Introduction to particle physics: experimental part

❑ Computing statistical results

- Estimating the value of a parameter
- Testing hypotheses
- Discovery
- Limits
- Confidence intervals

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

Prof. dr hab. Elżbieta Richter-Wąs

How to represent the data

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

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

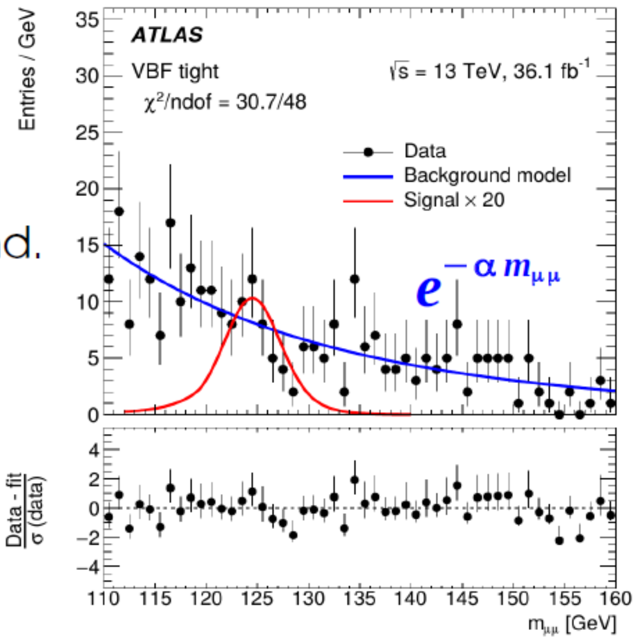
Model can include multiple **categories**, each with a separate description

Model parameters

Model typically includes:

- **Parameters of interest** (POIs) : what we want to measure
  → S, σ, m_W, …

- **Nuisance parameters** (NPs) : other parameters needed to define the model
  → B
  → For binned data, f^{sig}_i , f^{bkg}_i
  → For unbinned data, parameters needed to define P_{bkg}
  e.g. exponential slope α of H→μμ background.



Phys. Rev. Lett. 119 (2017) 051802 **4**

NPs must be either

→ **given a value "by hand"** (possibly within systematics) or

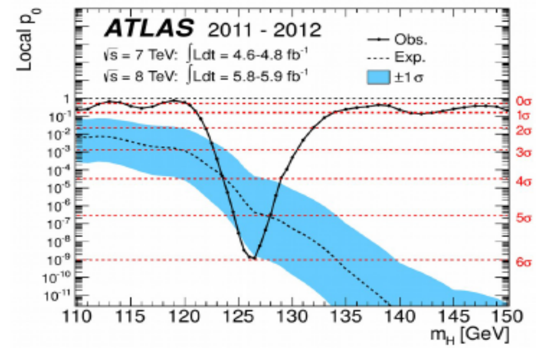
→ **constrained by the data** (e.g. in sidebands)

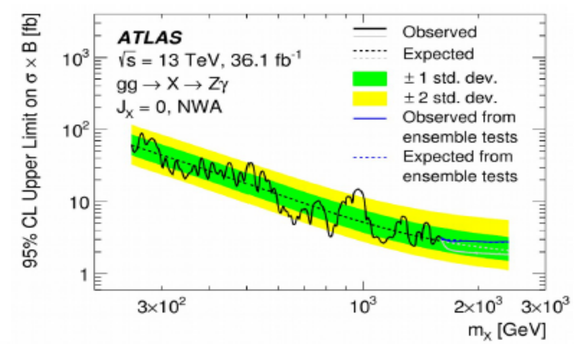
Statistical computations

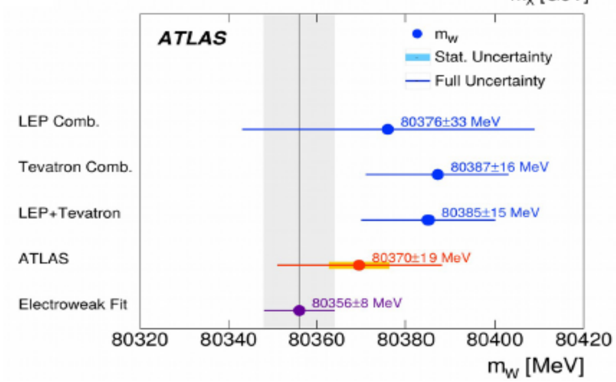
Now that we have a model, can use it to compute analysis results:

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

→ The Statistical Model already contains all the needed information – how to use it ?







Using the PDF

Model describes the distribution of the observable: **P(data; parameters)**

⇒ Possible outcomes of the experiment, for given parameter values

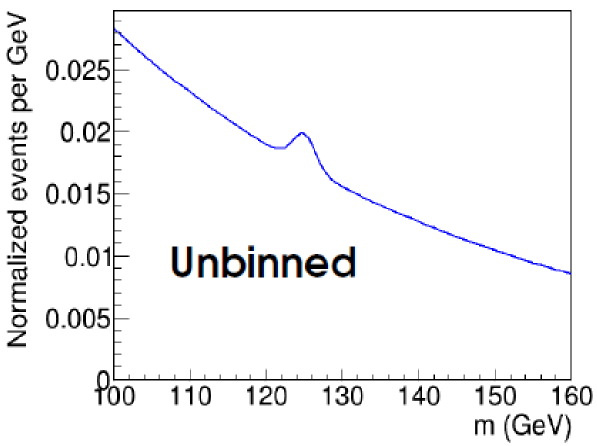
Can draw random events according to PDF : **generate** ***pseudo-data***

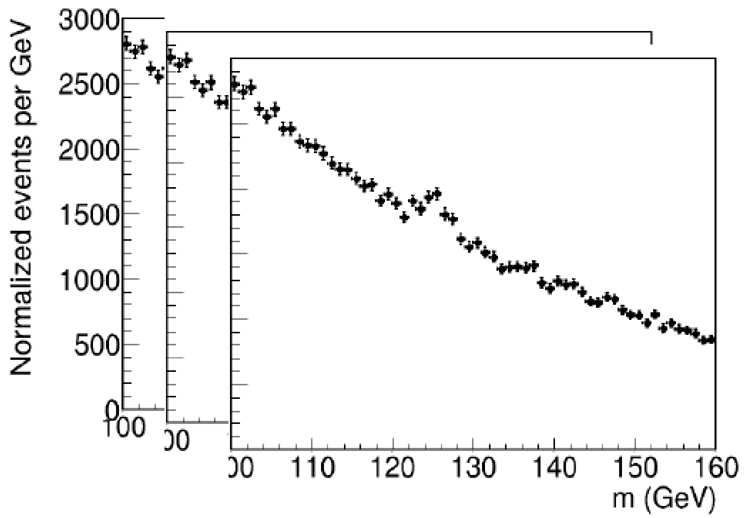
$P(\lambda=5)$

2, 5, 3, 7, 4, 9, ….

Each entry = separate "experiment"

Generate





Likelihood

Model describes the distribution of the observable: **P(n; λ), P(data; parameters)**

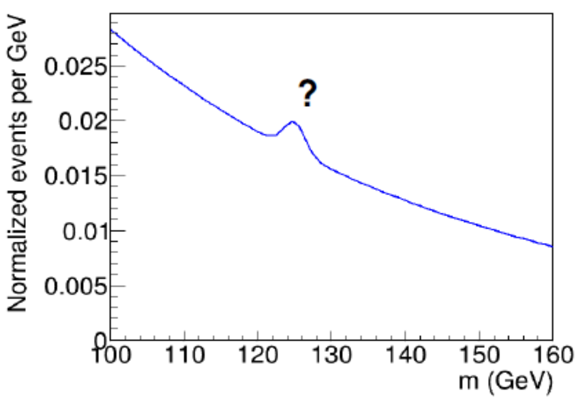
⇒ Possible outcomes of the experiment, for given parameter values

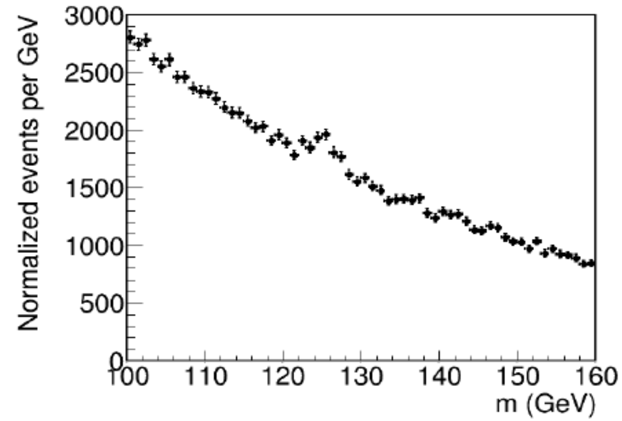
We want the **other** direction: **use data to get information on parameters**

$P(\lambda = ?)$

2

Estimate





Likelihood: **L(parameters) = P(data;parameters)**

→ **same as the PDF,** but seen as function of the parameters

9

Poisson example

Assume **Poisson distribution** with B = 0 :

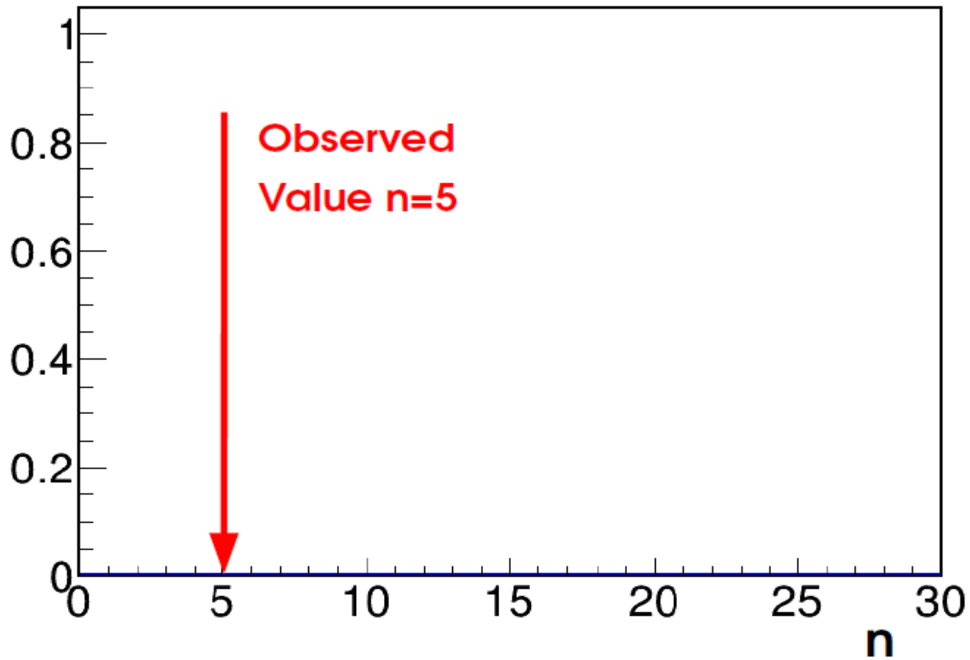
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

Say we **observe n=5**, want to infer information on the parameter **S**

→ Try different values of S for a fixed data value n=5

→ Varying parameter, fixed data: **likelihood**

$$L(S;n=5) = e^{-S}\frac{S^5}{5!}$$



Poisson example

Assume **Poisson distribution** with B = 0 :

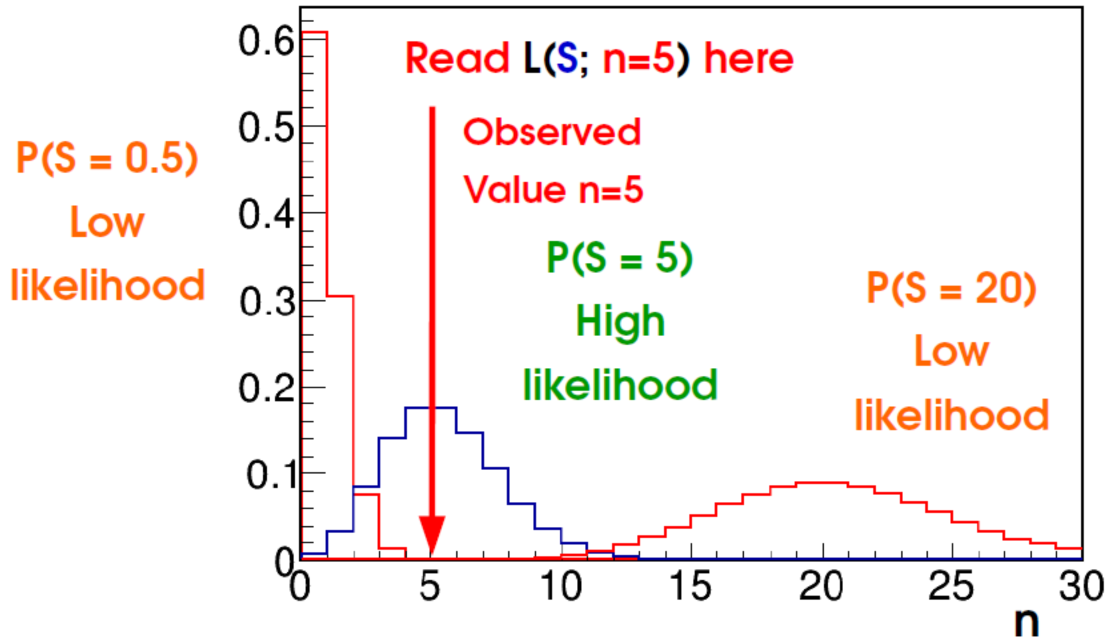
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

Say we **observe** **n=5**, want to infer information on the parameter **S**

→ Try different values of S for a fixed data value n=5

→ Varying parameter, fixed data: **likelihood**

$$L(S;n=5) = e^{-S}\frac{S^5}{5!}$$



Poisson example

Assume **Poisson distribution** with B = 0 :

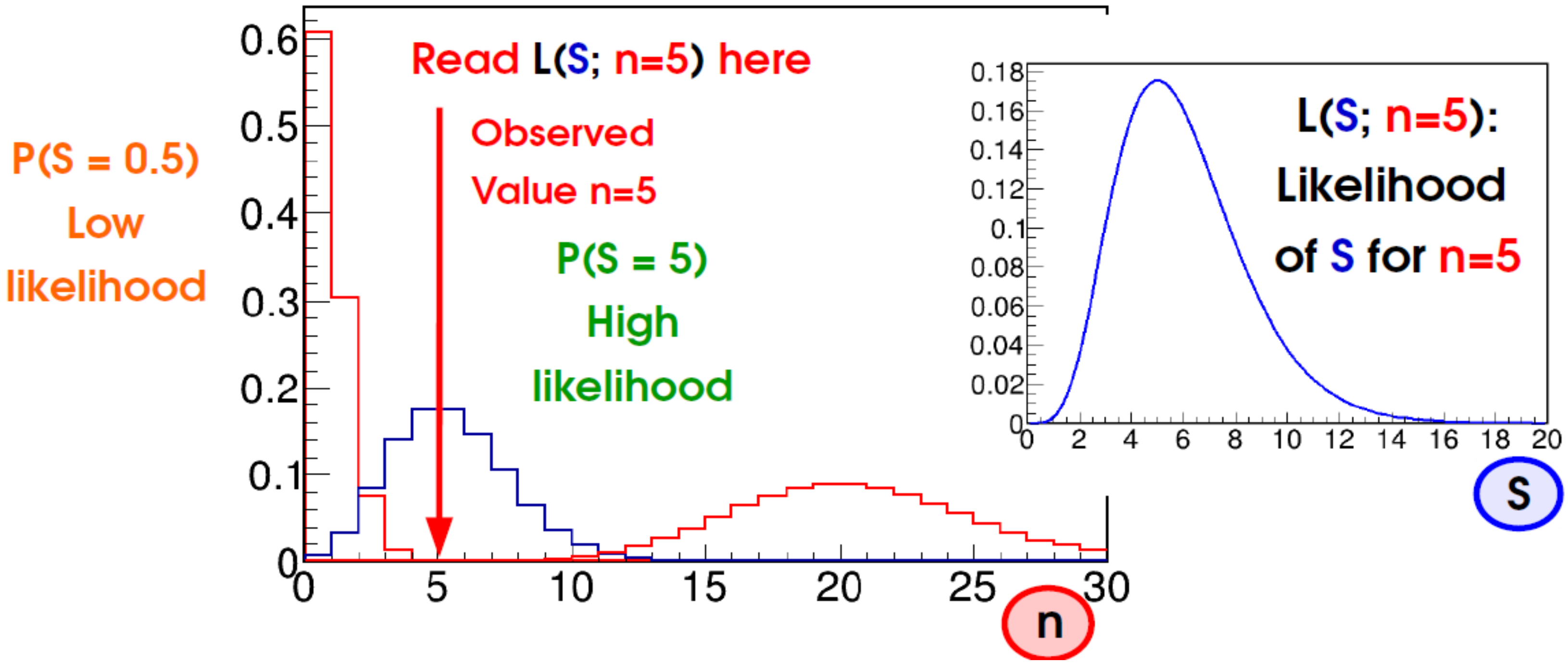
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

Say we **observe** n=5, want to infer information on the parameter S

→ Try different values of S for a fixed data value n=5

→ Varying parameter, fixed data: **likelihood**

$$L(S;n=5) = e^{-S}\frac{S^5}{5!}$$

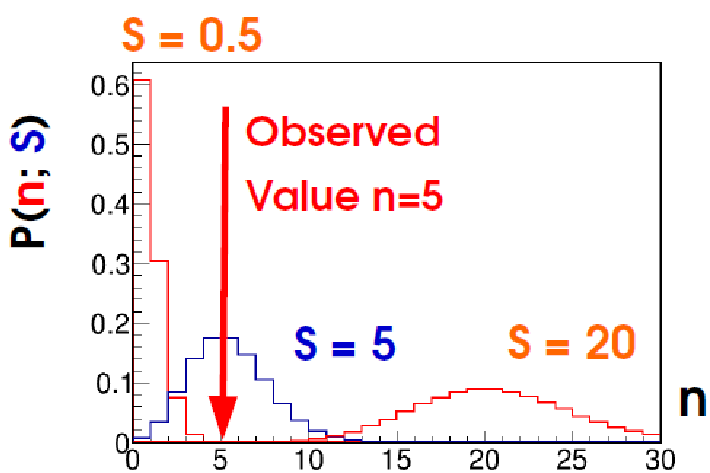


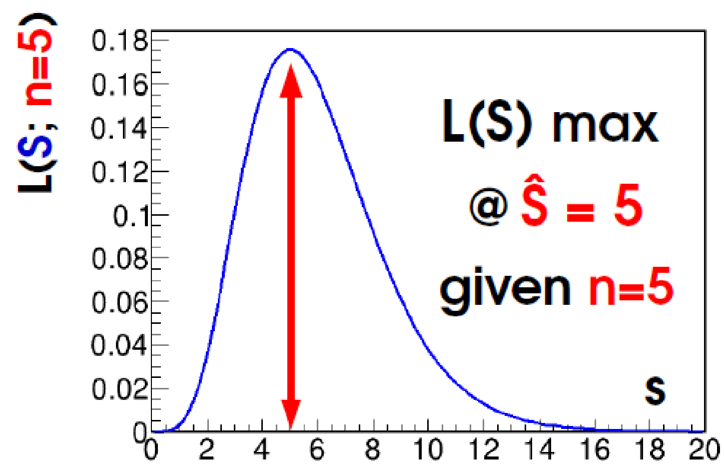
Maximum Likelihood Estimation

To estimate a parameter μ, find the **value $\hat{\mu}$ that maximizes** L(μ)

Maximum Likelihood Estimator (MLE) $\hat{\mu}$:

$$\hat{\mu} = \arg\max L(\mu)$$



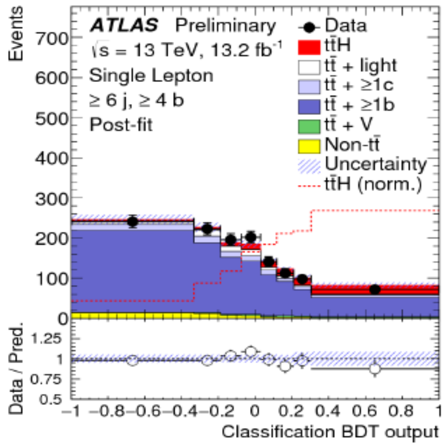


MLE: the value of μ for which **this data** was ***most likely to occur***

The MLE is a function of the data – itself an **observable**

No guarantee it is the true value (data may be "unlikely") but sensible estimate

MLEs in shape analyses



Binned shape analysis:

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

Maximize global L(S) (each bin may prefer a different **S**)

In practice easier to minimize

$$\lambda_{\mathrm{Pois}}(S) = -2 \log L(S) = -2 \sum_{i=1}^{N} \log \mathrm{Pois}(n_i; S f_i + B_i)$$

Needs a computer...

In the Gaussian limit

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

χ^2 formula!

→ **Gaussian MLE** (min χ^2 or min λ_{Gaus}) : ***Best fit value*** in a χ^2 (Least-squares) fit

→ **Poisson MLE** (min λ_{Pois}) : ***Best fit value*** in a *likelihood* fit (in `ROOT`, fit option "`L`")

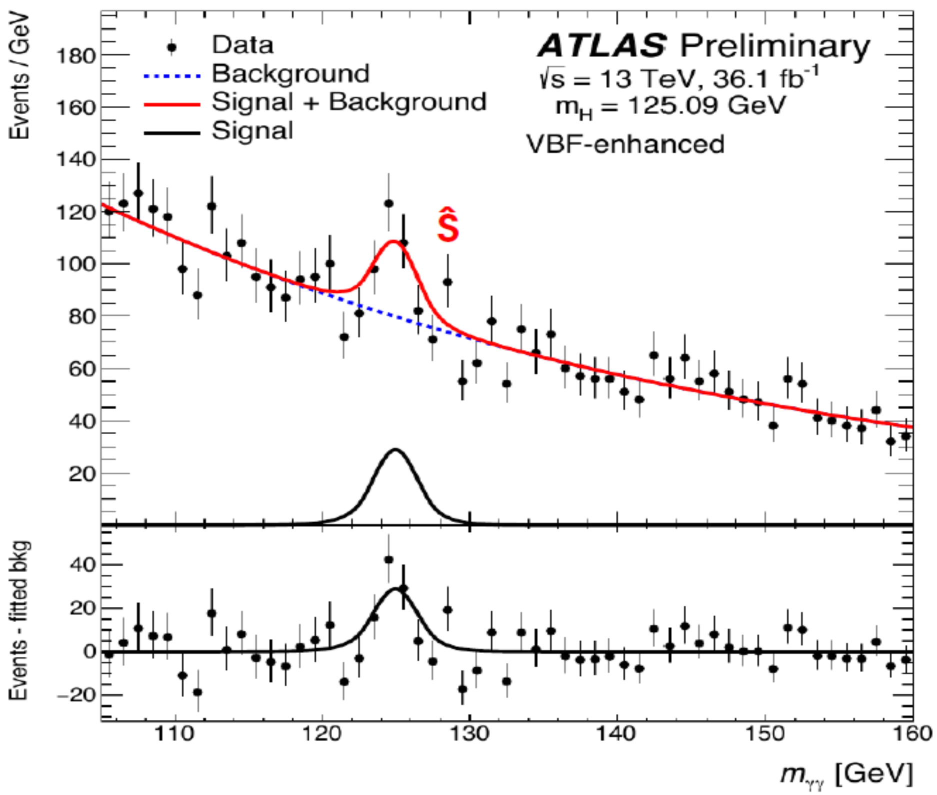
In RooFit, $\boldsymbol{\lambda}_{\mathrm{Pois}}$ ⇒ `RooAbsPdf::fitTo()`, $\boldsymbol{\lambda}_{\mathrm{Gaus}}$ ⇒ `RooAbsPdf::chi2FitTo()`.

In both cases, MLE ⇔ *Best Fit*

MLEs in shape analyses

H→γγ

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



Estimate the MLE $\hat{S}$ *of* S *?*

→ Perform (likelihood) best-fit of model to data

⇒ fit result for S is the desired $\hat{S}$.

In particle physics, often use the *MINUIT* minimizer within ROOT.

ATLAS-CONF-2017-045

MLE Properties

Asymptotically Gaussian and unbiased :

$$P(\hat{\mu}) \propto \exp\left(-\frac{(\hat{\mu}-\mu^*)^2}{2\sigma_{\hat{\mu}}^2}\right) \quad \text{for } n \rightarrow \infty$$

Standard deviation of the distribution of $\hat{\mu}$

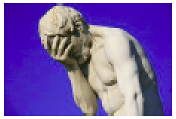
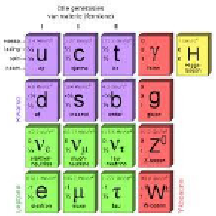
for large enough datasets

- **Asymptotically Efficient** : σ_{μ} is the **lowest possible value** (in the limit $n \rightarrow \infty$) among consistent estimators.
  → MLE captures all the available information in the data
- Also **consistent**: $\hat{\mu}$ converges to the true value for large n, $\hat{\mu} \xrightarrow{n\rightarrow\infty} \mu^*$
- **Log-likelihood : ** Can also **minimize** $\lambda = -2 \log L$
  → Usually more efficient numerically
  → For Gaussian L, λ is parabolic: $$\lambda(\mu) = \left(\frac{\hat{\mu}-\mu}{\sigma_{\mu}}\right)^2$$
- Can **drop multiplicative constants in L** (additive constants in λ)

Hypothesis Testing

Hypothesis: assumption on model parameters, say value of S (e.g. **H_0 : S=0**)

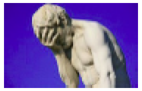
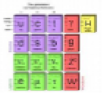
→ **Goal** : decide if H_0 is favored or disfavored using a test based on the data

Possible outcomes:	**Data disfavors H_0 (Discovery claim)**	**Data favors H_0 (Nothing found)**
H_0 is false (New physics!)	Discovery!  	Missed discovery Type-II error (1 - Power) 
H_0 is true (Nothing new)	False discovery claim Type-I error (→ p-value, significance) 	No new physics, none found 

"... the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis." – R. A. Fisher

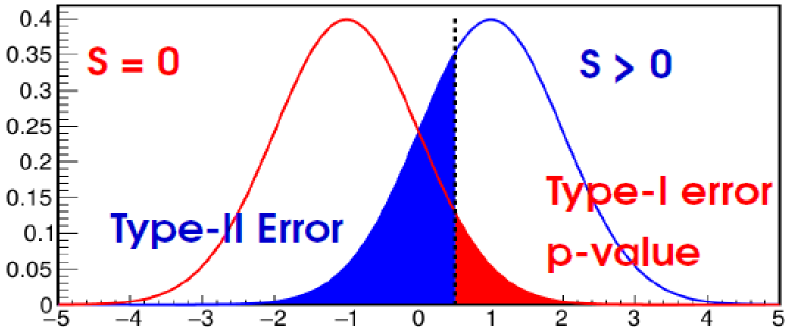
Hypothesis Testing

Hypothesis: assumption on model parameters, say value of S (e.g. $\mathbf{H_0}$ **: S=0**)

	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	**Discovery!** 	Type-II error (Missed discovery) 
H_0 is true (Nothing new)	Type-I error (False discovery) 	No new physics, none found 

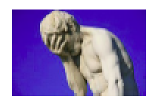
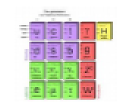
Lower Type-I errors ⇔ **Higher Type-II errors** and vice versa: cannot have everything!

→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.



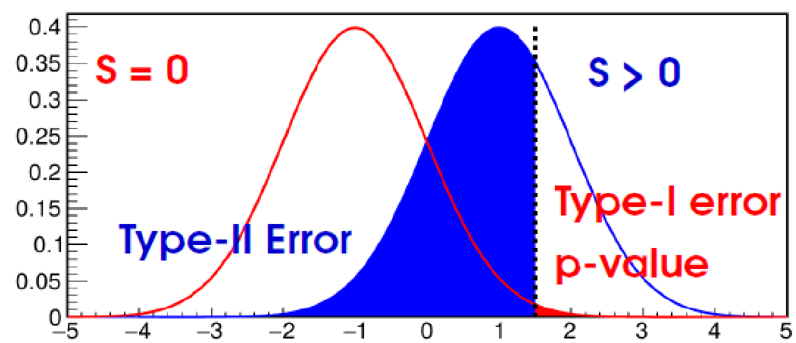
Hypothesis Testing

Hypothesis: assumption on model parameters, say value of S (e.g. $\mathbf{H_0 : S=0}$)

	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	Discovery! 	Type-II error (Missed discovery) 
H_0 is true (Nothing new)	Type-I error (False discovery) 	No new physics, none found 

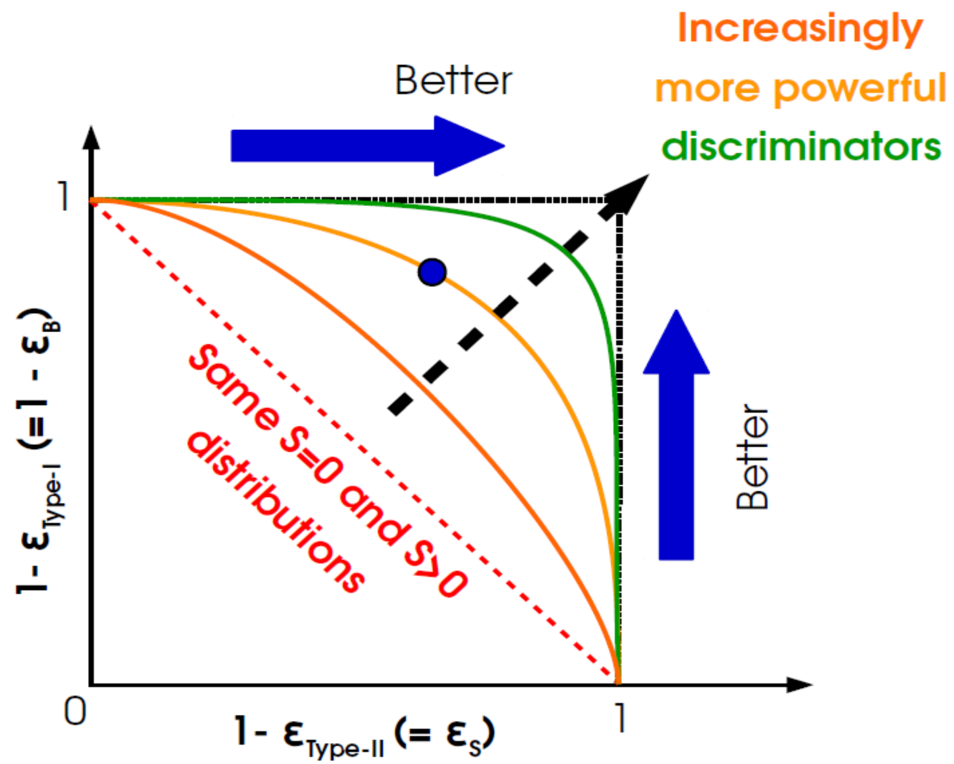
Lower Type-I errors ⇔ Higher Type-II errors and vice versa: cannot have everything!

→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.



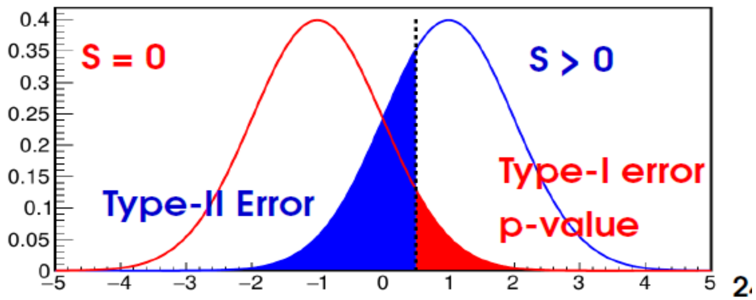
ROC Curves

"Receiver operating characteristic" (ROC) Curve:
→ Plot Type-I vs Type-II rates for different cut values
→ All curves monotonically decrease from (0,1) to (1,0)
→ Better discriminators more bent towards (1,1)



→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.

→ Usually set predefined level of **acceptable Type-I error** (e.g. "5σ")



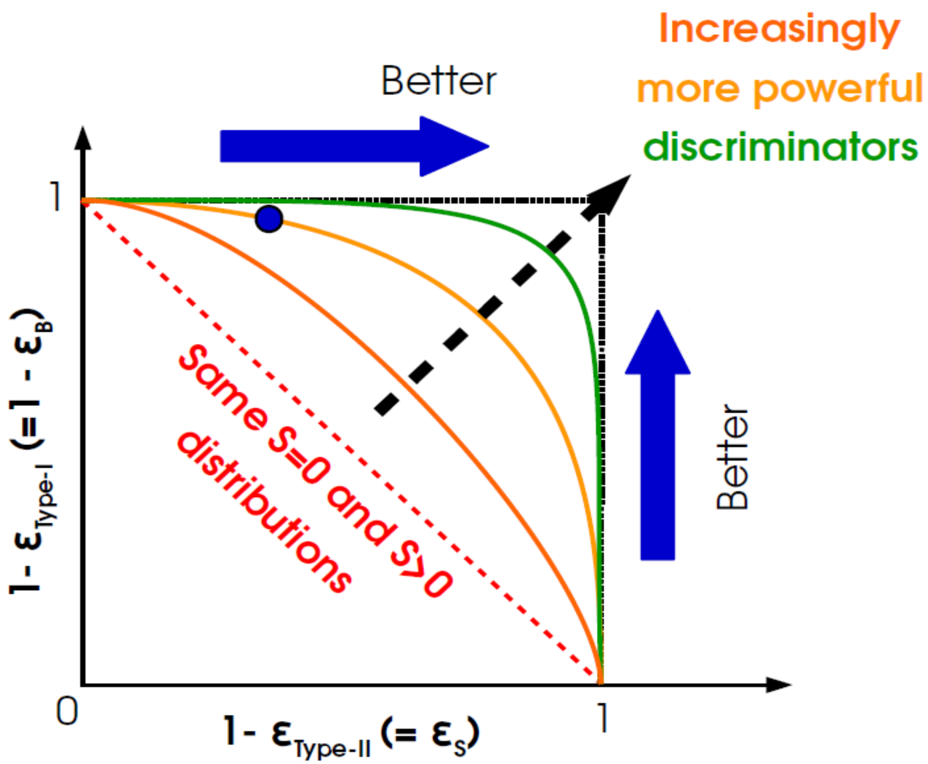
ROC Curves

"Receiver operating characteristic" (ROC) Curve:

→ Plot Type-I vs Type-II rates for different cut values

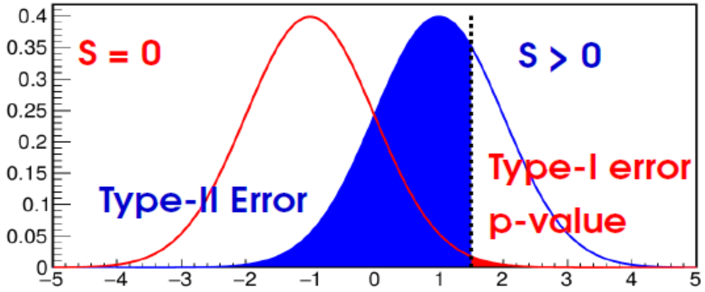
→ All curves monotonically decrease from (0,1) to (1,0)

→ Better discriminators more bent towards (1,1)



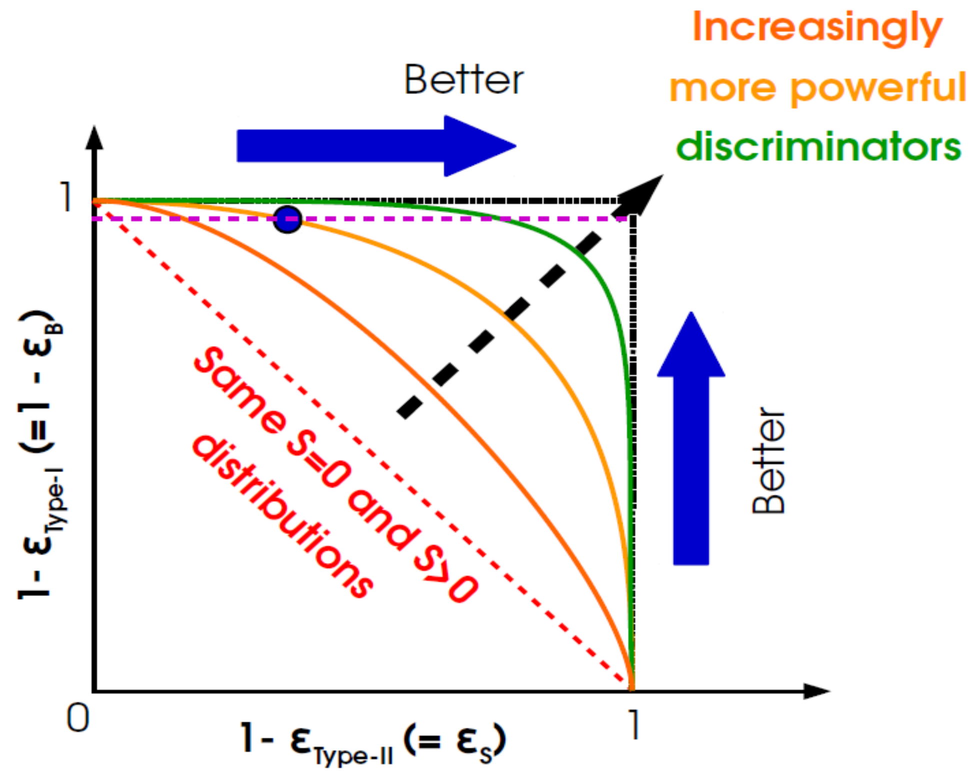
→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.

→ Usually set predefined level of **acceptable Type-I error** (e.g. "5σ")



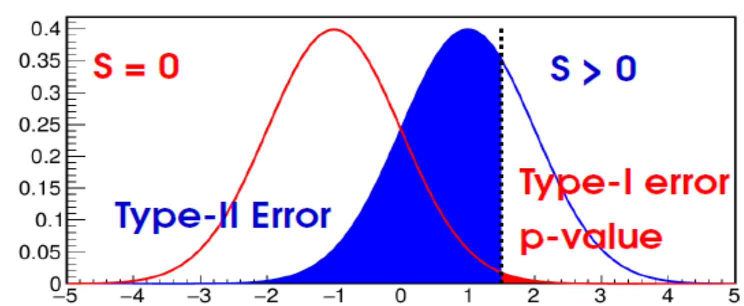
ROC Curves

"Receiver operating characteristic" (ROC) Curve:
→ Plot Type-I vs Type-II rates for different cut values
→ All curves monotonically decrease from (0,1) to (1,0)
→ Better discriminators more bent towards (1,1)



→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.

→ Usually set predefined level of **acceptable Type-I error** (e.g. "5σ")



Hypothesis testing with Likelihoods

Neyman-Pearson Lemma

When comparing two hypotheses H_0 *and* H_1*, the optimal discriminator is the* ***Likelihood ratio*** *(LR)*

$$\frac{L(H_1; data)}{L(H_0; data)}$$

e.g. $$\frac{L(S = 5; data)}{L(S = 0; data)}$$

As for MLE, choose the hypothesis that is more likely **given the data we have**.

→ **Minimizes Type-II uncertainties** for given level of Type-I uncertainties

→ Always need an **alternate hypothesis** to test against.

Caveat: Strictly true only for *simple hypotheses* (no free parameters)

→ **In the following**: all tests based on LR, will focus on p-values (Type-I errors), trusting that Type-II errors are anyway as small as they can be...

Discovery: Test Statistic

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

Discovery :

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

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

⇒ Use LR $$\frac{L(S=0)}{L(\hat{S})}$$

→ In fact use the **test statistic**

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

→ Set $q_0=0$ for $\hat{S} < 0$, same as for $\hat{S} = 0$: negative signal is same as no signal

→ *one-sided* test statistic

Discovery: p-value

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

data prefer S = 0 ⟷ data prefer S > 0

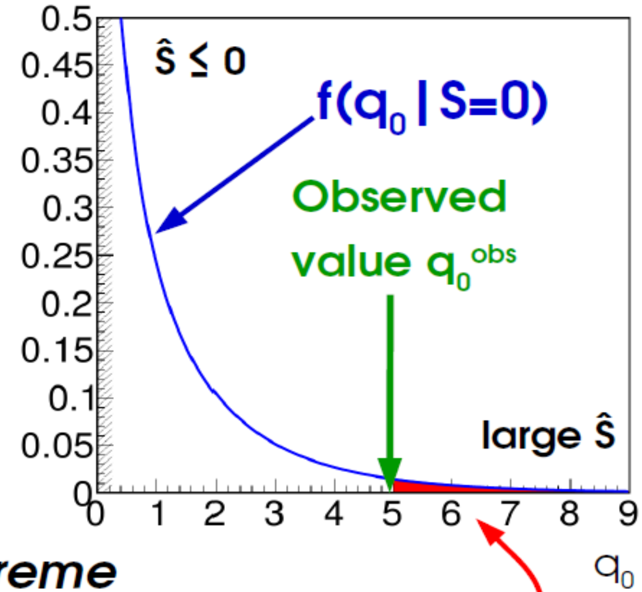
⇒ observed $\hat{S}$ is far from 0

⇒ $H_0(S=0)$ ***disfavored*** compared to $H_1(S\neq0)$.

How large q_0 before we can exclude H_0 ?
(and **claim a discovery!**)

→ Need small Type-I rate (falsely accepting H_0)

→ Type-I rate also known as the ***p-value*** p_0 **:**



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

→ Compute from the distribution $f(q_0 | S=0)$: $p_0 = \int_{q_0^{obs}}^{\infty} f(q_0|S=0)\, dq_0$

→ Smaller p-value ⇒ Stronger case for discovery

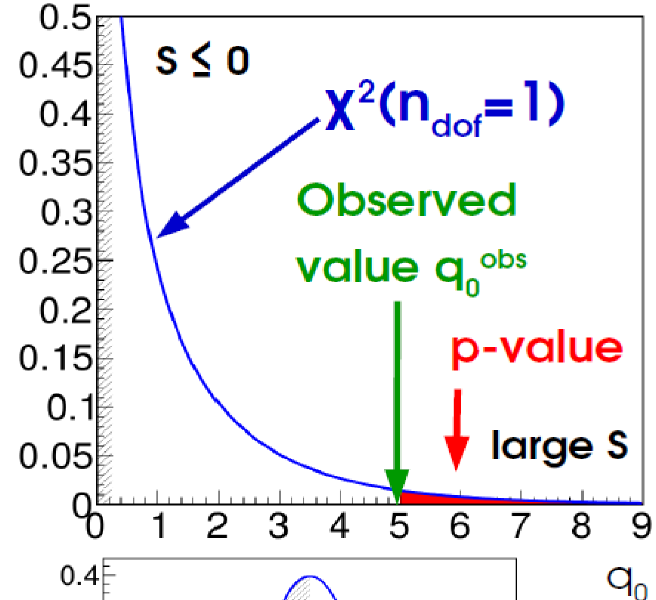
Asymptotic distribution of q_0

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

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

⇒ **q_0 is distributed as a χ^2** under $H_0(S=0)$, for $\hat{S} \geq 0$: ***Wilk's Theorem*** (*)

$$f\left(q_0 \mid H_0, \hat{S} \geq 0\right) = f_{\chi^2(n_{dof}=1)}\left(q_0\right)$$



⇒ Can compute p-values from Gaussian quantiles

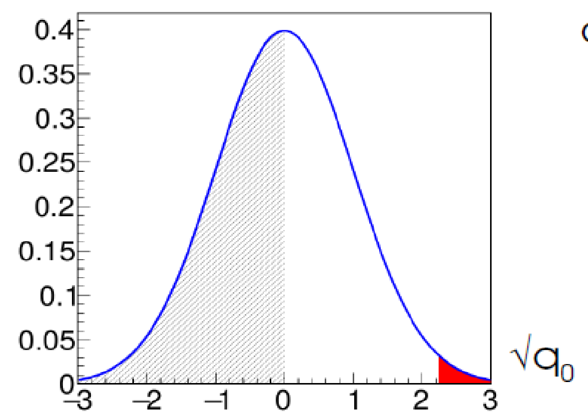
$$p_0 = 1 - \Phi\left(\sqrt{q_0}\right)$$

By definition, $q_0 \sim \chi^2 \Rightarrow \sqrt{q_0} \sim G(0,1)$

⇒ Even more simply, the significance is:

$$Z = \sqrt{q_0}$$

Typically works well already for for event counts of O(5) and above ⇒ Widely applicable



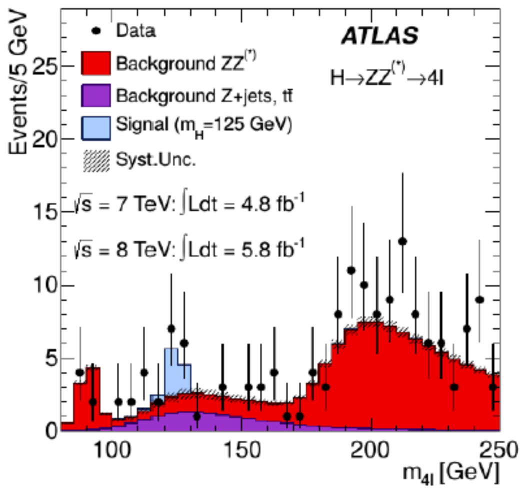
(*) 1-line "proof" : asymptotically L and S are Gaussian, so

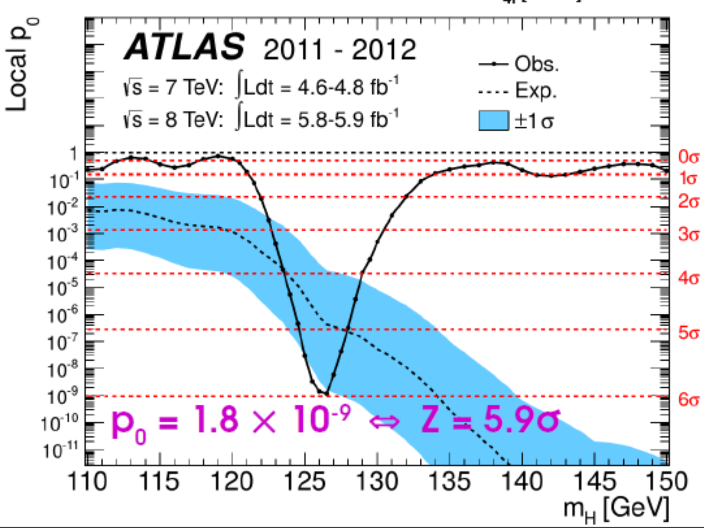
$$L(S) = \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}}{\sigma}\right)^2\right] \Rightarrow q_0 = \left(\frac{\hat{S}}{\sigma}\right)^2 \Rightarrow \sqrt{q_0} = \frac{\hat{S}}{\sigma} \sim G(0,1) \Rightarrow q_0 \sim \chi^2(n_{dof}=1)$$

Some exmples

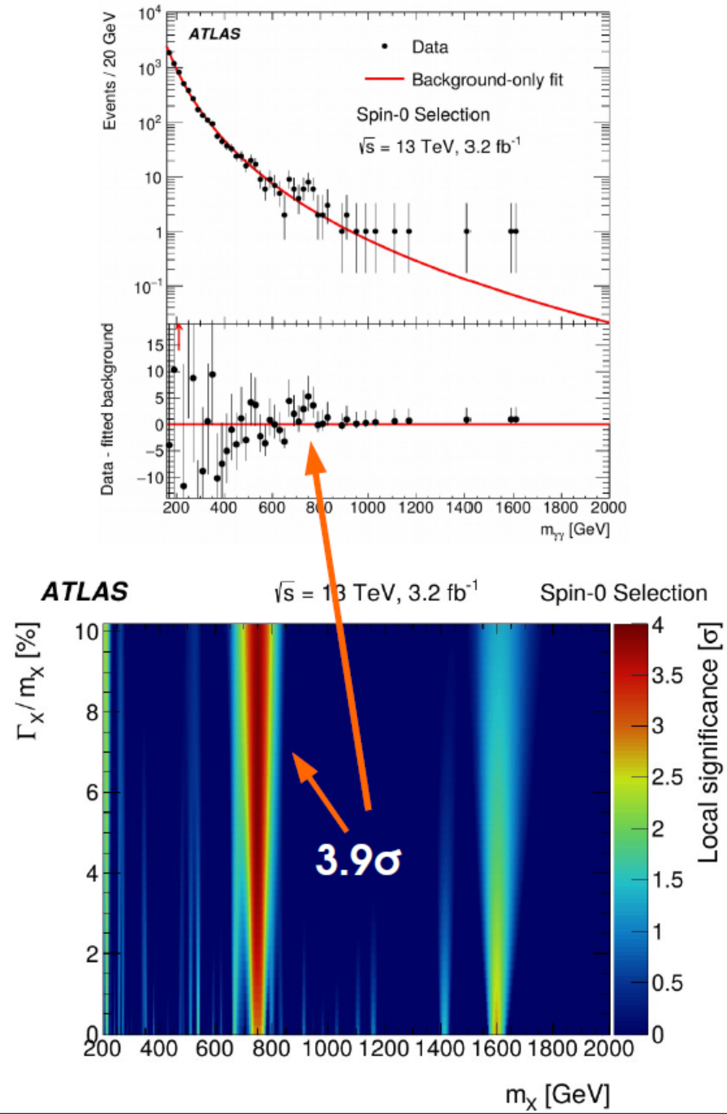
Some Examples

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 the value $\hat{\mu}$ that maximizes L(μ) → best-fit value

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

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

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

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

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

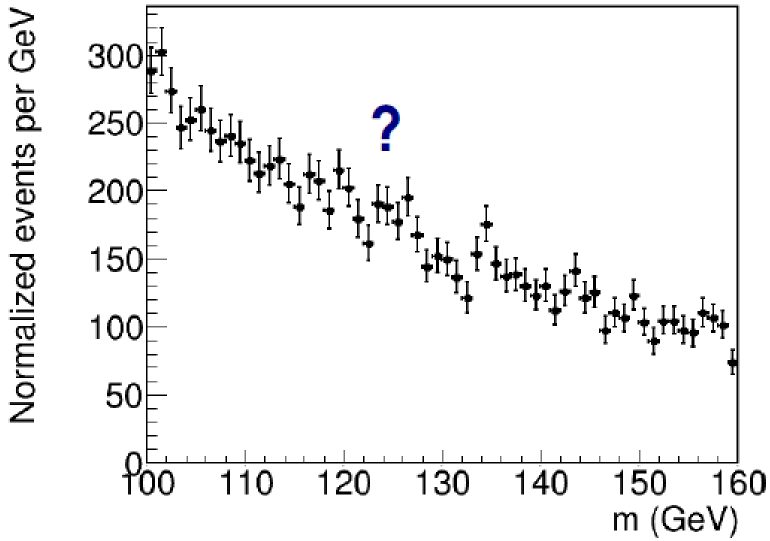
Hypothesis tests for Limits

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**

→ Typically report **95% CL** upper limit (p-value = 5%) : "$S < S_0$ @ 95% CL"



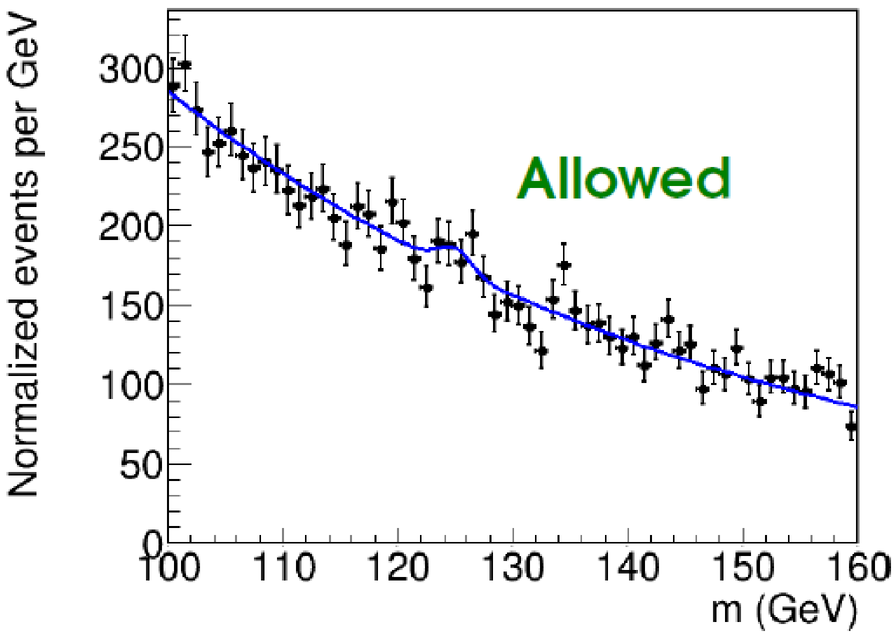
Hypothesis tests for Limits

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**

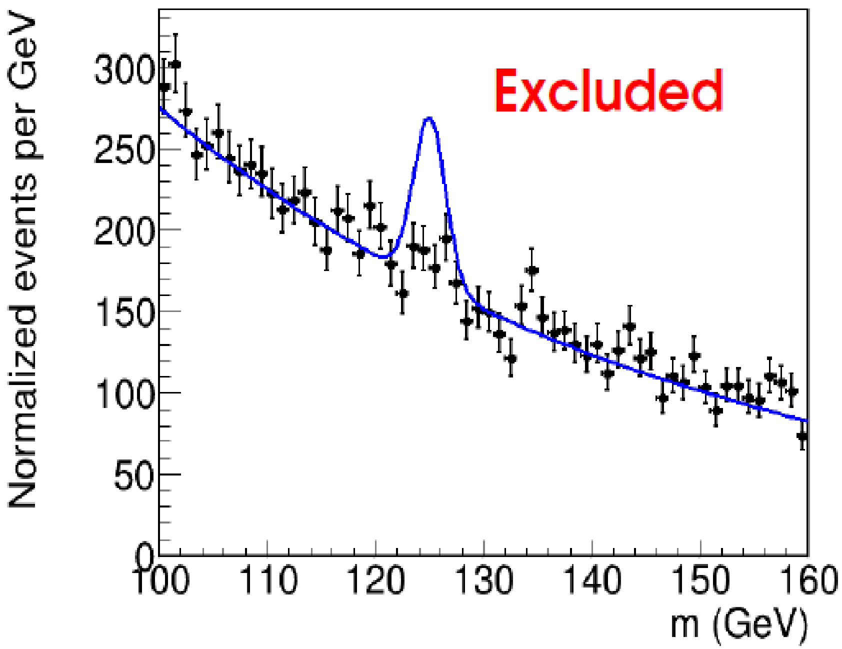
→ Typically report **95% CL** upper limit (p-value = 5%) : "$S < S_0$ @ 95% CL"



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**

→ Typically report **95% CL** upper limit (p-value = 5%) : "$S < S_0$ @ 95% CL"



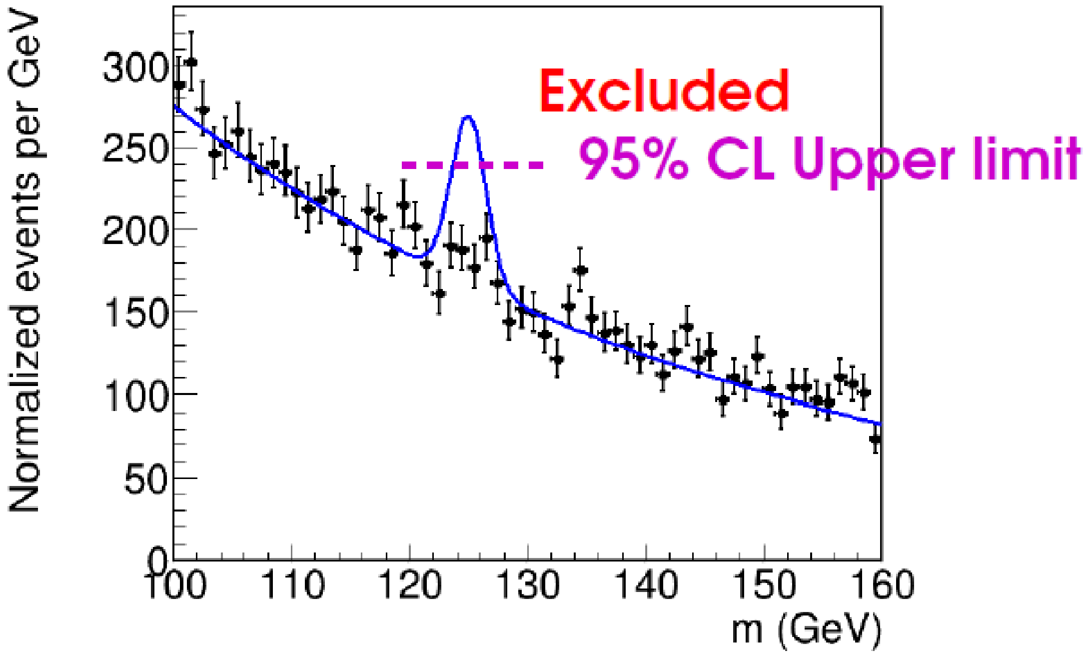
Hypothesis tests for Limits

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**

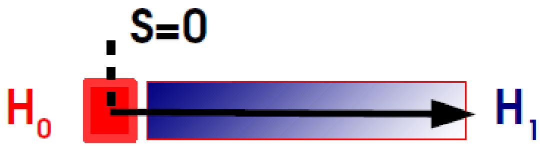
→ Typically report **95% CL** upper limit (p-value = 5%) : "$S < S_0$ @ 95% CL"



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

$(\hat{S} > 0)$

Limit-setting

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

Compare

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

← Likelihood of H_0

← Likelihood of H_1

$(\hat{S} < S_0)$

Same as q_0 :

→ large values ⇒ good rejection of H_0.

⇒ Can compute p-value from q_{S0}.

Inversion: Getting the limit for a given CL

Procedure:

→ Compute q_{S0} for some S_0, get the **exclusion p-value p_{S0}**.

Asymptotic case: can use $p_{S_0} = 1 - \Phi(\sqrt{q_{S_0}})$

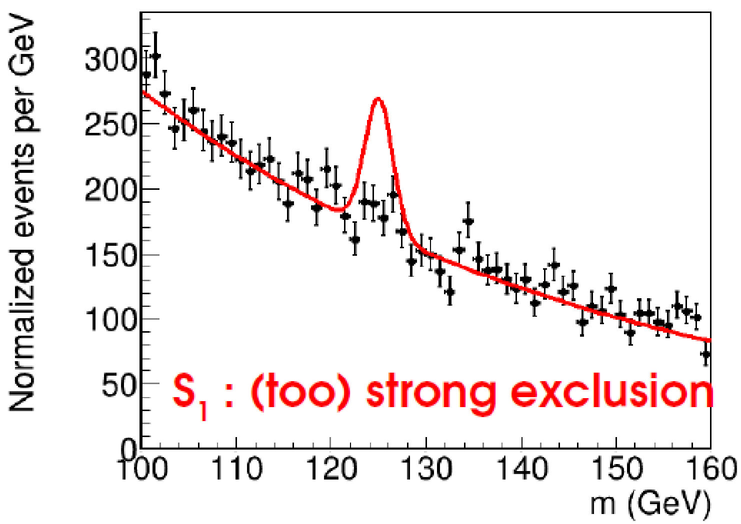
→ Adjust S_0 until 95% CL exclusion (p_{S0} = 5%) is reached

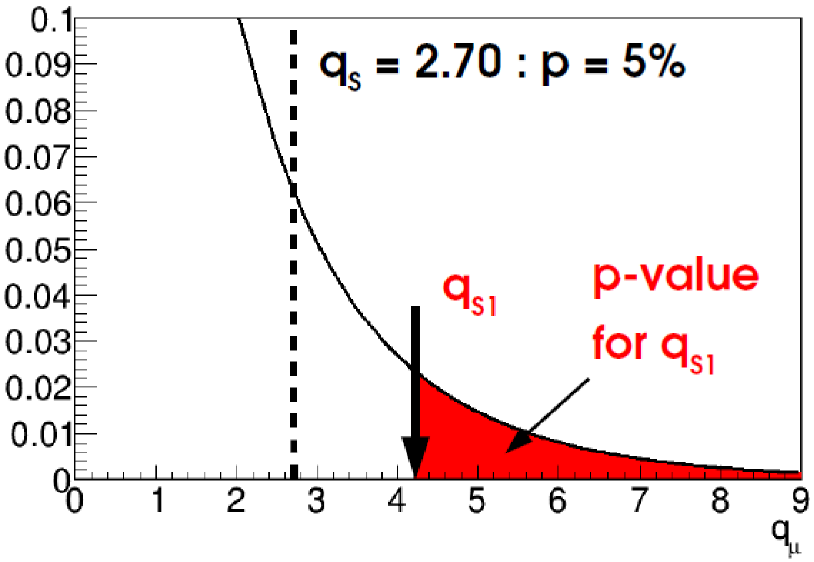
Asymptotic case: need q_{S0} = 2.70

Asymptotics

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

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:

→ Compute q_{S0} for some S_0, get the **exclusion p-value p_{S0}**.

Asymptotic case: can use $p_{S_0} = 1 - \Phi\left(\sqrt{q_{S_0}}\right)$

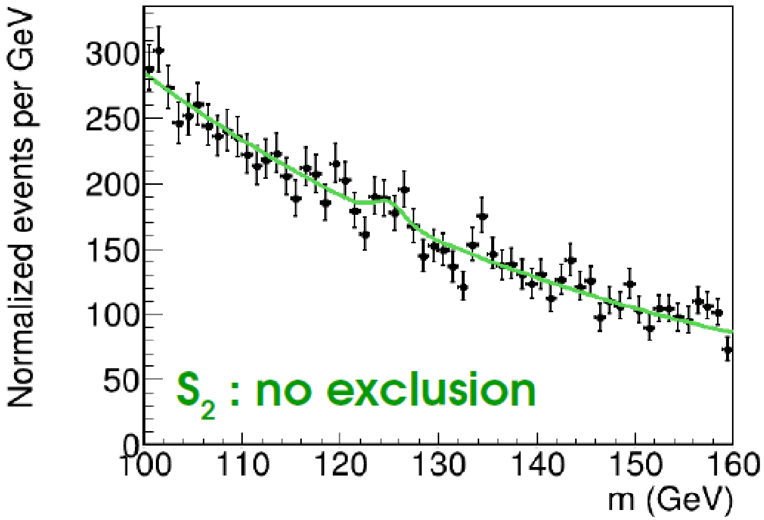
→ Adjust S_0 until 95% CL exclusion (p_{S0} = 5%) is reached

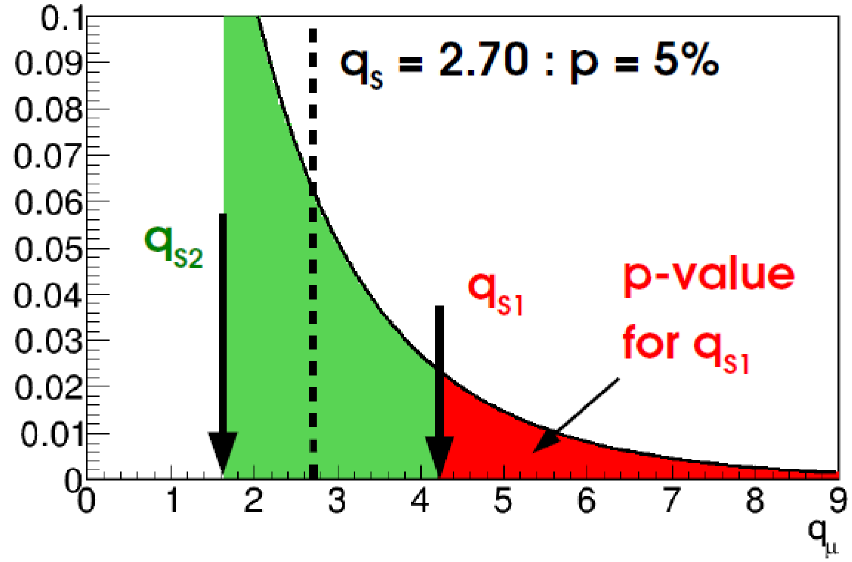
Asymptotic case: need q_{S0} = 2.70

Asymptotics

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

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:

→ Compute q_{S0} for some S_0, get the **exclusion p-value p_{S0}**.

Asymptotic case: can use $p_{S_0} = 1 - \Phi(\sqrt{q_{S_0}})$

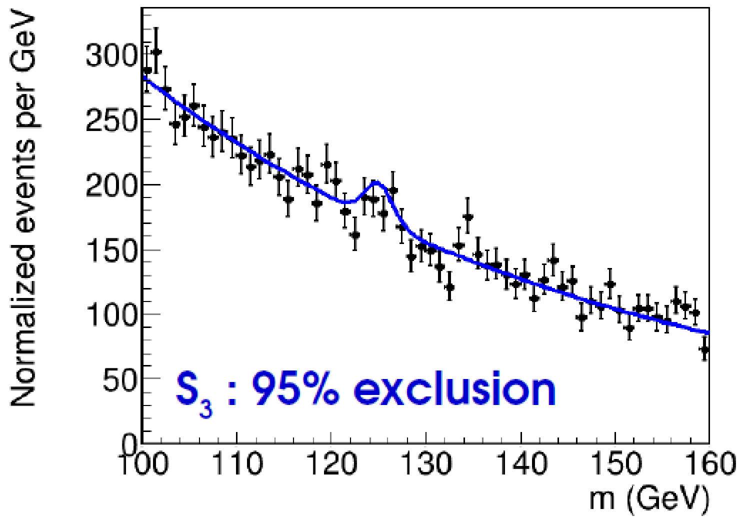
→ Adjust S_0 until 95% CL exclusion (p_{S0} = 5%) is reached

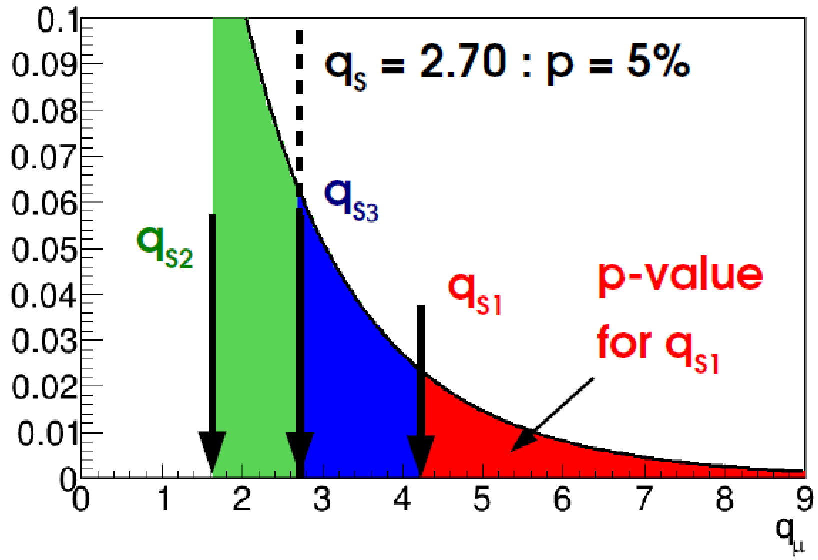
Asymptotic case: need q_{S0} = 2.70

Asymptotics

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

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





CL_s

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

Usual solution in HEP : **CL_s**.

→ Compute modified p-value

$$p_{CL_s} = \frac{p_{S_0}}{1 - p_B}$$

The usual p-value under $H(S=S_0)$ (=5%)

The p-value computed under H(S=0)

⇒ **Rescale** exclusion at S_0 by exclusion at S=0.

→ Somewhat ad-hoc, but good properties…

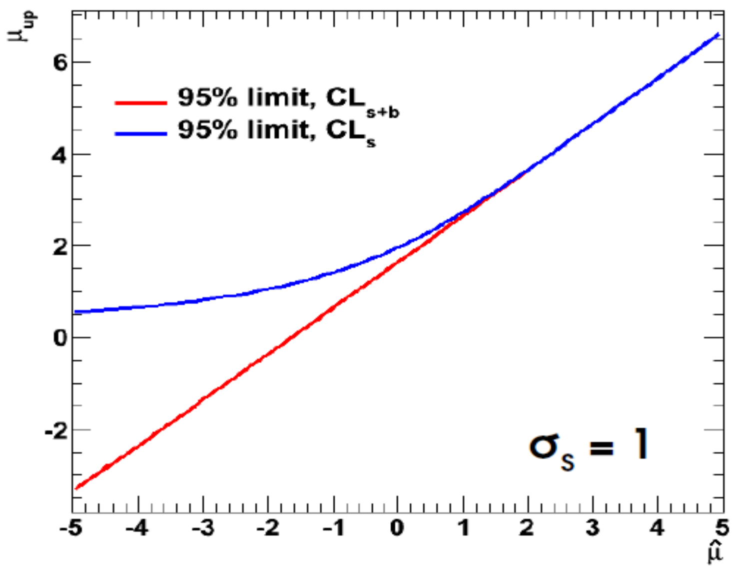
$\hat{S}$ compatible with 0 : $p_B \sim O(1)$

$p_{CLs} \sim$ **p_{S0} ~ 5%, no change**.

Far-negative $\hat{S}$: $1 - p_B \ll 1$

$p_{CLs} \sim p_{S0}/(1-p_B) \gg 5\%$

→ lower exclusion ⇒ higher limit,
usually >0 as desired

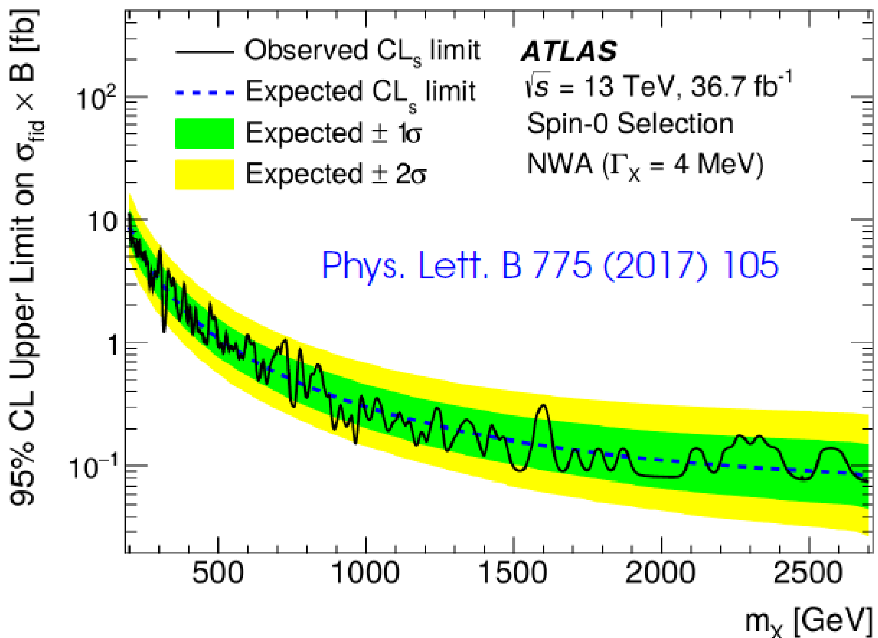


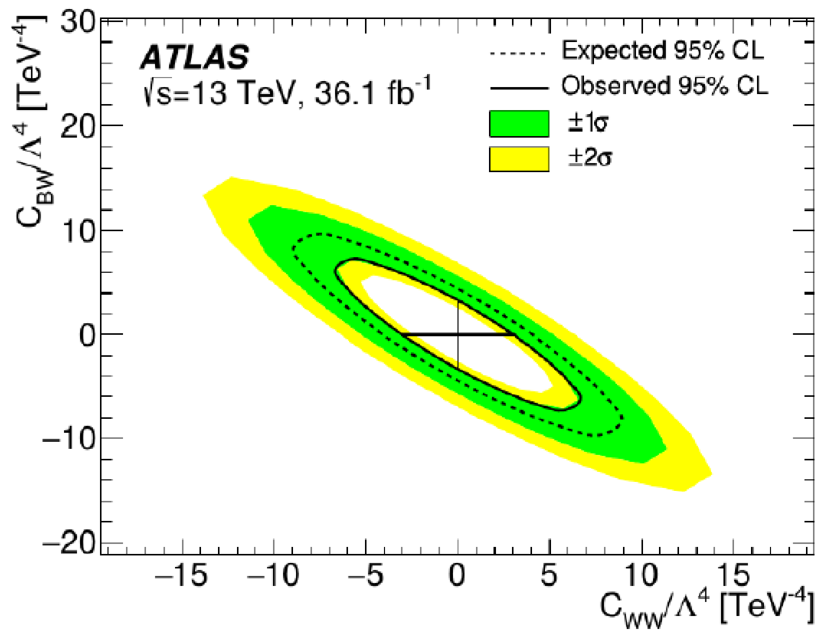
Drawback: ***overcoverage***

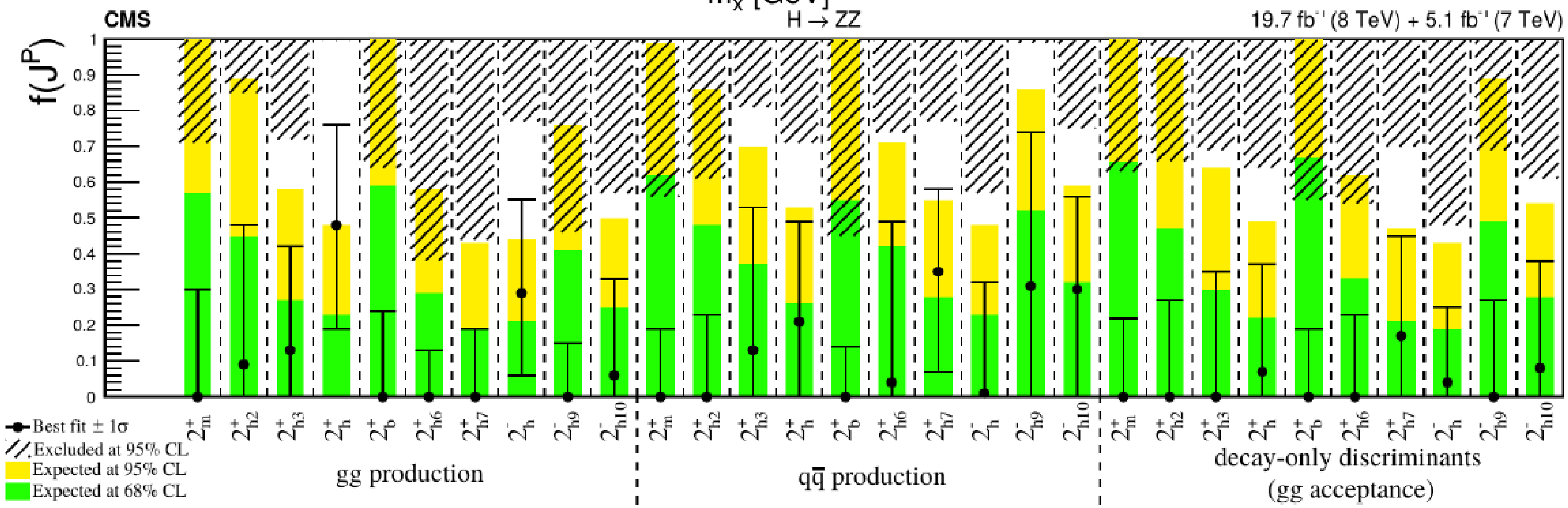
→ limit is claimed to be 95% CL, but actually >95% CL for small $1-p_B$.

Upper Limit Examples

ATLAS 2015-2016 4l aTGC Search







Phys. Rev. D 92 (2015) 012004

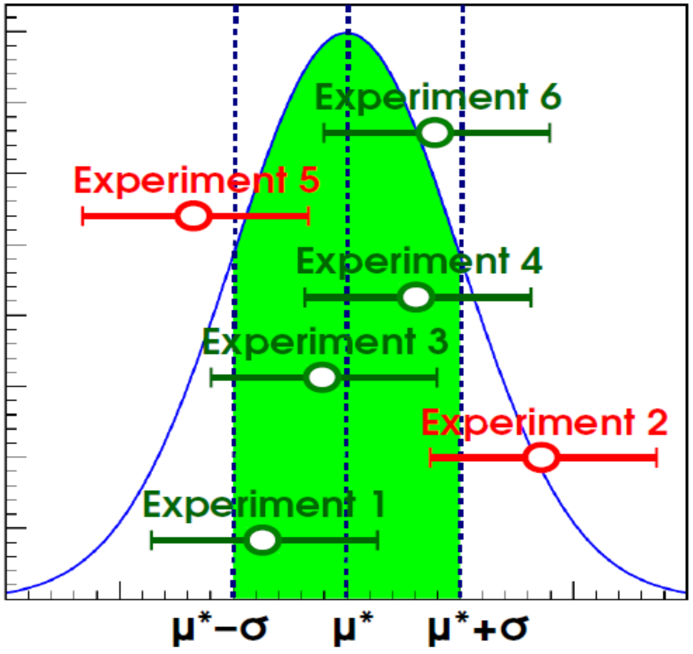
Gaussian Intervals

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

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

This is a probability for $\hat{\mu}$, not μ^*!

→ μ^* is a **fixed number**, **not a random variable**



But we can invert the relation:

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

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

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

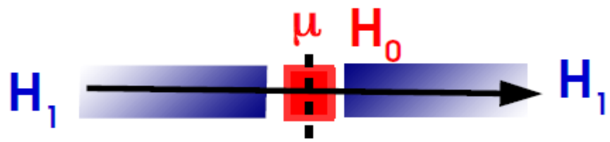
→ This gives the desired statement on μ^* : *if we repeat the experiment many times, [$\hat{\mu} - \sigma$, $\hat{\mu} + \sigma$] will **contain the true value** 68.3% of the time:* $\mu^* = \hat{\mu} \pm \sigma$

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

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

Z	1	1.96	2
CL	0.683	0.95	0.955

Likelihood Intervals

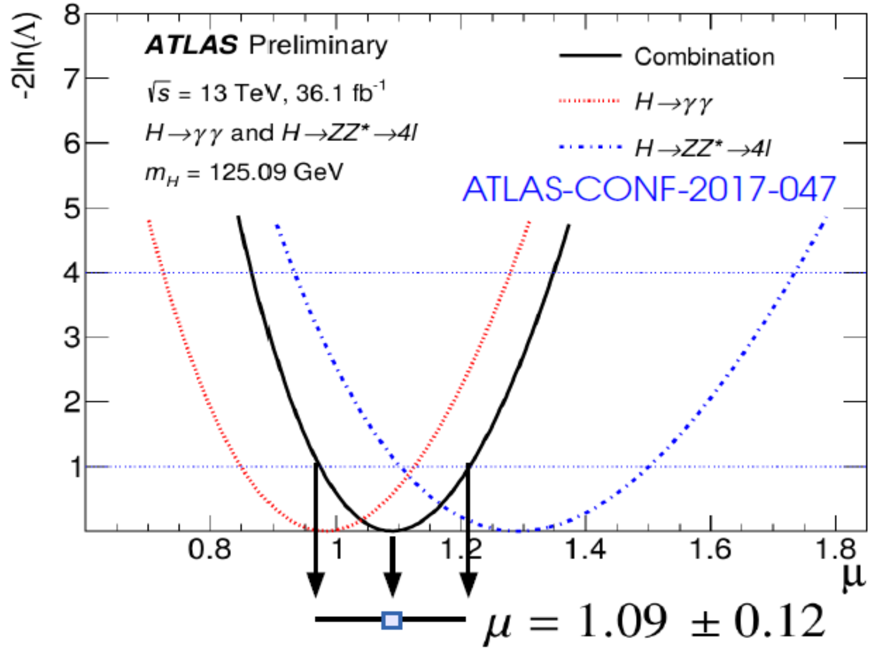


Confidence intervals from L:

- Test $H(\mu_0)$ against alternative using $t_{\mu_0} = -2\log\frac{L(\mu=\mu_0)}{L(\hat{\mu})}$
- Two-sided test since true value can be higher or lower than observed

Gaussian L:

- $t_{\mu_0} = \left(\frac{\hat{\mu}-\mu_0}{\sigma_\mu}\right)^2$: parabolic in μ_0.
- Minimum occurs at $\mu = \hat{\mu}$
- Crossings with $t_\mu = 1$ give the 1σ interval



General case:

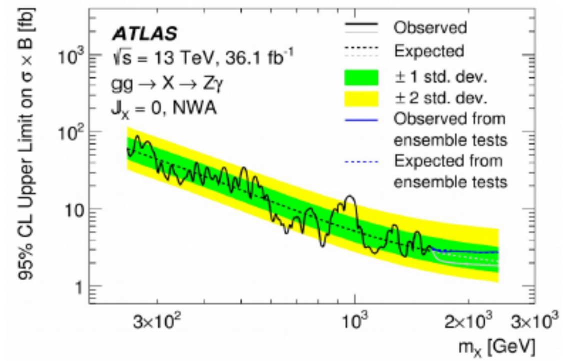
- Generally not a perfect parabola
- Minimum still occurs at $\mu = \hat{\mu}$
- Still define 1σ interval from the $t_\mu = \pm 1$ crossings

Takeaways

Limits : use LR-based test statistic:

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

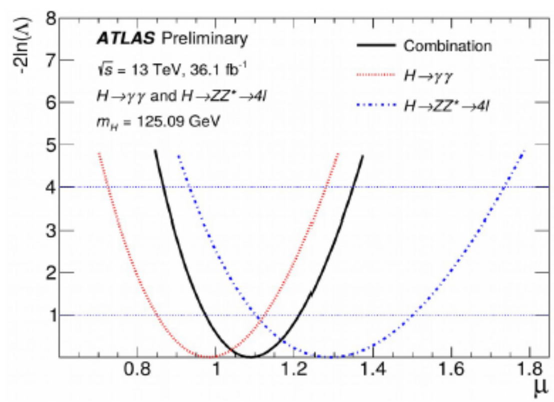
→ Use **CL_s procedure** to avoid negative limits



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

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

→ Crossings with $t_{\mu 0} = Z^2$ for $\pm Z\sigma$ intervals (in 1D)



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