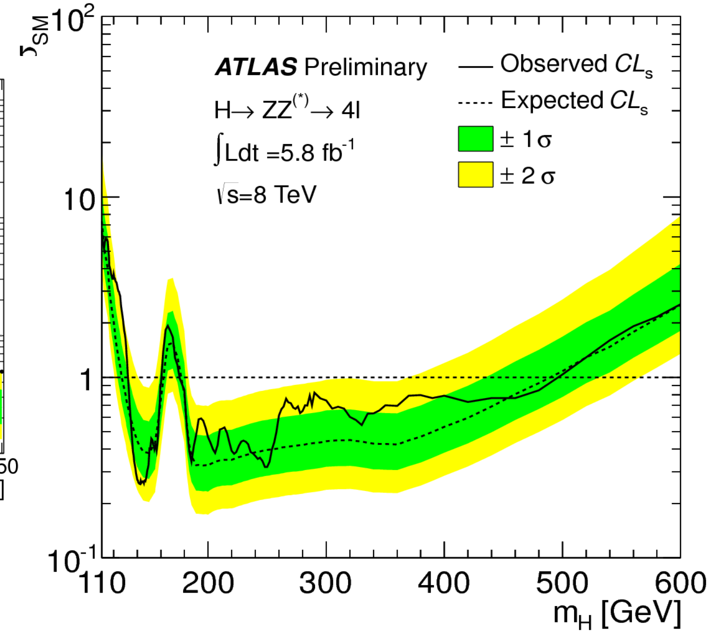




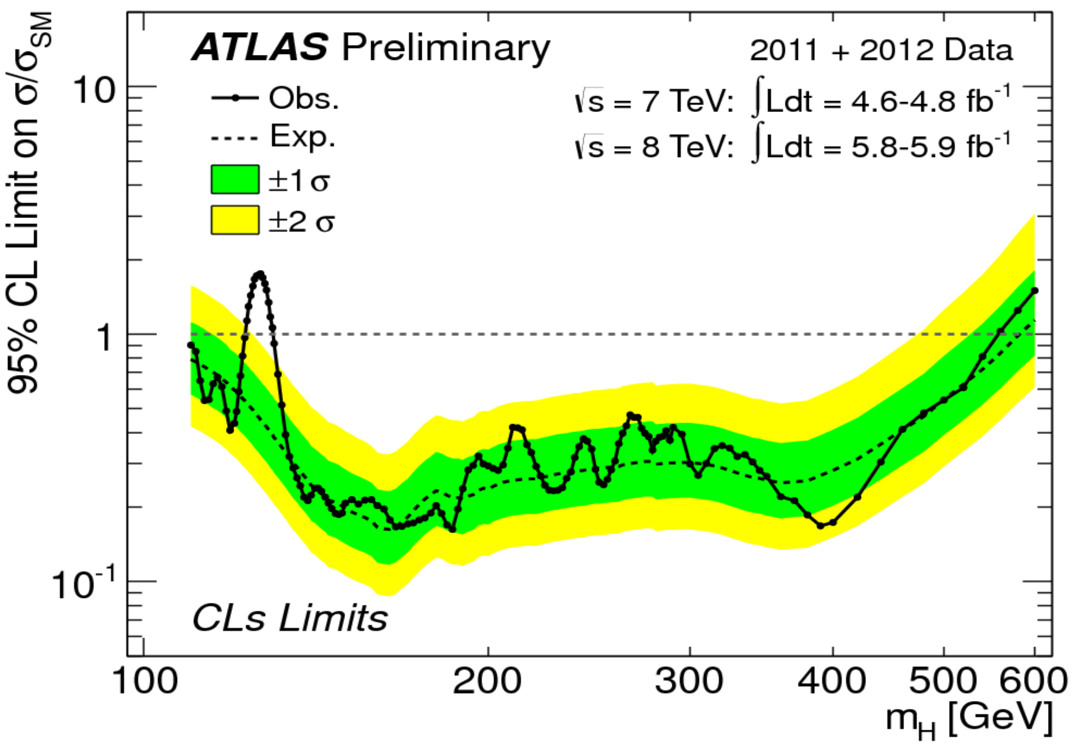
Limit Results







PLB 716 (2012) 1-29



As before, used combined model (78 categories) for the limit combination

Outline

What are the goals ?

Setting up the problem : Maximum likelihood and Likelihood ratios

Discovery

Additional wrinkles (categories, LEE)

Limit setting

Further topics

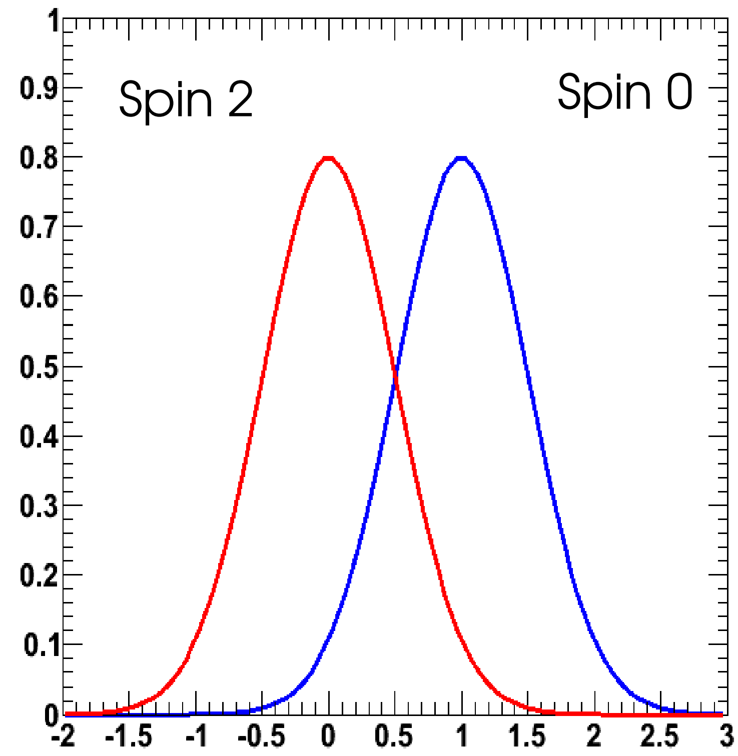
Spin measurement

What is the spin of "the boson" ?
Could be 0, could be 2. Less likely 3+.

Strategy:
Simple hypotheses, so LLR is optimal.
Use e.g.
q = -2 log L(spin 2; data)/L(spin 0; data).

Of course L should now include spin-sensitive information (decay angles, etc.) to have discrimination.

No results yet...

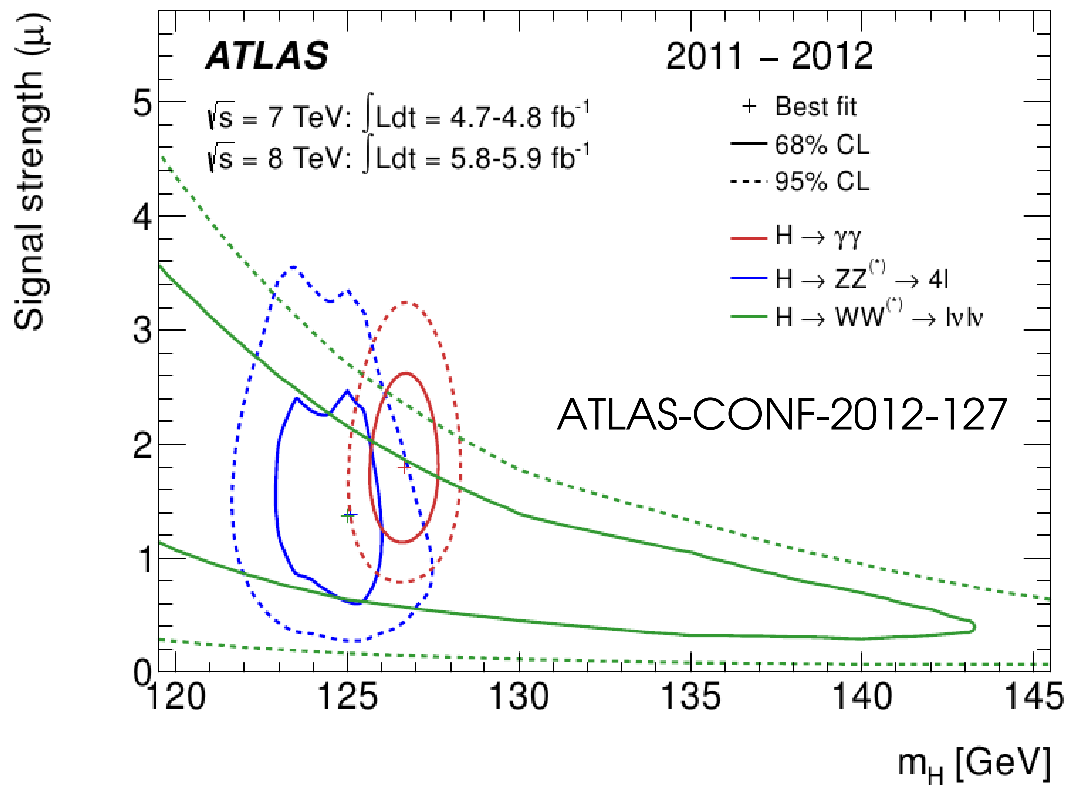


Mass measurement

Can leave mass free when fitting for μ :
→ Define a 2D version of the profile likelihood: $\lambda(\mu, m_H) = -2\log \frac{L(\mu, m_H)}{L(\hat{\mu}, \hat{m}_H)}$
→ Wilks' theorem: l distributed as $\chi^2(n_{dof}=2)$

Scan (μ, m_H) plane, compute λ at each point
→ Max at the best-fit value
→ Make contours of equal λ
→ Use asymptotic formula to convert λ values to CL (68%, 95%)

Could also do a 1D profile in m_H only
→ However error on m_H depends on μ, so a bit sensitive to chosen value of μ



In practice (both at μ=1 and at $\hat{\mu}$)
M_H = 126.0 ± 0.4 (stat) ± 0.4 (syst)

Couplings measurement

Idea: consider separately Higgs production modes: ggH, VBF, WH, ZH, ttH

Different contributions to categories:
→ 2-jet category is enriched in VBF production
→ High-pT categories enriched in VBF, VH
=> Can "solve" for separate productions

H→γγ category breakdown at 8 TeV

category	ggH	VBF	WH+ZH
low-pTt	93%	4%	3%
high-pTt	66%	16%	16%
2-jets	~30%	~70%	

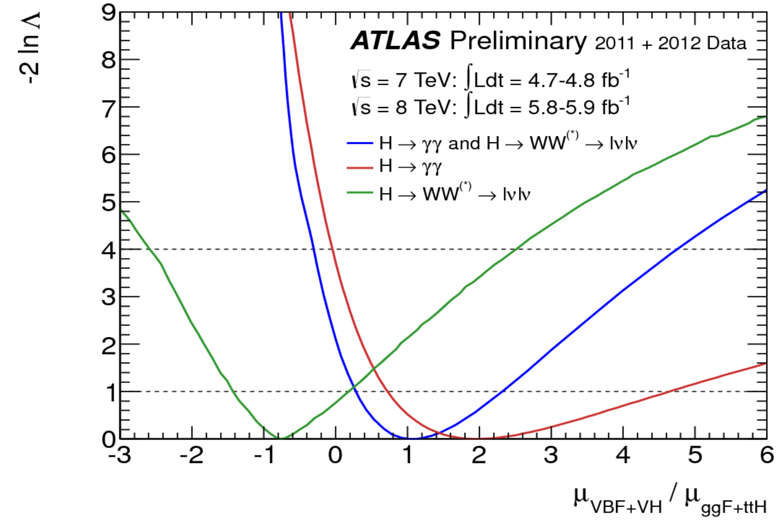
Technically:

Instead of a single μ, allow 2 separate μ:
→ μ_t which scales the numbers of ggH and ttH
→ μ_V which scales VBF, WH and ZH
Define a profile-likelihood statistic to test (μ_t, μ_V) hypotheses
→ By Wilks' theorem, distributed as a $\chi^2(n_{dof}=2)$

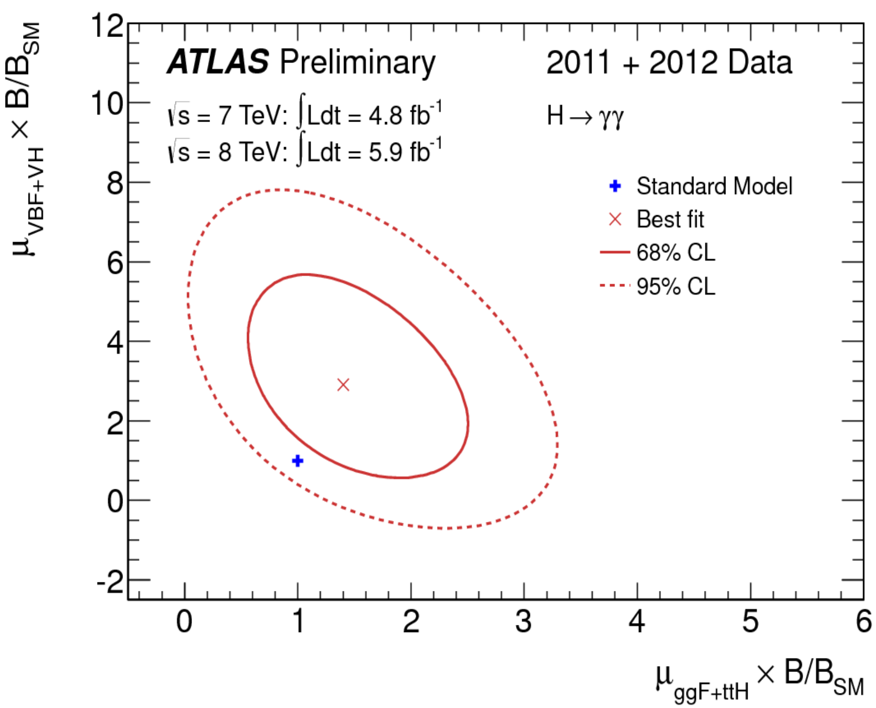
$$\lambda(\mu_t, \mu_V) = -2 \log \frac{L(\mu_t, \mu_V)}{L(\hat{\mu}_t, \hat{\mu}_t)}$$

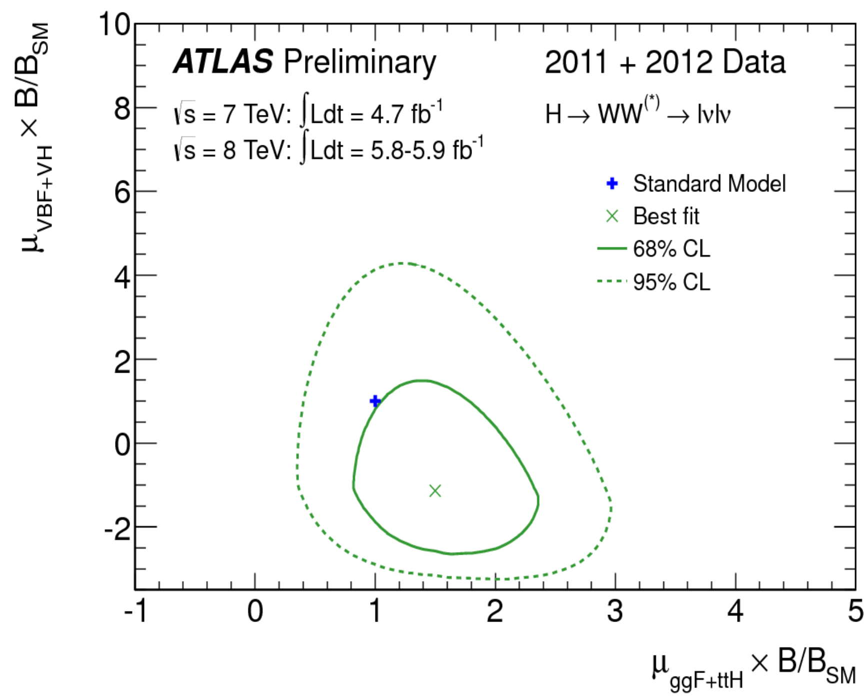
Coupling measurement (2)

Scan (μ_t, μ_V) plane, draw contours of $\lambda(\mu_t, \mu_V)$.
→ Max at best-fit value
→ Use $\chi^2(n_{dof}=2)$ quantiles to translate λ values to CL (68%, 95%)



ATLAS-CONF-2012-127





Coupling measurements (3)

μ not directly linked to couplings, since Couplings also affect H decay rates
Better parametrization: define

→ κ_F : correction to Higgs fermion couplings

→ κ_V : correction to Higgs vector boson couplings

→ SM : $\kappa_F=\kappa_V=1$

ATLAS-CONF-2012-127

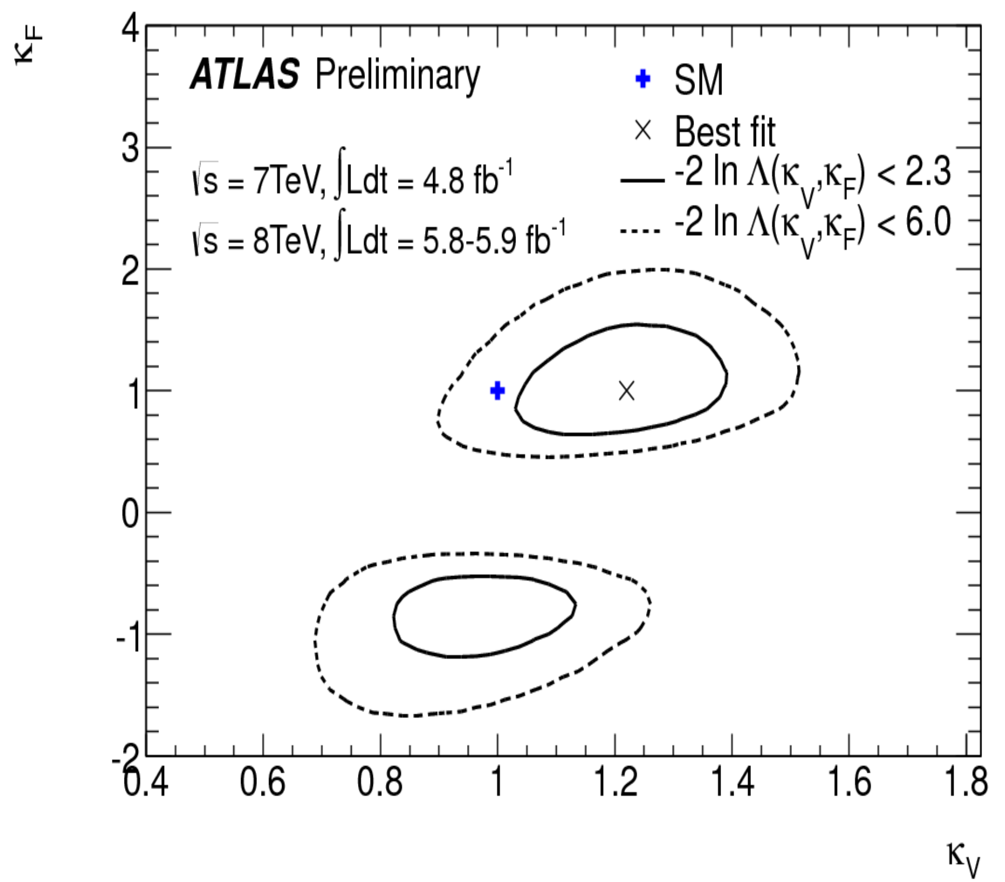
Express μ_t, μ_V as functions of κ_F, κ_V , including both production and decay.Use

$$\lambda(\kappa_F,\kappa_V)=-2\log\frac{L(\kappa_F,\kappa_V)}{L(\hat{\kappa}_F,\hat{\kappa}_V)}$$

Since validity of Wilks' theorem not checked here, show λ values not CL:

$$\int_0^{2.3}\chi^2(\Lambda;ndof=2)\,d\Lambda\approx 0.68$$

$$\int_0^{6.0}\chi^2(\Lambda;ndof=2)\,d\Lambda\approx 0.95$$



Outlook

$\rightarrow$ The last few years have seen significant developments in statistical methods used in HEP

Moving towards:
$\rightarrow$ Standard methods that are well-suited to many HEP situations.
$\rightarrow$ Standard tools, e.g. RooFit, RooStats, distributed with ROOT.

Hopefully to be used for many discoveries to come!

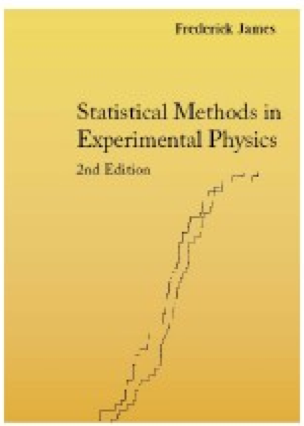
Further reading

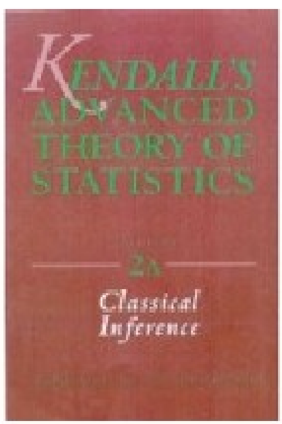
F. James, Statistical Methods in Experimental Physics, 2nd ed., World Scientific, 2006;

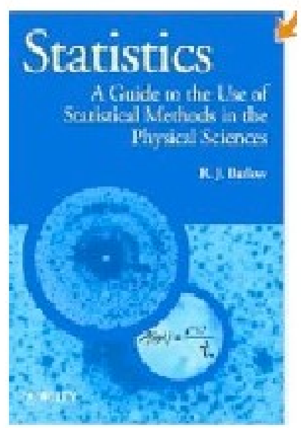
G. Cowan, Statistical Data Analysis, Clarendon Press, Oxford, 1998.

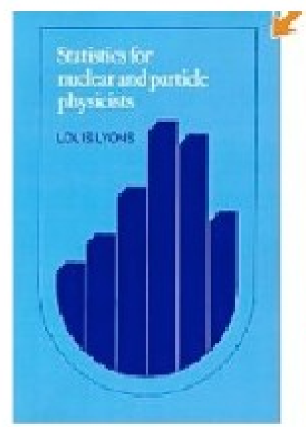
R.J.Barlow, A Guide to the Use of Statistical Methods in the Physical Sciences, John Wiley, 1989;

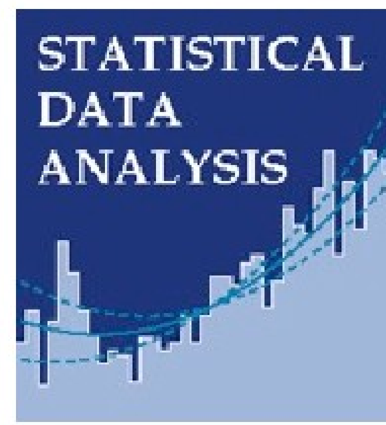
L. Lyons, Statistics for Nuclear and Particle Physics, CUP, 1986.











See also this lecture series by G. Cowan:
https://indico.cern.ch/conferenceDisplay.py?confId=173726