IN2P3 School of Statistics 2022

Computing Statistical Results

Classical interval estimation
Limits, Systematics
and beyond

Nicolas Berger (LAPP Annecy)

Lecture 1

Lecture Plan

Statistics basic concepts (Monday/Tuesday)

- **Basic ingredients** (PDFs, etc.)
- **Parameter estimation** (maximum likelihood, least-squares, …)
- **Model testing** (χ^2 tests, hypothesis testing, p-values, …)

These lectures: Computing statistical results

- **Statistical modeling**
- **Review of model testing**
- **Computing results**
  - **Confidence intervals**
  - **Discovery**
  - **Upper limits**
- **Systematics and profiling**
- **Bayesian techniques**

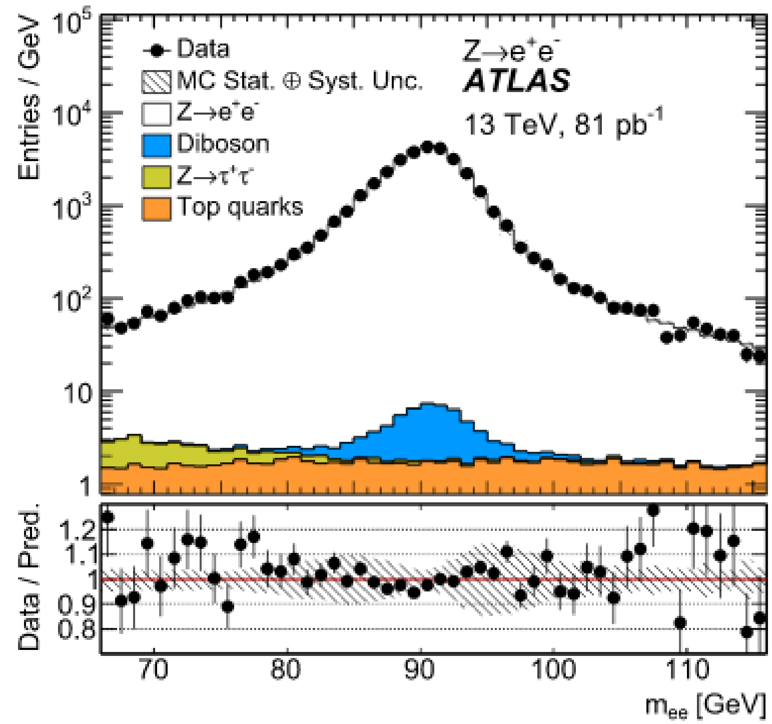
See also the Hands-on tutorial yesterday covering both sets of lectures.

Statistical Modeling

Example 1: Z counting

Phys. Lett. B 759 (2016) 601

Measure the cross-section (event rate) of the $Z \to ee$ process



$$\sigma^{fid} = \frac{n_{data} - N_{bkg}}{C_{fid}\, L}$$

n_{data}: 35000 ± 187

N_{bkg}: 175 ± 8

L: (81 ± 2) pb^{-1}

C_{fid}: 0.552 ± 0.006

σ^{fid} = 0.781 ± 0.004 (stat) ± 0.018 (syst) nb

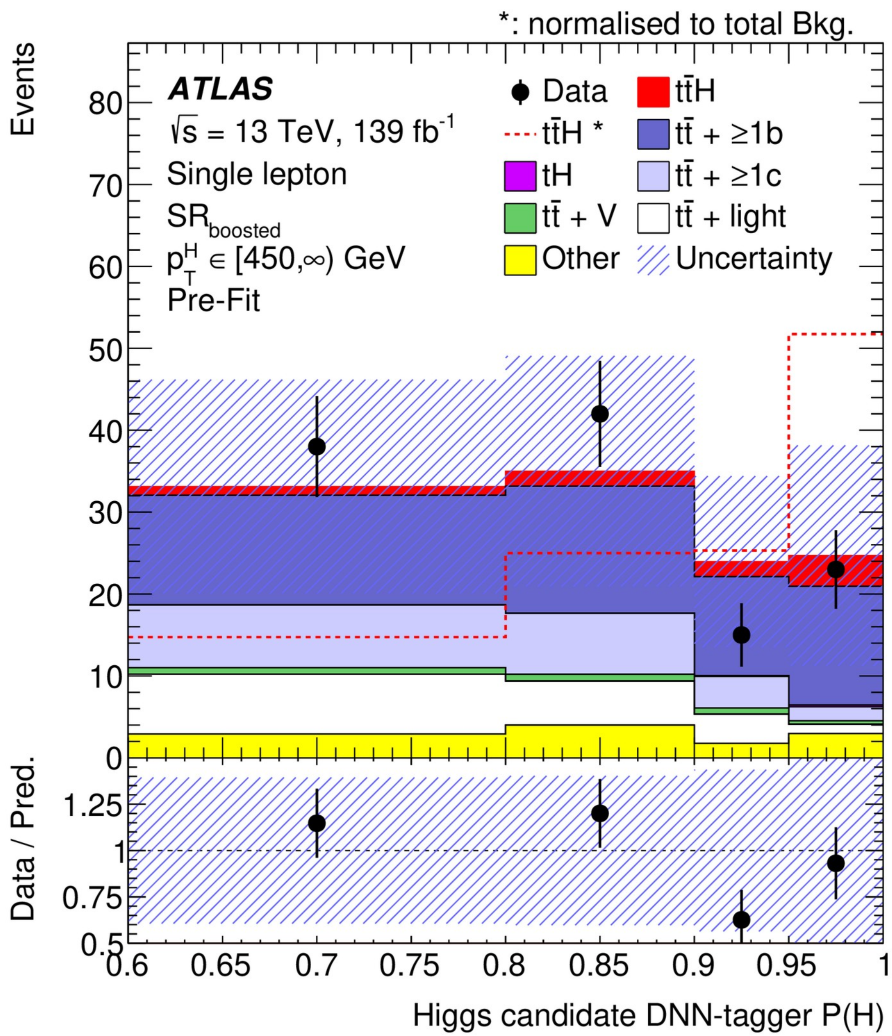
Fluctuations in the data counts

Other uncertainties (assumptions, parameter values)

"Single bin counting" : only data input is $\mathbf{N_{data}}$.

Example 2: ttH→bb

arXiv:2111.06712



Event counting in different regions:

Multiple-bin counting

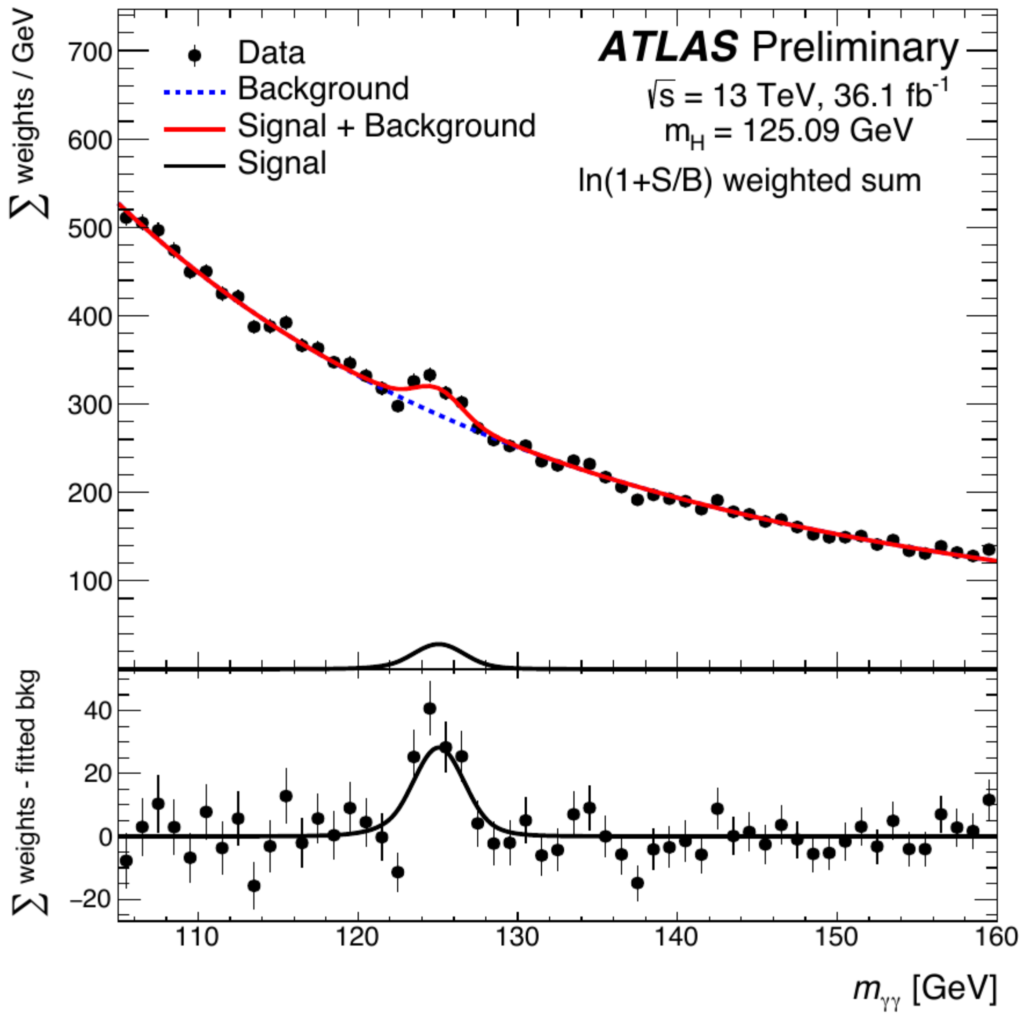
Lots of information available

→ Potentially higher sensitivity

→ How to make optimal use of it ?

Example 3: unbinned modeling

ATLAS-CONF-2017-045



All modeling done using continuous distributions:

$$P_{\text{total}}(m_{\gamma\gamma}) = \frac{S}{S+B} P_{\text{signal}}(m_{\gamma\gamma}; m_H) + \frac{B}{S+B} P_{\text{bkg}}(m_{\gamma\gamma})$$

How to count

Common situation: produce many events N, select a (very) small fraction P

→ In principle, binomial process

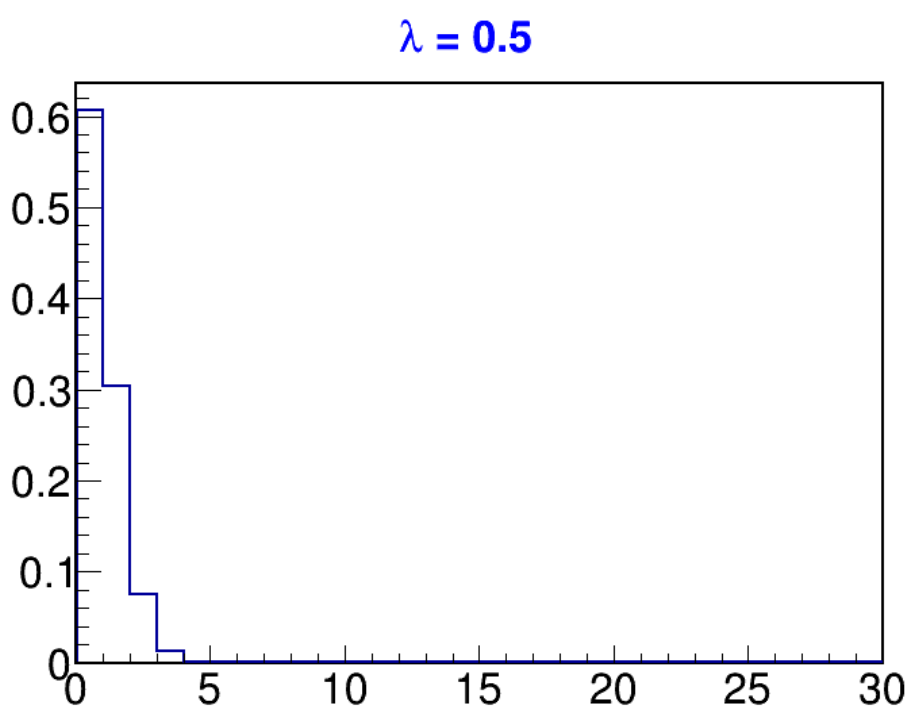
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution

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

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = √λ

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \overset{\lambda \to \infty}{\to} G(\lambda, \sqrt{\lambda})$$

How to count

Common situation: produce many events N, select a (very) small fraction P

→ In principle, binomial process

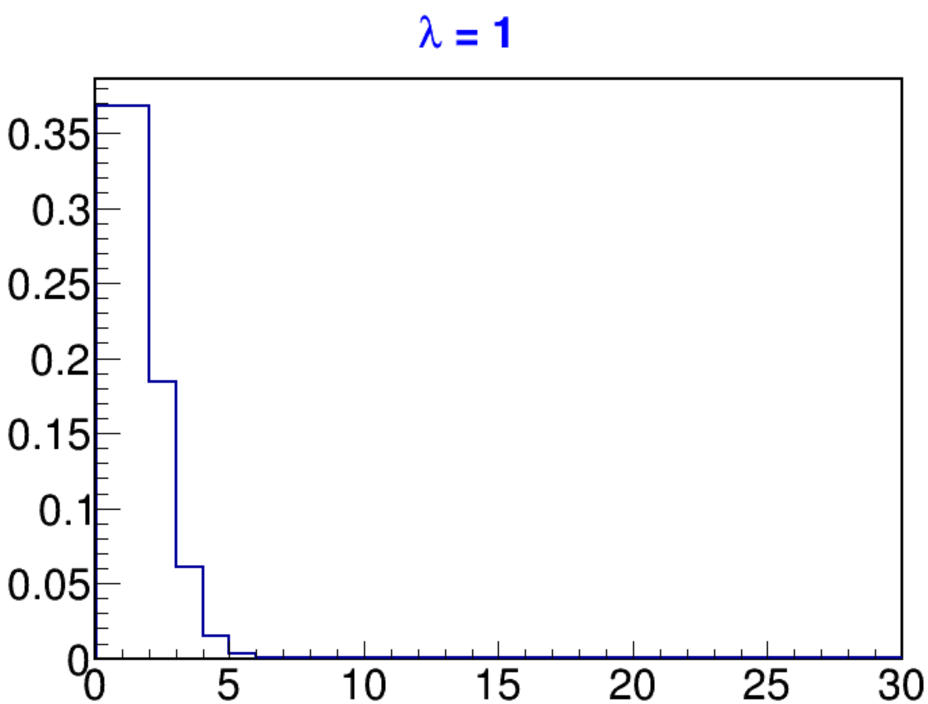
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution

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

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = √λ

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \xrightarrow{\lambda \to \infty} G(\lambda, \sqrt{\lambda})$$

How to count

Common situation: produce many events N, select a (very) small fraction P

→ In principle, binomial process

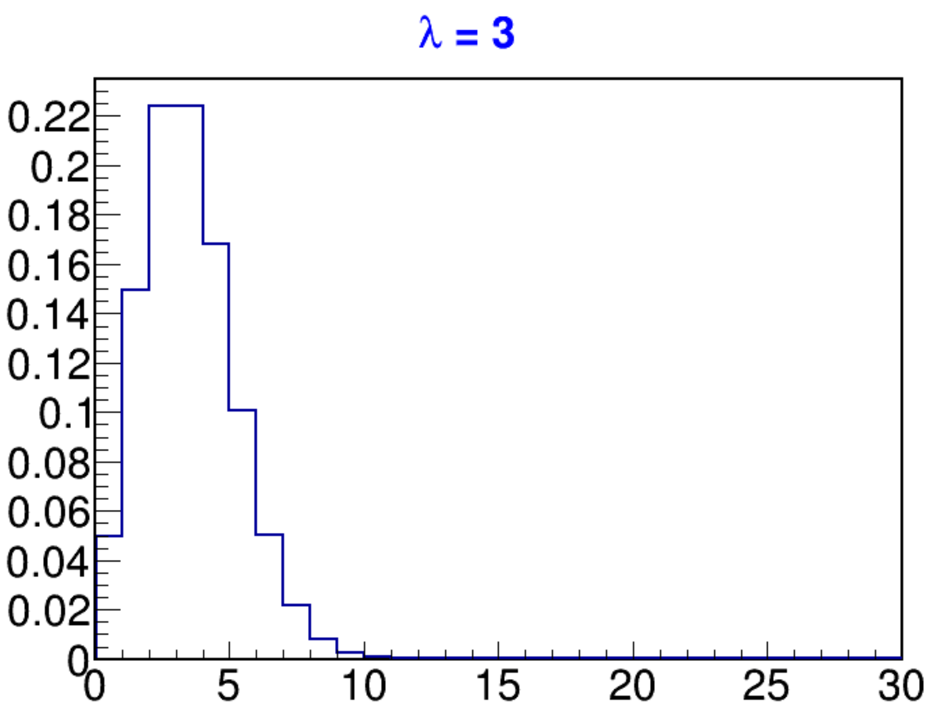
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution

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

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = √λ

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \overset{\lambda \to \infty}{\to} G(\lambda, \sqrt{\lambda})$$

How to count

Common situation: produce many events N, select a (very) small fraction P

→ In principle, binomial process

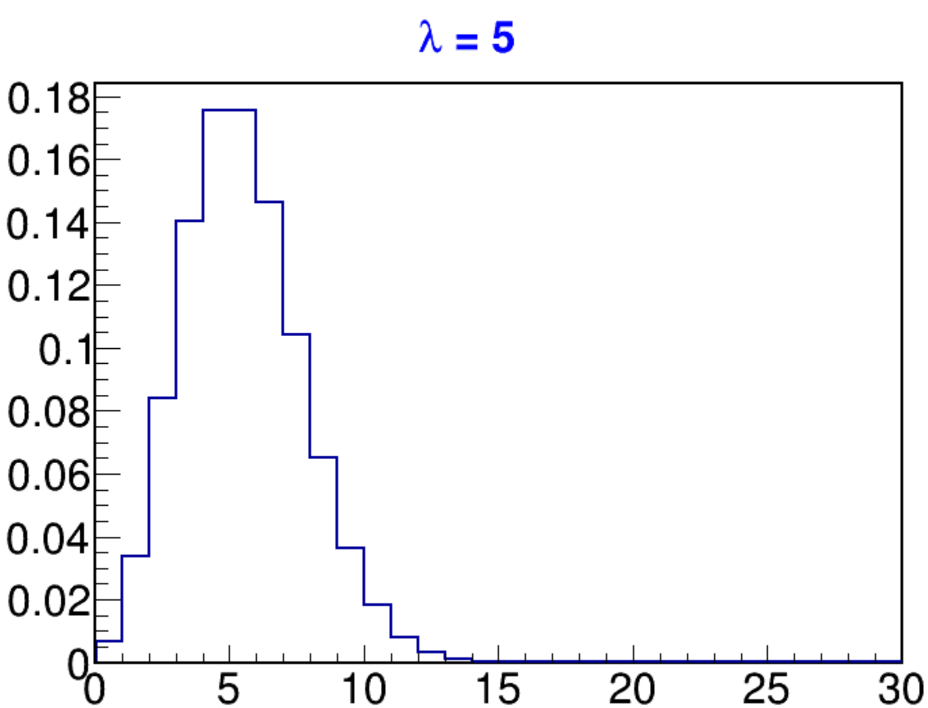
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* very rare process, but very many trials so still expect to see good events

Poisson distribution

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

$$(1-P)^{N-n} \overset{n\ll N}{\simeq} \left(1-\frac{\lambda}{N}\right)^N \overset{N\gg 1}{\simeq} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = √λ

Central limit theorem :

becomes Gaussian for large λ :

$$P(\lambda) \overset{\lambda\to\infty}{\to} G(\lambda,\sqrt{\lambda})$$

How to count

Common situation: produce many events N, select a (very) small fraction P

→ In principle, binomial process

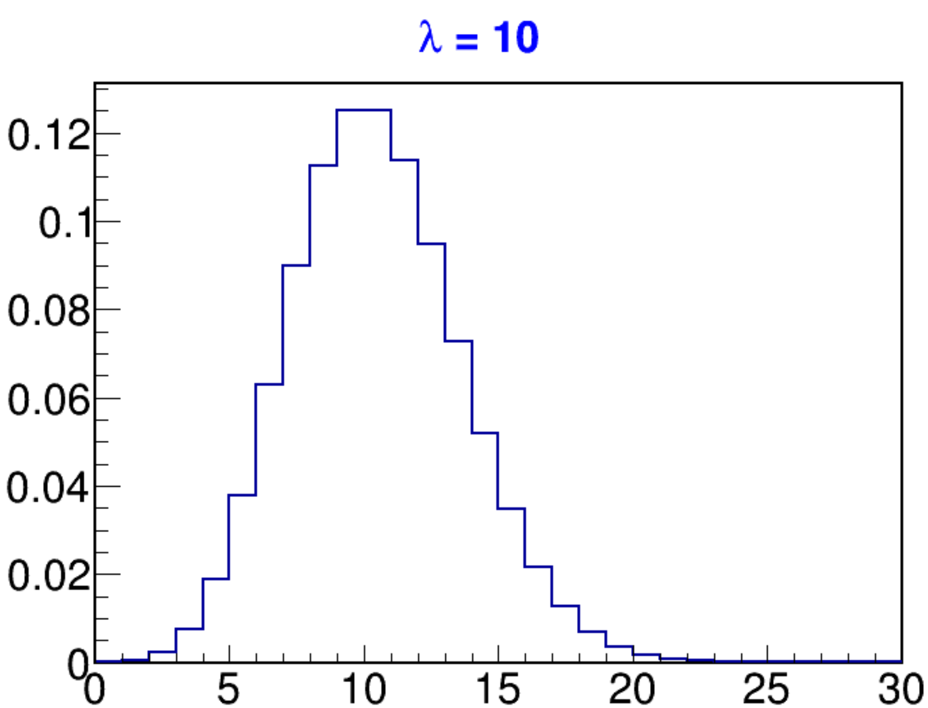
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* very rare process, but very many trials so still expect to see good events

Poisson distribution

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

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = √λ

Central limit theorem :

becomes Gaussian for large λ :

$$P(\lambda) \overset{\lambda \to \infty}{\to} G(\lambda, \sqrt{\lambda})$$

How to count

Common situation: produce many events N, select a (very) small fraction P

→ In principle, binomial process

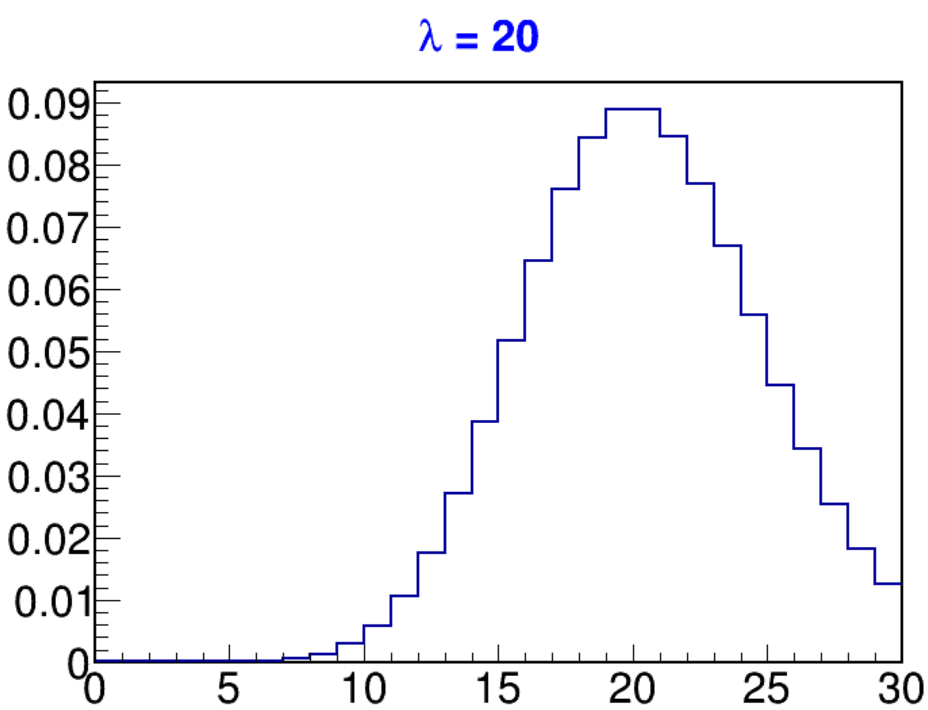
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* very rare process, but very many trials so still expect to see good events

Poisson distribution

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

$$(1-P)^{N-n} \overset{n\ll N}{\simeq} \left(1-\frac{\lambda}{N}\right)^N \overset{N\gg 1}{\simeq} e^{-\lambda}$$



Mean = λ

Variance = λ

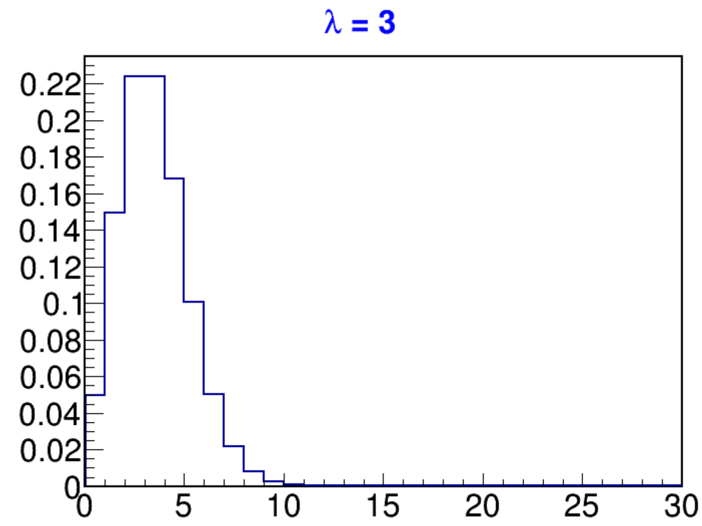
σ = √λ

Central limit theorem :

becomes Gaussian for large λ :

$$P(\lambda) \overset{\lambda\to\infty}{\rightarrow} G(\lambda,\sqrt{\lambda})$$

Statistical Model for Counting



Observable: number of events n

Typically both **S**ignal and **B**ackground present:

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

S : # of events from signal process

B : # of events from bkg. process(es)

Model has **parameters S** and **B.**

B can be known a priori or not (S usually not...)

→ Example: **assume B is known,** use **measured n** to find out about **S.**

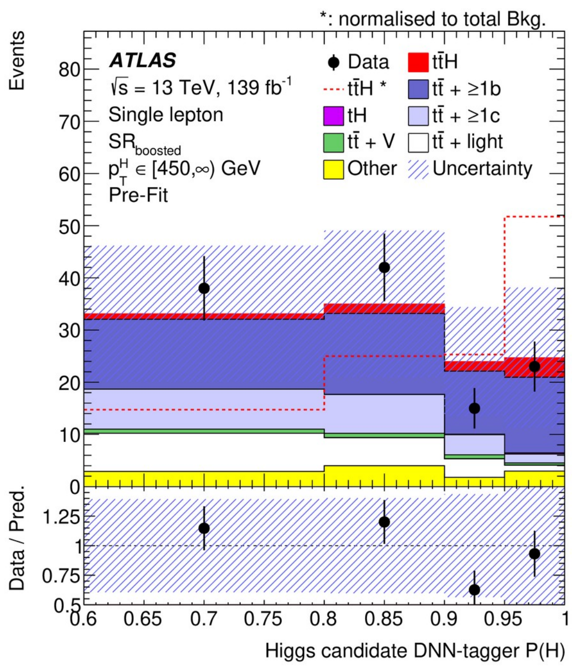
Multiple counting bins

Count in bins of a variable ⇒ *histogram* $n_1 \dots n_N$.
(N : number of bins)

Per-bin fractions (=shapes) of Signal and Background

$$P(\{n_i\};S,B) = \prod_{i=1}^{N} e^{-(S f_{S,i} + B f_{B,i})} \frac{(S f_{S,i} + B f_{B,i})^{n_i}}{n_i!}$$

Poisson distribution in each bin



Shapes f typically obtained from simulated events (*Monte Carlo*)

→ HEP: typically excellent modeling from simulation, although some uncertainties need to be accounted for.

However not always possible to generate sufficiently large MC samples
MC stat fluctuations can create artefacts, especially for S ≪ B.

Model Parameters

Model typically includes:

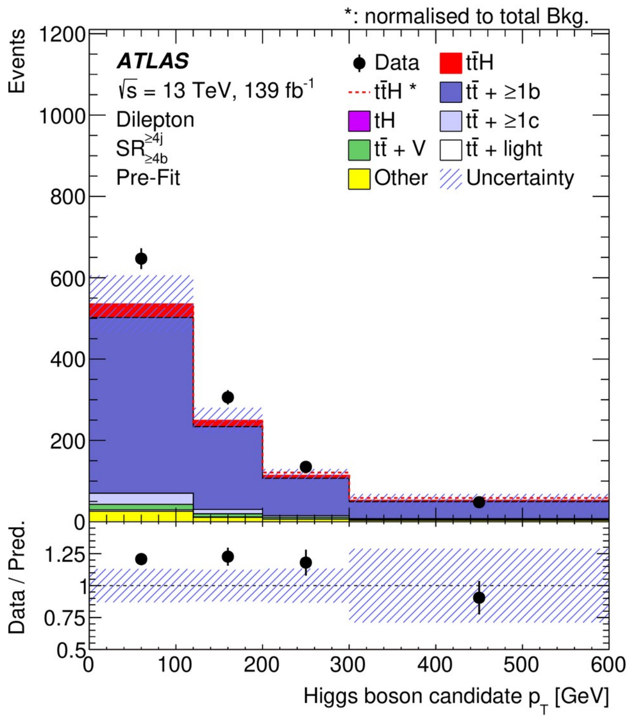
- **Parameters of interest** (POIs) : what we want to measure

  → **S**, **m_W**, …

- **Nuisance parameters** (NPs) : other parameters needed to define the model

  → Background levels (**B**)

  → For binned data, f^{sig}_i , f^{bkg}_i

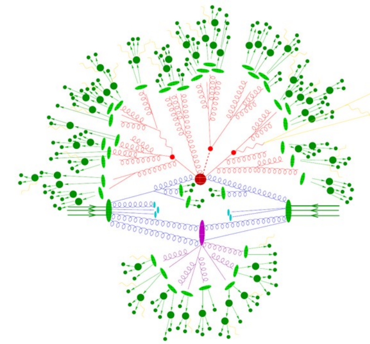


NPs must be either:

→ **Known a priori** (within uncertainties) or

→ **Constrained by the data**

Takeaways



Random data must be described using a statistical model:

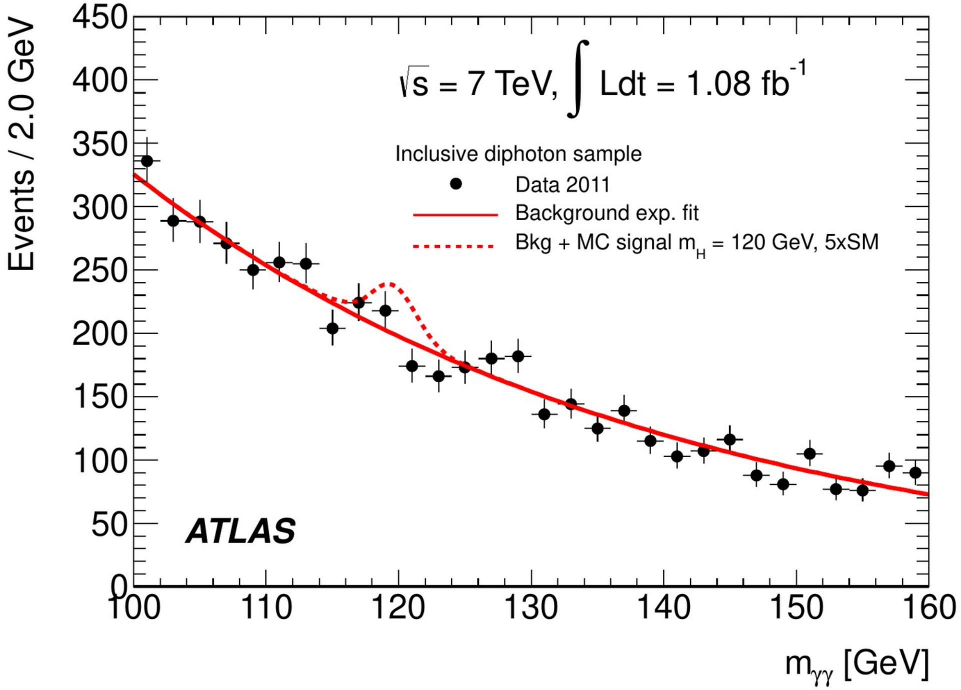
Description	Observable	Likelihood
Counting	**n**	**Poisson** $P(n; S, B) = e^{-(S+B)} \frac{(S+B)^n}{n!}$
Binned shape analysis	n_i, i = 1 .. N_{bins}	**Poisson product** $P(n_i; S, B) = \prod_{i=1}^{n_{bins}} e^{-(S f_i^{sig} + B f_i^{bkg})} \frac{(S f_i^{sig} + B f_i^{bkg})^{n_i}}{n_i!}$
Unbinned shape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

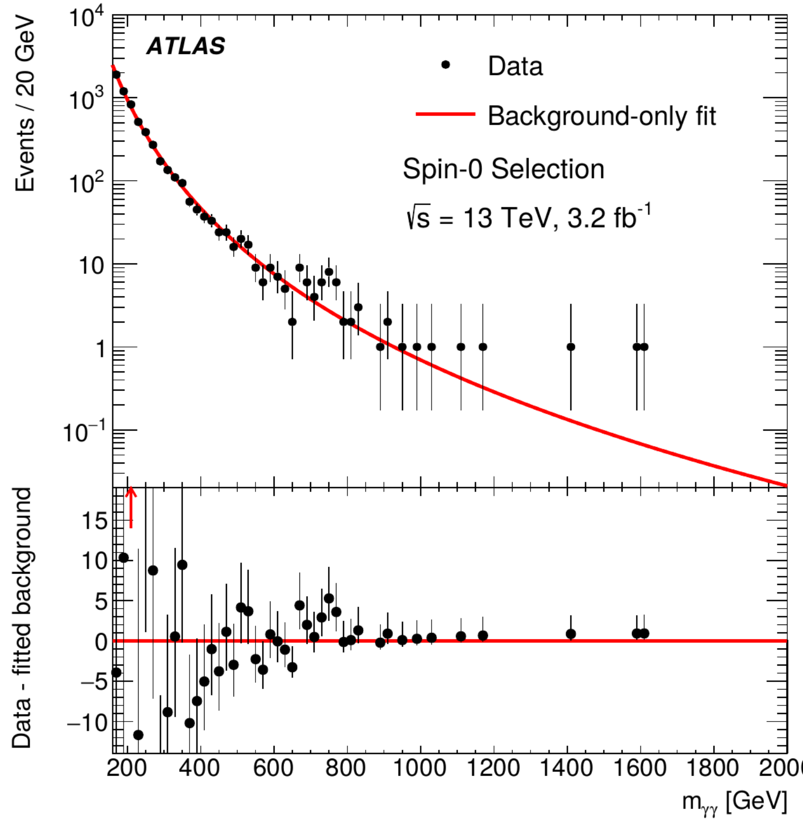
Model can include multiple **categories**, each with a separate description

Includes **parameters of interest** (POIs) but also **nuisance parameters** (NPs)

Next step: use the model to obtain information on the POIs

Hypothesis Testing and discovery

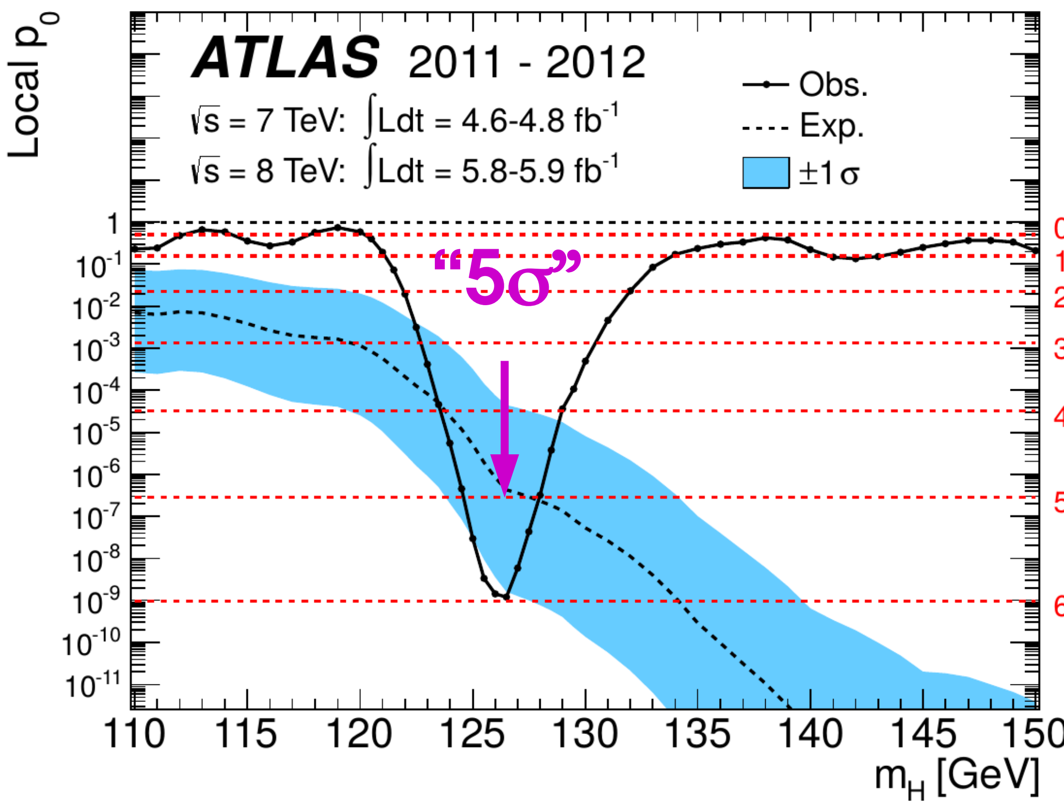


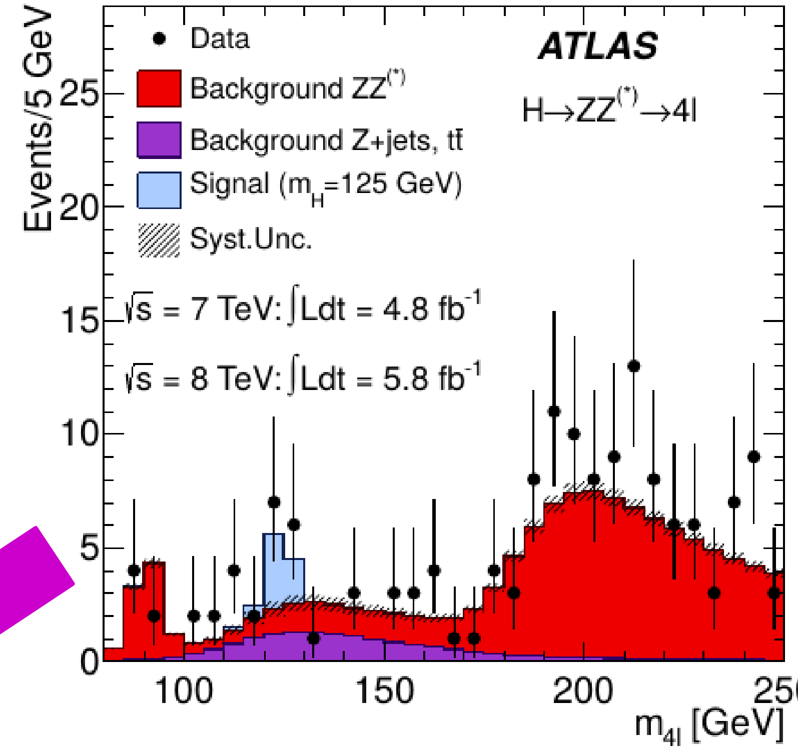


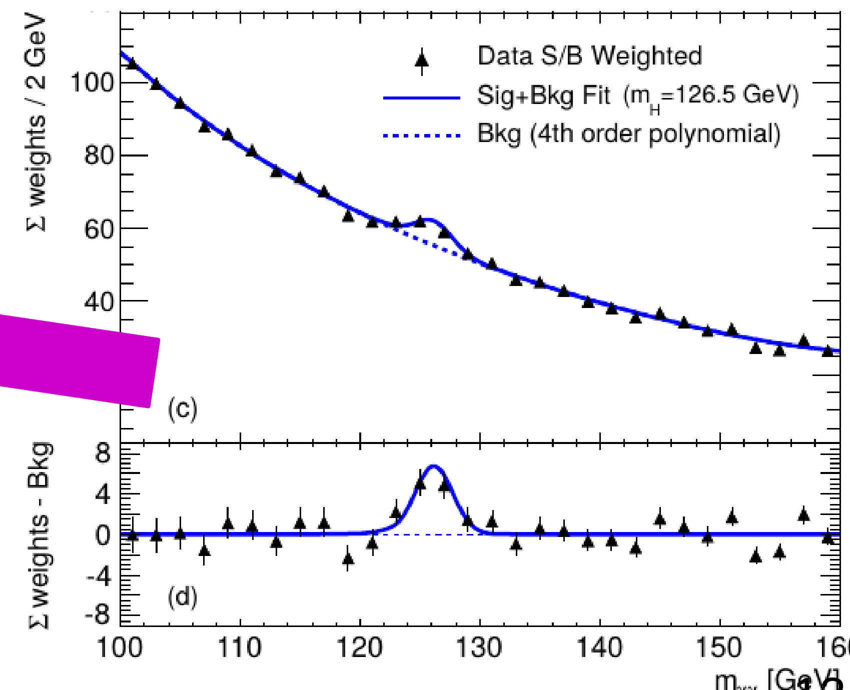
Discovery Testing

We see an unexpected feature in our data, is it a signal for new physics or a fluctuation ?

e.g. Higgs discovery : **"We have 5σ" !**



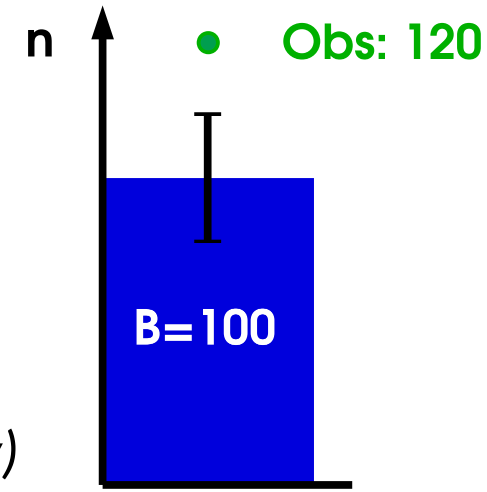




Phys. Lett. B 716 (2012) 1-29

Discovery Testing

Say we have a Gaussian measurement with a background **B=100**, and we measure **n=120**



Did we just discover something ? ***Maybe :-)*** *(but not very likely)*

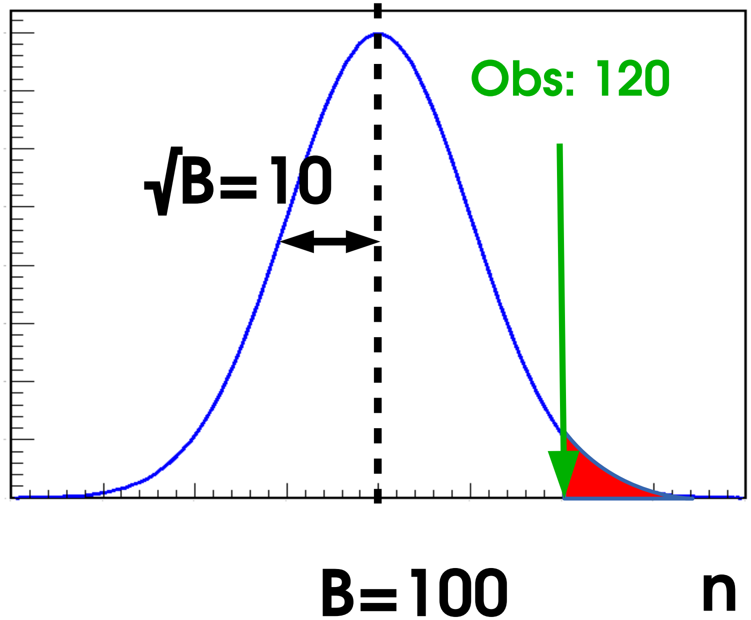
The measured signal is **S = 20**. $\mathbf{S = n_{obs} - B}$

Uncertainty on B is √B = 10

⇒ Significance Z = 2

⇒ we are **~2σ** away from S=0.

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



Gaussian quantiles :

Z = 2 happens **p_0 ~ 2.3%** of the time if S=0

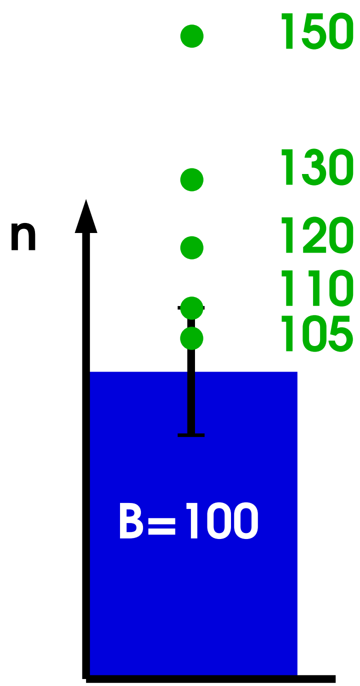
P-value:

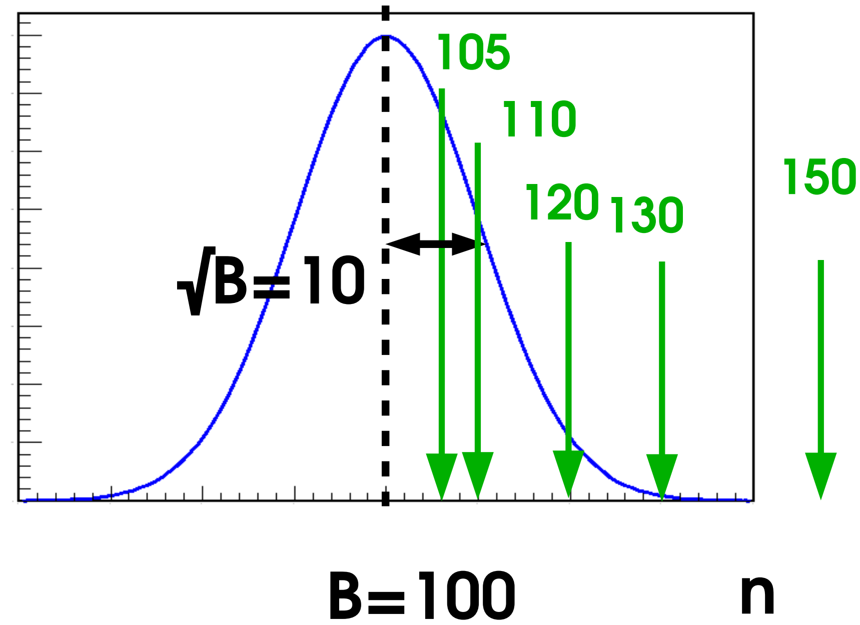
$$p_0 = 1 - \Phi(Z)$$

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

⇒ Rare, but not exceptional

Discovery Testing





n_{obs}	S	Z	p_0
105	5	0.5σ	31%
110	10	1σ	16%
120	20	2σ	2.3%
130	30	3σ	0.1%
150	50	5σ	$3\ 10^{-7}$

Straightforward in this Gaussian case

Need to be able to do the same in more complex cases:

- **Determine S**
- **Compute Z and p_0**

Evidence

Discovery

What is PDF is for

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

⇒ Possible outcomes of the experiment, for given parameter values

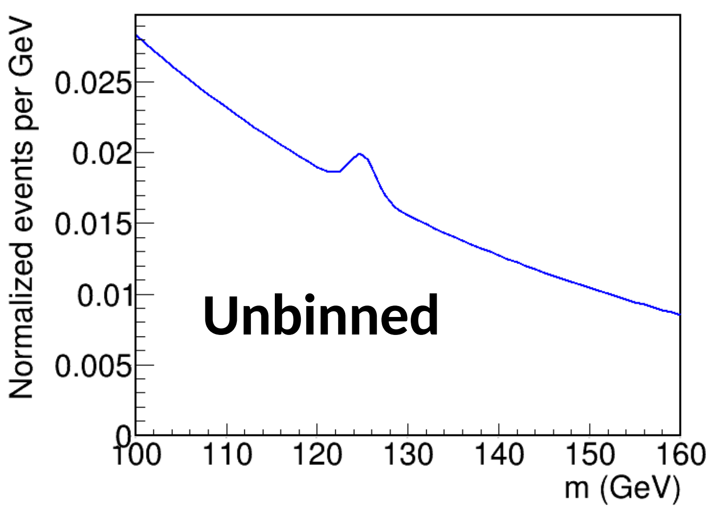
Can draw random events according to PDF : **generate** ***pseudo-data***

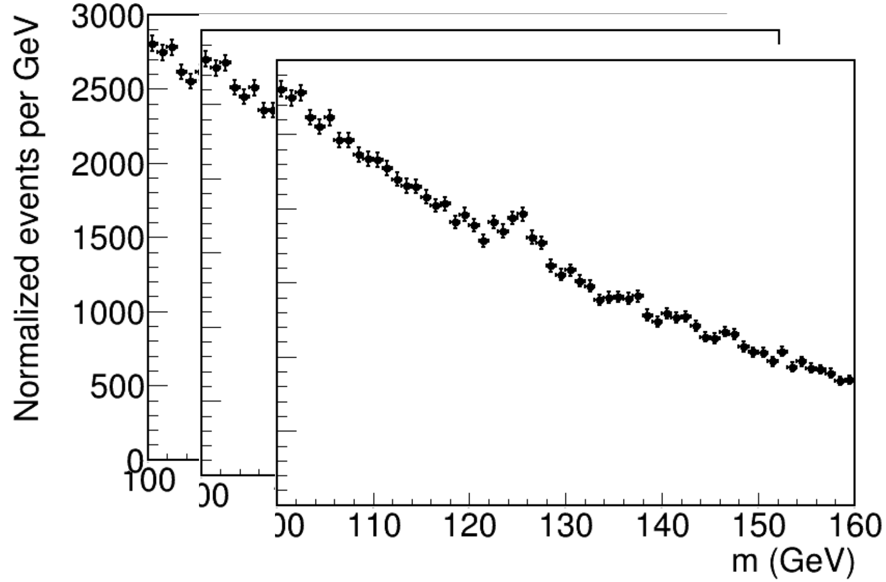
$$P(\lambda=5)$$

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

Each entry = separate “experiment”

Generate





What is PDF is also for: Likelihood

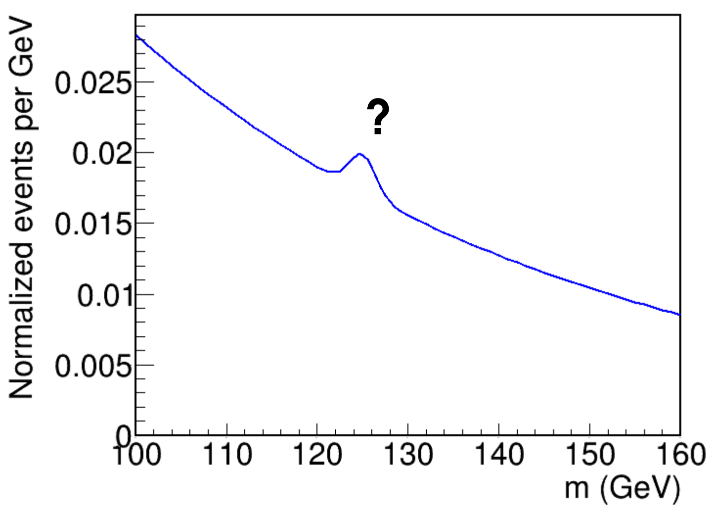
Model describes the distribution of the observable: **P(data; parameters)**

⇒ Possible outcomes of the experiment, for given parameter values

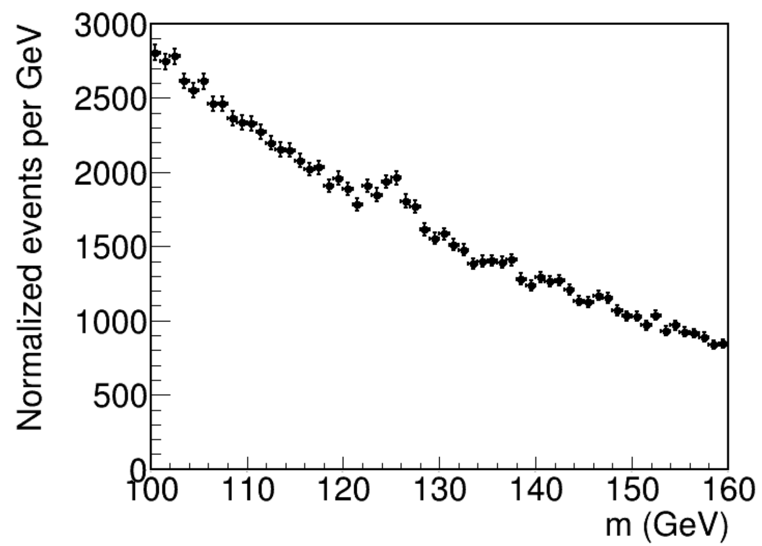
We want the other direction: **use data to get information on parameters**

$P(\lambda = ?)$

2



Estimate



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

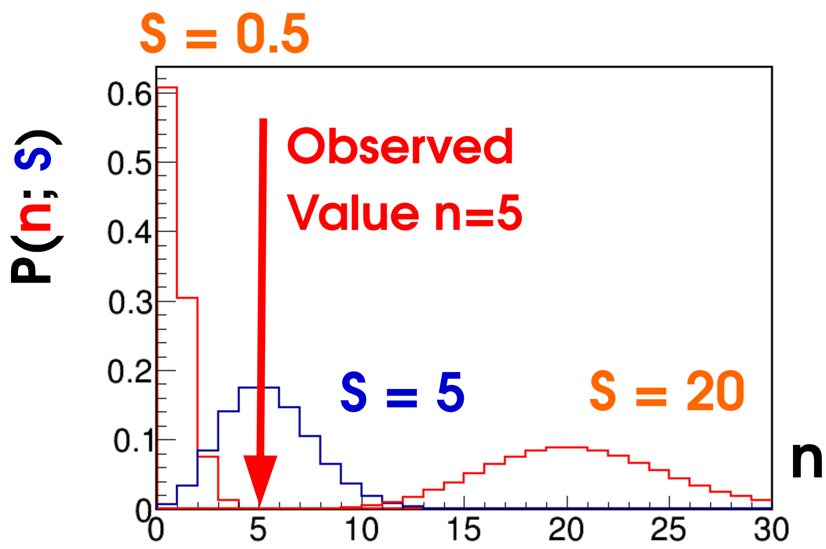
→ **same as the PDF,** but seen as function of the parameters

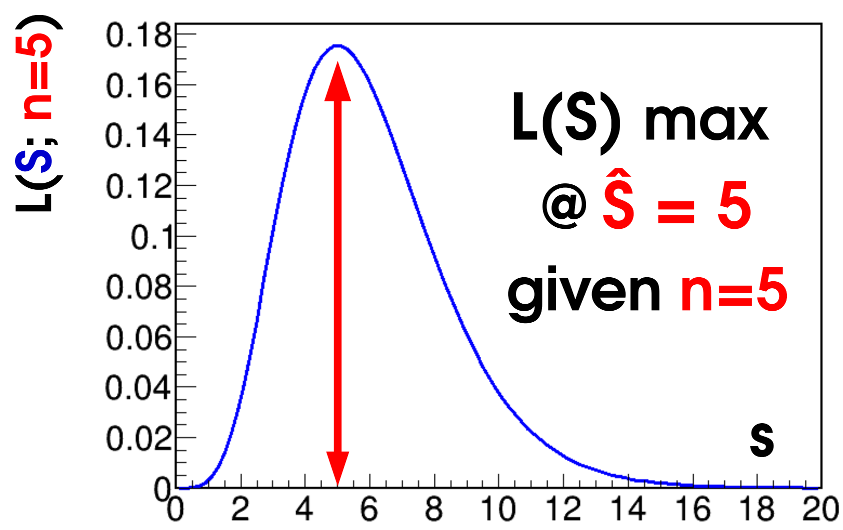
Maximum Likelihood Estimation

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

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

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



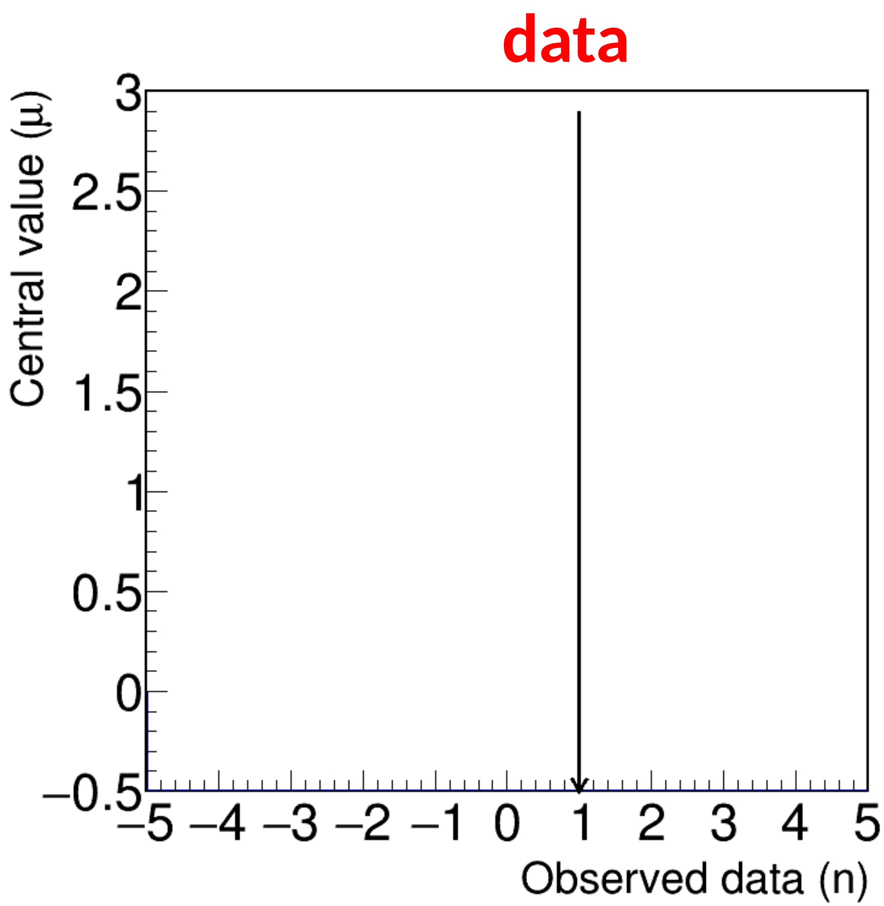


MLE: the value of μ for which **this data** was ***most likely to occur***

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

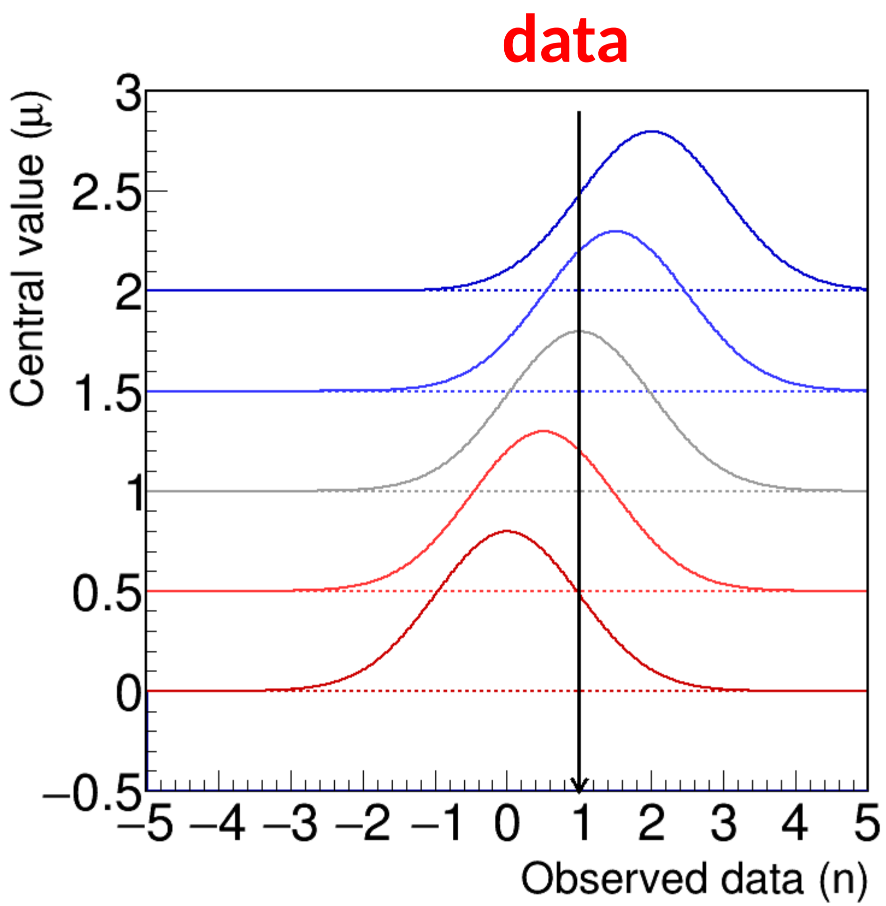
No guarantee it is the true value (data may be "unlikely") but sensible estimate

Gaussian case



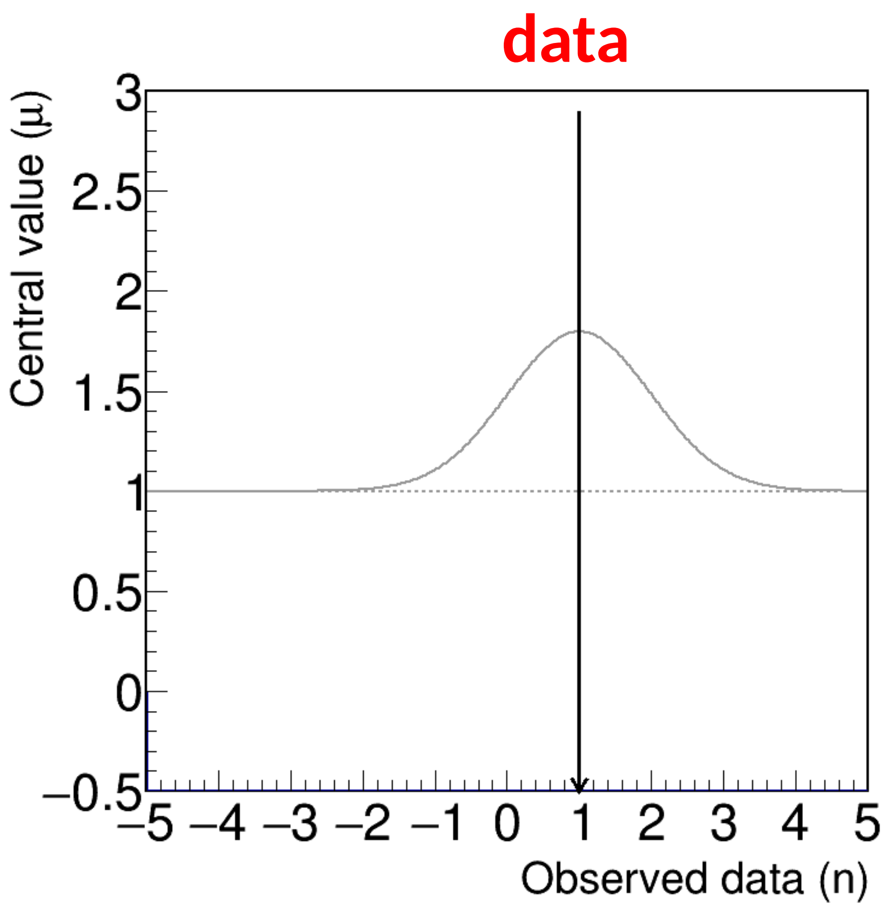
Best-fit of Gaussian PDF mean to observed data

Gaussian case



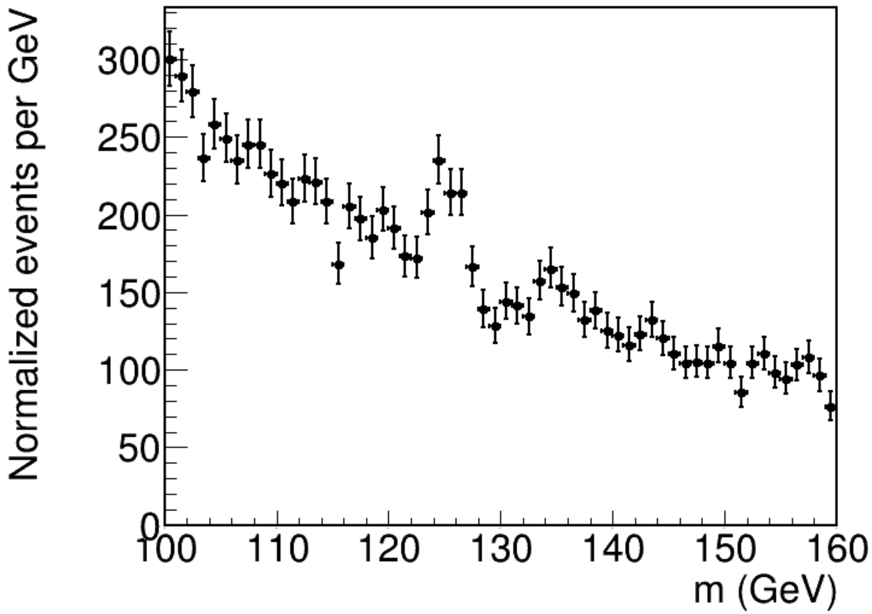
Best-fit of Gaussian PDF mean to observed data

Gaussian case



Best-fit of Gaussian PDF mean to observed data

Multiple Gaussian bins



-2 log Likelihood:

$$\lambda(\mu) = -2\log L(\mu) = \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i} \right)^2$$

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

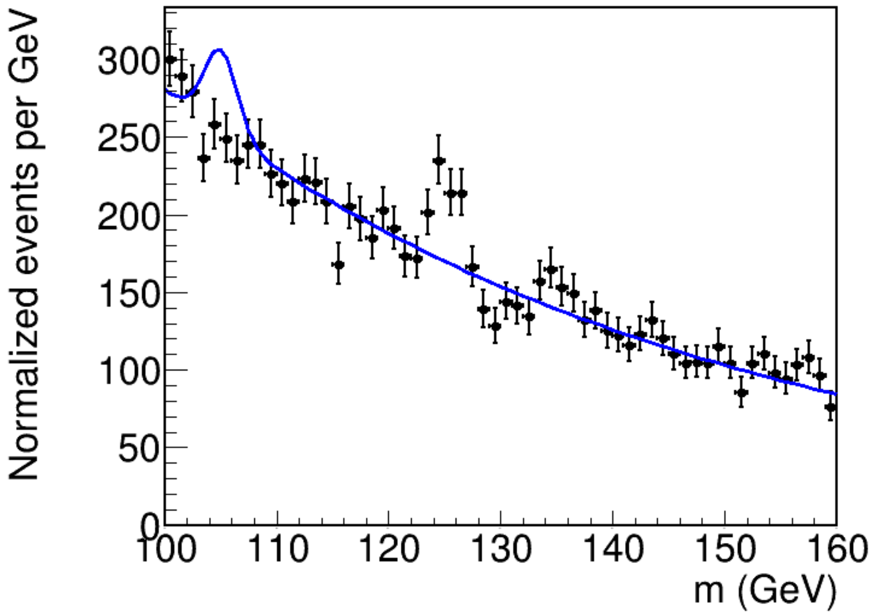
However typically need to perform non-linear minimization.

HEP practice:

- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/... backends

  → Usual methods – gradient-based, etc.

Multiple Gaussian bins



-2 log Likelihood:

$$\lambda(\mu) = -2 \log L(\mu) = \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i} \right)^2$$

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

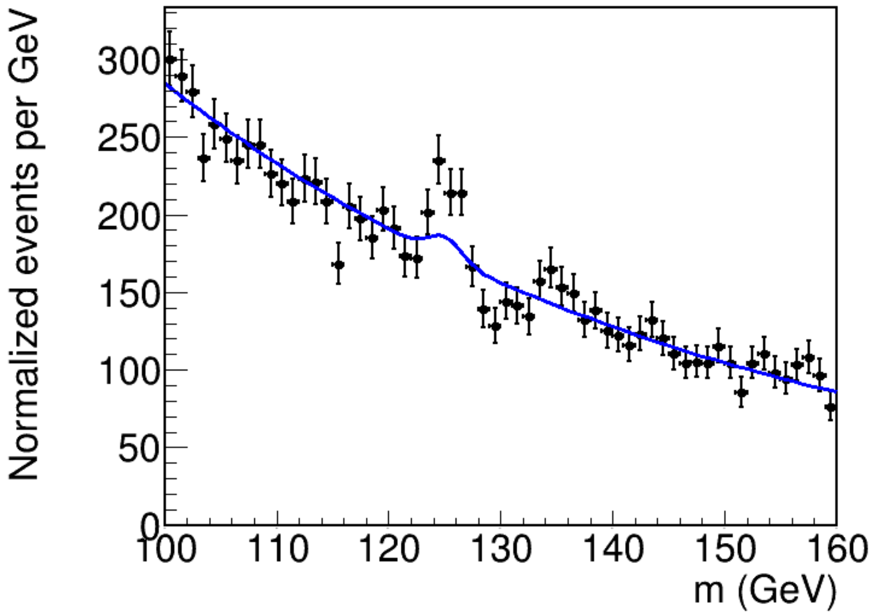
However typically need to perform non-linear minimization.

HEP practice:

- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/... backends

  → Usual methods – gradient-based, etc.

Multiple Gaussian bins



-2 log Likelihood:

$$\lambda(\mu) = -2 \log L(\mu) = \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i} \right)^2$$

Maximum likelihood ⇔ Minimum χ^2

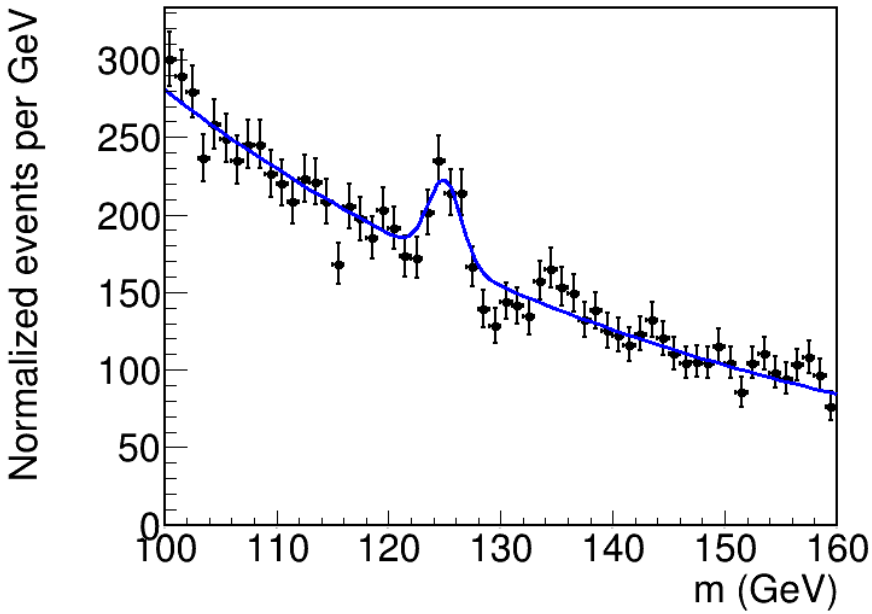
⇔ Least-squares minimization

However typically need to perform non-linear minimization.

HEP practice:

- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/… backends
  → Usual methods – gradient-based, etc.

Multiple Gaussian bins



-2 log Likelihood:

$$\lambda(\mu) = -2\log L(\mu) = \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i} \right)^2$$

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

However typically need to perform non-linear minimization.

HEP practice:

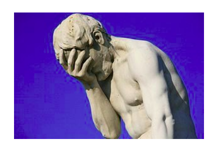
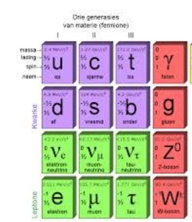
- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/... backends

  → Usual methods – gradient-based, etc.

Hypothesis Testing

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

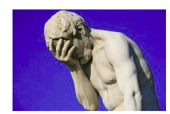
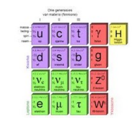
→ **Goal** : decide if H_0 is favored or disfavored using a test based on the data

Possible outcomes:	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	**Discovery!**  	**Missed discovery** 
H_0 is true (Nothing new)	**False discovery** 	**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

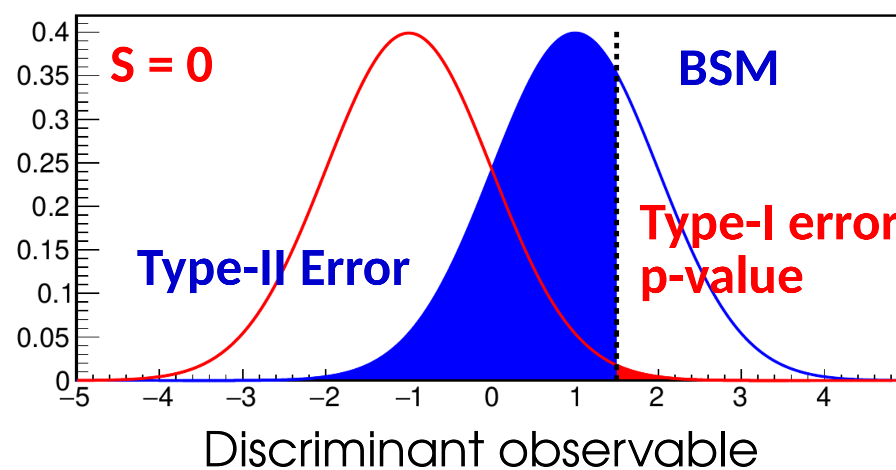
Hypothesis: assumption on model parameters, say value of S (e.g. **H_0 : S=0**)

	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** 

p-value, significance

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

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



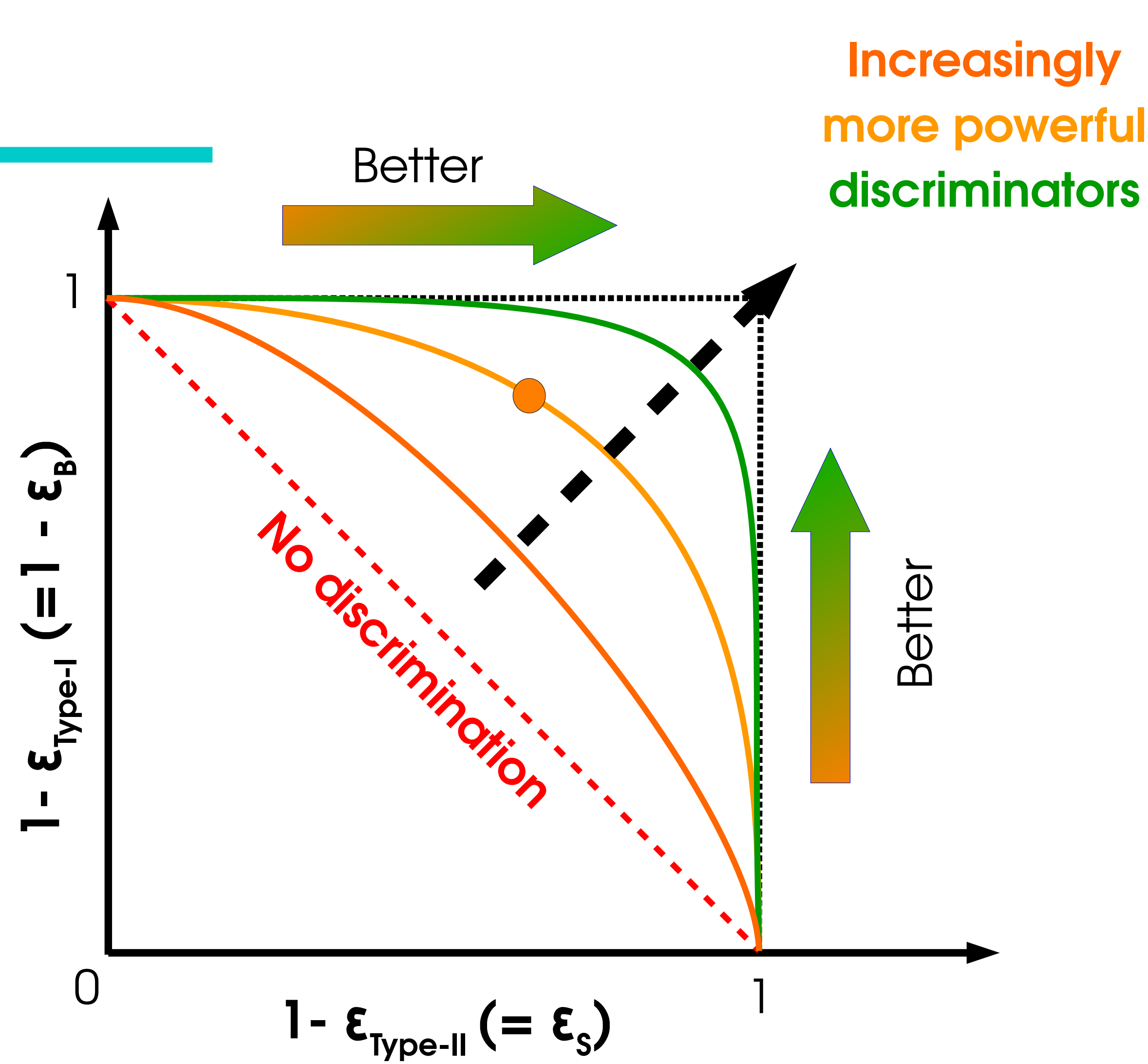
ROC Curves

“Receiver operating characteristic” (ROC) Curve:

→ Shows Type-I vs Type-II rates for different selections

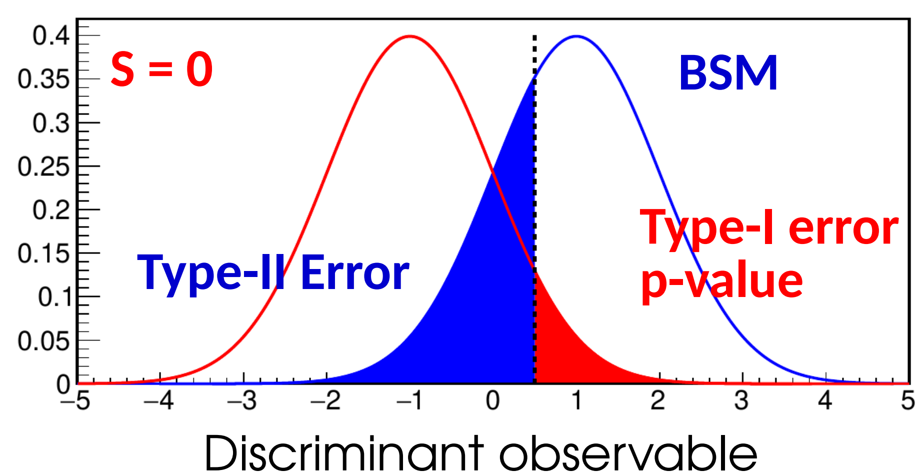
→ All curves monotonically decrease from (0,1) to (1,0)

→ 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”)



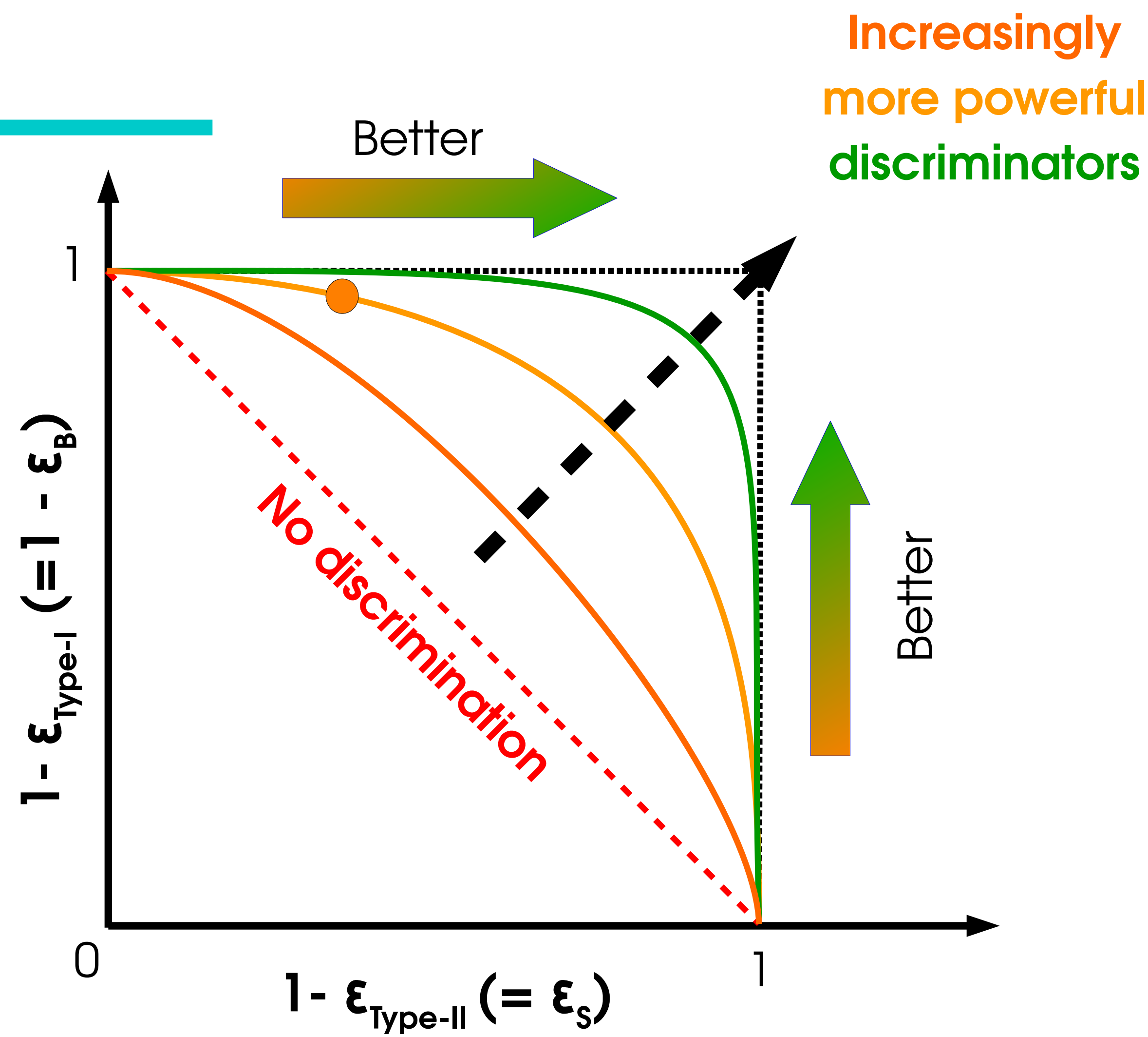
ROC Curves

"Receiver operating characteristic" (ROC) Curve:

→ Shows Type-I vs Type-II rates for different selections

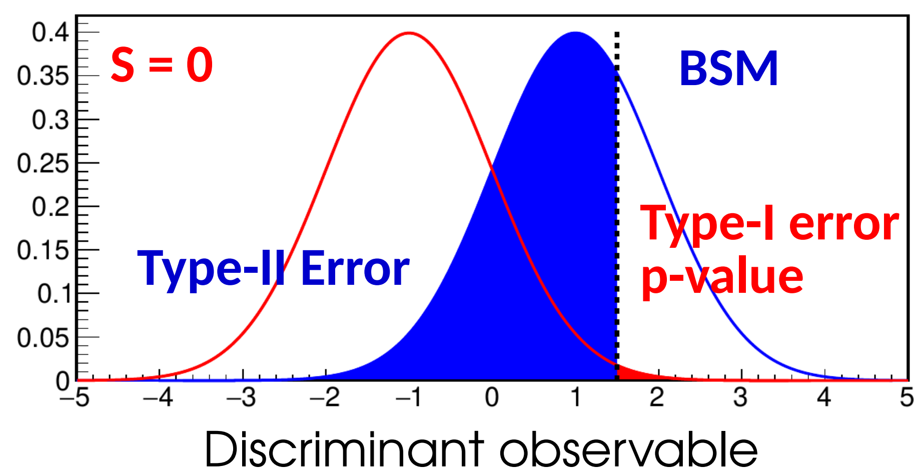
→ All curves monotonically decrease from (0,1) to (1,0)

→ 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")



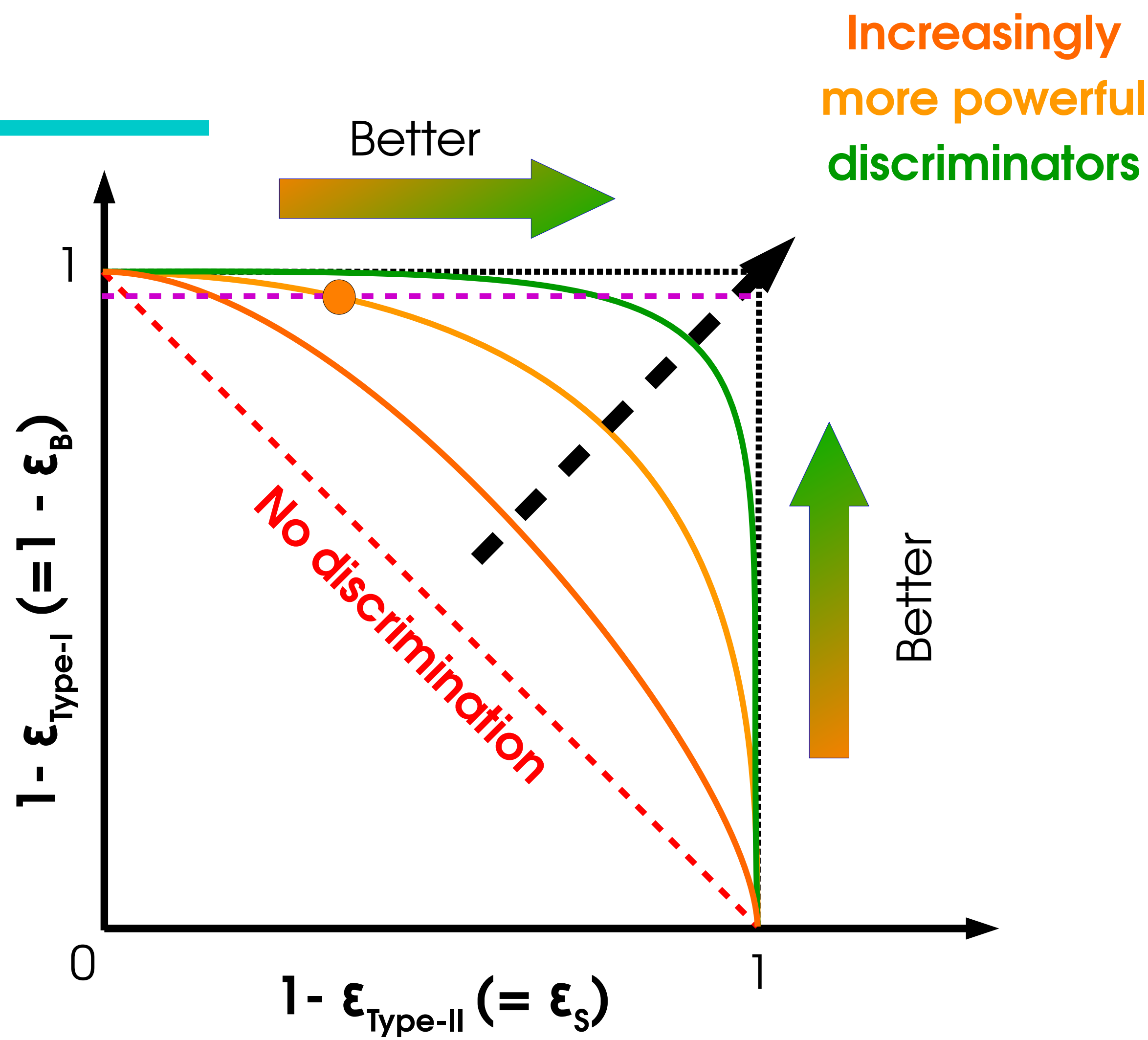
ROC Curves

“Receiver operating characteristic” (ROC) Curve:

→ Shows Type-I vs Type-II rates for different selections

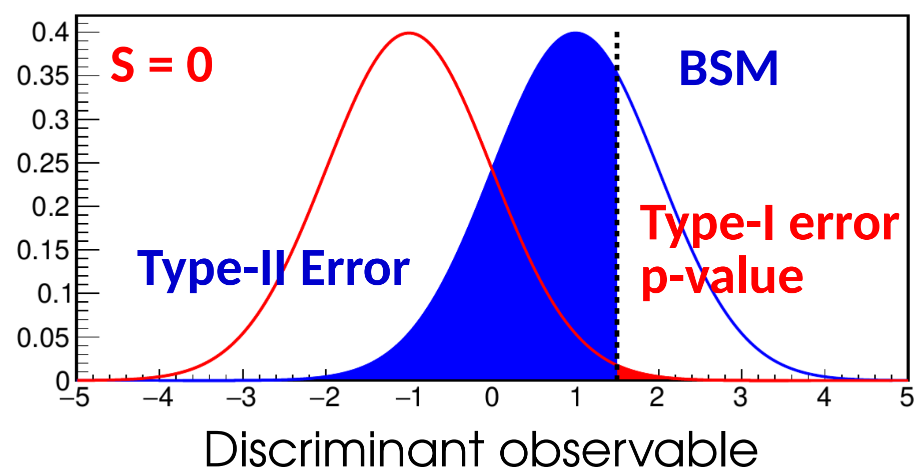
→ All curves monotonically decrease from (0,1) to (1,0)

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

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

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

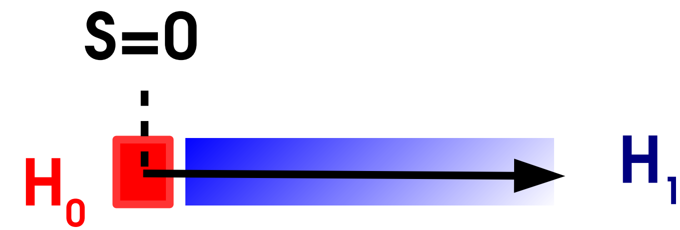
→ Always need an **alternate hypothesis** to test against the **null**.

→ **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, Ŝ.**

⇒ Use Likelihood ratio: $\frac{L(S=0)}{L(\hat{S})}$

→ In fact use the **test statistic** $q_0 = -2\log\frac{L(S=0)}{L(\hat{S})}$

Note: for Ŝ < 0, set q_0=0 to reject negative signals ("*one-sided* test statistic")

Discovery p-value

data prefer S = 0 ⟺ data prefer S > 0

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

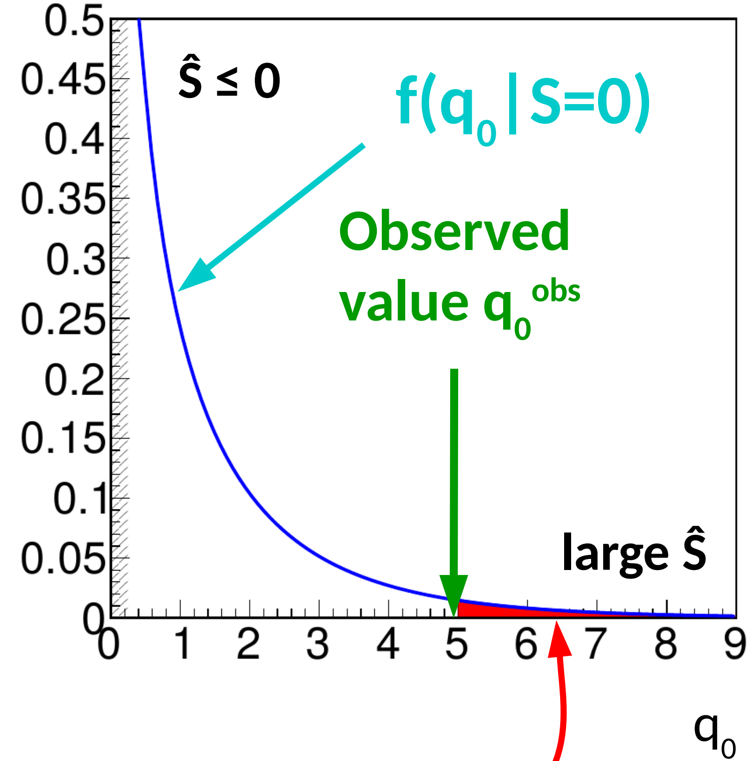
⇒ observed $\hat{S}$ is far from 0

⇒ $H_0(S=0)$ ***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 rejecting H_0)



→ Type-I error rate, a.k.a. the ***p-value :*** $p_0 = \int_{q_0^{obs}}^{\infty} f(q_0|S=0)\, dq_0$

= *Fraction of outcomes that are*

***At least as extreme** (signal-like) **as data**, when H_0 **is true** (no signal).*

Asymptotic distribution of q_0

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

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

Wilk's Theorem: **q_0 distributed as χ^2 (n_{par}) for $S = 0$**

⇒ **n_{par} = 1 : $\sqrt{q_0}$ is distributed as a Gaussian**

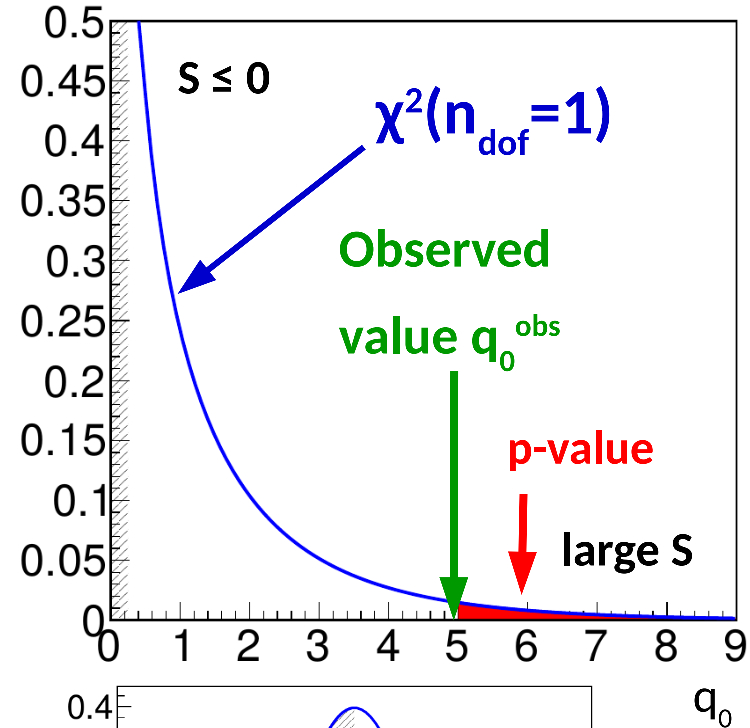
⇒ Can compute p-values from Gaussian quantiles

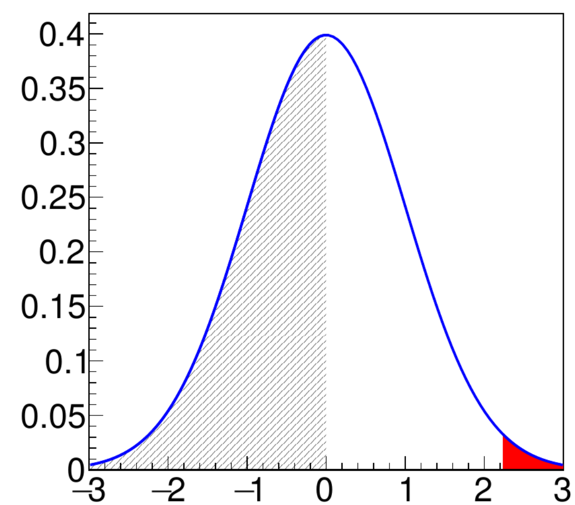
$$p_0 = 1 - \Phi\left(\sqrt{q_0}\right)$$

⇒ 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

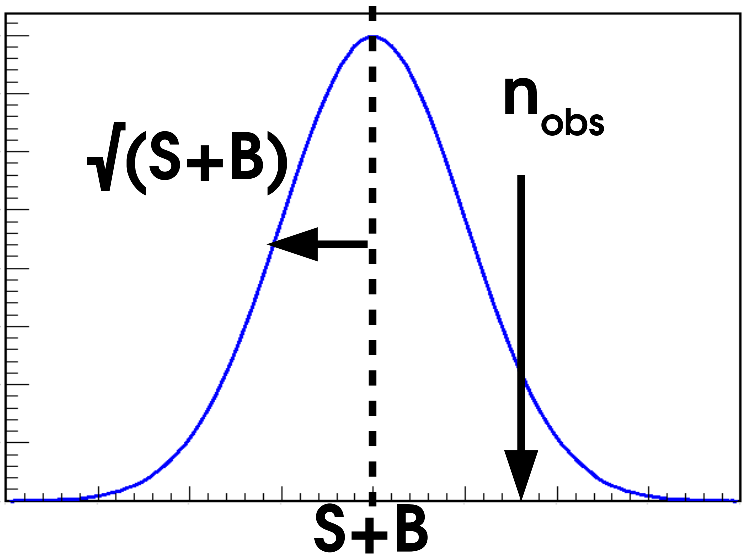
$$L(S) = \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}}{\sigma}\right)^2\right] \Rightarrow q_0 = \left(\frac{\hat{S}}{\sigma}\right)^2 \Rightarrow \sqrt{q_0} = \frac{\hat{S}}{\sigma} \sim G(0,1) \Rightarrow q_0 \sim \chi^2(n_{dof}=1)$$

Homework 1: Gaussian Counting

Count number of events n in data

→ Assume n large enough so process is Gaussian

→ Assume B is known, and we measure S



Likelihood : $$L(S;n_{\rm obs}) = e^{-\frac{1}{2}\left(\frac{n_{\rm obs}-(S+B)}{\sqrt{S+B}}\right)^2}$$

→ Find the best-fit value (MLE) Ŝ for the signal

(can use λ = -2 log L instead of L for simplicity)

→ Find the expression of q_0 for Ŝ > 0.

→ Find the expression for the significance

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

Homework 2: Poisson Counting

Same problem but now ***not*** assuming Gaussian behavior:

$$L(S;n) = e^{-(S+B)}(S+B)^n$$

(Can remove the n! constant since we're only dealing with L ratios)

→ As before, compute $\hat{S}$, and q_0

→ Compute $Z = \sqrt{q_0}$, assuming asymptotic behavior

Solution:

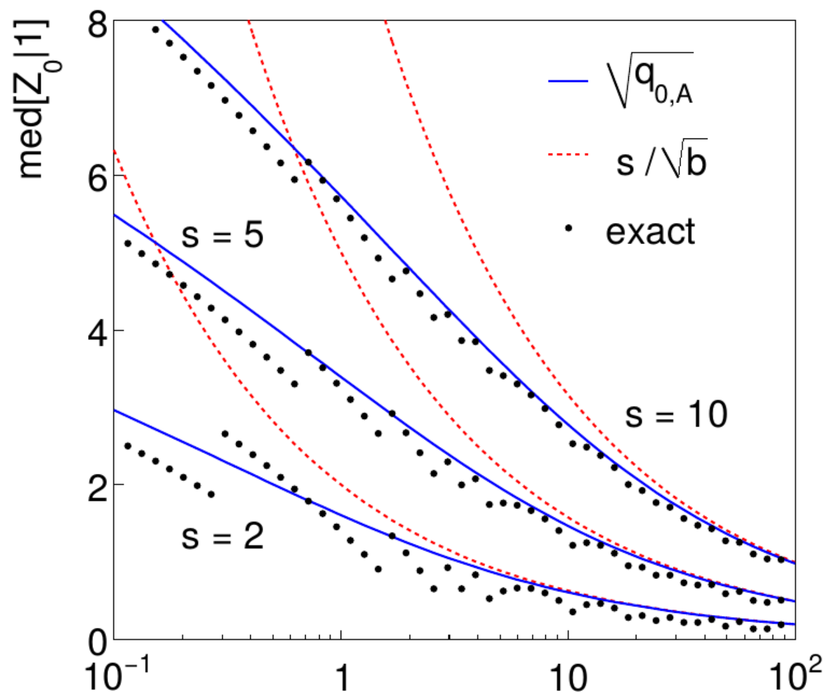
$$Z = \sqrt{2\left[(\hat{S}+B)\log\left(1+\frac{\hat{S}}{B}\right)-\hat{S}\right]}$$

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 (down to 5 events!)

Eur.Phys.J.C71:1554,2011



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

Discovery Thresholds

Evidence : **3σ ⇔ p_0 = 0.3% ⇔ 1 chance in 300**

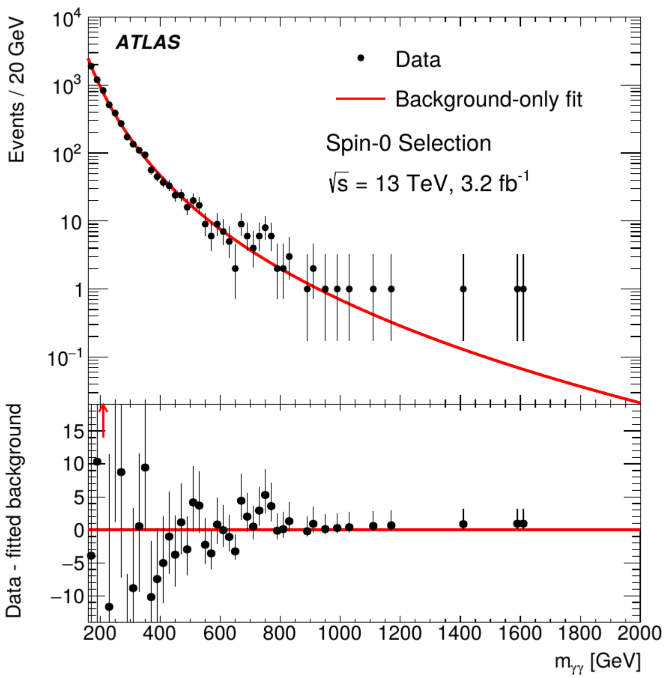
Discovery: 5σ ⇔ p_0 = 3 10^{-7} ⇔ 1 chance in 3.5M

Why so high thresholds ? (from Louis Lyons):

- **Look-elsewhere effect**: searches typically cover multiple independent regions ⇒ Higher chance to have a fluctuation "somewhere"

  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

Extraordinary claims require extraordinary evidence!

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

Confidence Intervals

Confidence Intervals

Last lecture we saw how to estimate (=compute) the value of a parameter

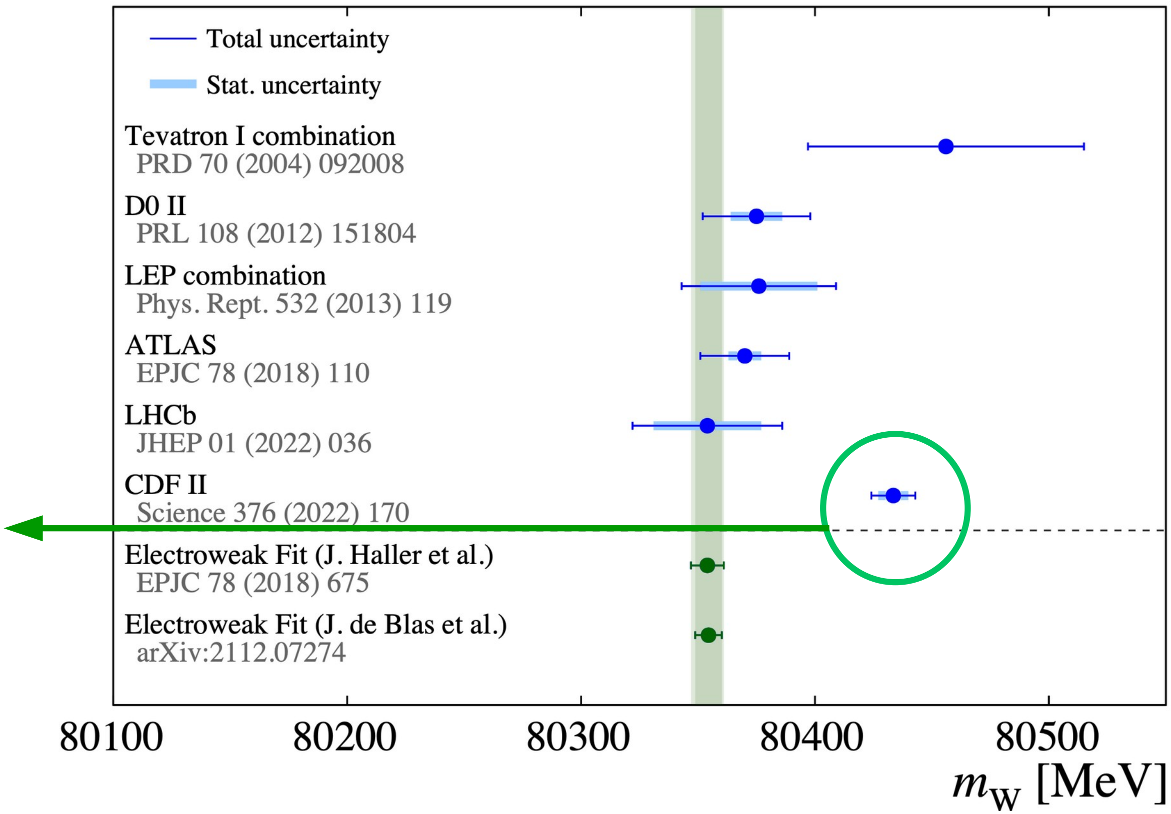
Maximum Likelihood Estimator (MLE) $\hat{\mu}$:

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

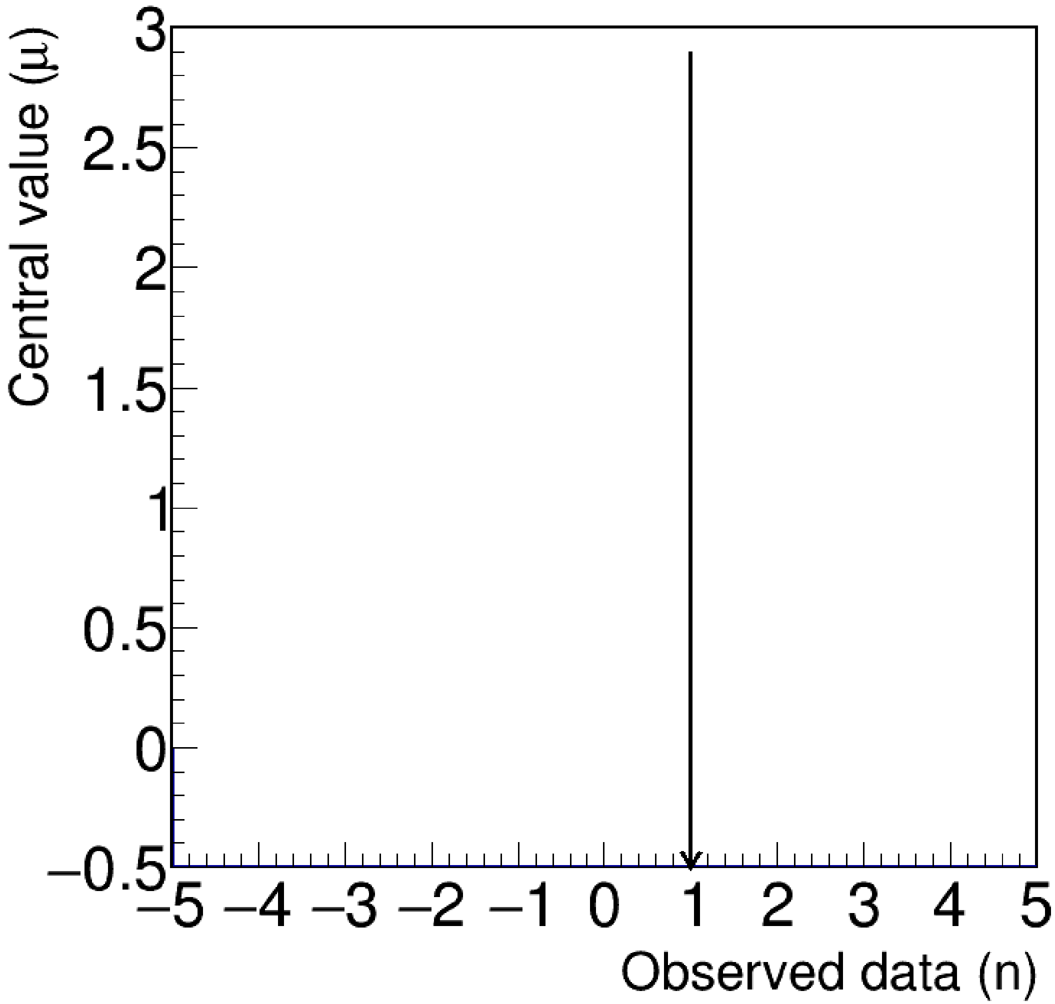
However we also need to estimate the associated uncertainty.

What is the meaning of an uncertainty ?

We don't know what the true value is, but **there is a 68% chance that it is within the error bar**



Gaussian confidence intervals



Consider a Gaussian likelihood:

$$L(\mu) = \exp\left[-\frac{1}{2}\left(\frac{n-\mu}{\sigma}\right)^2\right]$$

$$P(\mu - \sigma < n < \mu + \sigma) \geq 68.3\,\%$$

$$\Updownarrow$$

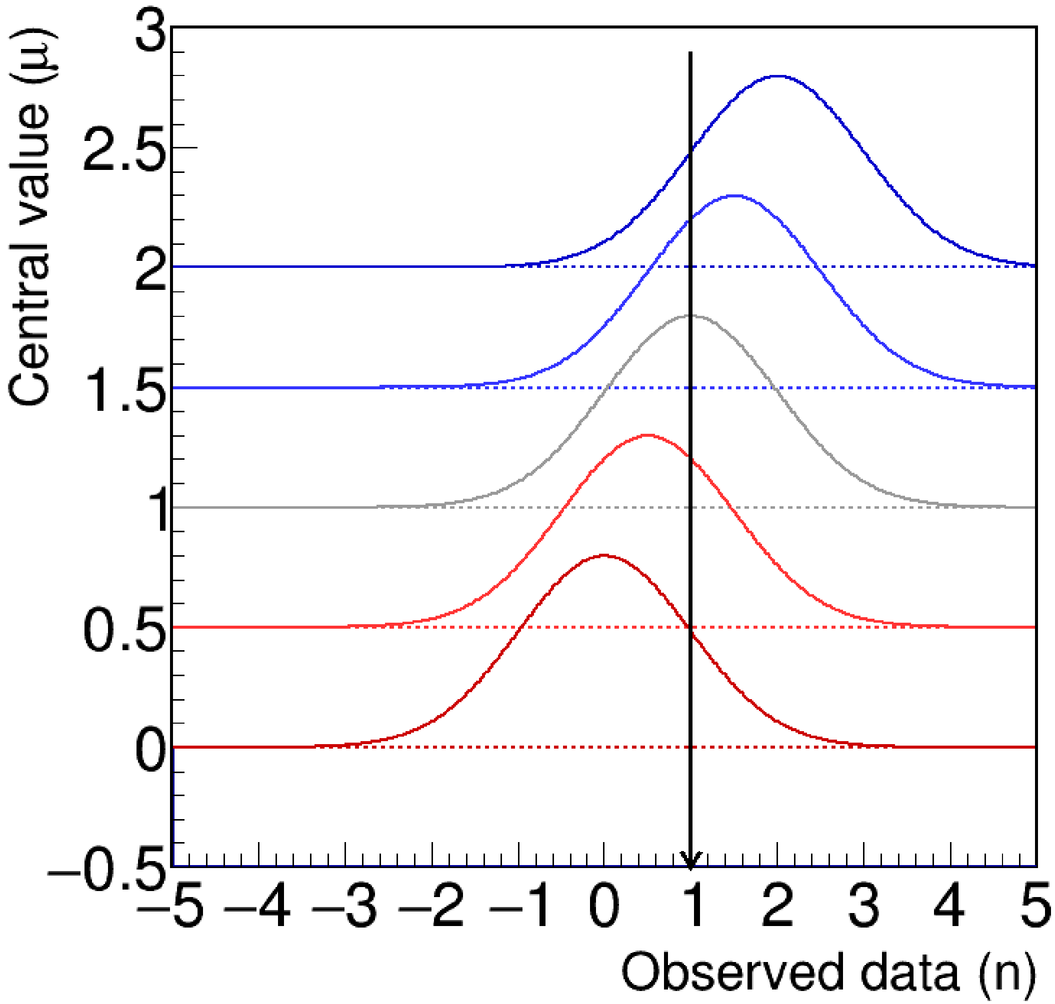
$$P(n - \sigma < \mu < n + \sigma) \geq 68.3\,\%$$

Still a statement on n!

μ = n ± σ at 68% CL ("1σ")

The reported interval n ± σ will contain the true value of μ 68.3% of the time

Gaussian confidence intervals



Consider a Gaussian likelihood:

$$L(\mu) = \exp\left[-\frac{1}{2}\left(\frac{n-\mu}{\sigma}\right)^2\right]$$

$$P(\mu - \sigma < n < \mu + \sigma) \geq 68.3\,\%$$

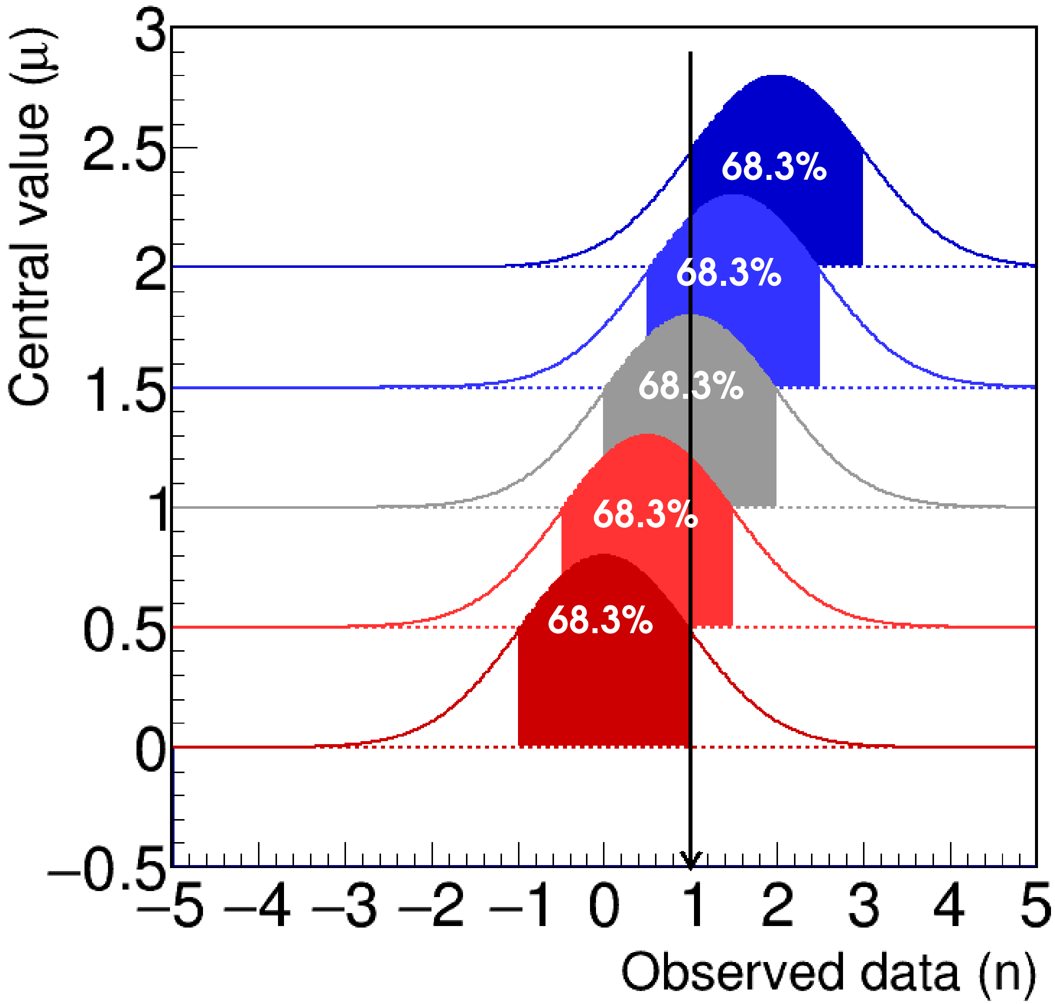
$$P(n - \sigma < \mu < n + \sigma) \geq 68.3\,\%$$

Still a statement on n!

μ = n ± σ at 68% CL ("1σ")

The reported interval n ± σ will contain the true value of μ 68.3% of the time

Gaussian confidence intervals



Consider a Gaussian likelihood:

$$L(\mu) = \exp\left[-\frac{1}{2}\left(\frac{n-\mu}{\sigma}\right)^2\right]$$

$$P(\mu - \sigma < n < \mu + \sigma) \geq 68.3\,\%$$

$$\Updownarrow$$

$$P(n - \sigma < \mu < n + \sigma) \geq 68.3\,\%$$

Still a statement on n!

μ = n ± σ at 68% CL ("1σ")

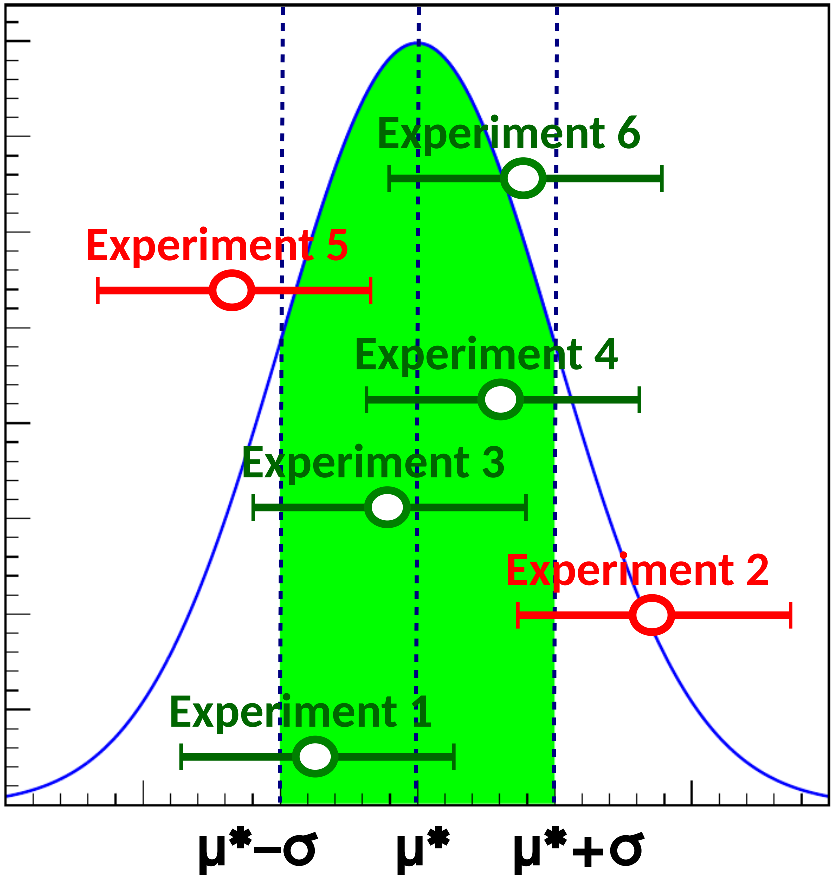
The reported interval n ± σ will contain the true value of μ 68.3% of the time

Gaussian confidence intervals

Frequentist interpretation

If we would repeat the same experiment multiple times, with true value μ^*, then 68.3% of the 1σ intervals would contain μ^*.

→ Crucially, this works even if we do not know μ^* !



For each experiment, get the interval

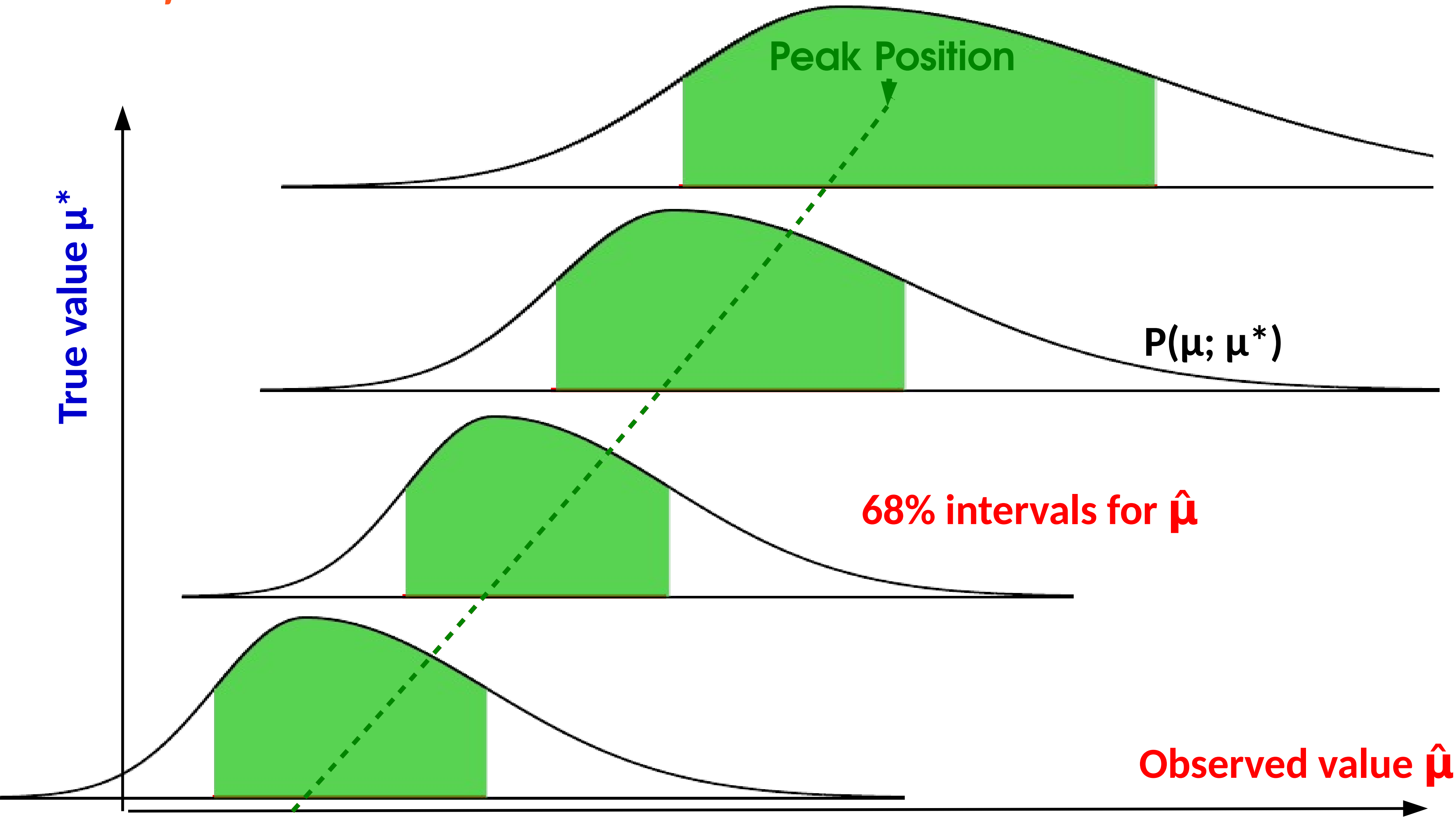
$\mu = n \pm \sigma$ at 68% CL ("1σ")

The reported interval $n \pm \sigma$ will contain the true value of μ 68.3% of the time

Neyman Construction

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

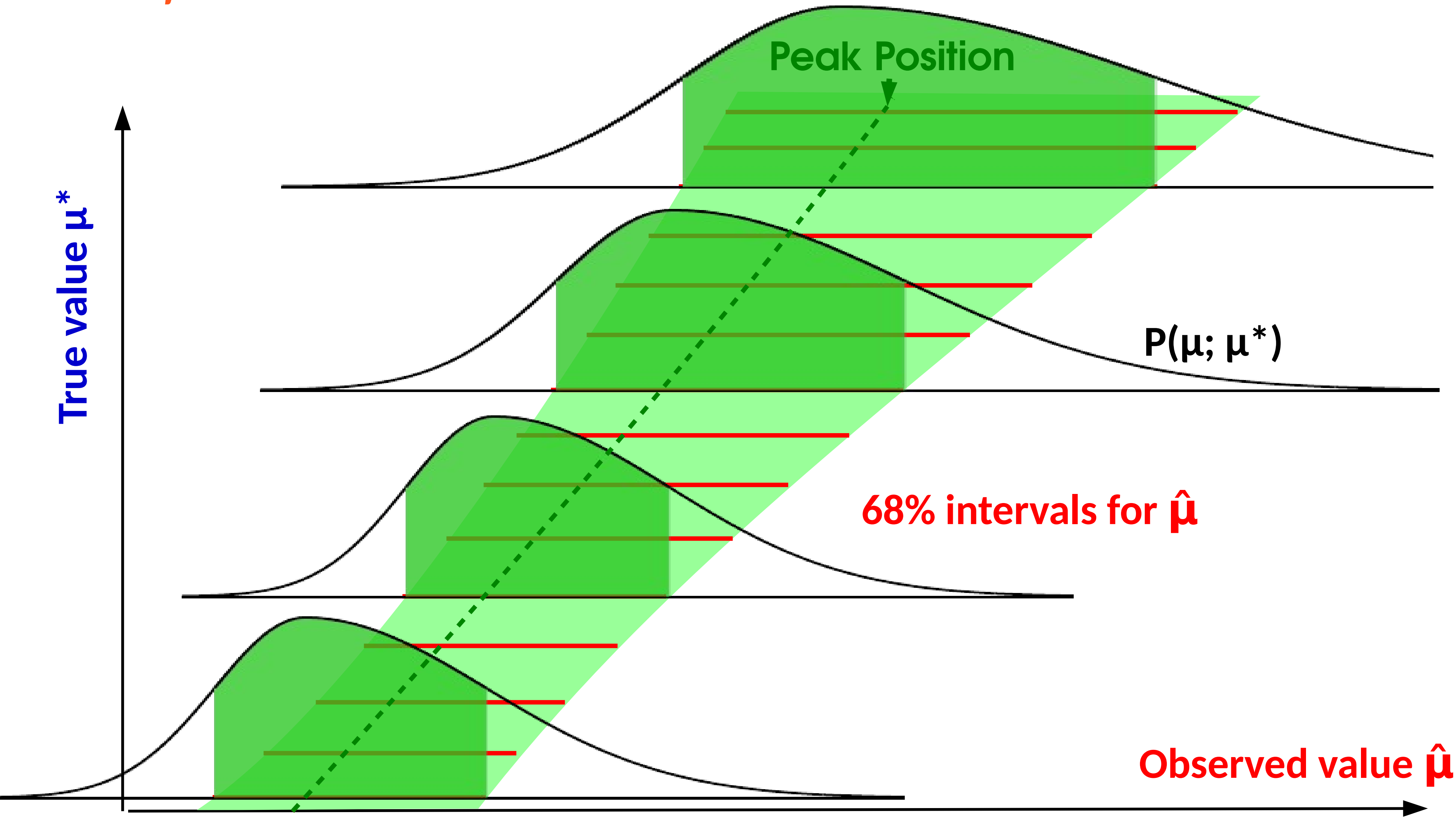
⇒ Confidence belt



Neyman Construction

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

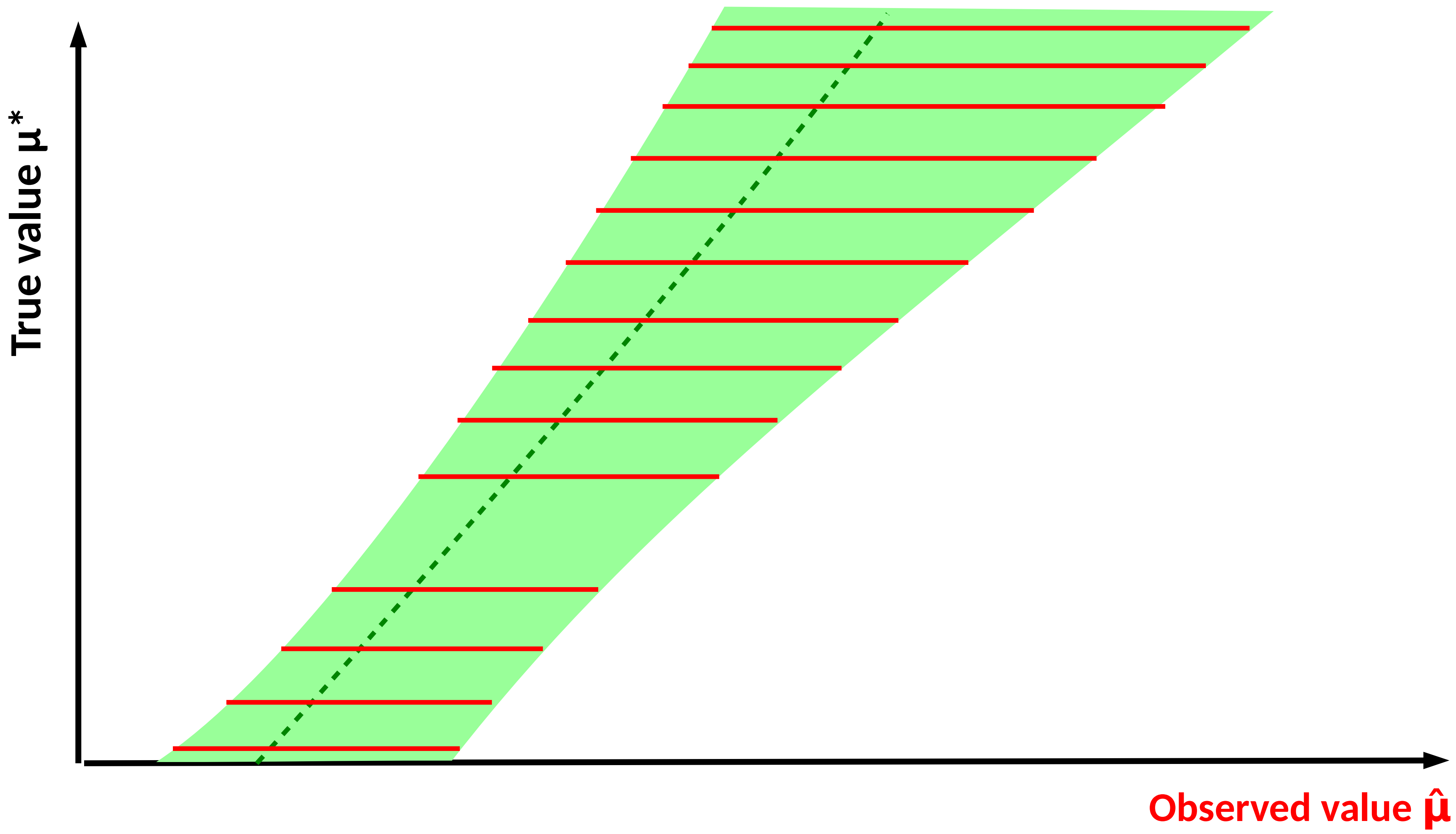
⇒ Confidence belt



Inversion using the Confidence Belt

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

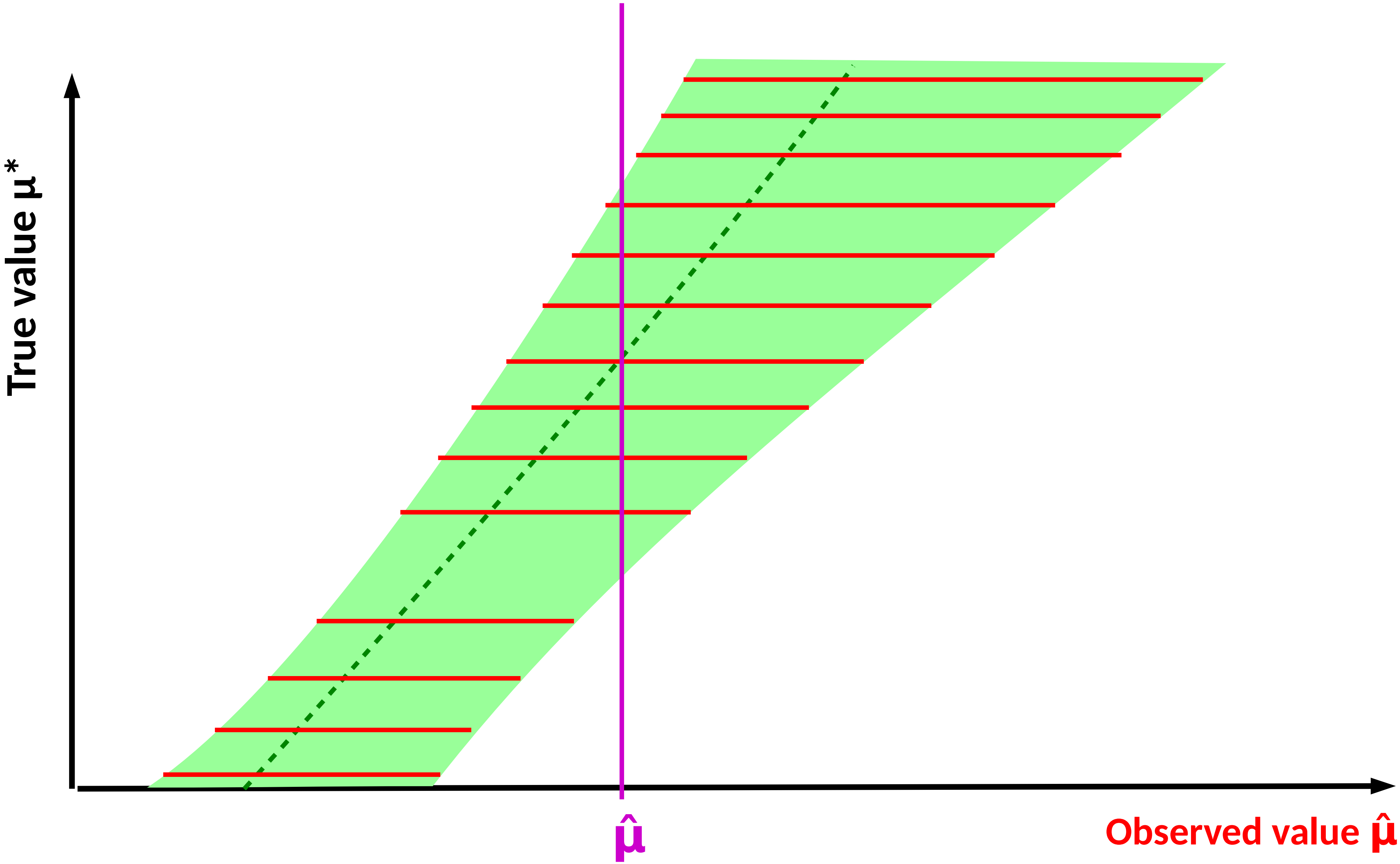
→ Same as before for Gaussian, works also when $P(\mu^{obs}|\mu)$ varies with μ.



Inversion using the Confidence Belt

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

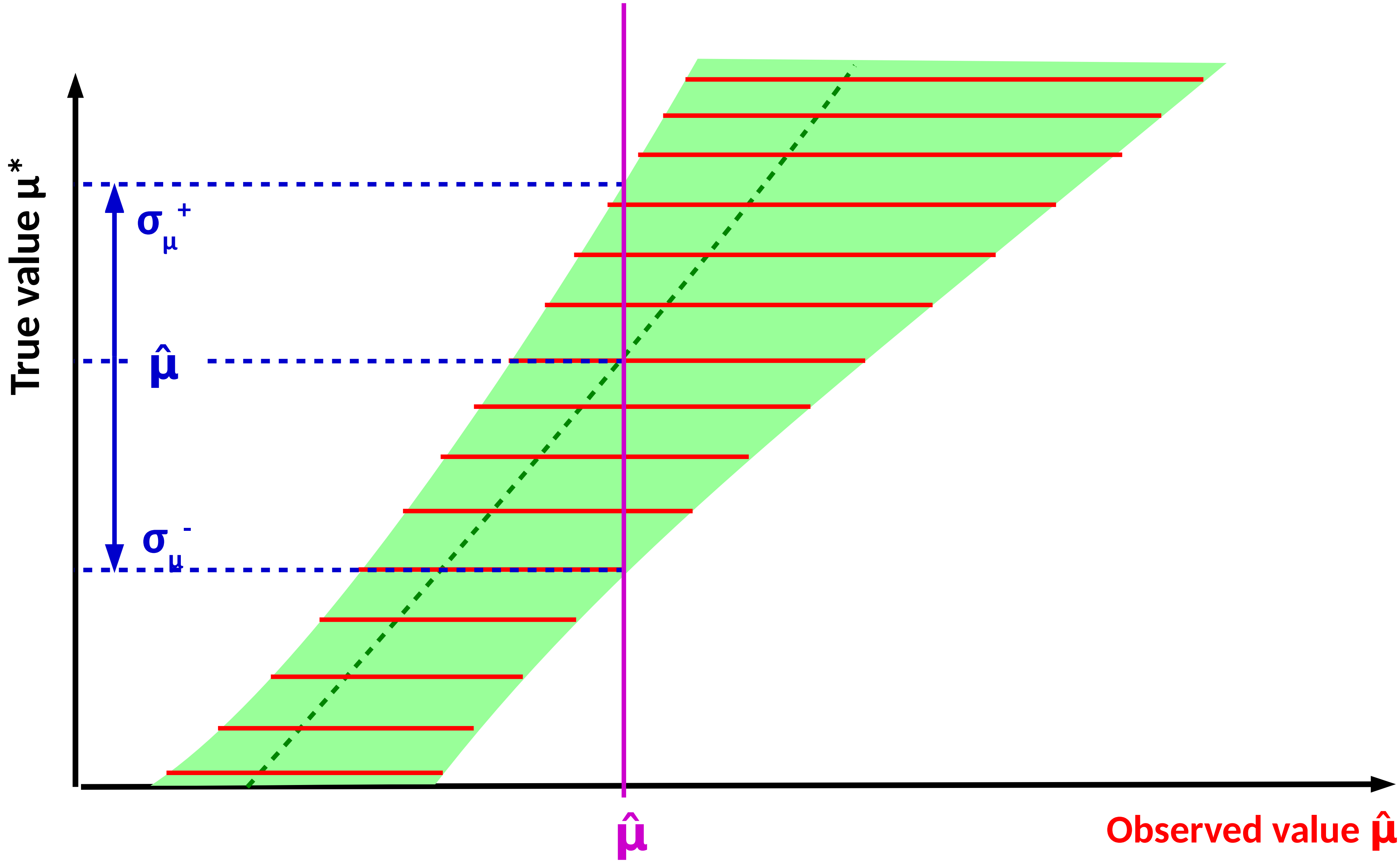
→ Same as before for Gaussian, works also when $P(\mu^{obs}|\mu)$ varies with μ.



Inversion using the Confidence Belt

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

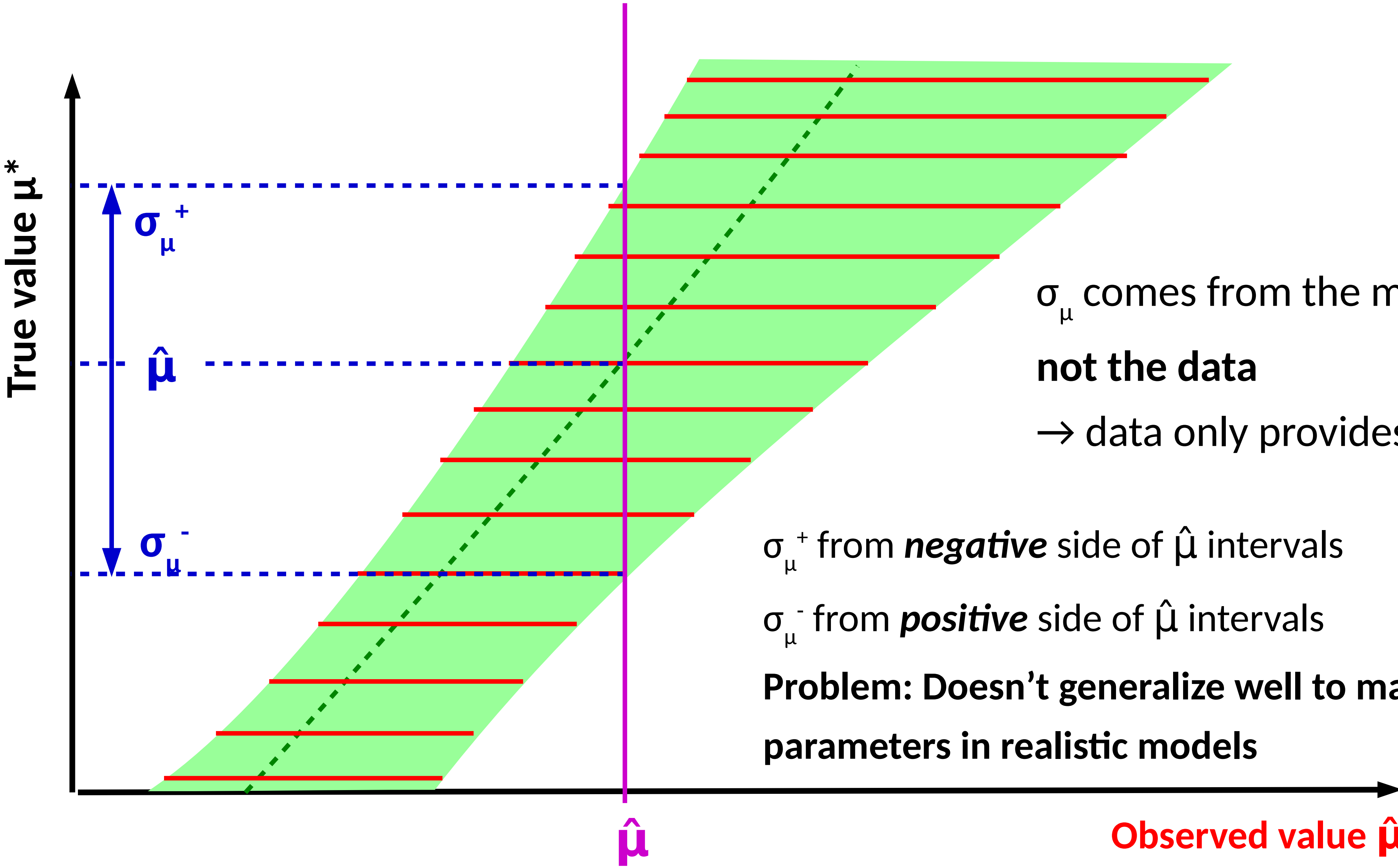
→ Same as before for Gaussian, works also when $P(\mu^{obs}|\mu)$ varies with μ.



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 $P(\mu^{obs}|\mu)$ 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

Problem: Doesn't generalize well to many parameters in realistic models

General case: Likelihood Intervals

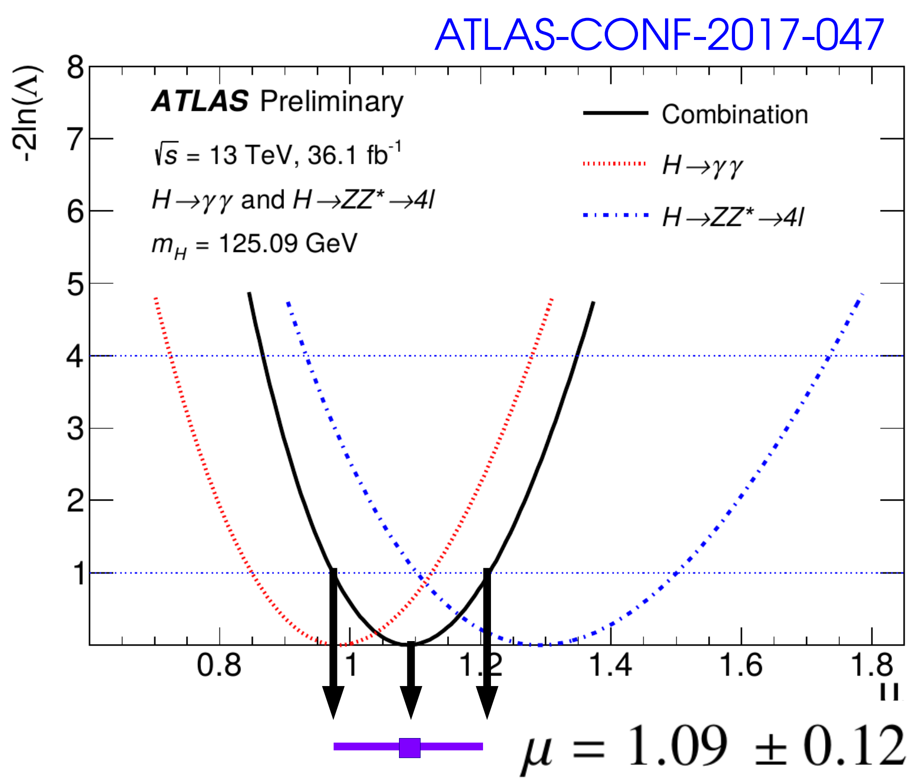
Confidence intervals from L(μ):

- Test various values μ using the **Profile Likelihood Ratio t(μ)**
- Minimum (=0) for μ=μ̂, rises away from μ̂.
- Good properties thanks to the Neyman-Pearson lemma.

Probability to observe the data for a given μ.

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

Probability to observe the data for best-fit μ̂.



Gaussian L(μ):

$$L(\mu) = \exp\left[-\frac{1}{2}\left(\frac{n-\mu}{\sigma}\right)^2\right]$$

$$t(\mu) = \left(\frac{n-\mu}{\sigma}\right)^2$$

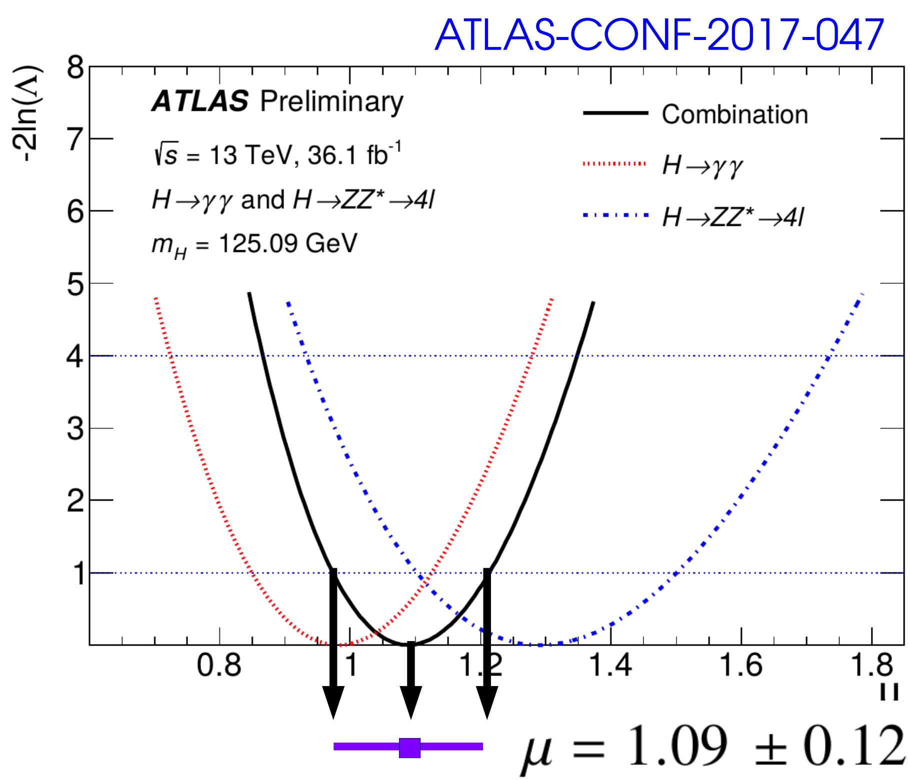
- t(μ) is parabolic, distributed as a χ^2
- Minimum occurs at **μ = μ̂**
- 1σ interval $[\mu_-, \mu_+]$ given by **$t(\mu_\pm) = 1$**

General case: Likelihood Intervals

Confidence intervals from L(μ):

- Test various values μ using the **Profile Likelihood Ratio t(μ)**
- Minimum (=0) for μ=μ̂, rises away from μ̂.
- Good properties thanks to the Neyman-Pearson lemma.

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



General case:

- Generally not a perfect parabola
- Minimum still at **μ = μ̂**

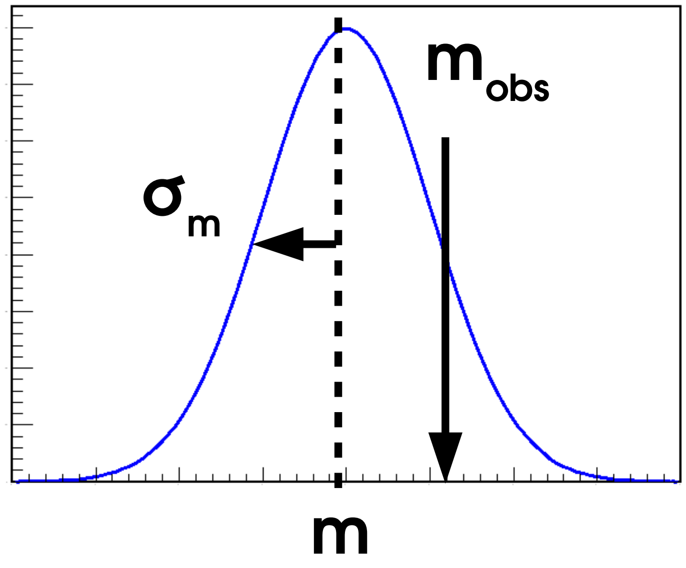
Asymptotic approximation

- Compute t(μ) using the exact L(μ)
- Assume t(μ) ~ χ^2 as for Gaussian (*"Wilks' Theorem"*)

1σ interval [μ_- , μ_+] given by $t(\mu_\pm)= 1$

Homework 3: Gaussian Case

Consider a parameter m (e.g. Higgs boson mass) whose measurement is Gaussian with known width σ_m, and we measure m_{obs}:



$$L(m; m_{obs}) = e^{-\frac{1}{2}\left(\frac{m - m_{obs}}{\sigma_m}\right)^2}$$

→ Compute the best-fit value (MLE) $\hat{m}$

→ Compute t_m

→ Compute the 1-σ (Z=1, ~68% CL) interval on m

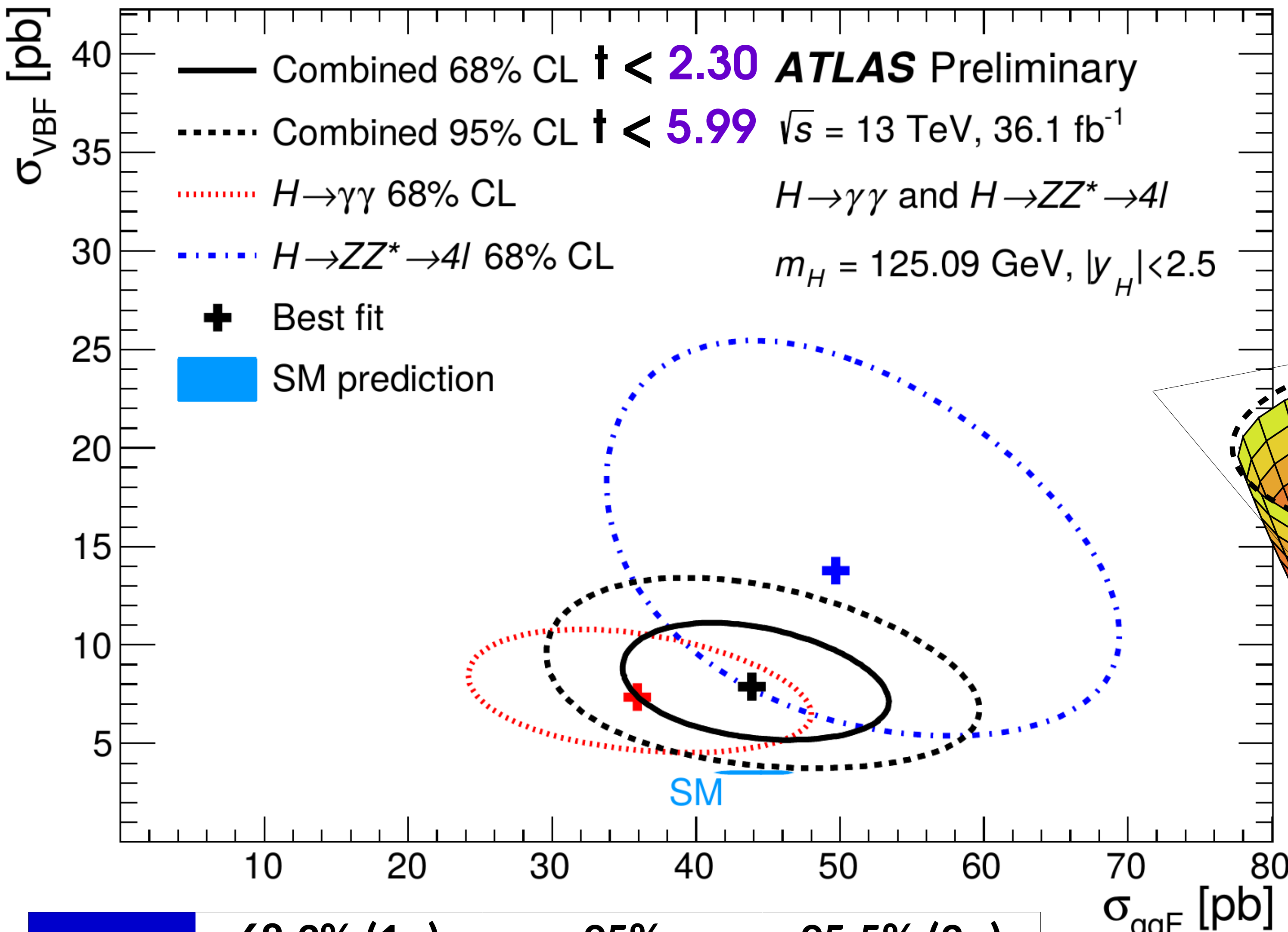
Solution: $m = m_{obs} \pm \sigma_m$

→ As expected!

→ General method can be applied in the same way to more complex cases

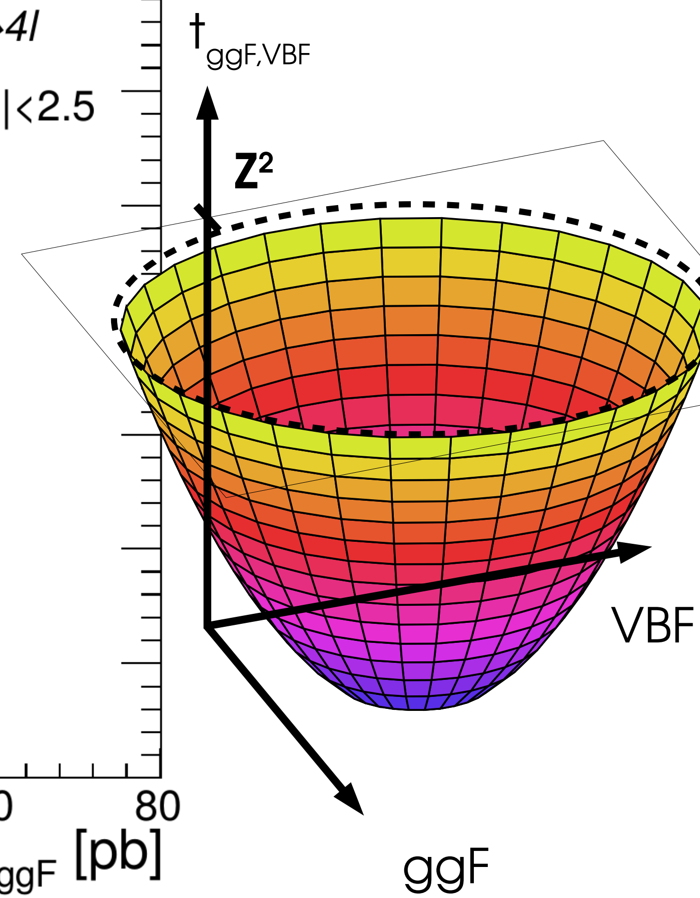
2D Example: Higgs σ_{VBF} vs. σ_{ggF}

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	68.3% (1σ)	95%	95.5% (2σ)
1D Z^2	1.00	3.84	4.00
2D Z^2	2.30	5.99	6.18

Gaussian case: elliptic paraboloid surface

Reparameterization

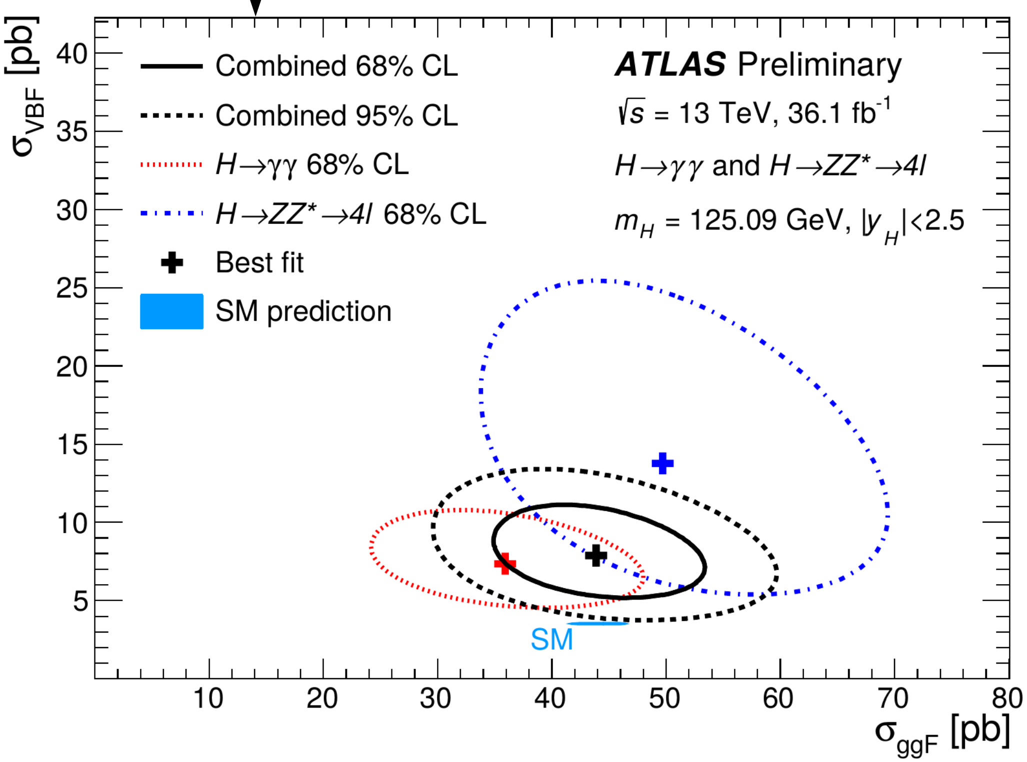
Start with basic measurement in terms of e.g. **σ×B**

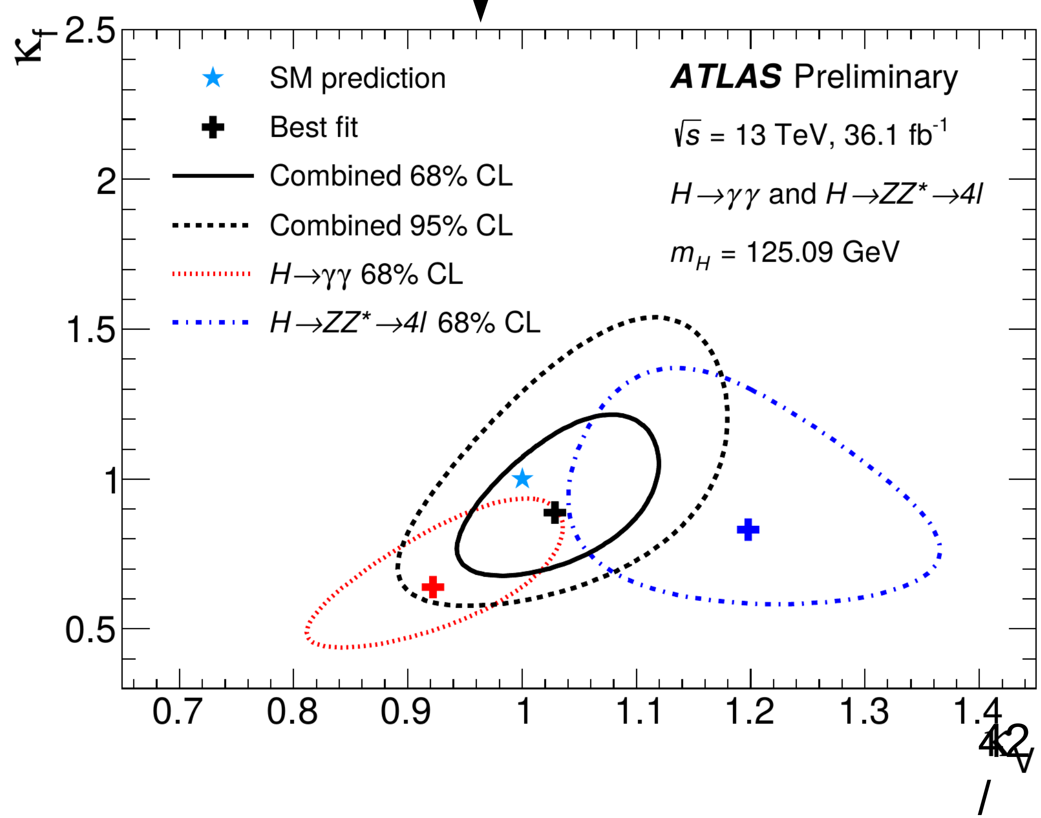
→ How to measure derived quantities (couplings, parameters in some theory model, etc.) ?

→ just reparameterize the likelihood:

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

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





Upper Limits

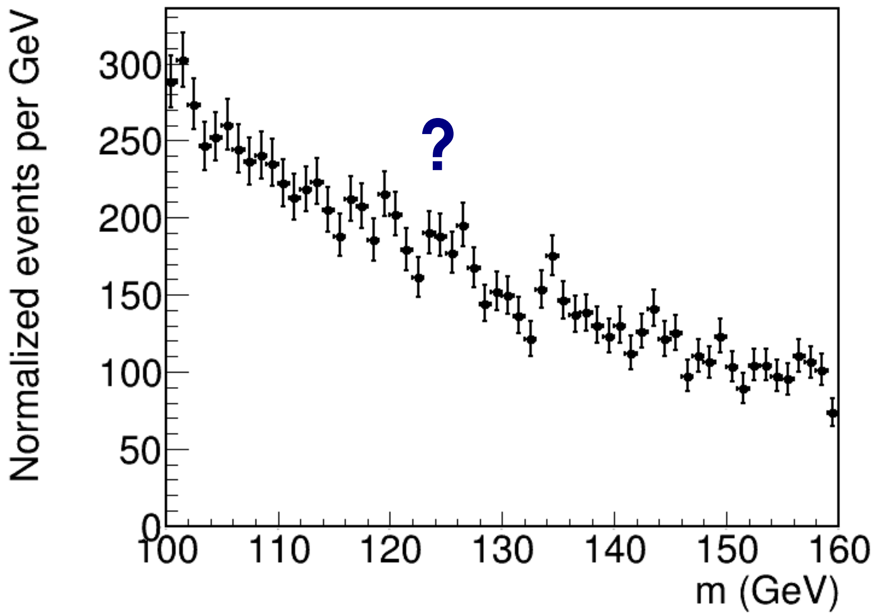
Hypothesis tests for Limits

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**

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



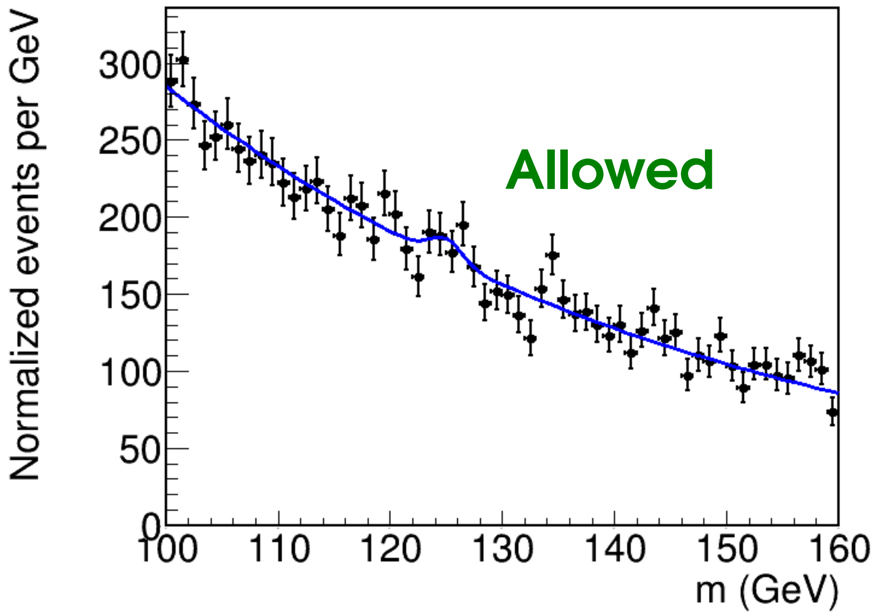
Hypothesis tests for Limits

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”



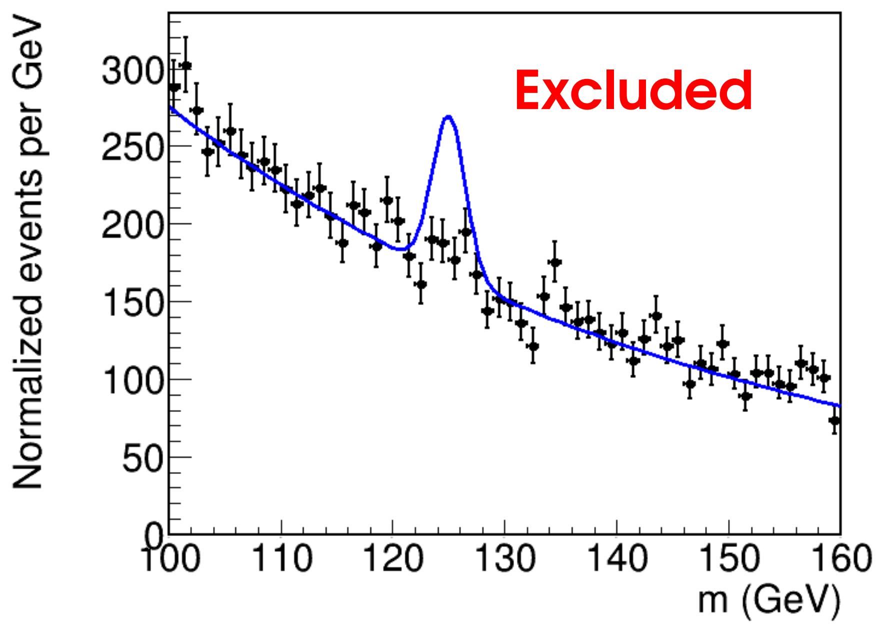
Hypothesis tests for Limits

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”



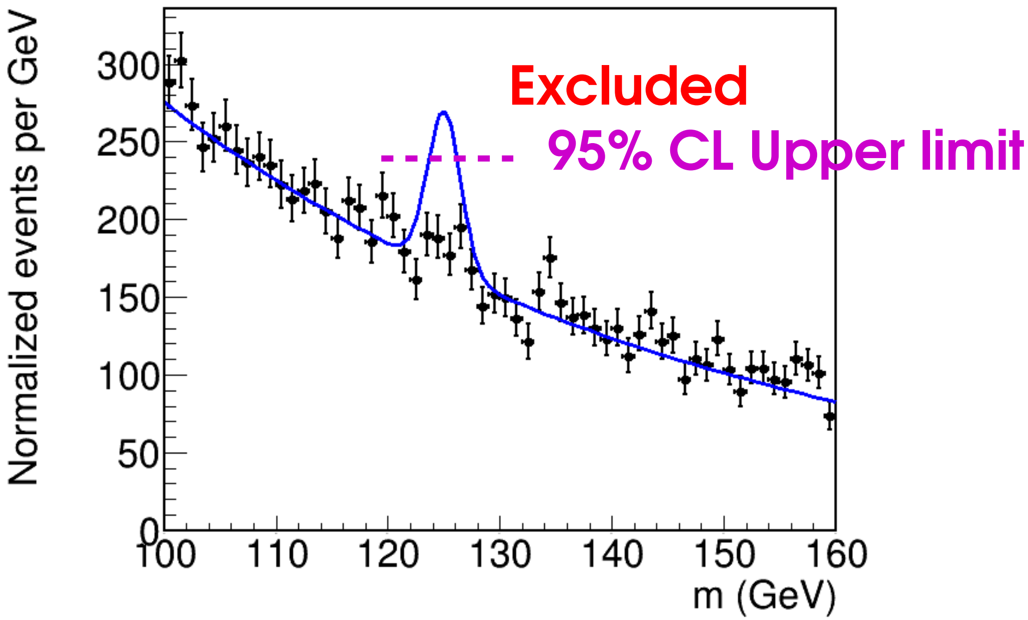
Hypothesis tests for Limits

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”

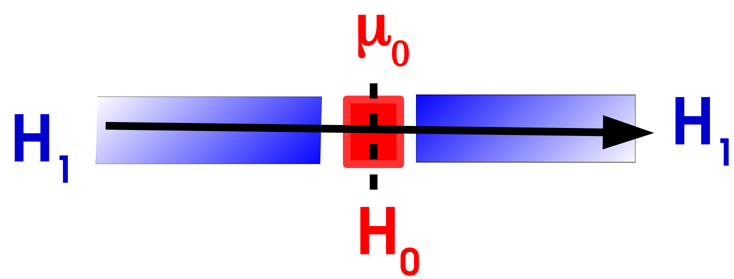


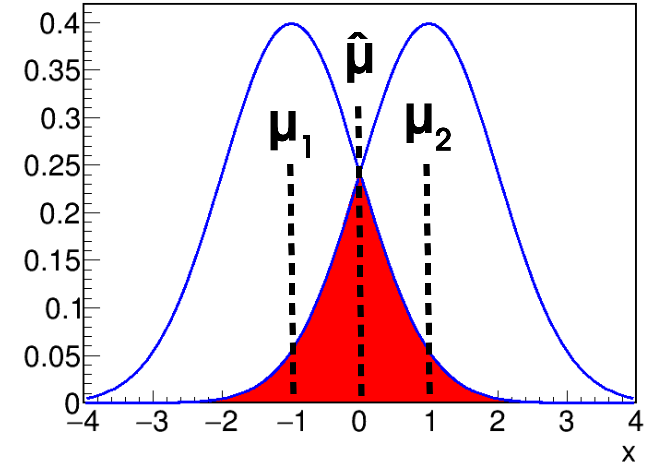
Test Statistics for Limit-Setting

Interval :

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$





Try to exclude μ values away from $\hat{\mu}$.

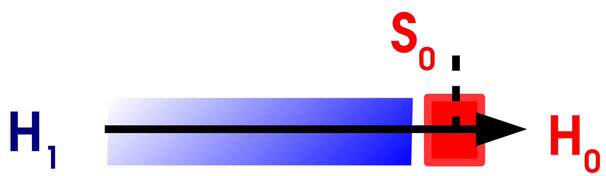
$$t(\mu_0) = -2\log \frac{L(\mu=\mu_0)}{L(\hat{\mu})}$$

"**Two-sided**" test

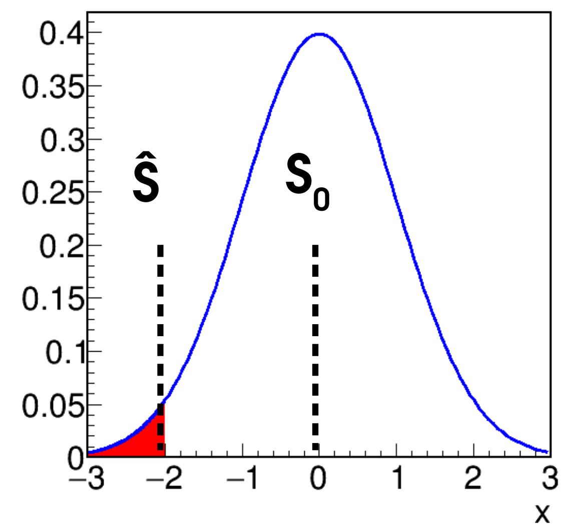
Limit-setting

$H_0 : S = S_0$

$H_1 : S < S_0$



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



Try to exclude values of S that are above $\hat{S}$.

⇒ "**One-sided**" test : only interested in excluding above

Discovery is also one-sided, for S>0 !

Inversion : Getting the limit for a given CL

Procedure:

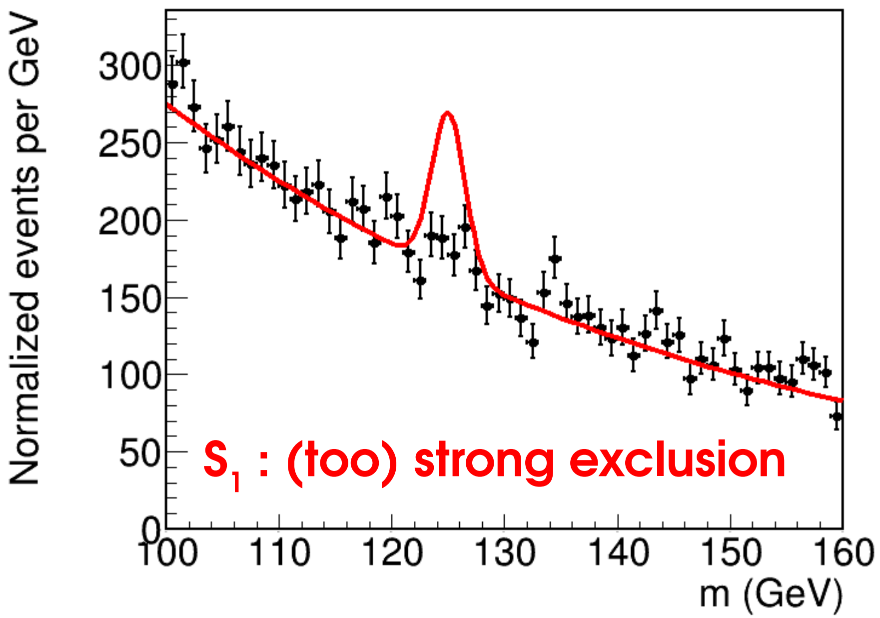
CL	p	Region
90%	10%	$\sqrt{q(S)} > 1.28$
95%	5%	$\sqrt{q(S)} > 1.64$
99%	1%	$\sqrt{q(S)} > 2.33$

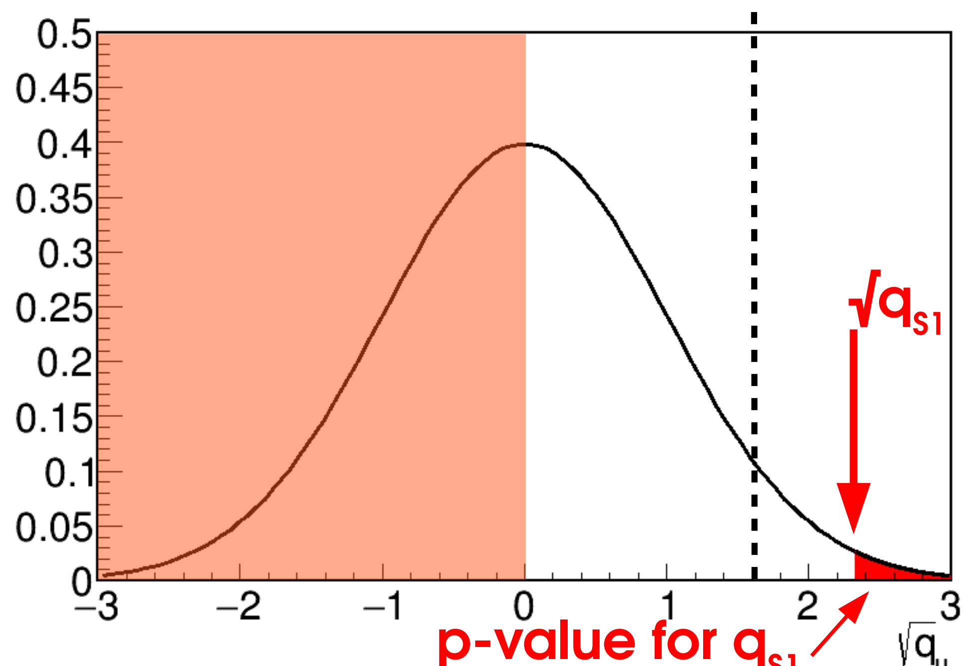
→ Compute $q(S_0)$ for some S_0, get the exclusion p-value $p(S_0)$.

Asymptotics: $$p(S_0) = 1 - \Phi\left(\sqrt{q(S_0)}\right)$$

→ **Adjust S_0** to get the desired exclusion

Asymptotics: need $\sqrt{q(S_{95})}$ **= 1.64** for **95% CL**





Inversion : Getting the limit for a given CL

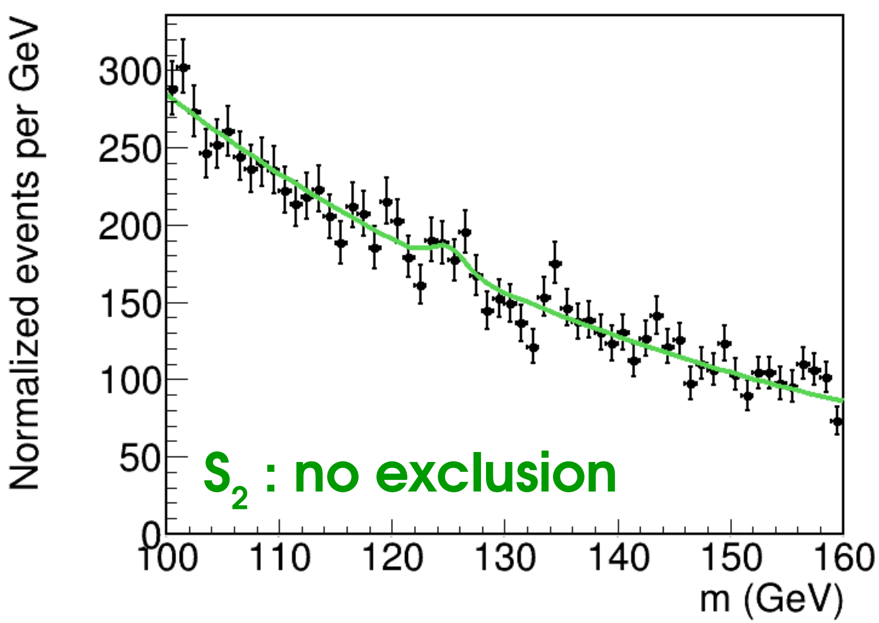
Procedure:

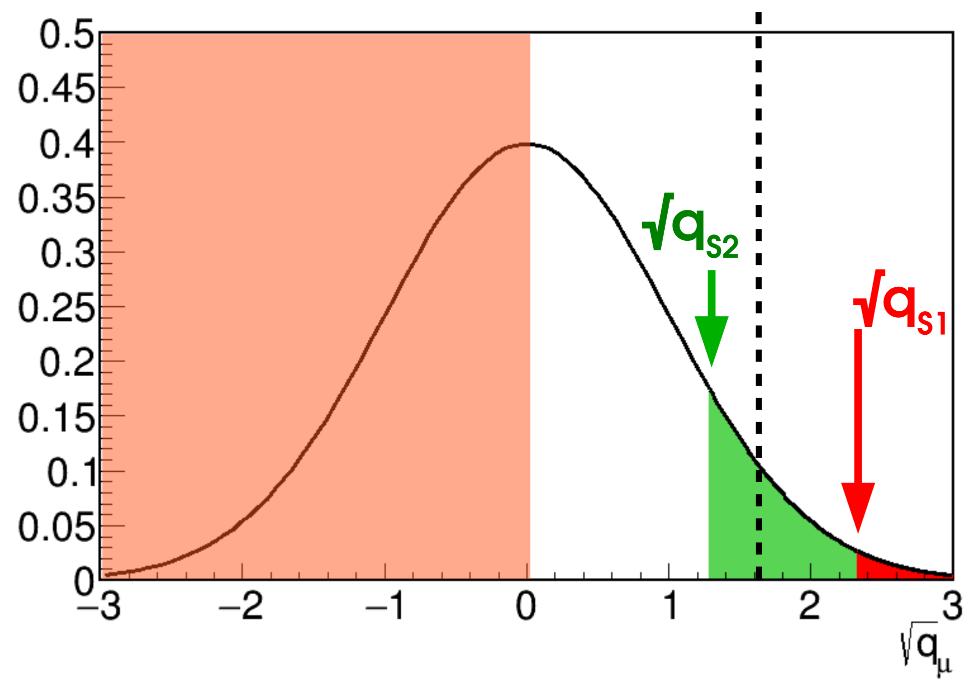
→ Compute $q(S_0)$ for some S_0,
get the exclusion p-value $p(S_0)$.

Asymptotics: $$p(S_0) = 1 - \Phi\left(\sqrt{q(S_0)}\right)$$

CL	p	Region
90%	10%	$\sqrt{q}(S) > 1.28$
95%	5%	$\sqrt{q}(S) > 1.64$
99%	1%	$\sqrt{q}(S) > 2.33$

→ **Adjust S_0** to get the desired exclusion
Asymptotics: need $\sqrt{q}(S_{95}) = 1.64$ for **95% CL**





Inversion : Getting the limit for a given CL

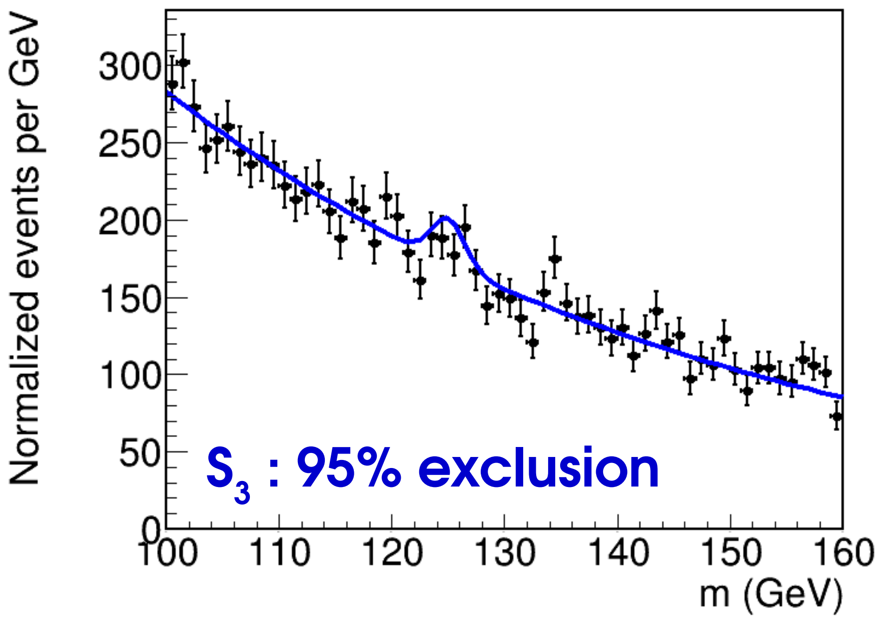
Procedure:

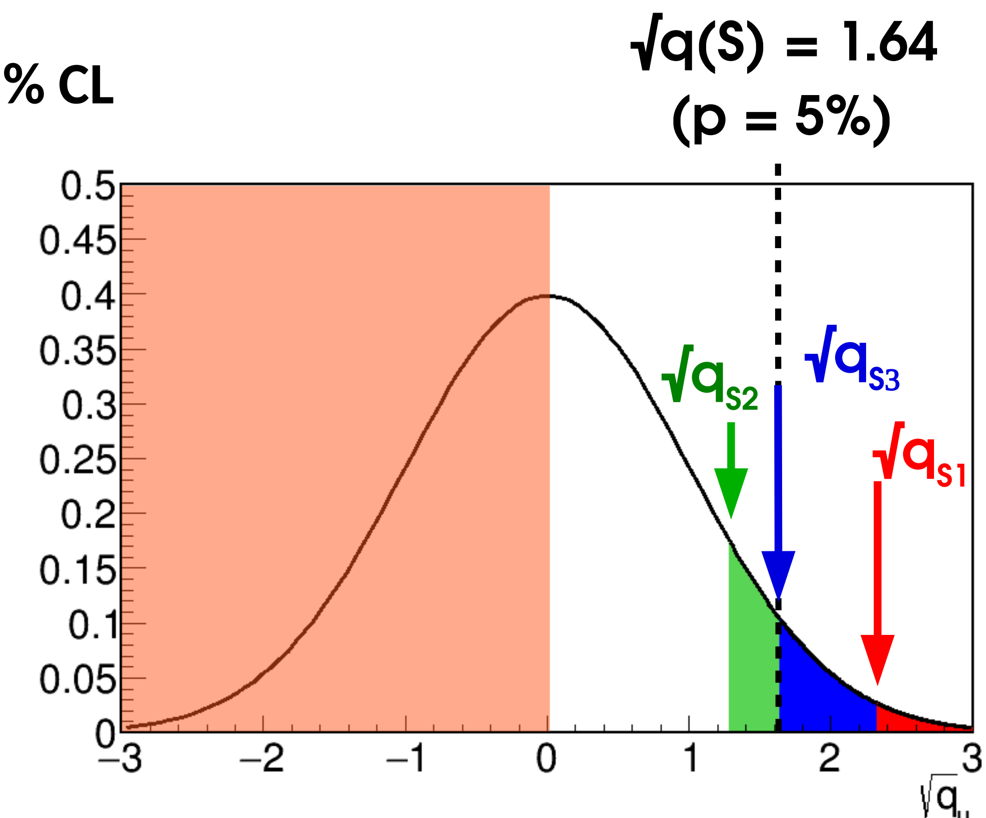
→ Compute $q(S_0)$ for some S_0,
get the **exclusion p-value $p(S_0)$**.

Asymptotics: $$p(S_0) = 1 - \Phi\left(\sqrt{q(S_0)}\right)$$

CL	p	Region
90%	10%	$\sqrt{q(S)} > 1.28$
95%	5%	$\sqrt{q(S)} > 1.64$
99%	1%	$\sqrt{q(S)} > 2.33$

→ **Adjust S_0** to get the desired exclusion
Asymptotics: need $\sqrt{q(S_{95})} = 1.64$ for **95% CL**



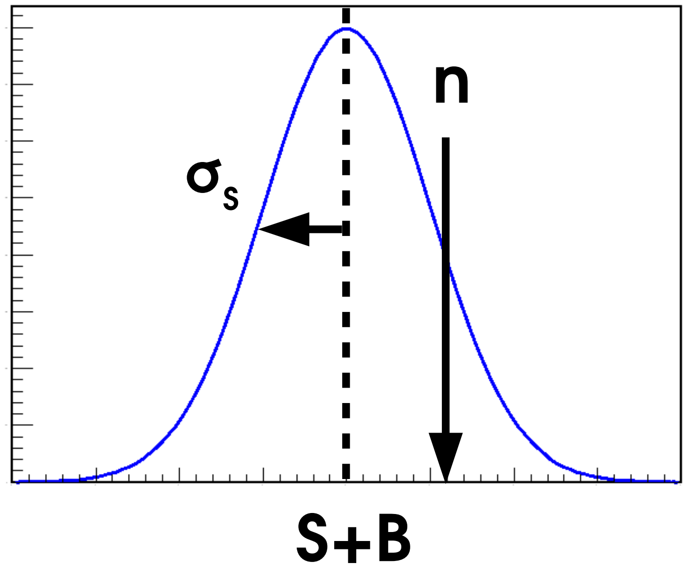


Homework 4: Gaussian Example

Usual Gaussian counting example with known B:

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

$\sigma_S \sim \sqrt{B}$ for small S



Reminder: Significance: $Z = \hat{S}/\sigma_S$

→ Compute q_{S0}

→ Compute the 95% CL upper limit on S, S_{up}, by solving $\sqrt{q_{S0}} = 1.64$.

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

CL_s

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

Upper limits sometimes take negative values (exclude all S>0 !)

Known feature – to avoid, usual solution in HEP is to use **CL_s** ”modified p-value”

$$p_{CL_s} = \frac{p(S_0)}{p_B}$$

Usual p-value for $S=S_0$

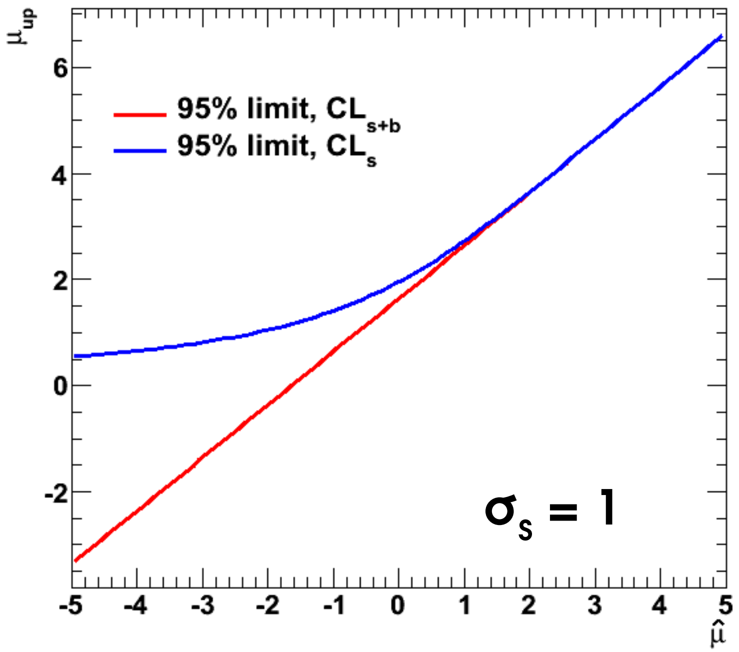
P-value for S=0

⇒ Compute exclusion relative to that of S=0

→ Somewhat ad-hoc, but good properties…

$\hat{S} \sim 0 \Rightarrow$ $p_B \sim O(1)$, p_{CLs} **~ $p(S_0)$ no change**

$\hat{S} \ll 0 \Rightarrow$ $p_B \ll 1$, $\mathbf{p_{CLs} \gg p(S_0)}$ **no exclusion at S=0**



Drawback: ***overcoverage***

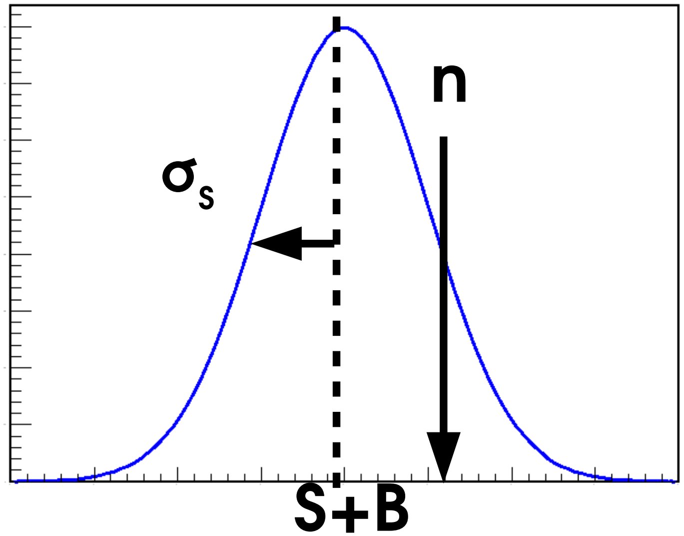
→ limit is claimed to be 95% CL, but actually >95% CL for small p_B.

Homework 5: CL_s : Gaussian Case

Usual Gaussian counting example with known B:

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

$\sigma_S \sim \sqrt{B}$ for small S



Reminder

CL_{s+b} limit: $$S_{up} = \hat{S} + 1.64\,\sigma_S \text{ at } 95\,\% \text{ CL}$$

CL_s upper limit :

→ Compute p_{S0} (same as for CLs+b)

→ Compute $1-p_B$ (hard!)

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

for $\hat{S} \sim 0$, $$S_{up} = \hat{S} + 1.96\,\sigma_S \text{ at } 95\,\% \text{ CL}$$

Homework 6: CL_s Rule of Thumb for n_{obs}=0

Same exercise, for the Poisson case with n_{obs} = 0. Perform an exact computation of the 95% CLs upper limit based on the definition of the p-value:

p-value : *sum probabilities of cases at least as extreme as the data*

Hint: for n_{obs}=0, there are no “more extreme” cases (cannot have n<0 !), so

p_{S0} = Poisson(n=0 | S_0+B) and 1 - p_B = Poisson(n=0 | B)

Solution: $$S_{up}(n_{obs}=0) = \log(20) = 2.996 \approx 3$$

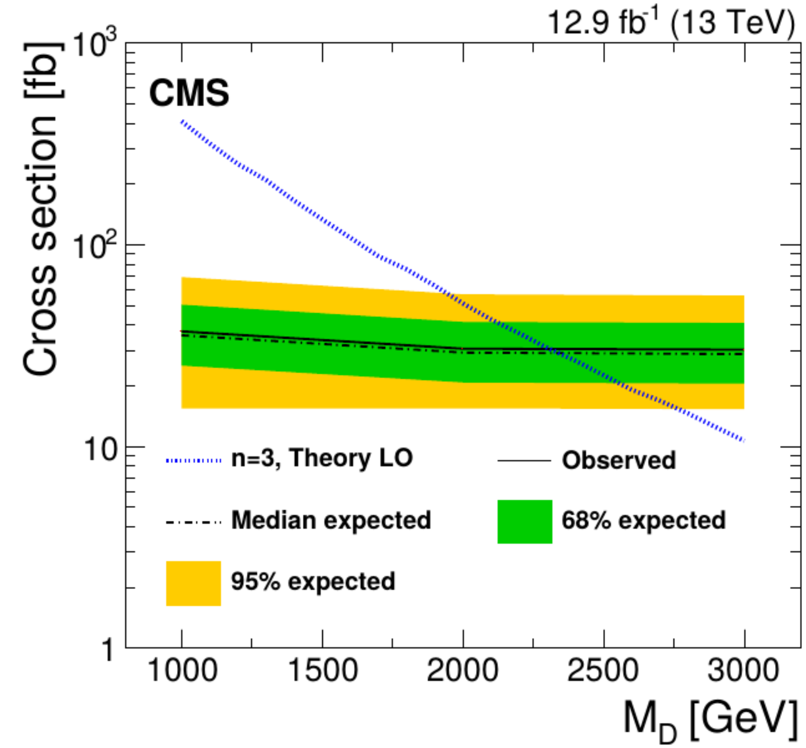
⇒ Rule of thumb: when n_{obs} = 0, the 95% CL_s limit is 3 events (for any B)

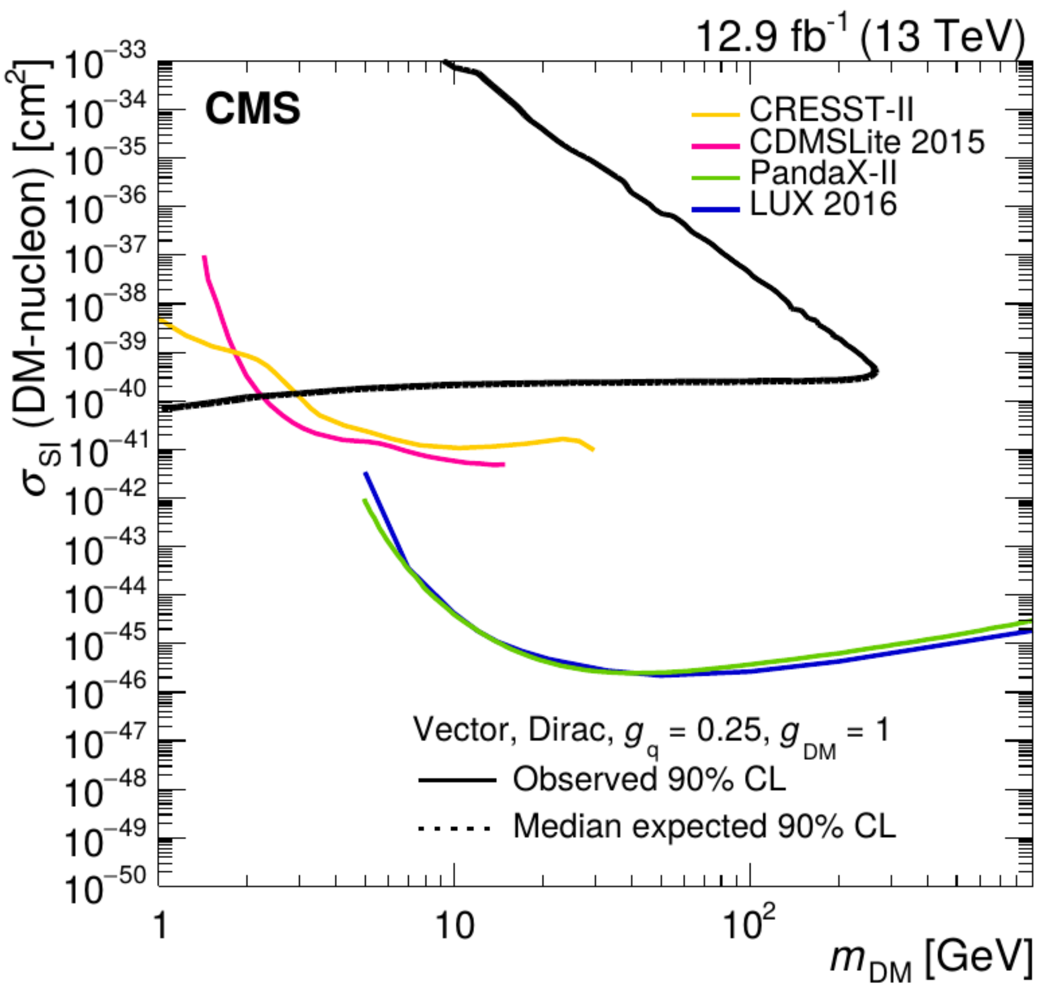
Reparameterization: Limits

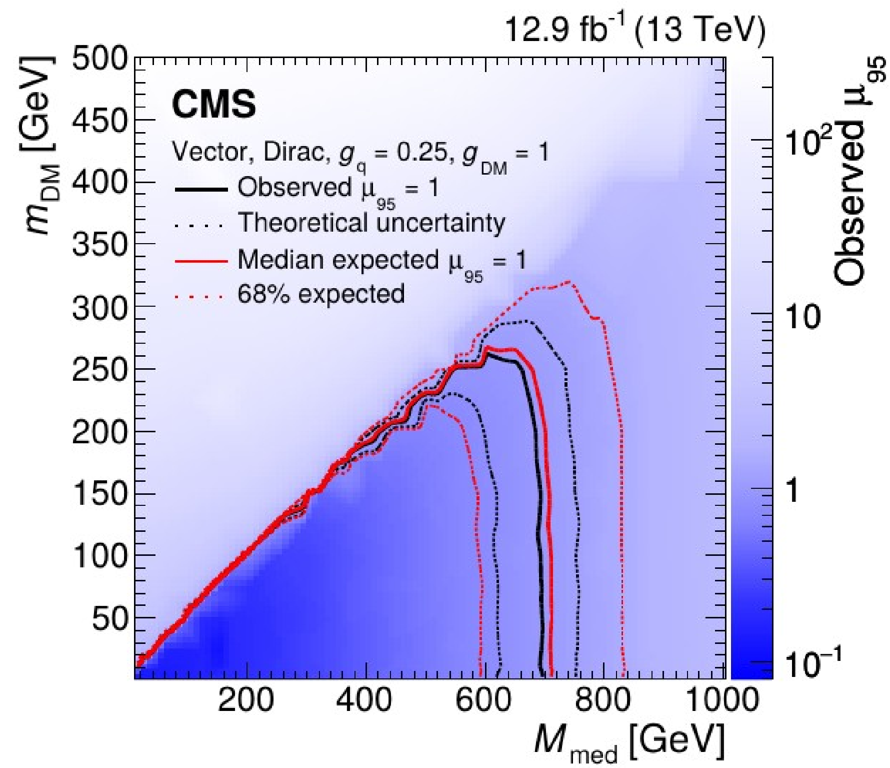
CMS Run 2 Monophoton Search: measured

N_s in a counting experiment reparameterized

according to various DM models







Generating Pseudo-data

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

⇒ Possible outcomes of the experiment, for given parameter values

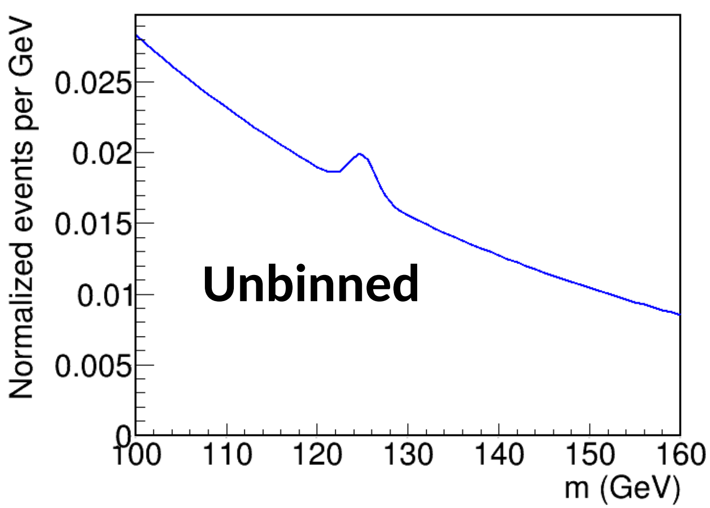
Can draw random events according to PDF : **generate *pseudo-data***

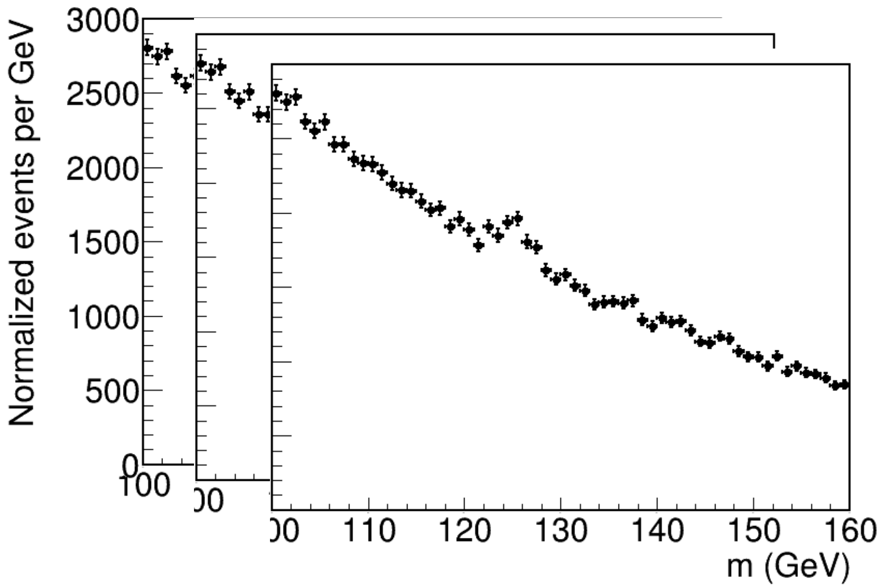
$$P(\lambda = 5)$$

2, 5, 3, 7, 4, 9, ….

Each entry = separate "experiment"

Generate





Expected Limits: Toys

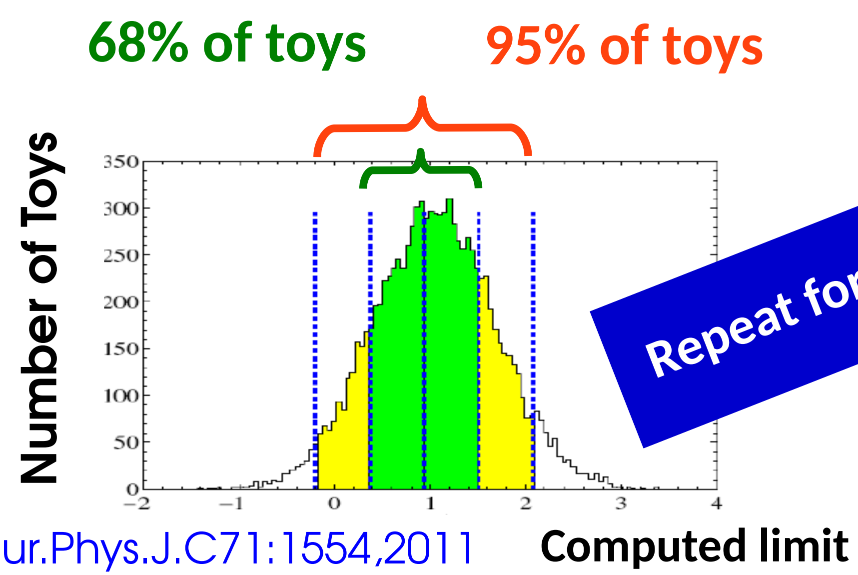
***Expected* results**: median outcome under a given hypothesis
→ usually B-only for searches, but other choices possible.

Two main ways to compute:

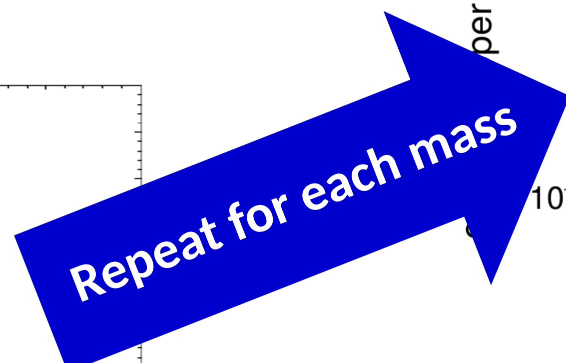
→ Pseudo-experiments (*toys*):

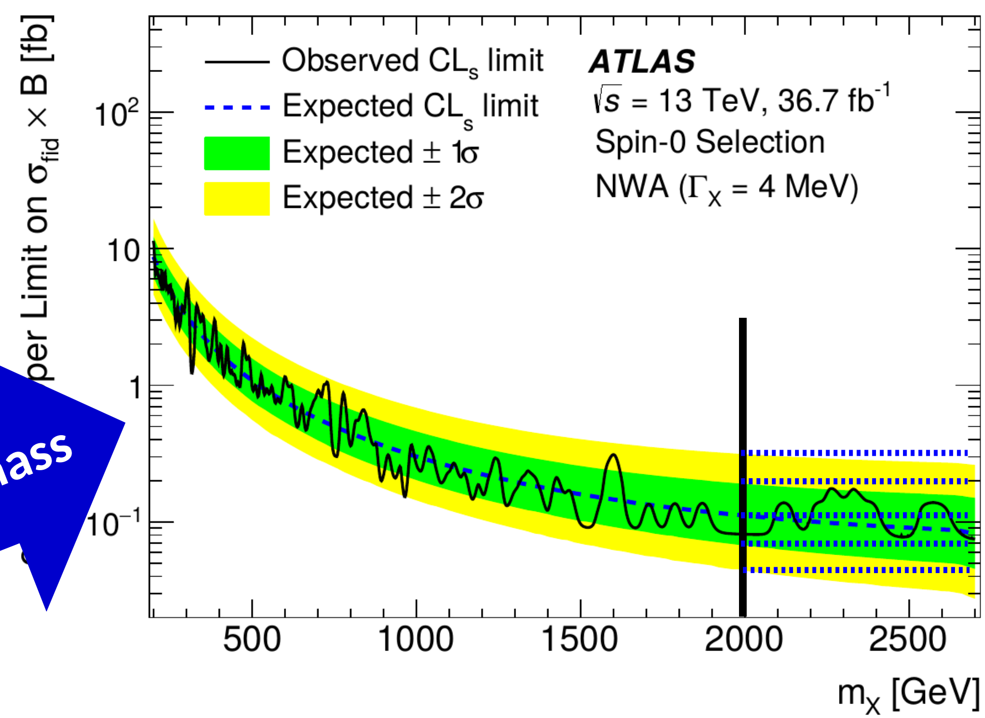
- Generate a pseudo-dataset in B-only hypothesis
- Compute limit
- Repeat and histogram the results
- Central value = median, bands based on quantiles

Phys. Lett. B 775 (2017) 105



Eur.Phys.J.C71:1554,2011





Expected Limits: Asimov Datasets

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

Two main ways to compute:

Strictly speaking, Asimov dataset if

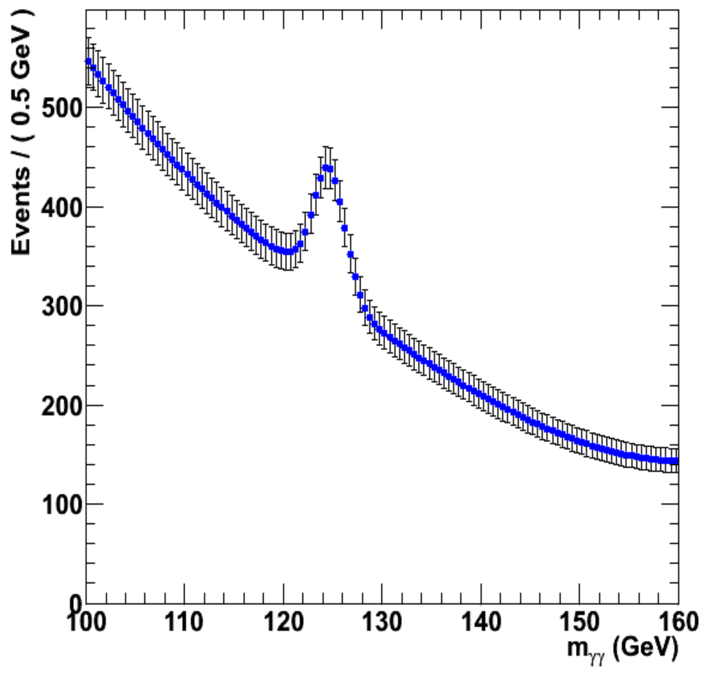
$\hat{X} = X_0$ for all parameters X,

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")**

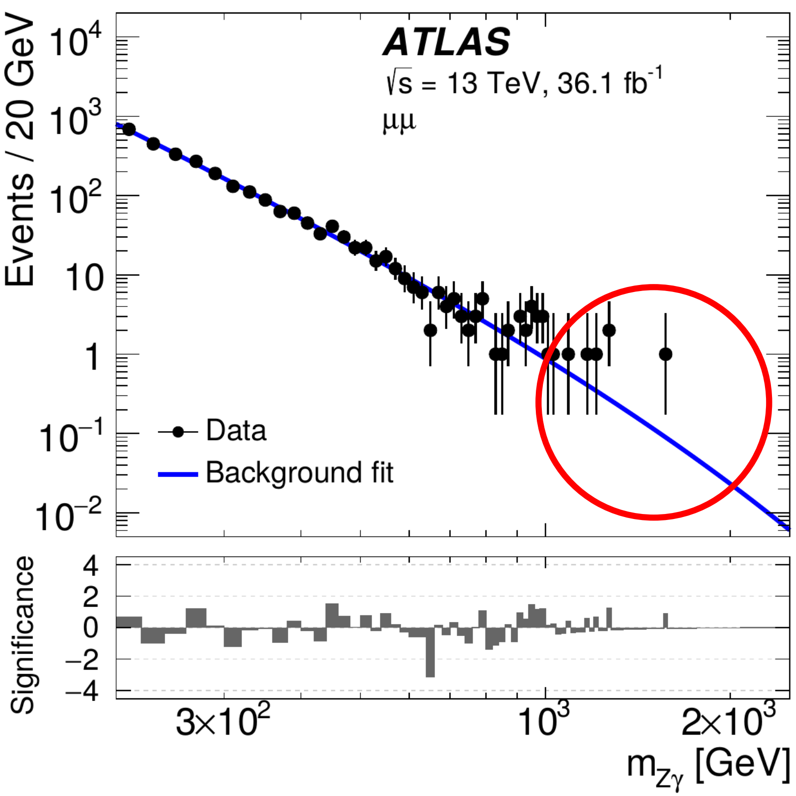
⊖ **Relies on Gaussian approximation**

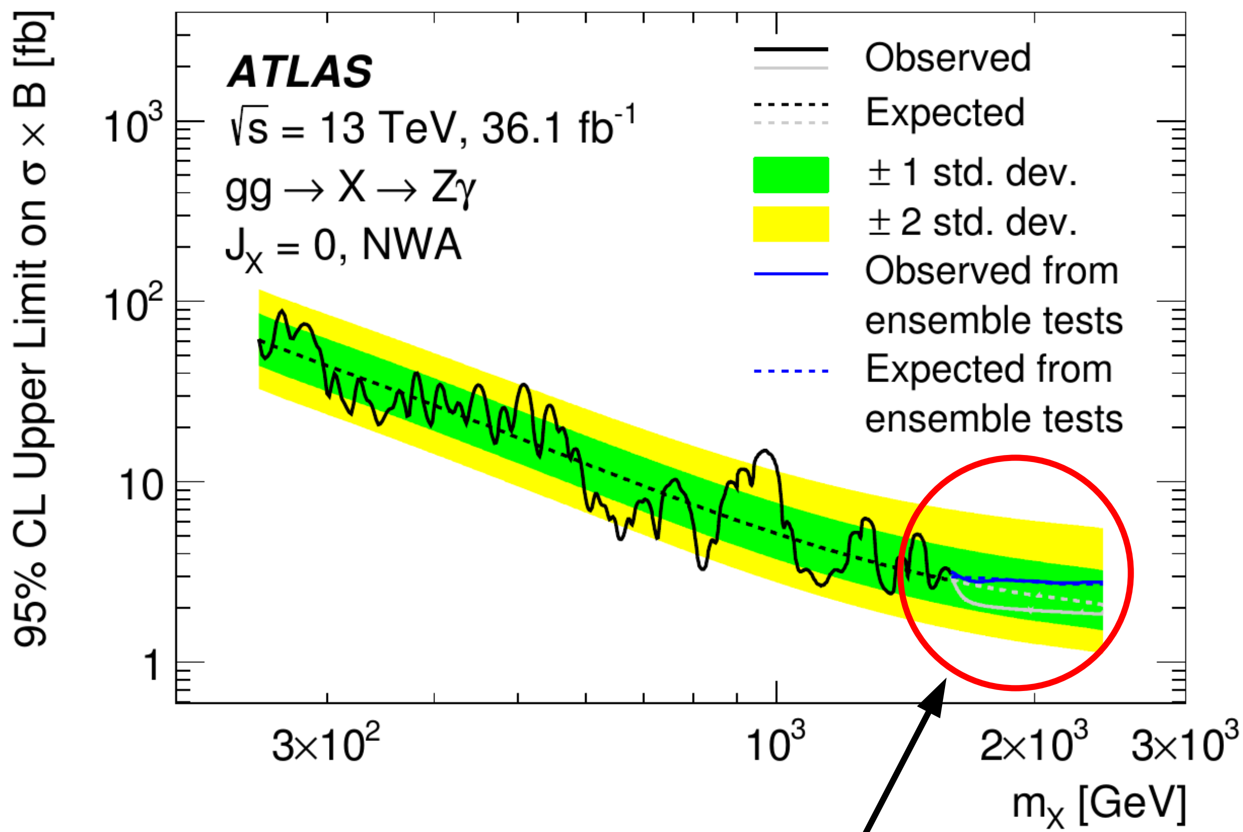
Toys: Example

JHEP 10 (2017) 112

ATLAS X→Zγ Search: covers 200 GeV < m_X < 2.5 TeV

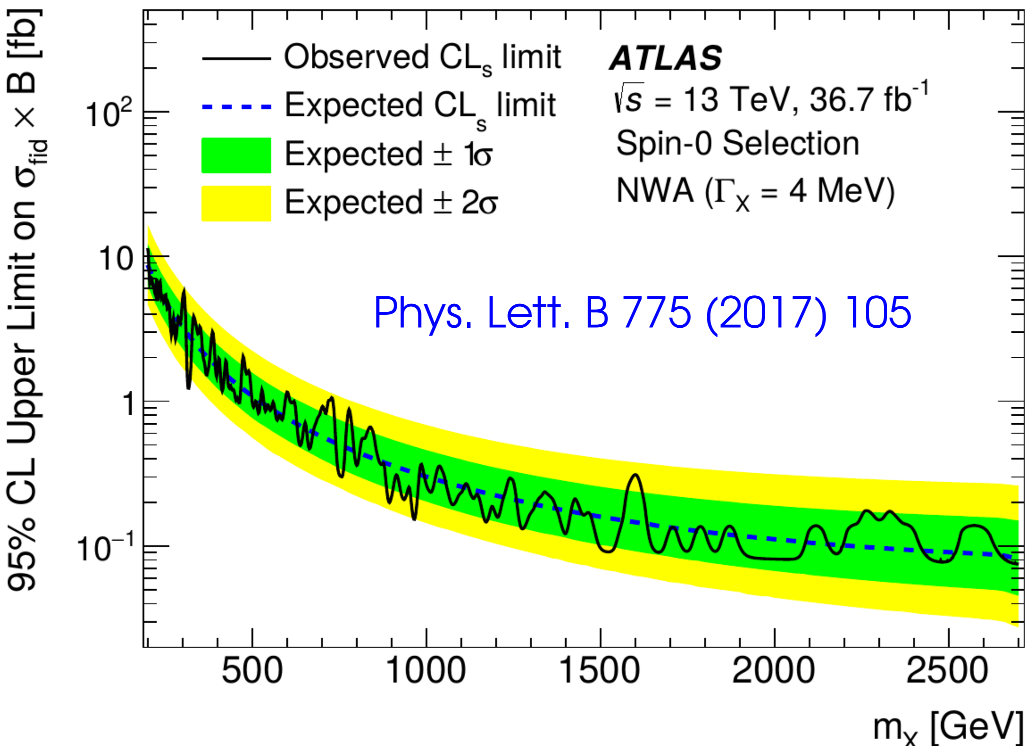
For m_X > 1.6 TeV, low event counts ⇒ derive results from toys





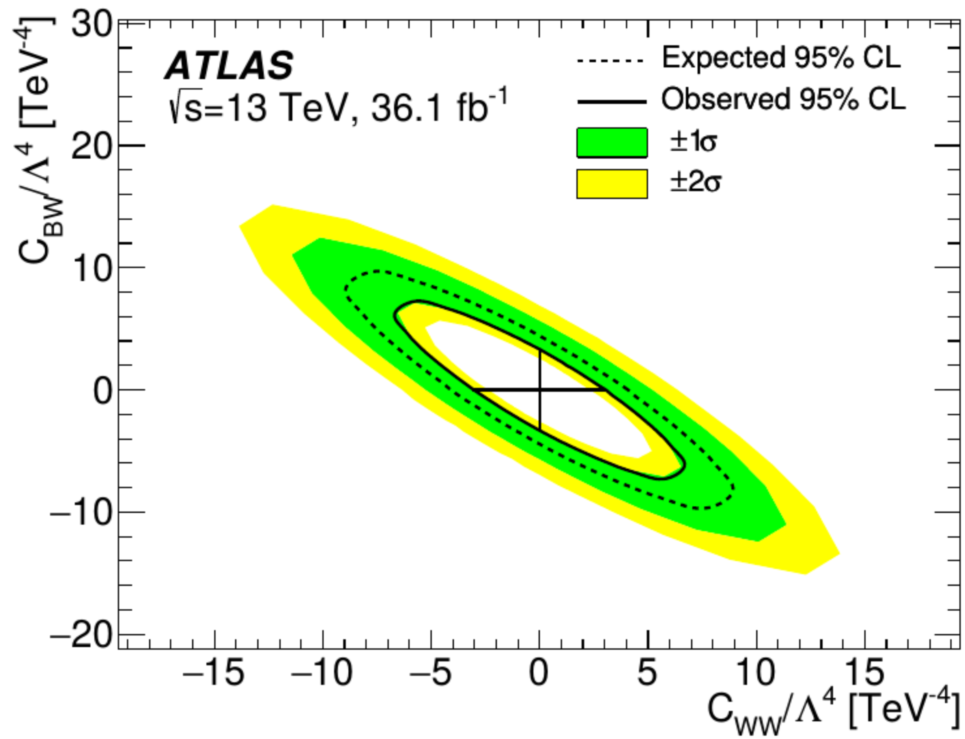
Asimov results (in gray) give optimistic result compared to toys (**in blue**)

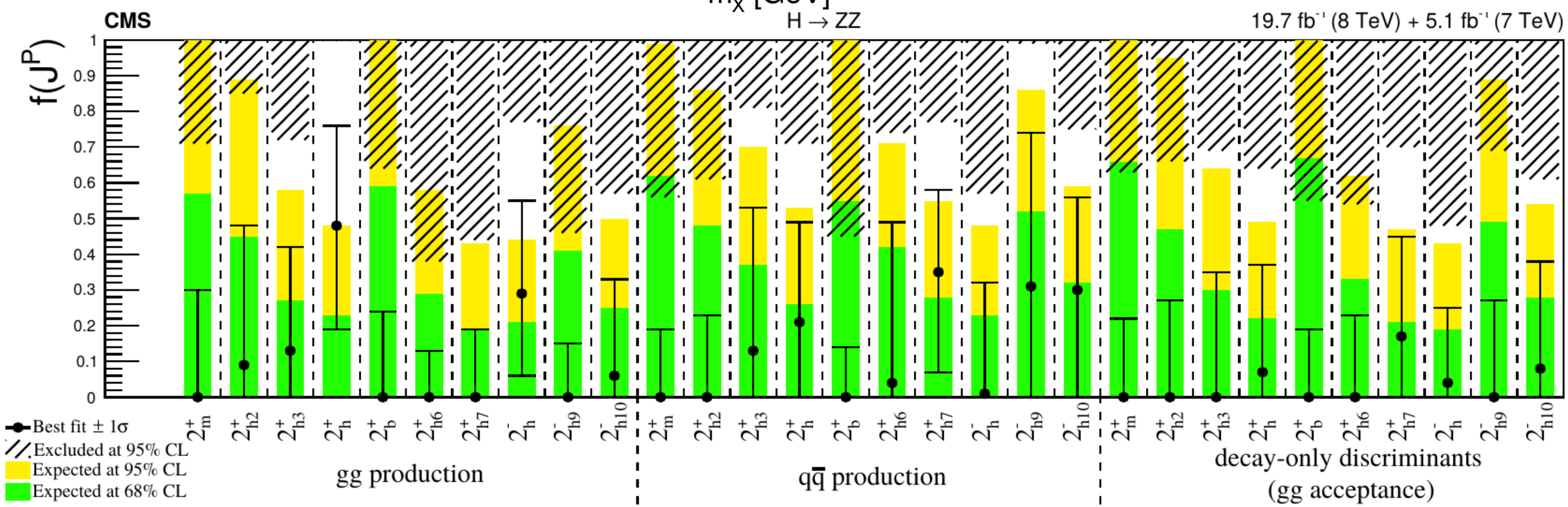
Upper Limit Examples



Phys. Lett. B 775 (2017) 105

ATLAS 2015-2016 4l aTGC Search

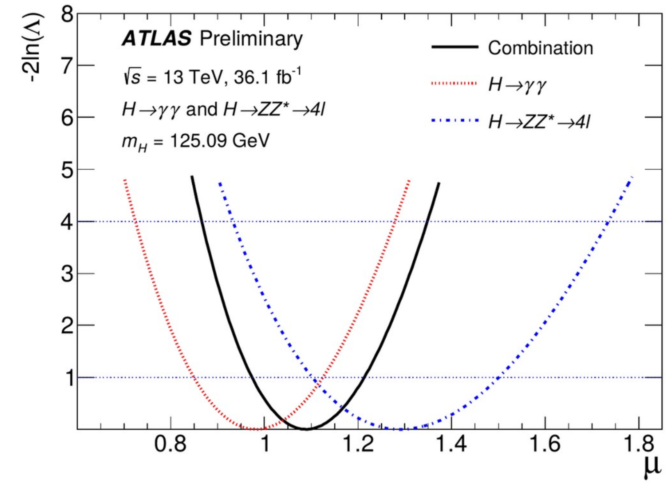




Phys. Rev. D 92 (2015) 012004

Takeaways

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



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

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

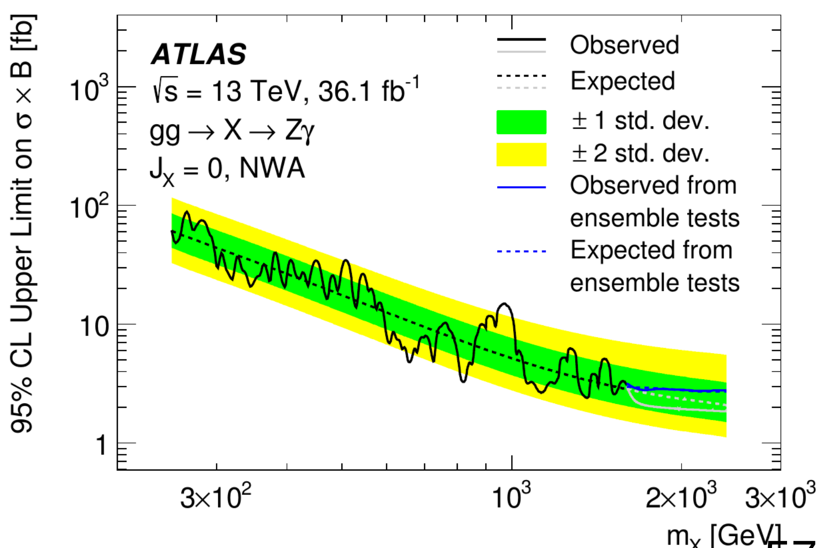
Limits : use LR-based test statistic:

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

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

Gaussian regime, n~0: **$S < \hat{S} + 1.96\sigma$ at 95% CL**

Poisson regime, n=0 : **S_{up} = 3 events at 95% CL**



Extra Slides

Rare Processes ?

HEP : almost always use Poisson distributions. Why ?

ATLAS :

- **Event rate ~ 1 GHz**

  ($L \sim 10^{34}\ cm^{-2}s^{-1} \sim 10\ nb^{-1}/s$, $\sigma_{tot} \sim 10^{8}$ nb,)

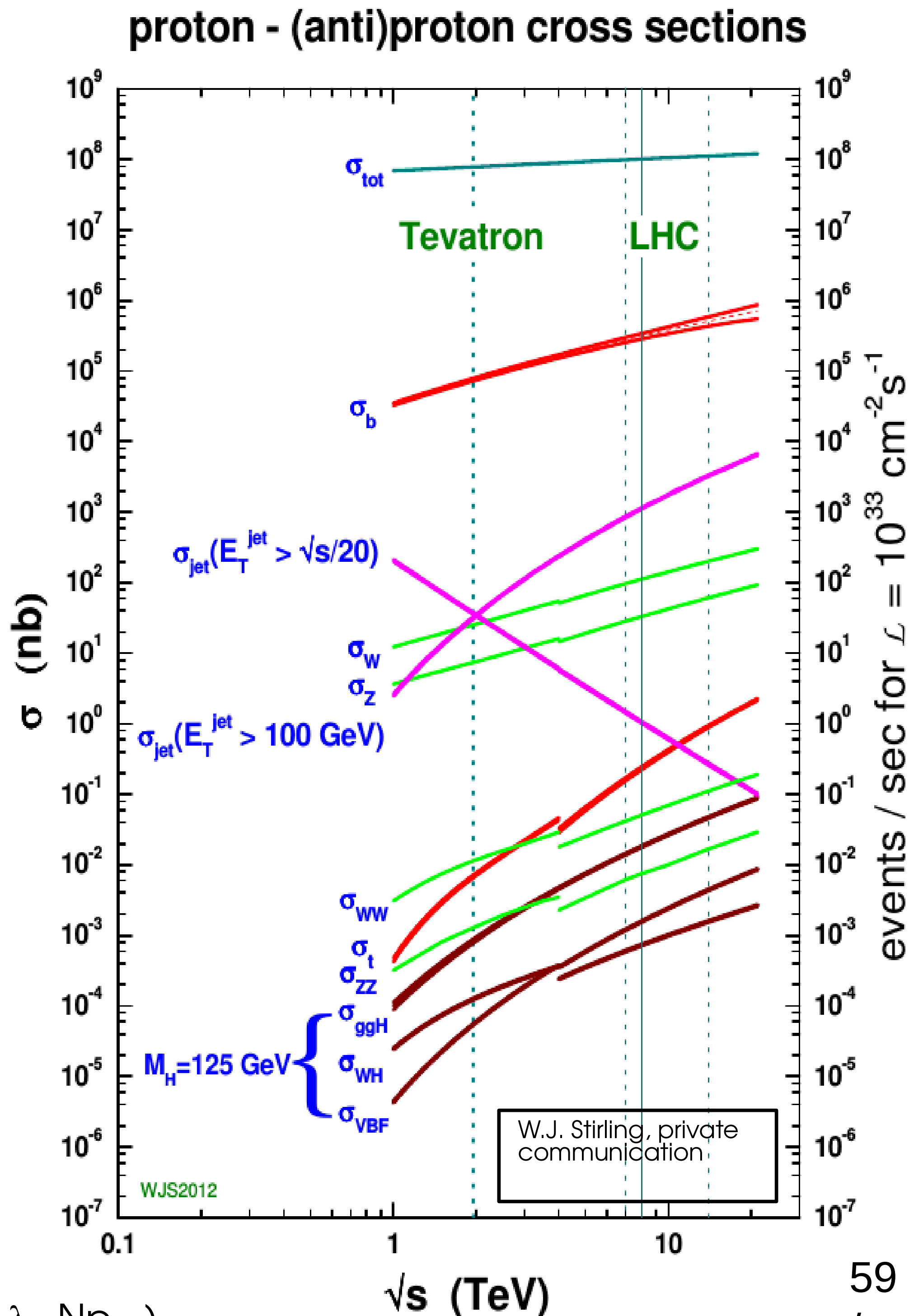
- **Trigger rate ~ 1 kHz**

  (Higgs rate **~ 0.1 Hz**)

⇒ **$p \sim 10^{-6} \ll 1$** ($p_{H\rightarrow\gamma\gamma} \sim$ **10^{-13}**)

A day of data: **$N \sim 10^{14} \gg 1$**

⇒ **Poisson regime! Similarly true in many other physics situations.**



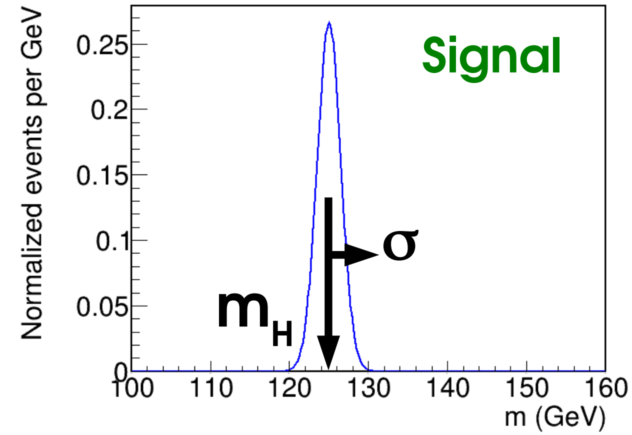
(Large N = design requirement, to get not-too-small $\lambda = Np$...)

Unbinned Shape Analysis

Observable: set of values $m_1 \ldots m_n$, one per event

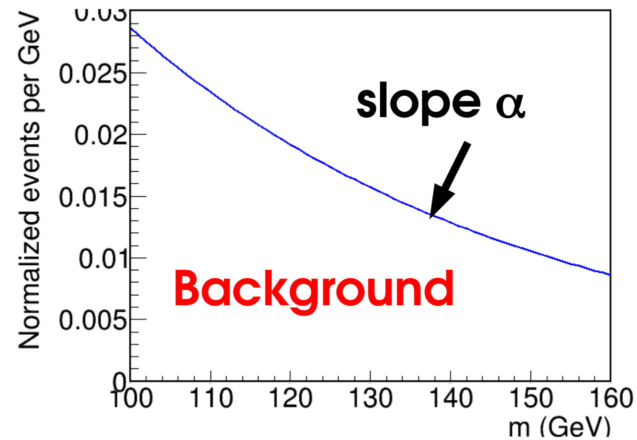
→ Describe shape of the **distribution of m**

→ Deduce the **probability to observe $m_1 \ldots m_n$**



H→γγ-inspired example:

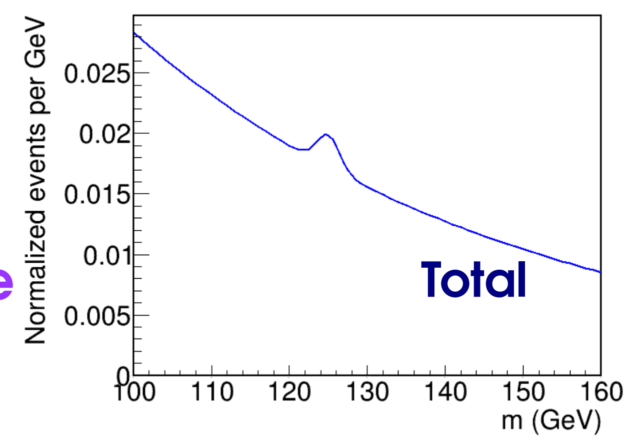
- **Gaussian signal** $P_{\text{signal}}(m) = G(m; m_H, \sigma)$
- **Exponential bkg** $P_{\text{bkg}}(m) = \alpha\, e^{-\alpha m}$



Expected yields : **S, B**

⇒ **Total PDF for a single event**:

$$P_{\text{total}}(m) = \frac{S}{S+B} G(m; m_H, \sigma) + \frac{B}{S+B} \alpha\, e^{-\alpha m}$$



⇒ **Total PDF for a dataset**

Probability to observe the value m_i

Probability to observe n events

$$P(\{m_i\}_{i=1\ldots n}) = e^{-(S+B)} \frac{(S+B)^n}{n!} \prod_{i=1}^{n} \frac{S}{S+B} G(m_i; m_H, \sigma) + \frac{B}{S+B} \alpha\, e^{-\alpha m_i}$$

Poisson Example

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

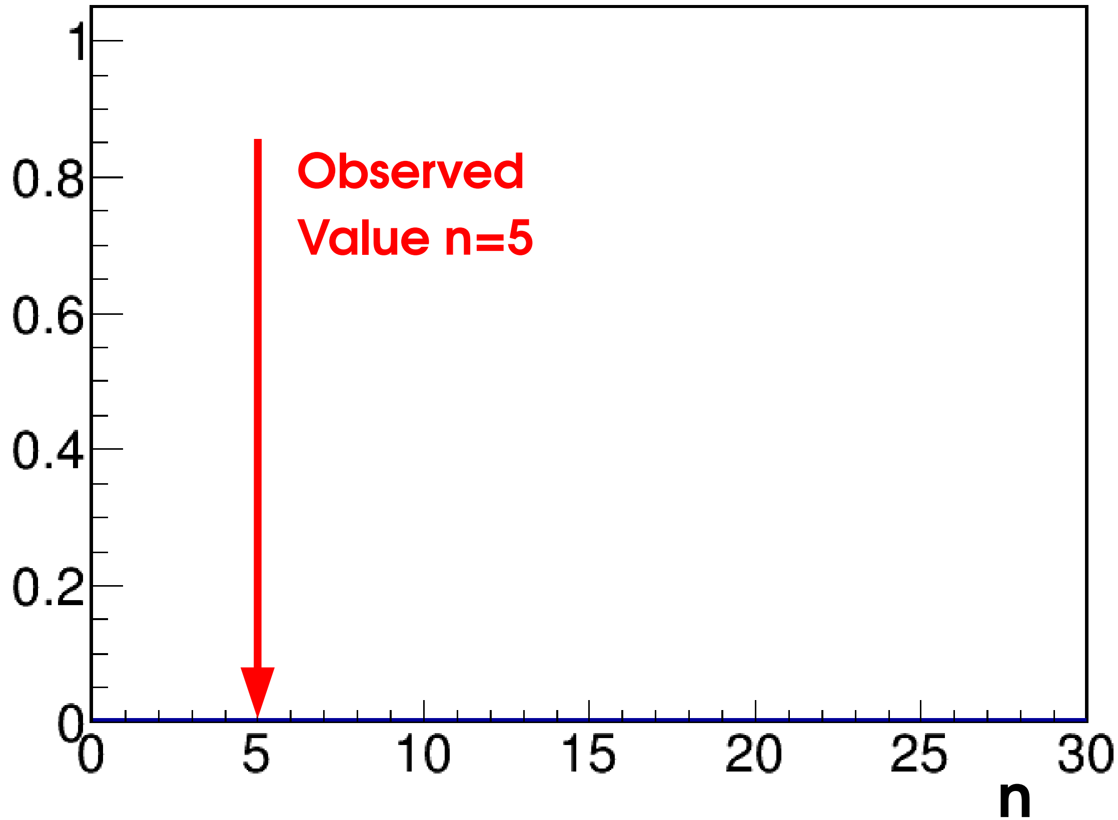
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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 :

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

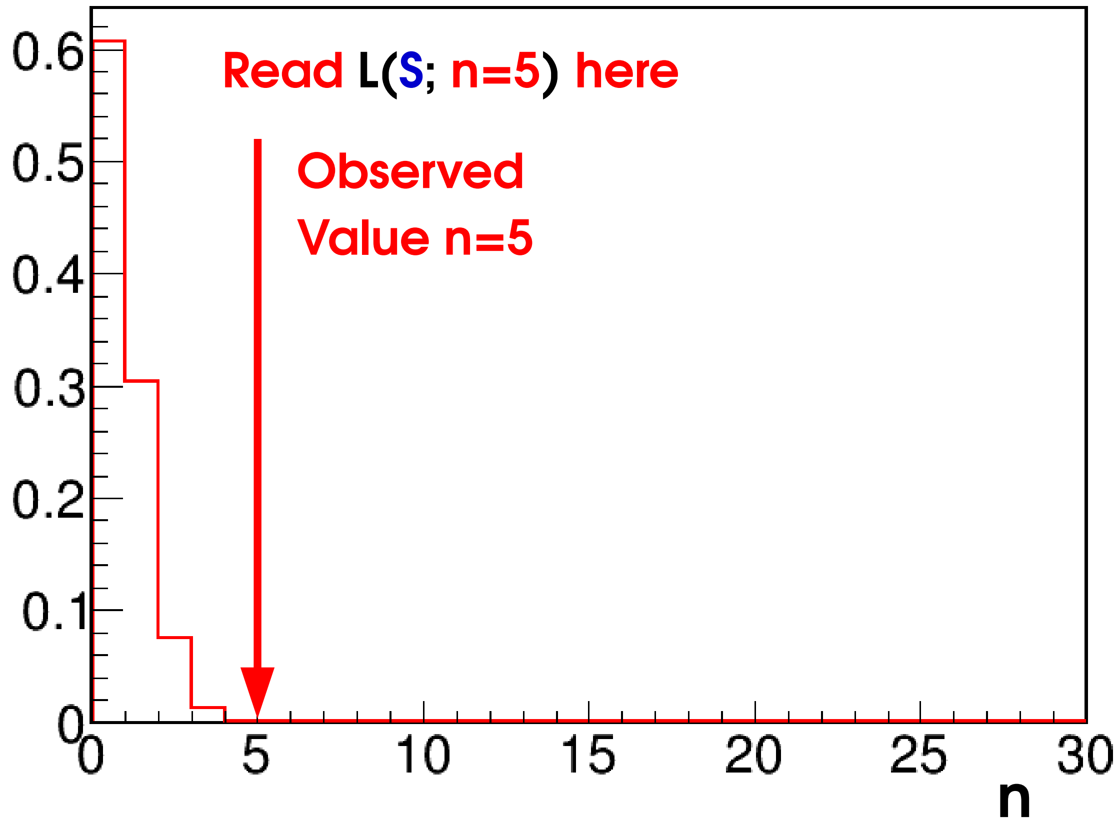
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

P(S = 0.5)
Low
likelihood



Poisson Example

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

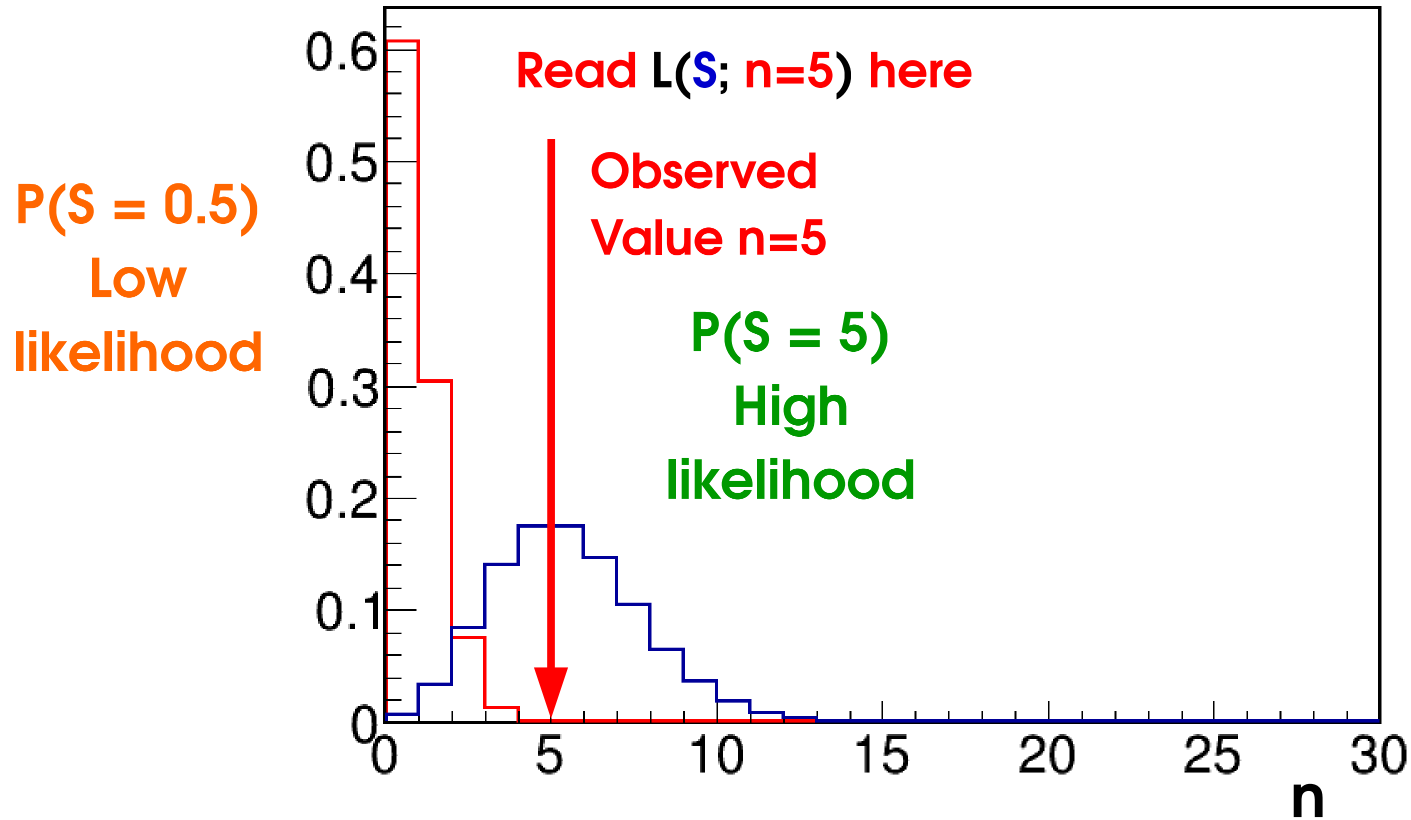
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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 :

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

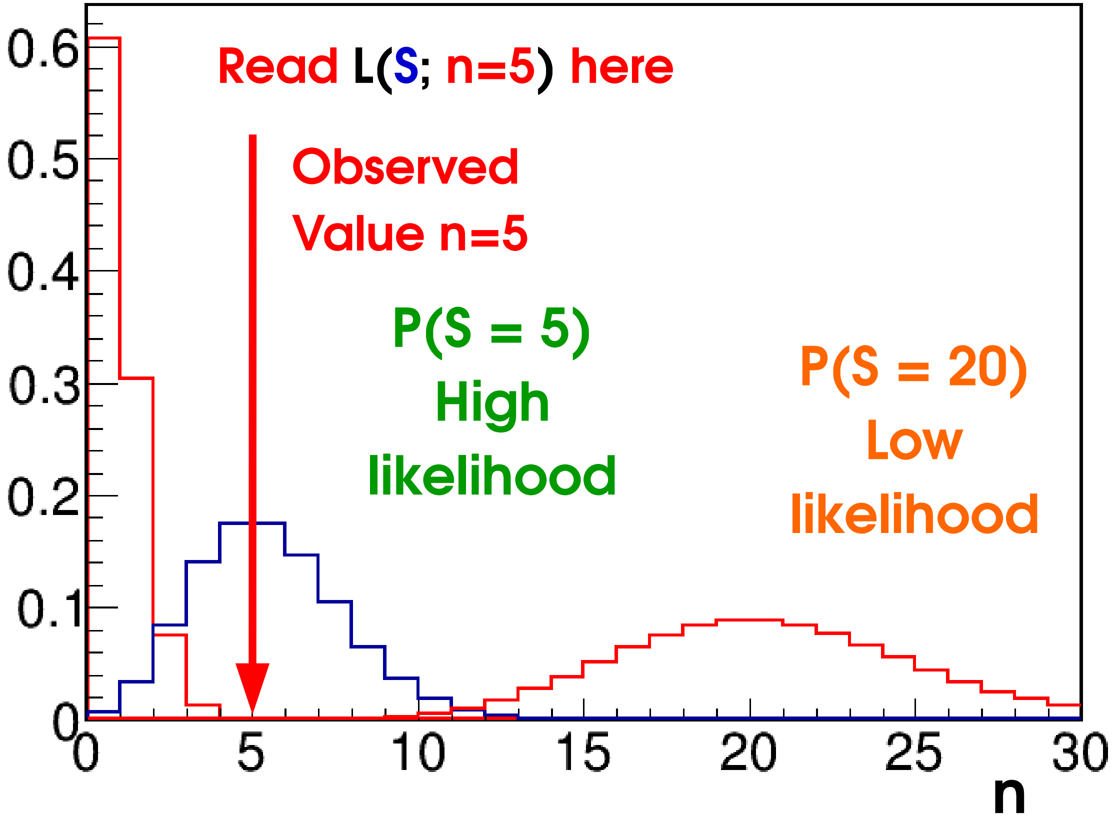
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

P(S = 0.5)
Low
likelihood



Poisson Example

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

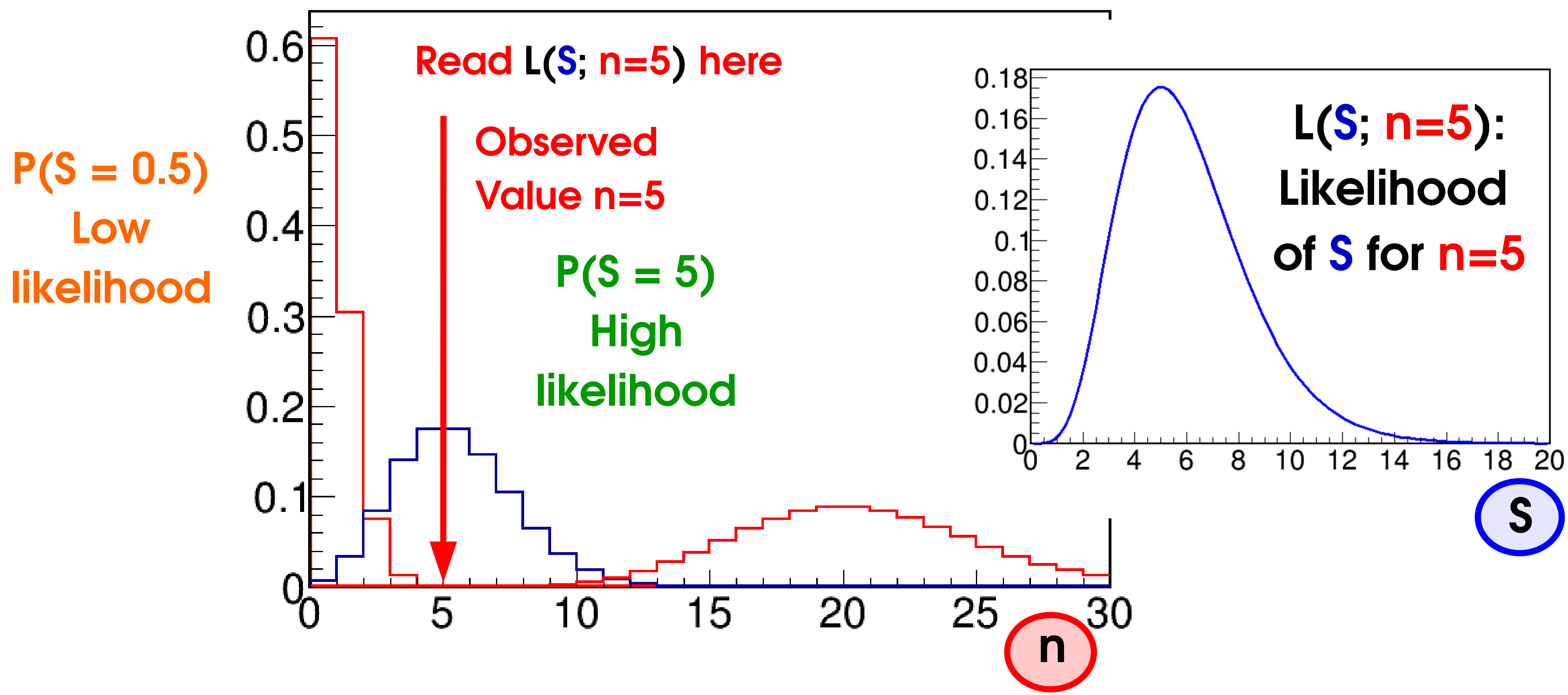
Assume **Poisson distribution** with B = 0 :

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

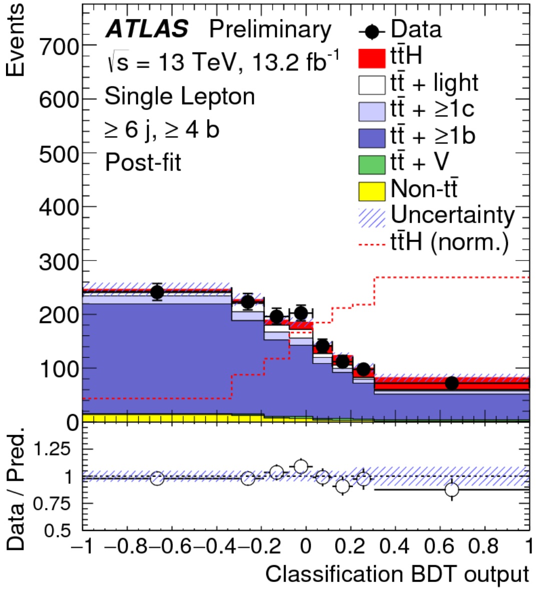
→ Try different values of S for a fixed data value n=5

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

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



MLEs in Shape Analyses



Binned shape analysis:

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

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

In practice easier to minimize

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

Needs a computer...

In the Gaussian limit

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

χ^2 formula!

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

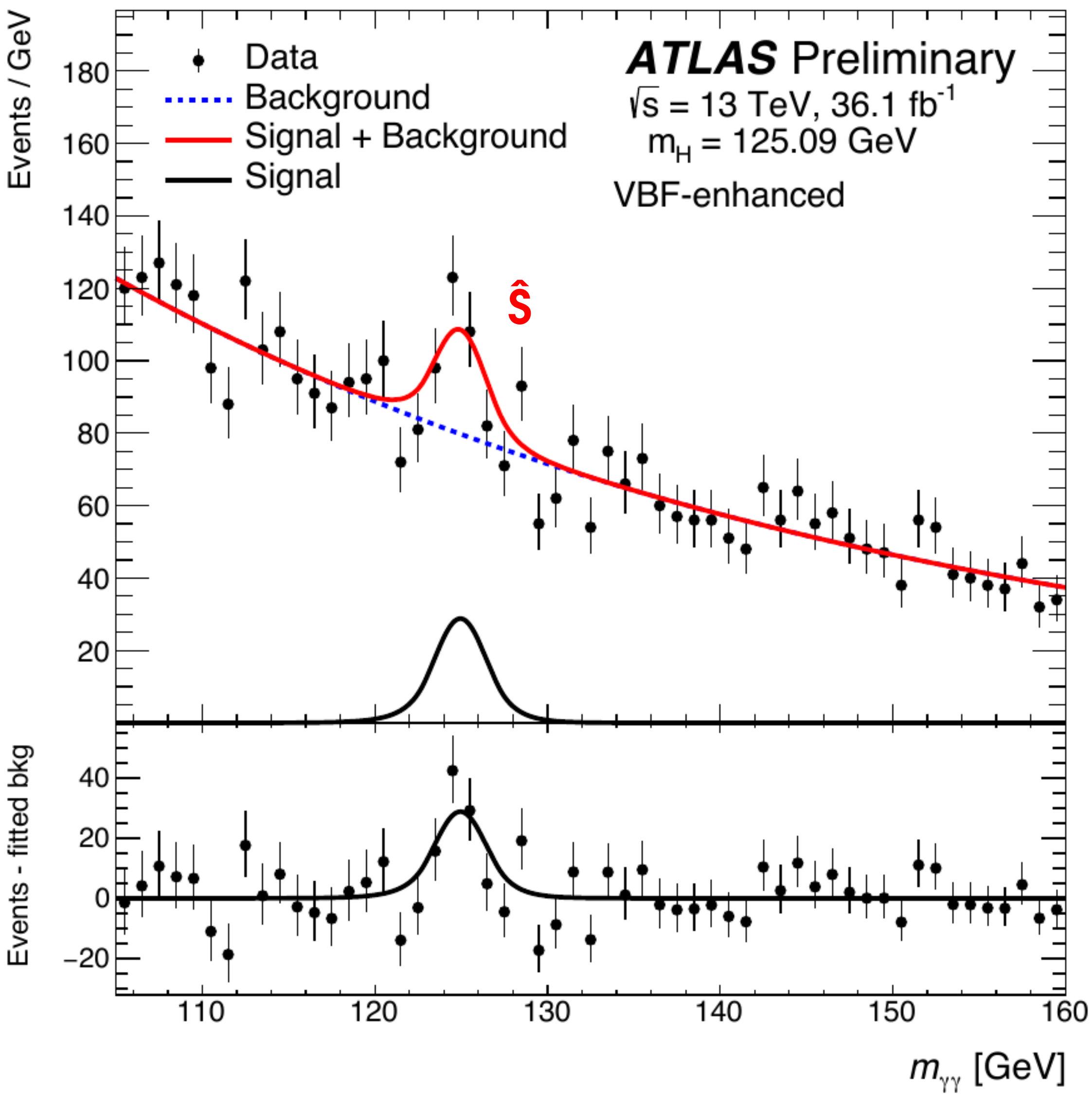
→ **Poisson MLE** (min λ_{Pois}) : ***Best fit value*** in a *likelihood* fit (in `ROOT`, fit option "`L`")

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

In both cases, MLE ⇔ *Best Fit*

H→γγ

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



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

  and unbiased $\langle\hat{\mu}\rangle = \mu^*$ for $n \to \infty$

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

  for large enough datasets

- **Asymptotically Efficient** : $\sigma_{\hat{\mu}}$ is the **lowest possible value** (in the limit $n\to\infty$) among consistent estimators.

  → MLE captures all the available information in the data

- Also **consistent**: $\hat{\mu}$ converges to the true value for large n, $\hat{\mu} \xrightarrow{n\to\infty} \mu^*$

- **Log-likelihood :** Can also **minimize** $\lambda = -2 \log L$

  → Usually more efficient numerically

  → For Gaussian L, λ is parabolic:

- Can **drop multiplicative constants in L** (additive constants in λ)

Extra: Fisher Information

Fisher Information:

$$I(\mu) = \left\langle \left(\frac{\partial}{\partial \mu} \log L(\mu) \right)^2 \right\rangle = -\left\langle \frac{\partial^2}{\partial \mu^2} \log L(\mu) \right\rangle$$

Measures the **amount of information** available in the measurement of μ.

Gaussian likelihood:

$$I(\mu) = \frac{1}{\sigma^2_{Gauss}}$$

→ smaller σ_{Gauss} ⇒ more information.

Gaussian case:

- For a Gaussian estimator $\tilde{\mu}$

$$P(\tilde{\mu}) \propto \exp\left(- \frac{(\tilde{\mu} - \mu^*)^2}{2\sigma^2_{\tilde{\mu}}} \right)$$

- **MLE**: $\mathrm{Var}(\hat{\mu}) = \sigma_{\hat{\mu}}^2$

Cramer-Rao: $\mathrm{Var}(\tilde{\mu}) \geq \sigma_{Gauss}^2 = \sigma_{\tilde{\mu}}^2$

Cramer-Rao bound:

$$\mathrm{Var}(\tilde{\mu}) \geq \frac{1}{I(\mu)}$$

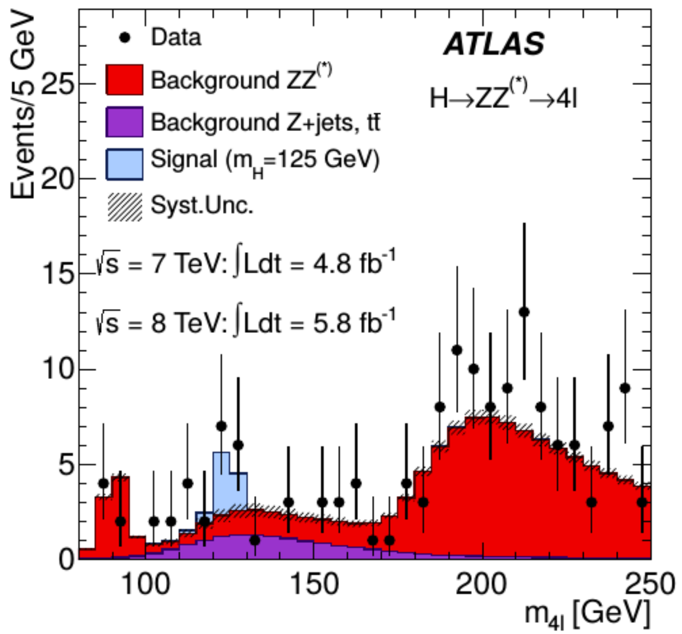
For any estimator $\tilde{\mu}$.

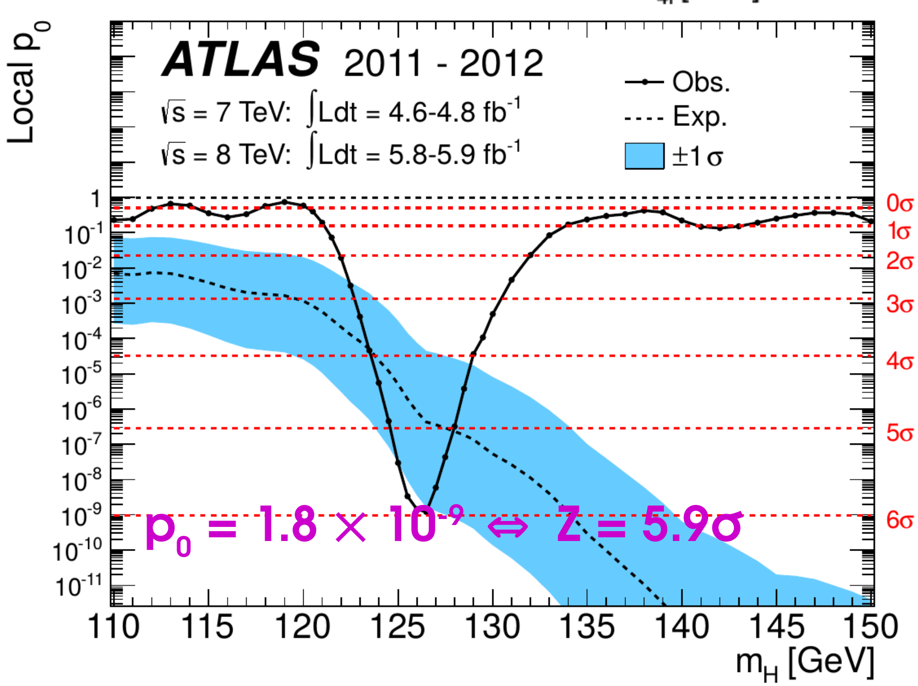
→ cannot be more precise than allowed by information in the measurement.

Efficient estimators reach the bound : **e.g. MLE in the large dataset limit.**

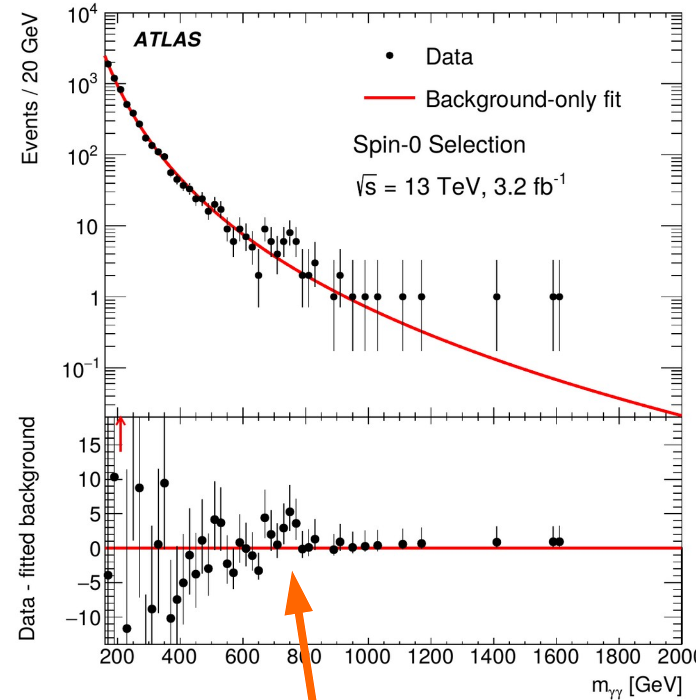
Some Examples

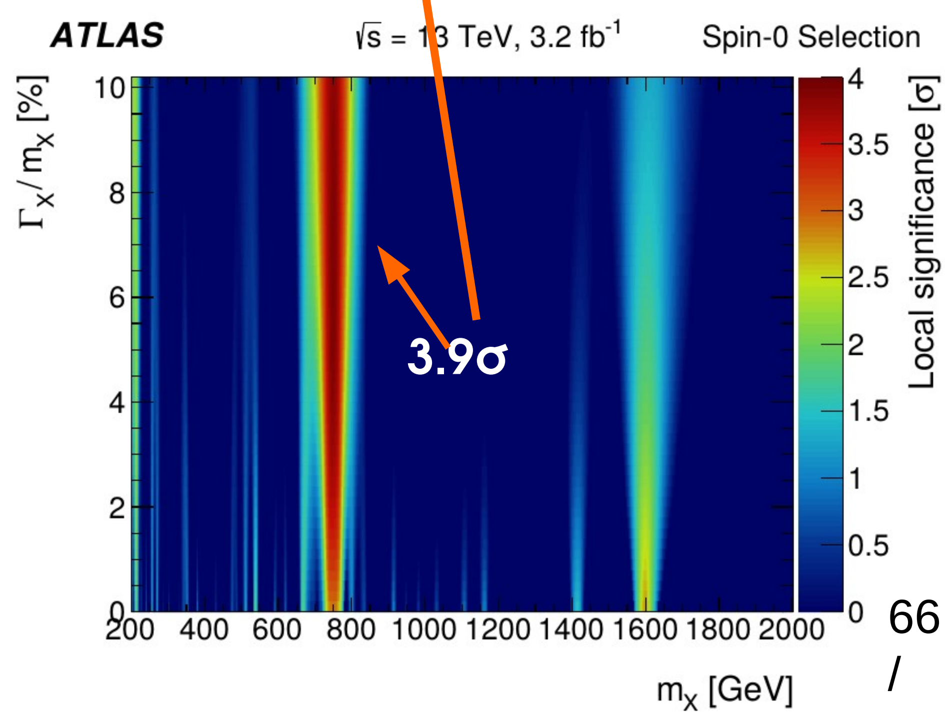
Higgs Discovery: Phys. Lett. B 716 (2012) 1-29





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





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 ?

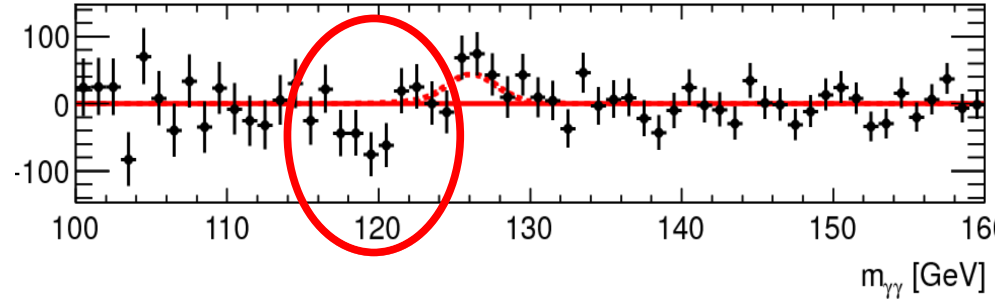
→ **Background modeling issue ?...** Or:

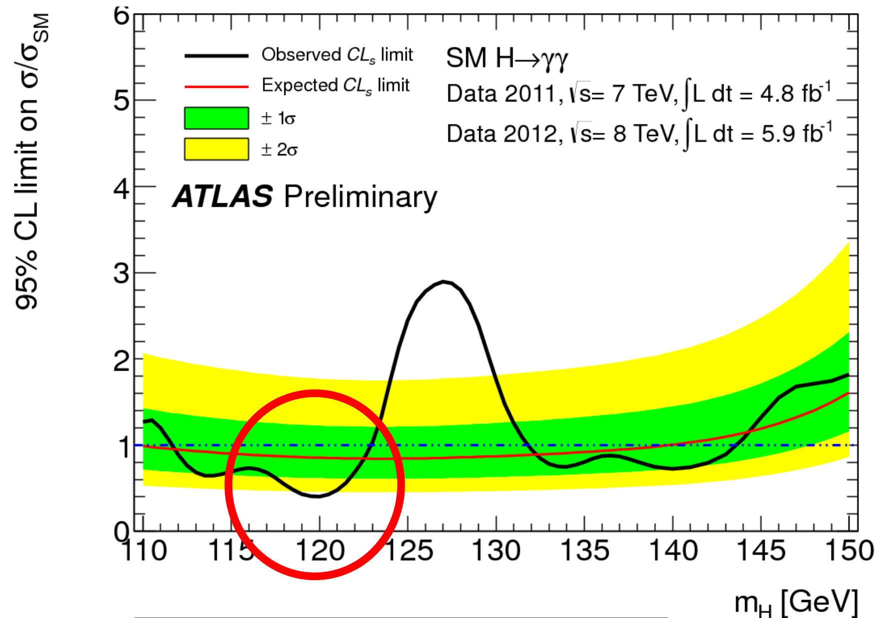
→ This is a **95%** limit ⇒ **5% of the time, the limit wrongly excludes the true value,** *e.g.* $S^*=0$.

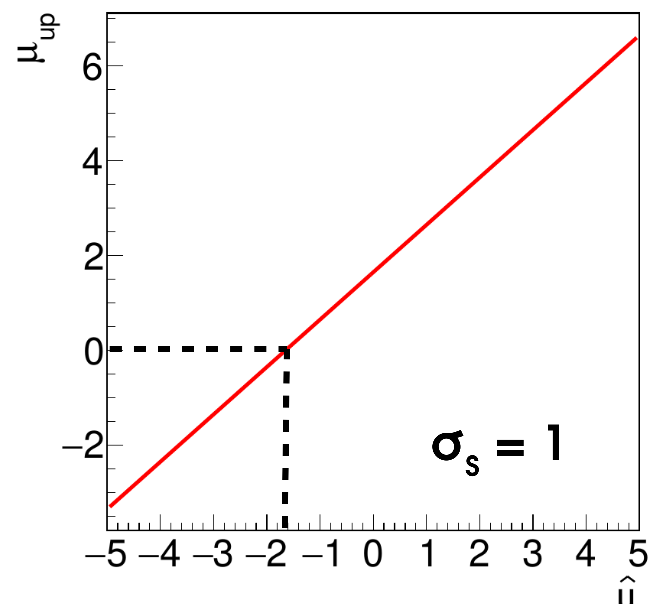
Options

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

→ **Special procedure to avoid these cases,** since if we assume S must be >0, we know a priori this is just a fluctuation.







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...

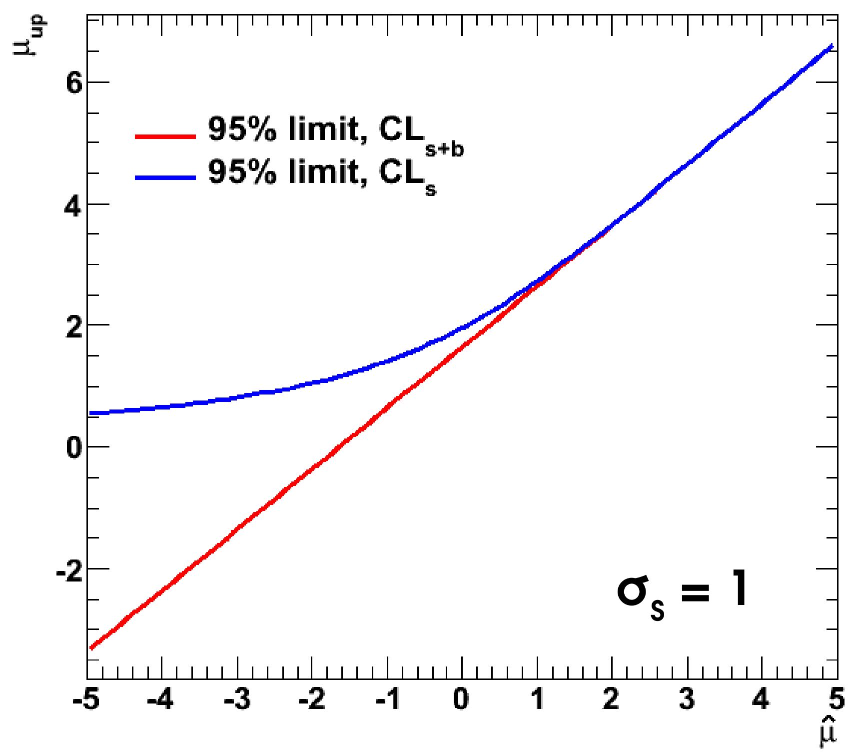
$\hat{S}$ compatible with 0 : $p_B \sim O(1)$

p_{CLs} **~ 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

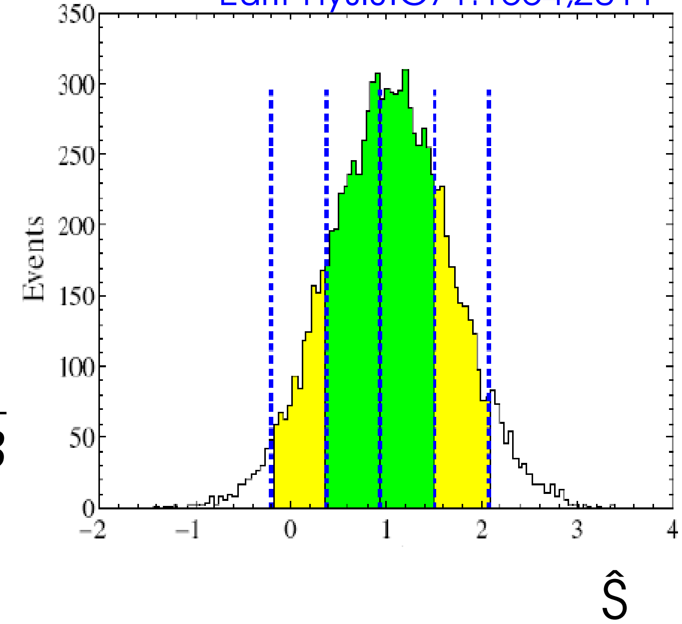


Drawback: ***overcoverage***

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

CL_s : Gaussian Bands

Eur.Phys.J.C71:1554,2011



Usual Gaussian counting example with known B:

95% CL_s upper limit on S:

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

with $\sigma_S = \sqrt{B}$

Compute expected bands for S=0:

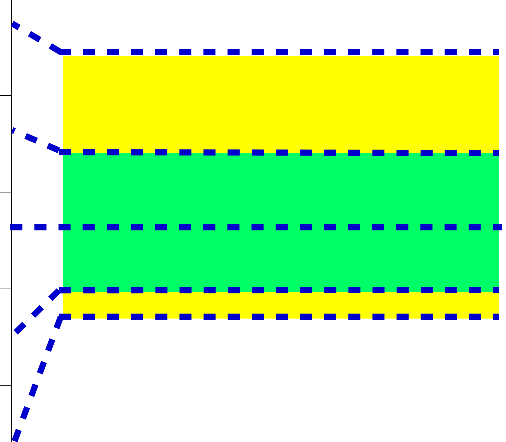
→ **Asimov dataset ⇔ $\hat{S} = 0$** :

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

→ **± nσ bands**:

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

n	$S_{exp}^{\pm n} / \sqrt{B}$
+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 $\sigma^2_{S,A} = \dfrac{S^2}{q_S(\text{Asimov})}$ depends on S, for non-Gaussian cases, different values for each band...

Comparison with LEP/TeVatron definitions

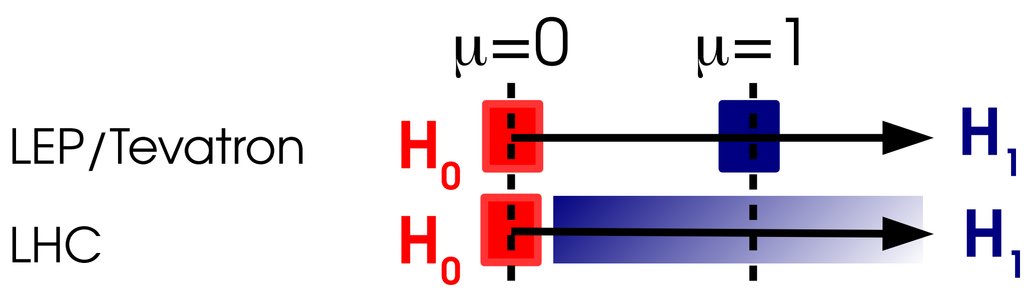
Likelihood ratios are not a new idea:

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

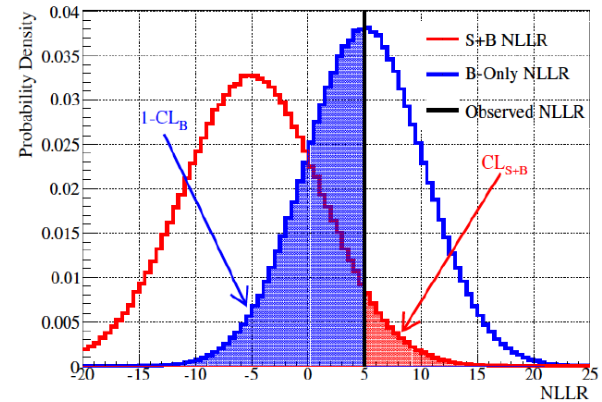
$$q_{LEP} = -2\log\frac{L(\mu=0,\tilde{\theta})}{L(\mu=1,\tilde{\theta})}$$

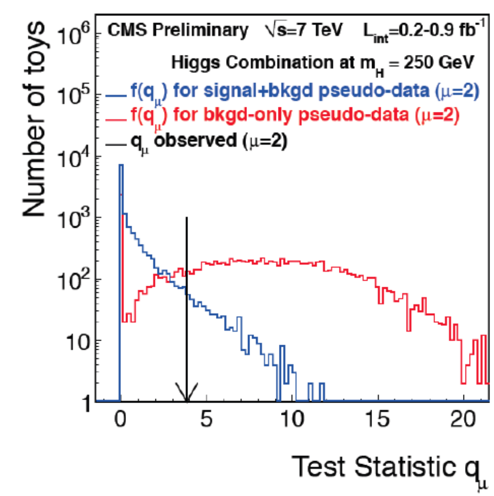
$$q_{Tevatron} = -2\log\frac{L(\mu=0,\hat{\hat{\theta}}_0)}{L(\mu=1,\hat{\hat{\theta}}_1)}$$

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



Example from Collie documentation (D0 note 5595)





→ 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