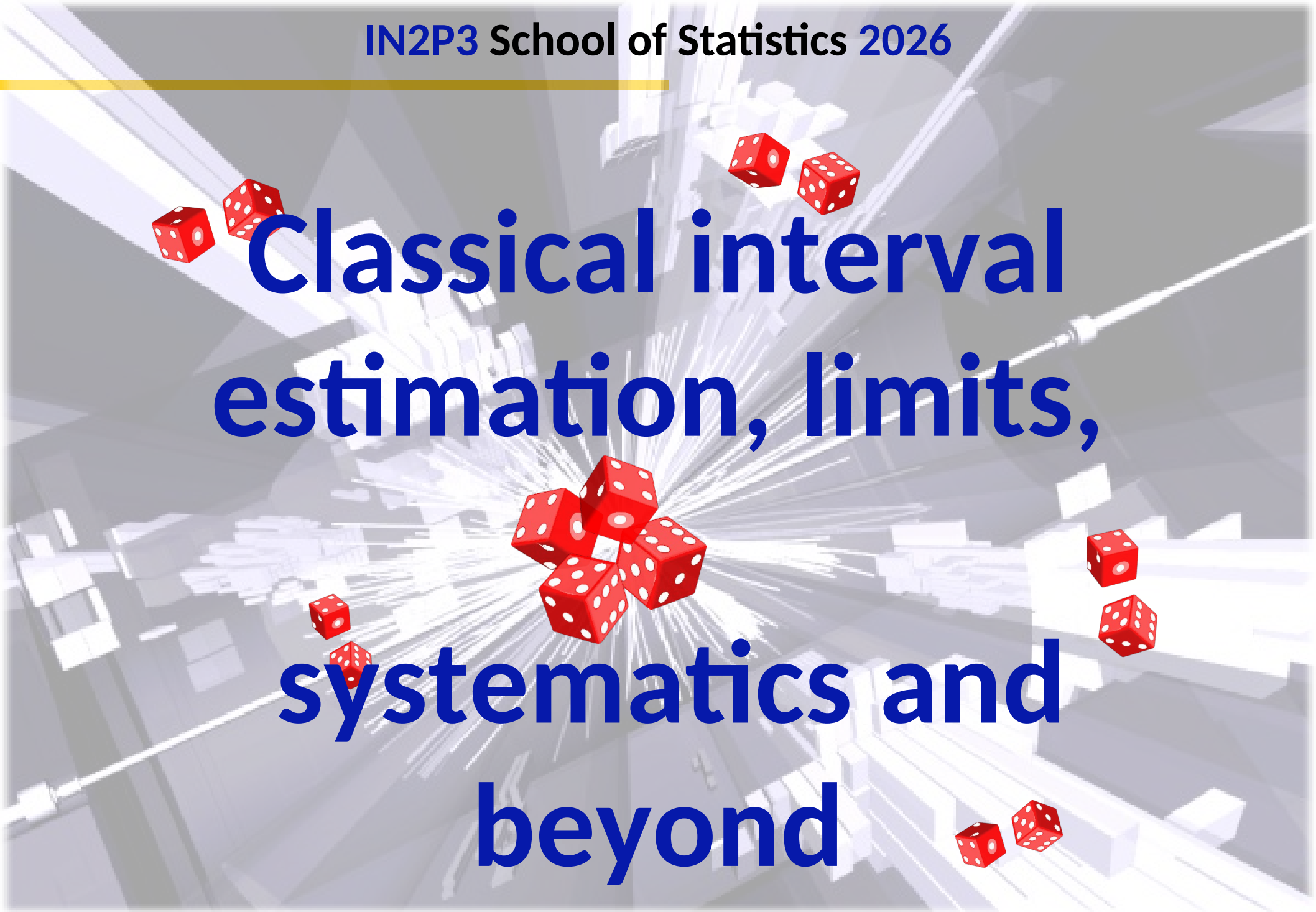


Classical interval estimation, limits, systematics and beyond

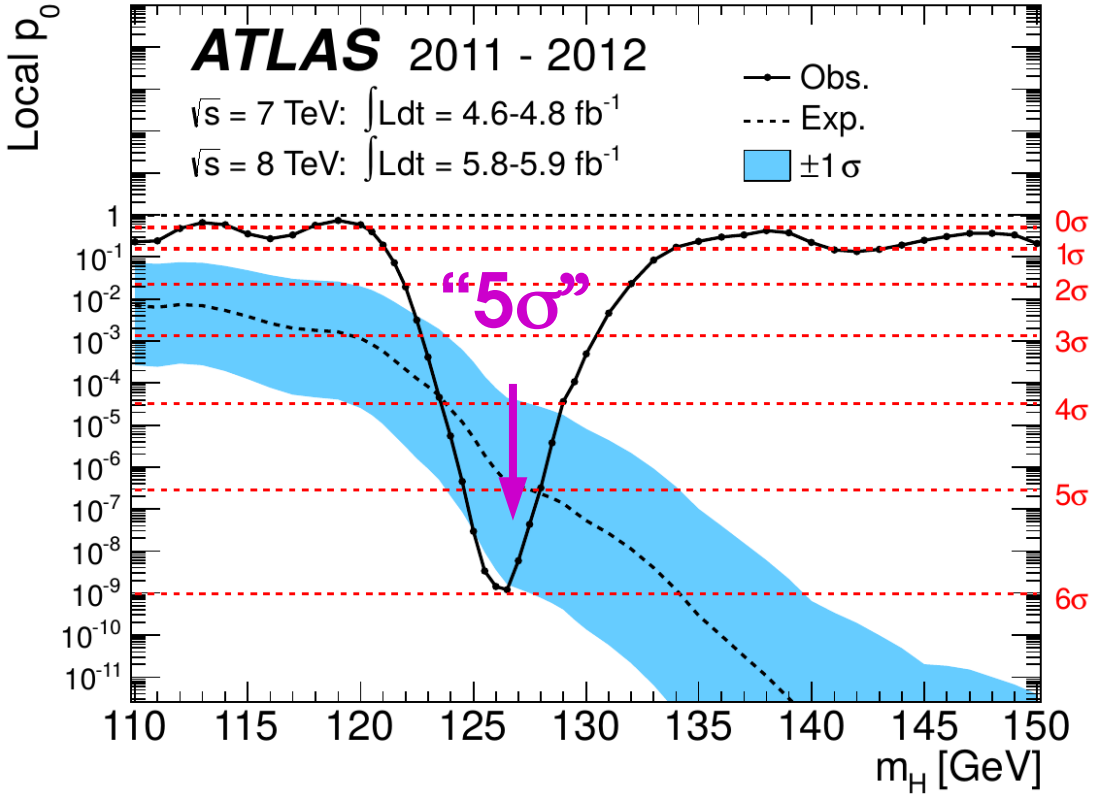
Lydia Brenner (Nikhef), Nicolas Berger (LAPP Annecy)

Lecture 1

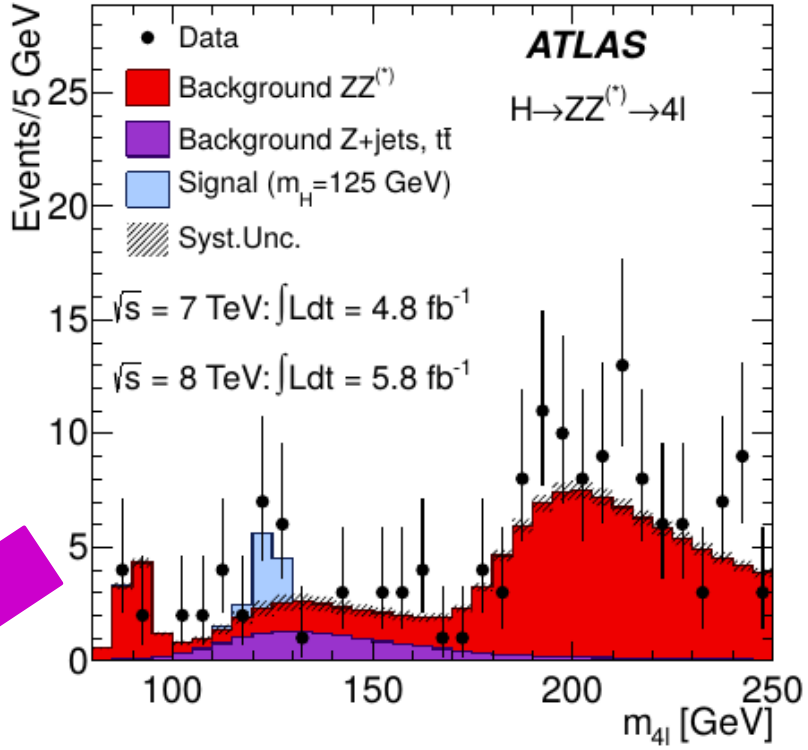
Introduction

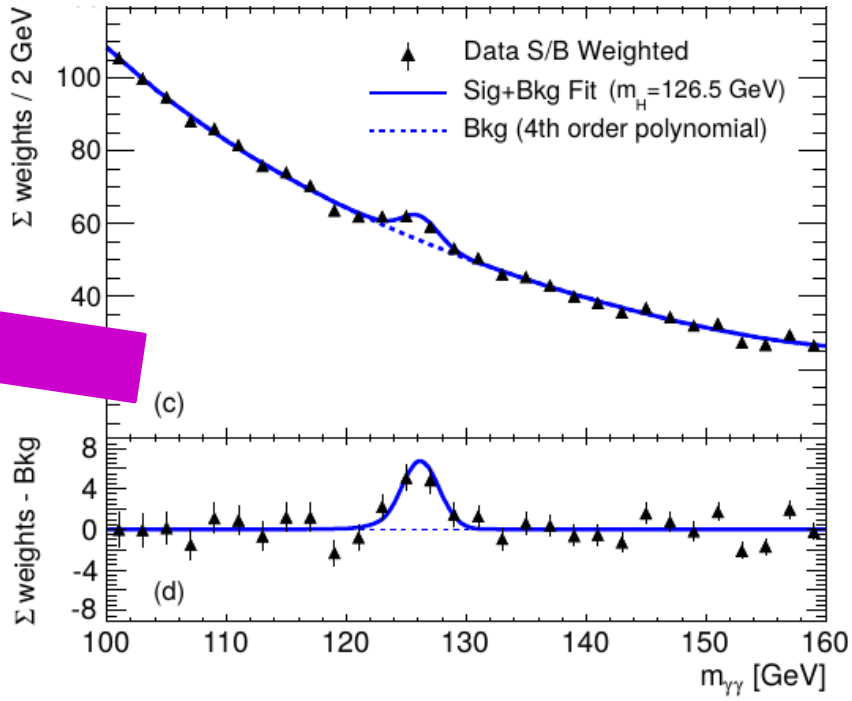
Statistical methods play a critical role in many areas of physics

Higgs discovery : **"We have 5σ" !**



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





Introduction

Precision measurements are another window into BSM effects

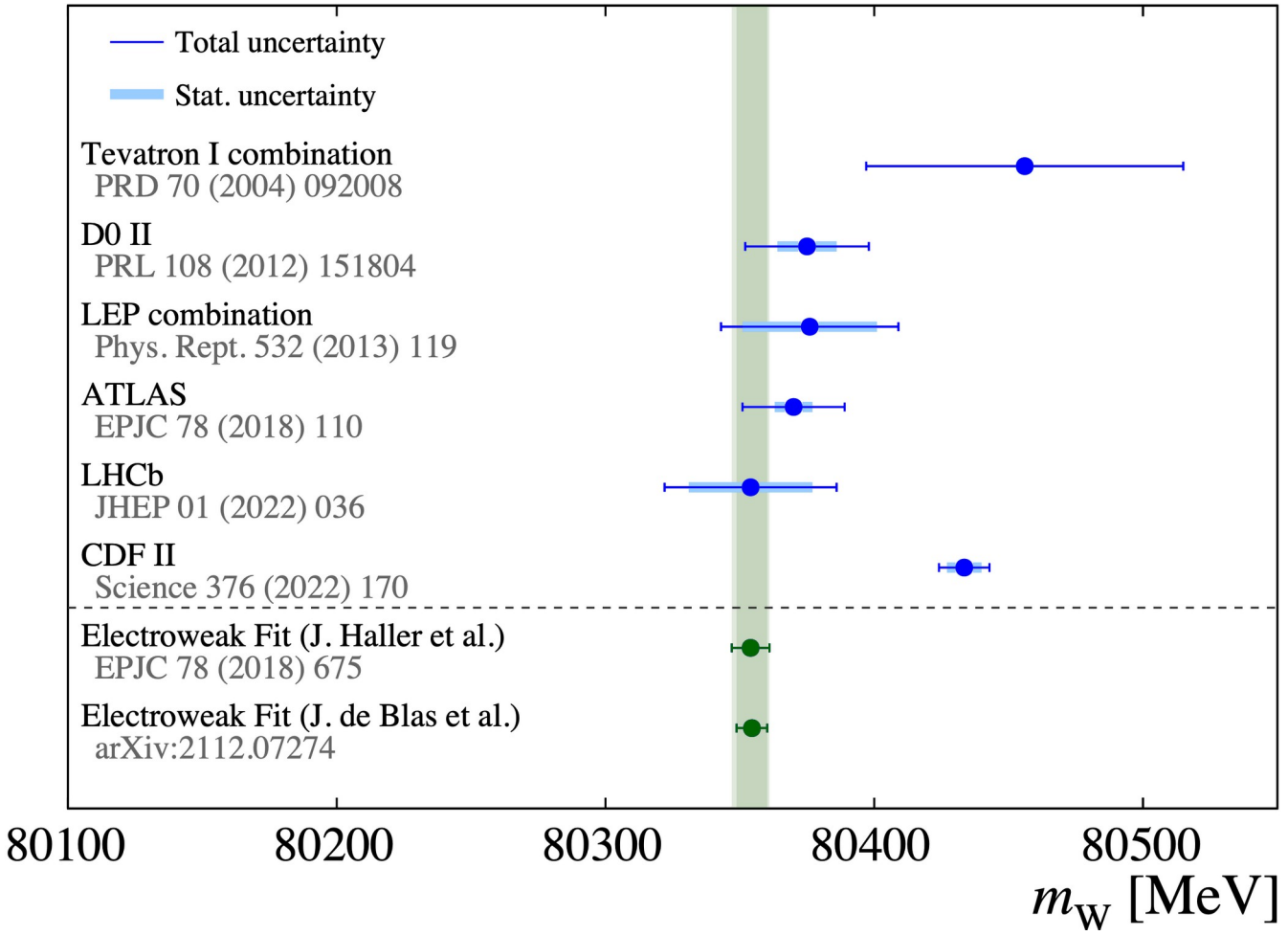


Image credits: CERN courier, LHCb

→ How to compute (and interpret) measurement intervals

→ How to model systematic uncertainties ?

→ How to get the **smallest achievable uncertainties ?**

Introduction

Precision measurements are another window into BSM effects

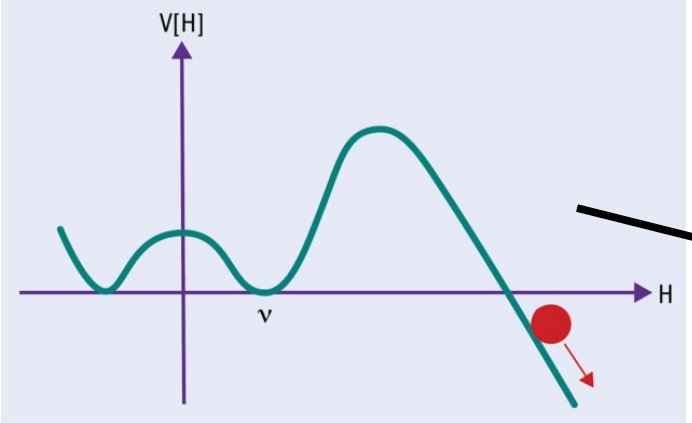
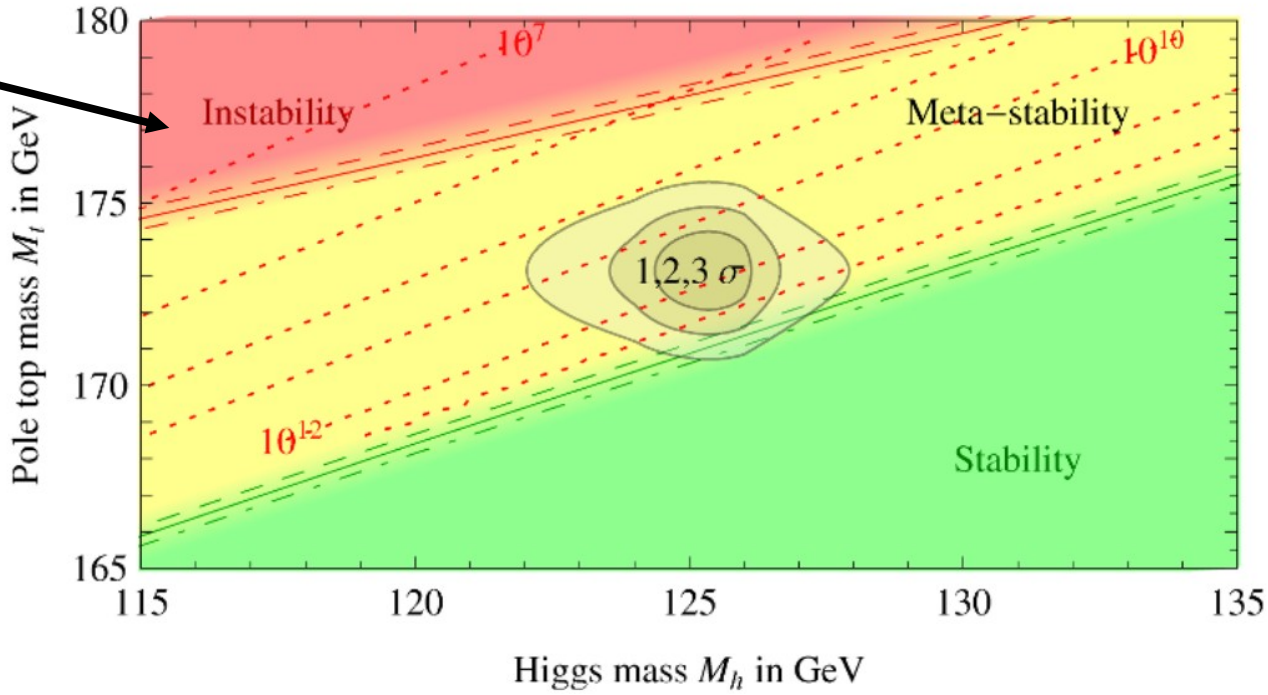


Image credits: CERN courier, JHEP 08 (2012) 098



→ How to compute (and interpret) measurement intervals

→ How to model systematic uncertainties ?

→ How to get the **smallest achievable uncertainties ?**

Course Plan

This morning:

Statistical Modeling (PDFs for HEP measurements)

Parameter estimation (max. likelihood, least-squares)

Model testing (hypothesis testing, p-values, …)

This afternoon (14:00)

Confidence intervals

Upper limits

Systematics and profiling

(Bayesian inference)

Disclaimer: examples and methods covered in the lectures will be biased towards LHC techniques (generally close to the state of the art anyway)

Each lecture will be followed by a **hands-on session**:

- This morning at **11:00**
- This afternoon at **16:00**

The class will also feature hands-on exercises using Jupyter notebooks

Hands-on exercises

The Statistics course will include Hands-on exercises on **jupyter notebooks** (built using the **`numpy/scipy/pyplot`** stack) are part of the course.

If you have a computer with you, **please install anaconda** as this provides a consistent installation of python, JupyterLab, etc.

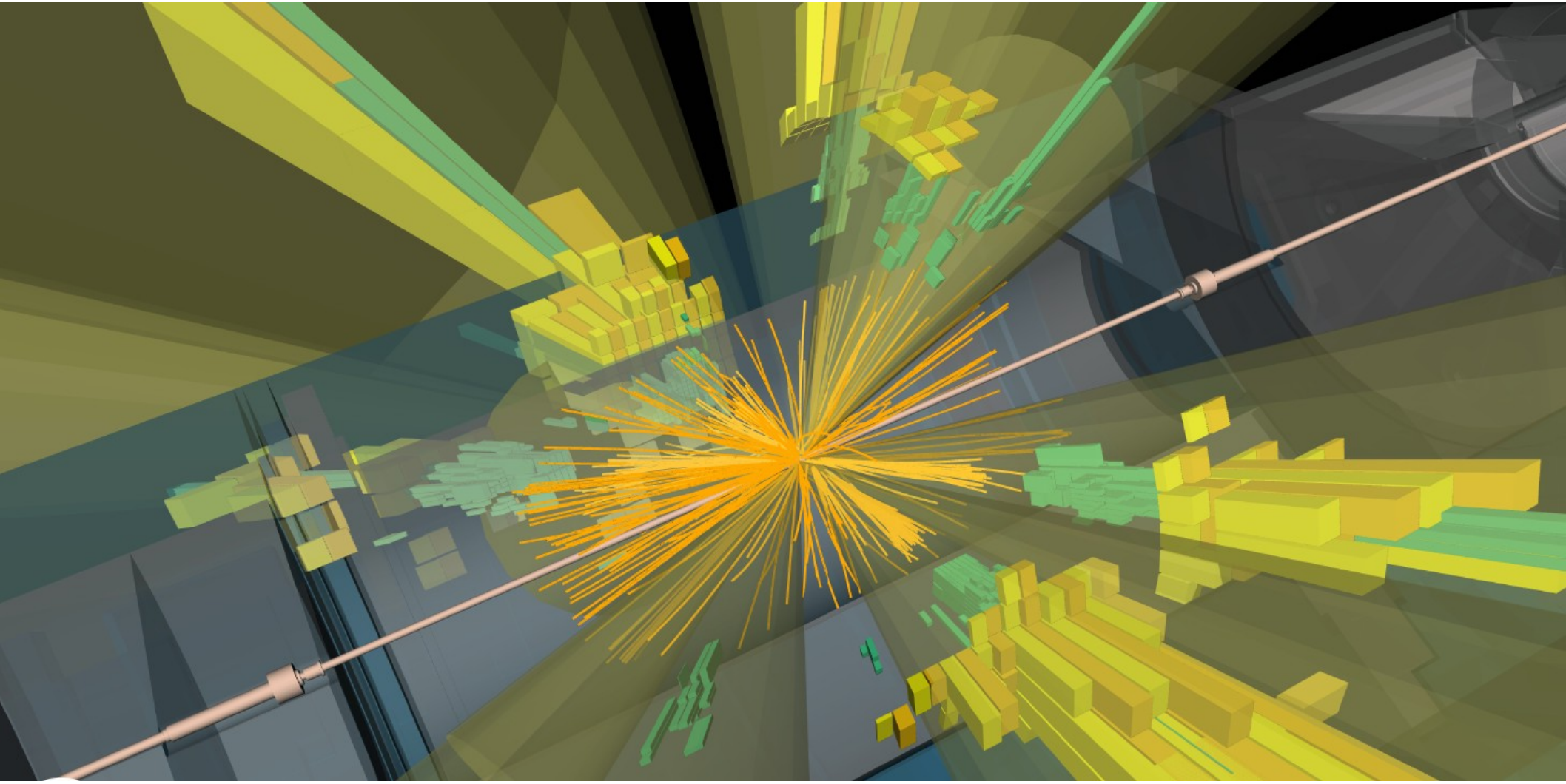
→ Alternatively, you can also install JupyterLab as a standalone package.

→ Another solution is to run on the public jupyter servers at mybinder.org. This will probably be slower but avoids a local install.

Lecture 1	Lecture Notes	notebook [solutions]	binder [solutions]
Lecture 2	Lecture Notes	notebook [solutions]	binder [solutions]

Randomness in High-Energy Physics

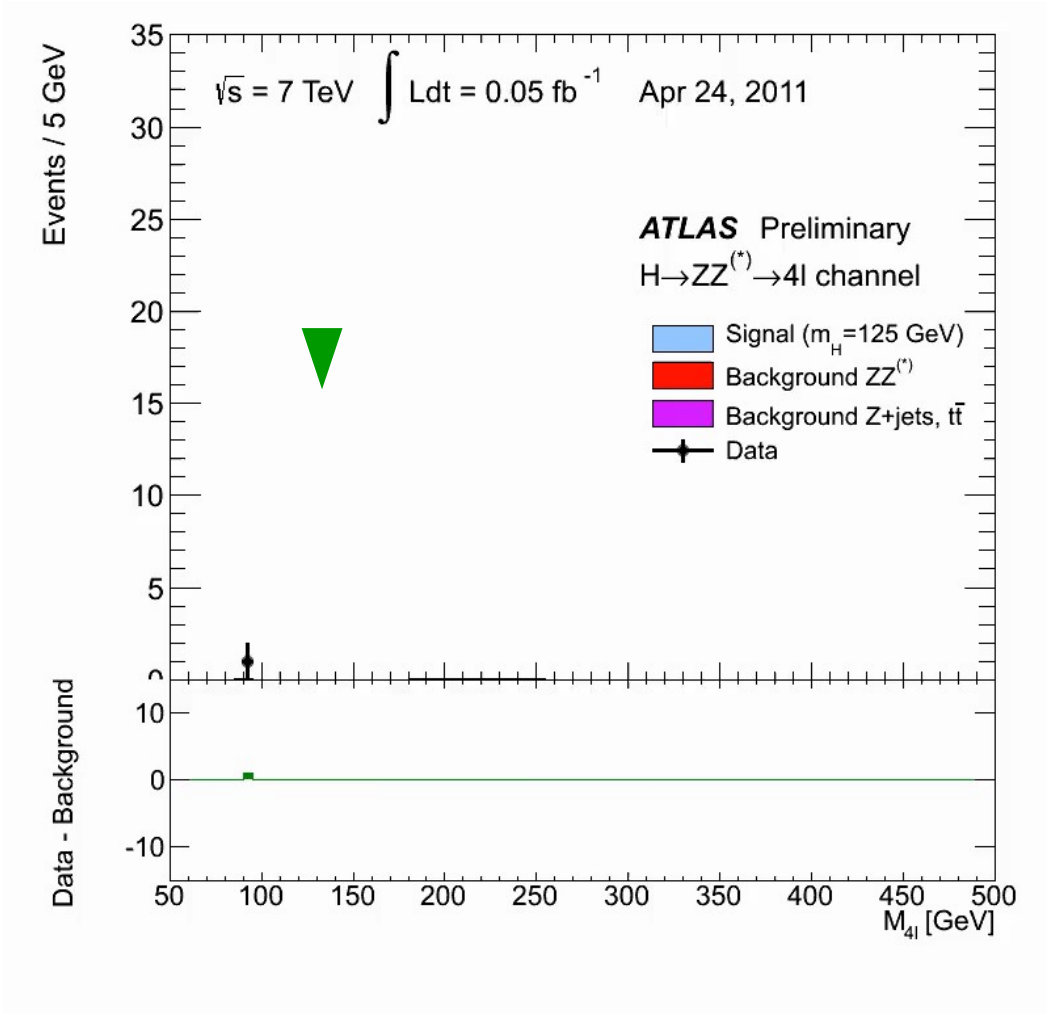
Experimental data is produced by **incredibly complex** processes



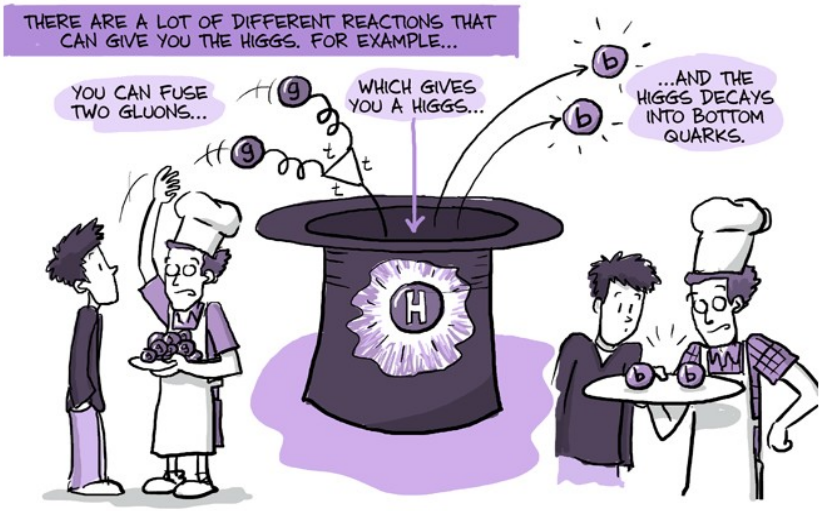
These processes are **random**: **classical randomness** in measurements + **quantum randomness** in physics processes

Quantum Randomness: H→ZZ*→4l

Phys. Rev. D **91**, 012006



Rare process: Expect 1 signal event every **~6 days**

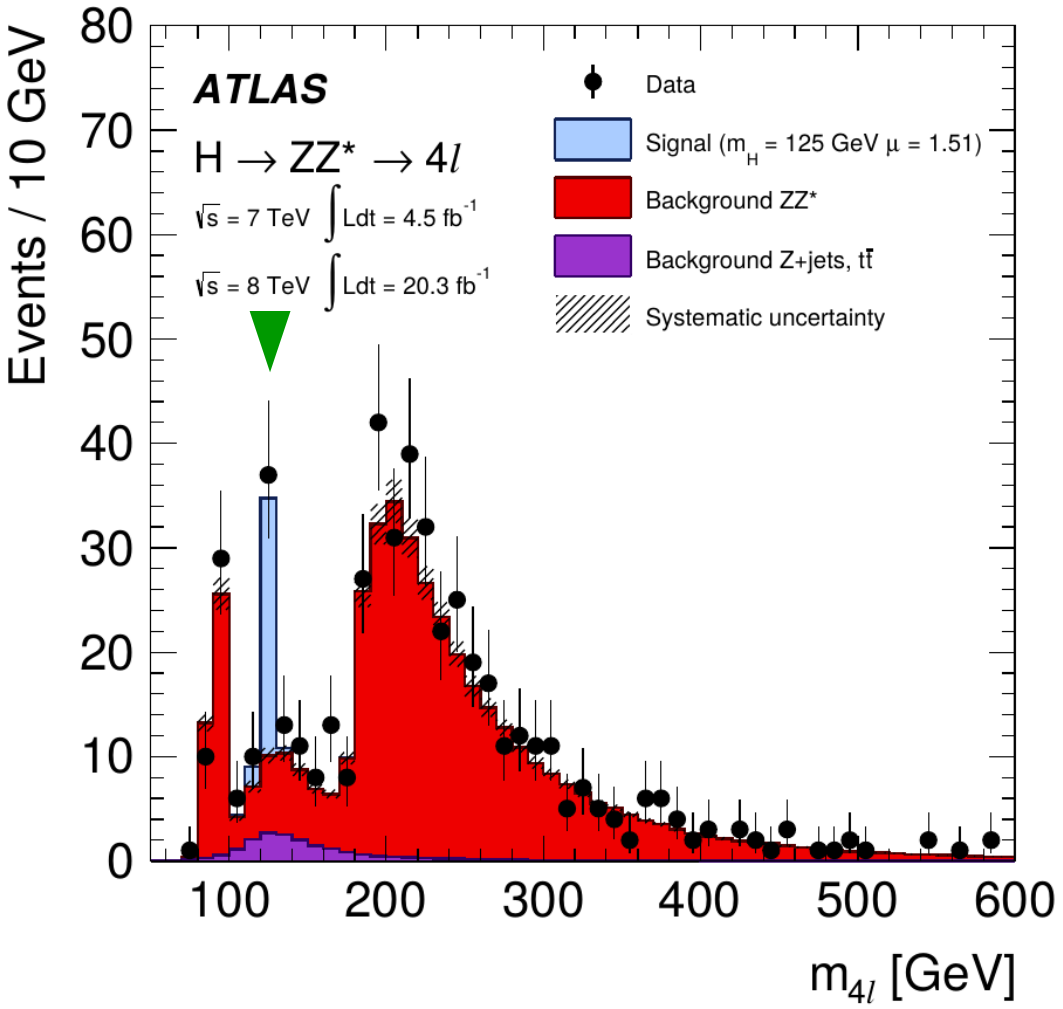


http://www.phdcomics.com/comics/archive.php?comicid=1489

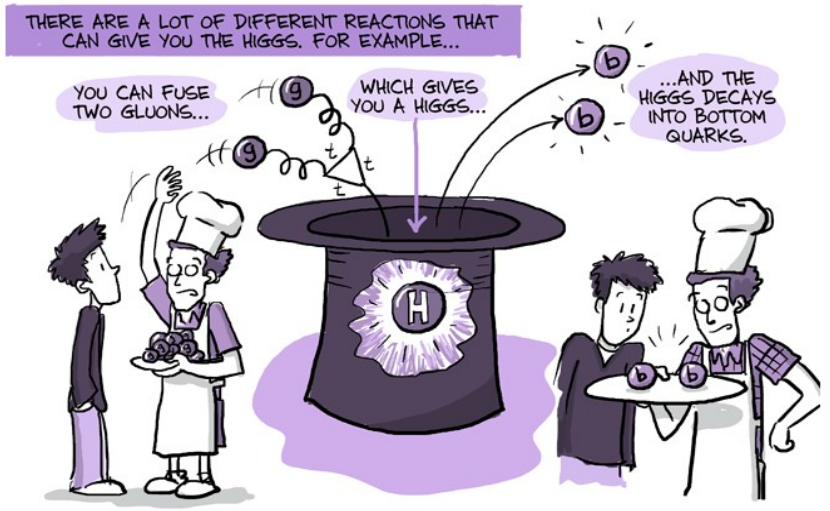
View online

Quantum Randomness: H→ZZ*→4l

Phys. Rev. D **91**, 012006



Rare process: Expect 1 signal event every **~6 days**

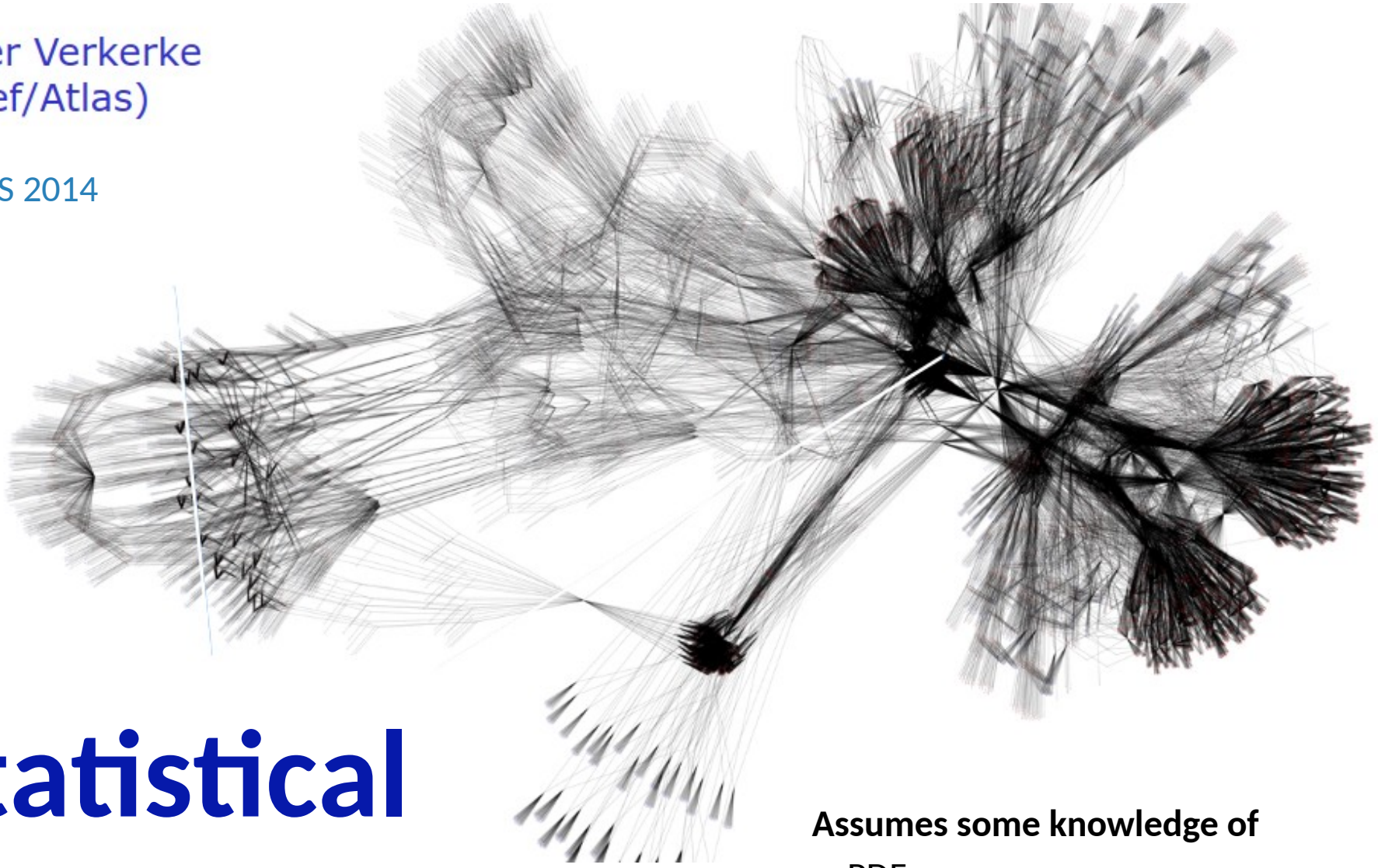


http://www.phdcomics.com/comics/archive.php?comicid=1489

"Will I get an event today ?" → only **probabilistic** answer

Wouter Verkerke
(Nikhef/Atlas)

SOS 2014



Statistical Modeling

Assumes some knowledge of

- PDFs
- Gaussian distributions
- χ^2 distributions
- Central limit theorem

Let us know if short reminders on any of these would be useful!

Example 1: Z counting

Phys. Lett. B 759 (2016) 601

Measure the cross-section (event rate) of the Z→ ee process

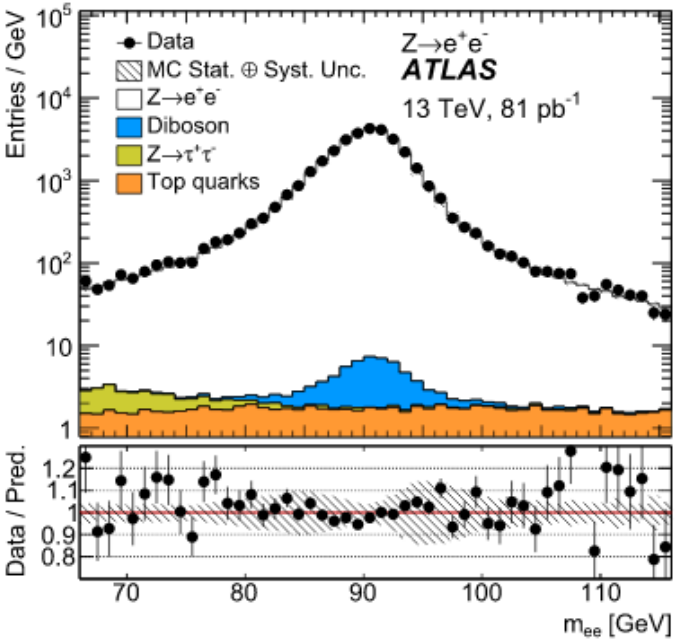
35000 ± 187

175 ± 8

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

(81 ± 2) pb^{-1}

0.552 ± 0.006



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

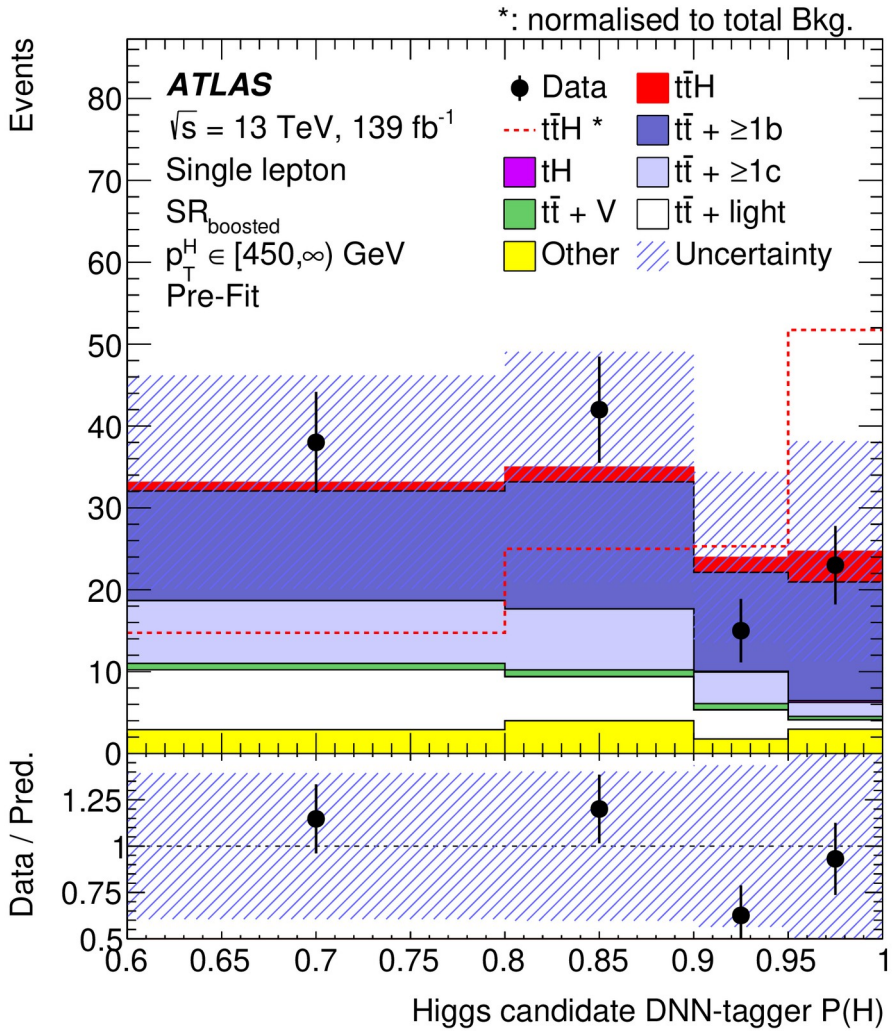
Statistical uncertainties
Fluctuations in the data counts

Systematic uncertainties
Everything else: assumptions, parameter values, …

"Single bin counting" : only data input is **n_{data}**.

Example 2: ttH→bb

arXiv:2111.06712



Event counting in different regions:

Multiple-bin counting

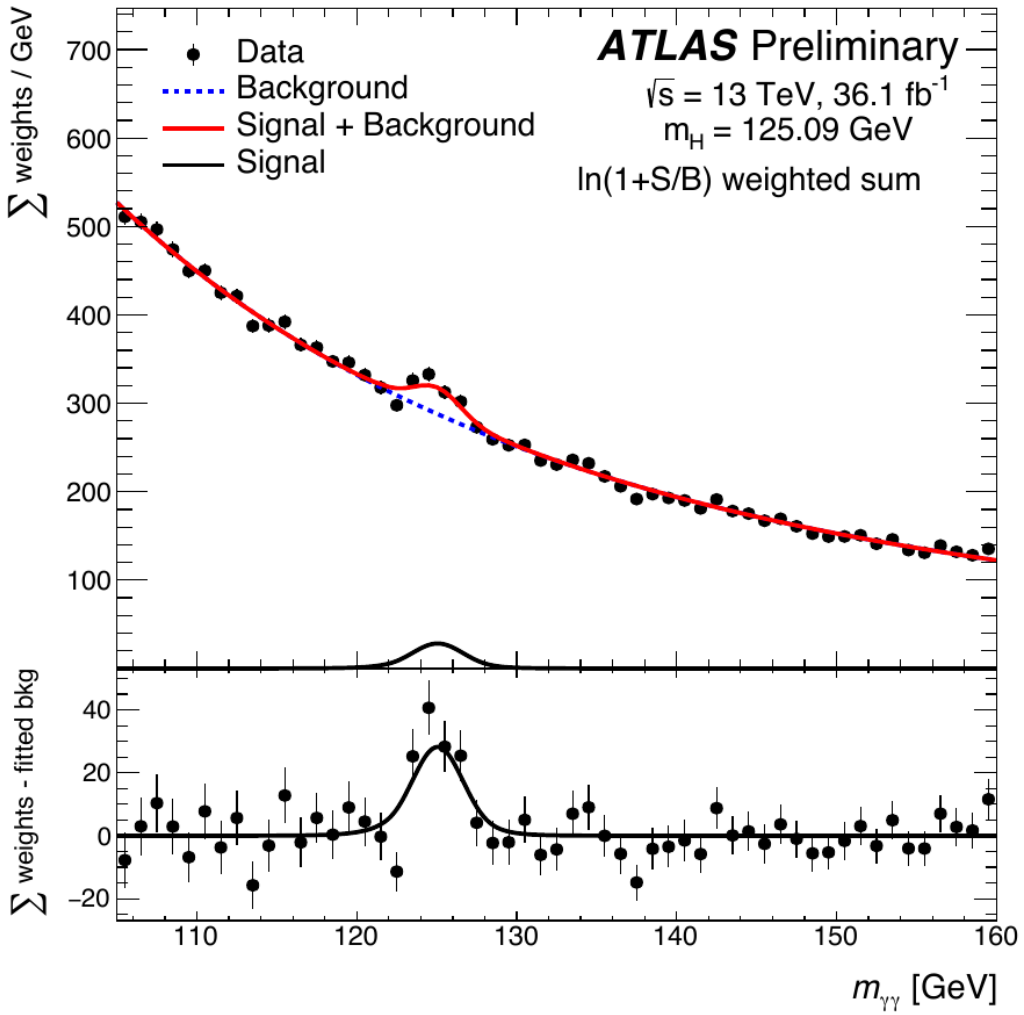
Lots of information available

→ 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

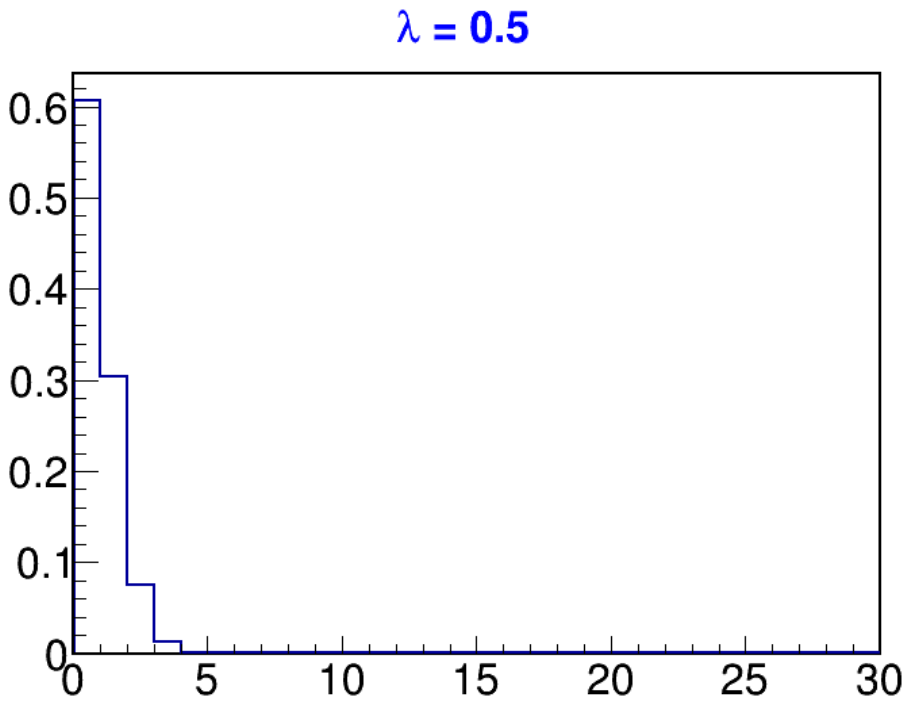
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ **Poisson approximation**.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

RMS (σ) = √λ

For a counting measurement, σ = √N

Central limit theorem : Poisson becomes **Gaussian for large λ** :

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

How to count

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

→ In principle, binomial process

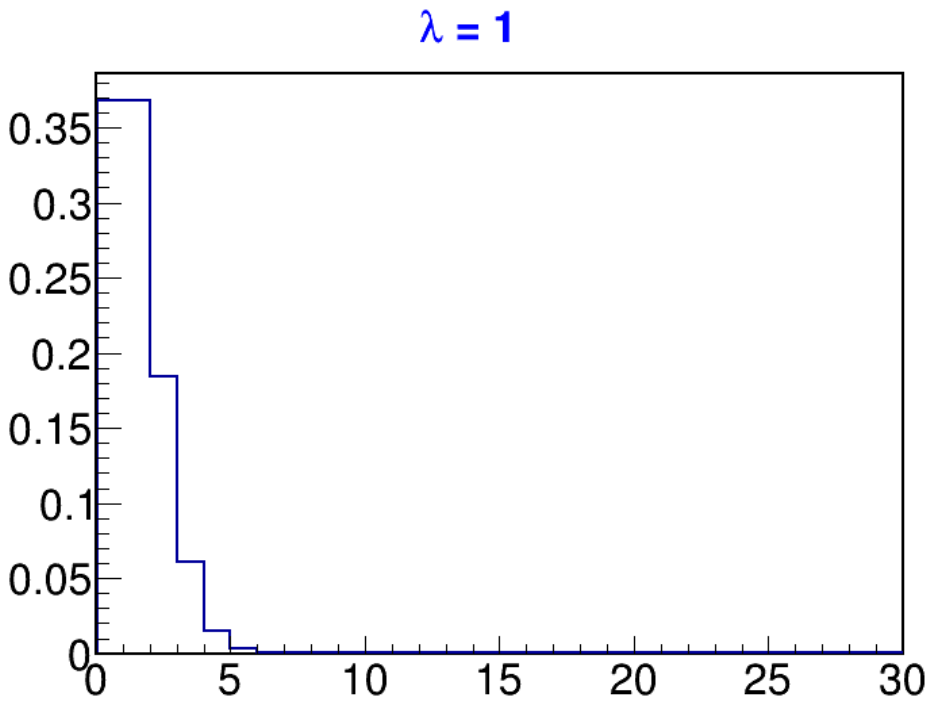
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ **Poisson approximation**.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

RMS (σ) = √λ

For a counting measurement, σ = √N

Central limit theorem : Poisson 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

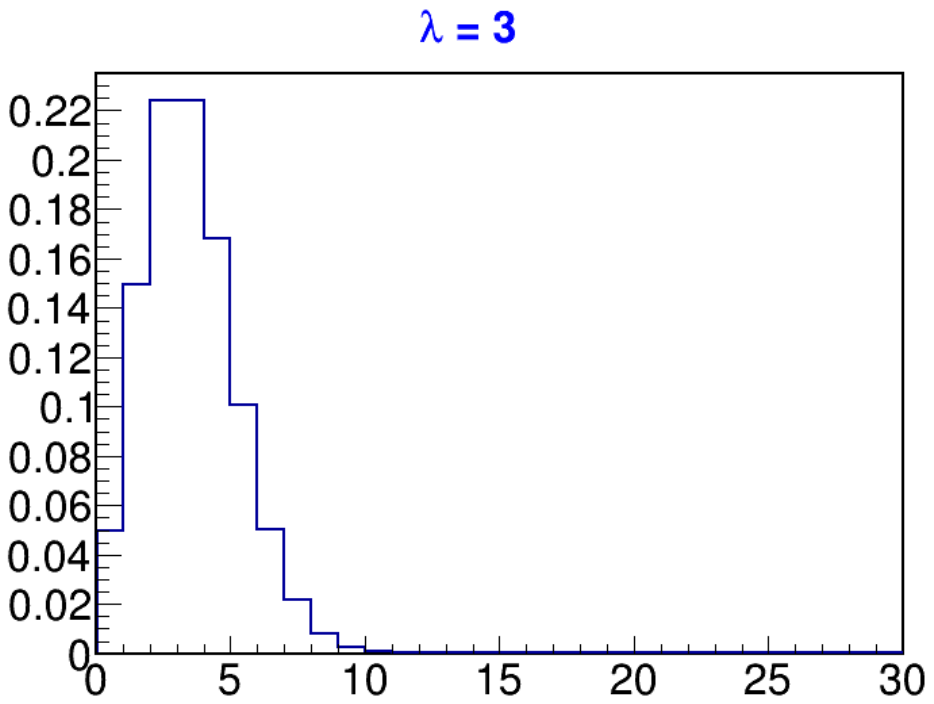
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ **Poisson approximation**.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

RMS (σ) = √λ

For a counting measurement, σ = √N

Central limit theorem : Poisson 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

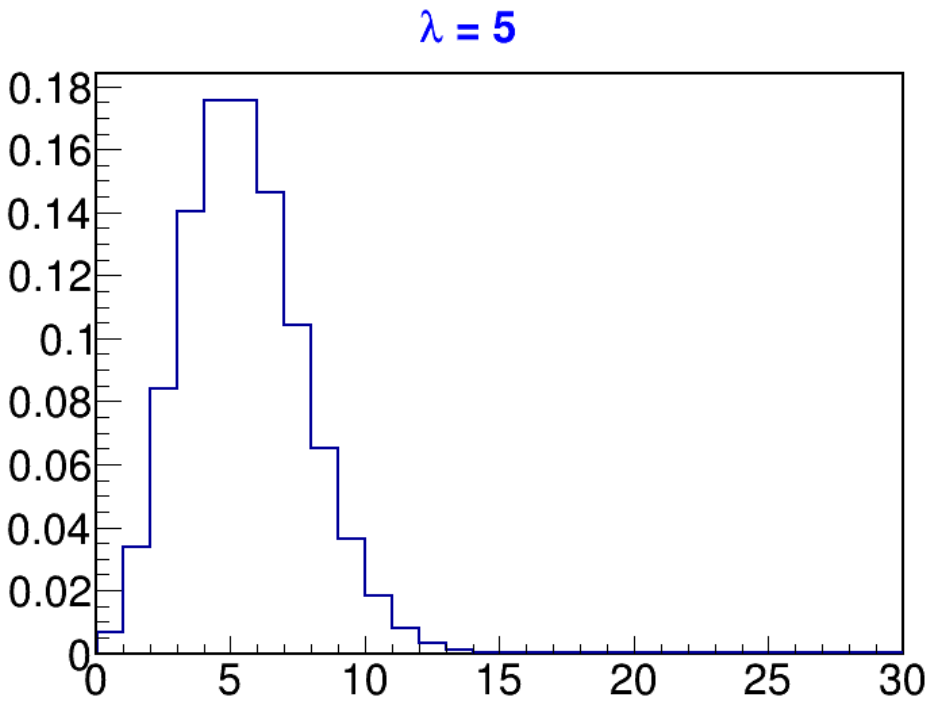
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ **Poisson approximation**.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

RMS (σ) = √λ

For a counting measurement, σ = √N

Central limit theorem : Poisson 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

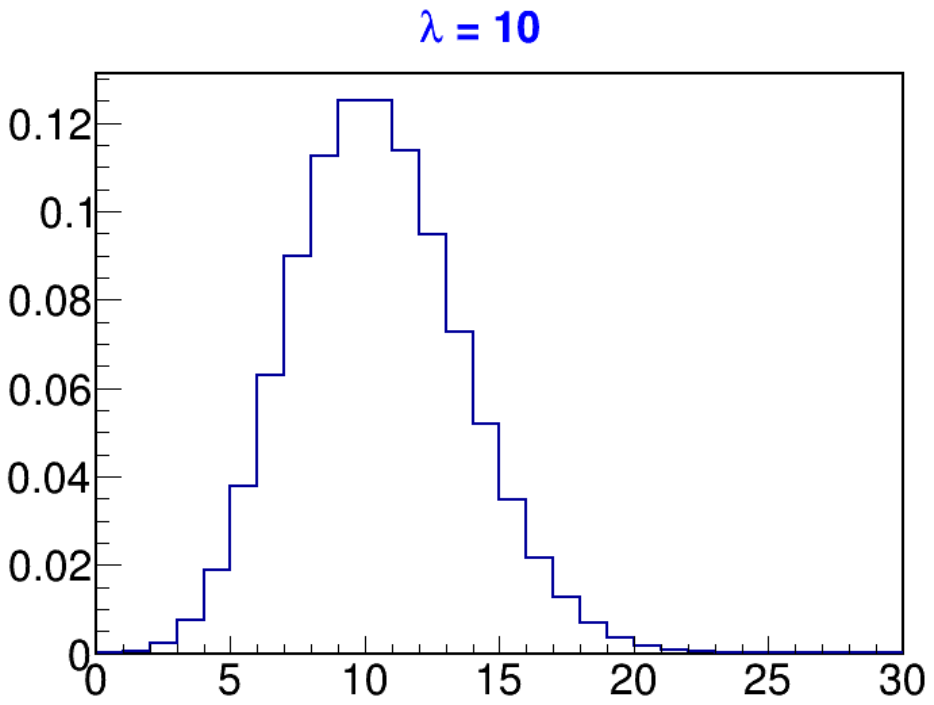
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ **Poisson approximation**.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

RMS (σ) = √λ

For a counting measurement, σ = √N

Central limit theorem : Poisson 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**.

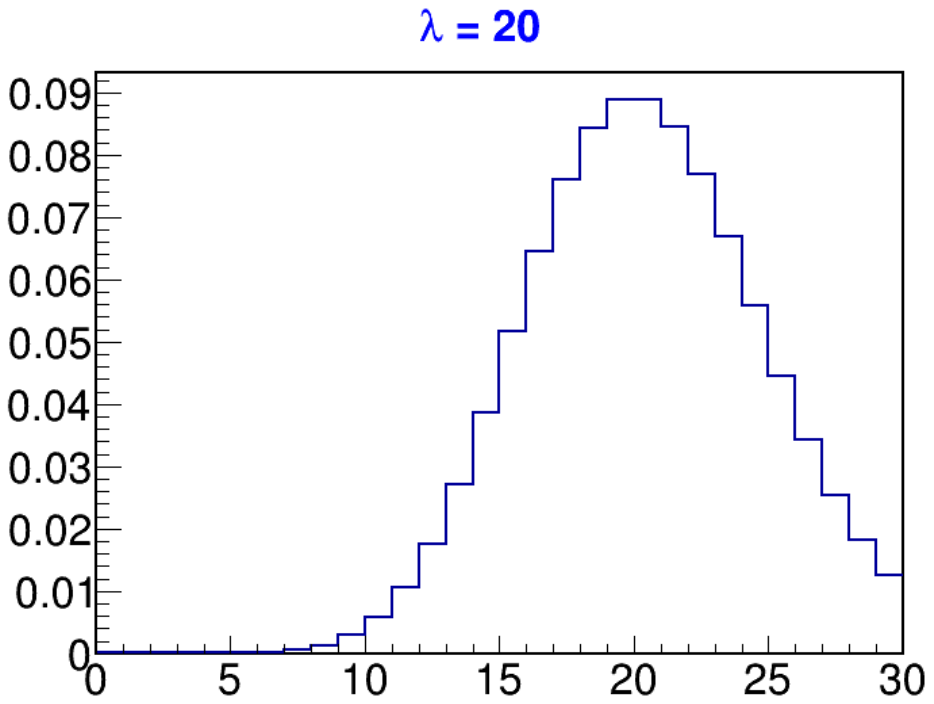
→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution

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

λ = NP

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



Mean = λ

Variance = λ

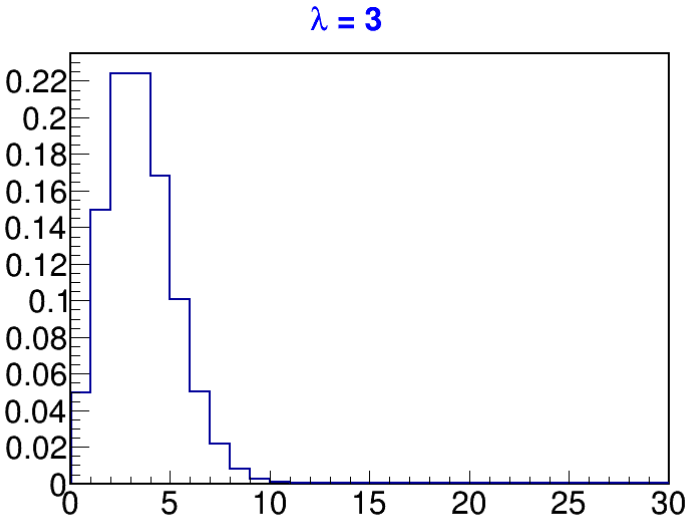
RMS (σ) = √λ

For a counting measurement, σ = √N

Central limit theorem : Poisson becomes **Gaussian for large λ** :

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

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

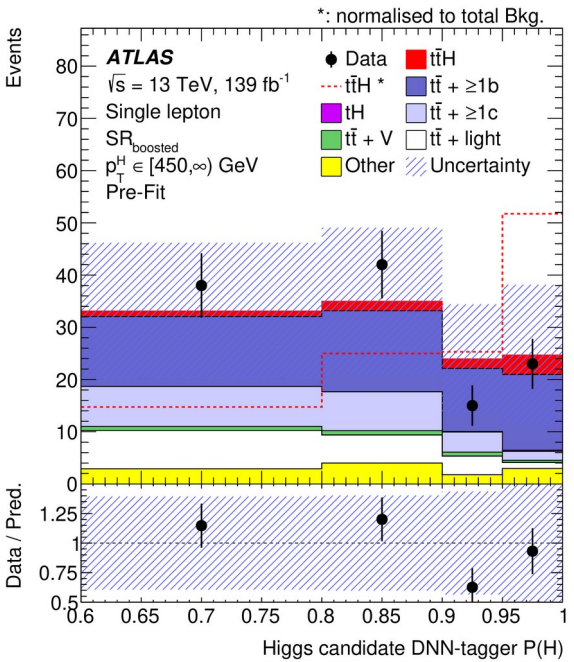
Count in bins of a variable ⇒ *histogram* $n_1 \dots n_N$.

(N : number of bins)

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

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

Poisson distribution in each bin



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

In HEP, generally good modeling from simulation (with some uncertainties)

Also 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**, $\mathbf{m_{top}}$, …

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

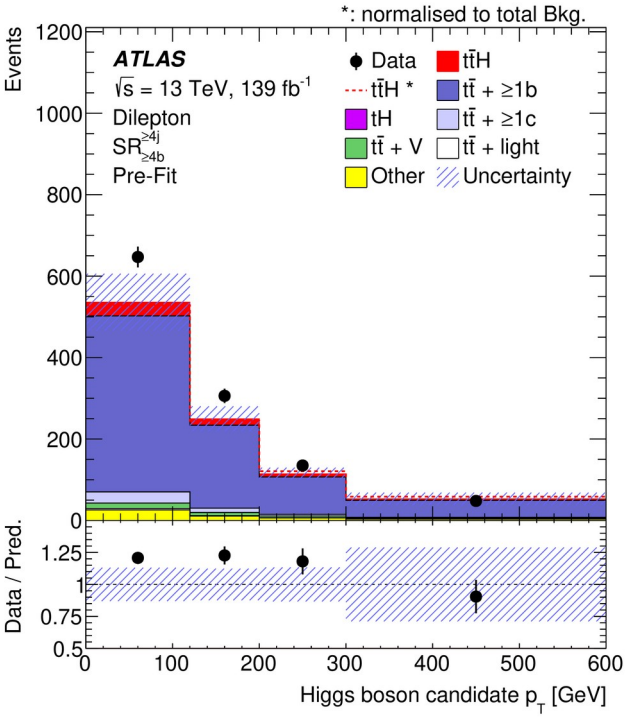
  → Background levels (**B**)

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

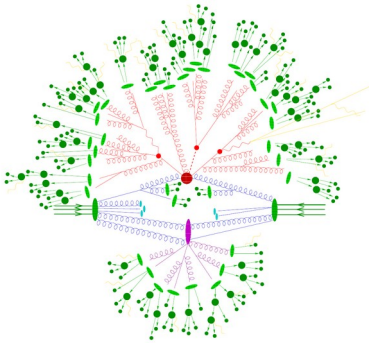
NPs must be either:

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

→ **Constrained by the data**



Takeaways



Random data must be described using a **statistical model**. Usual cases:

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

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

How do we obtain the values of the POIs ?

Parameter estimation

What a PDF is for

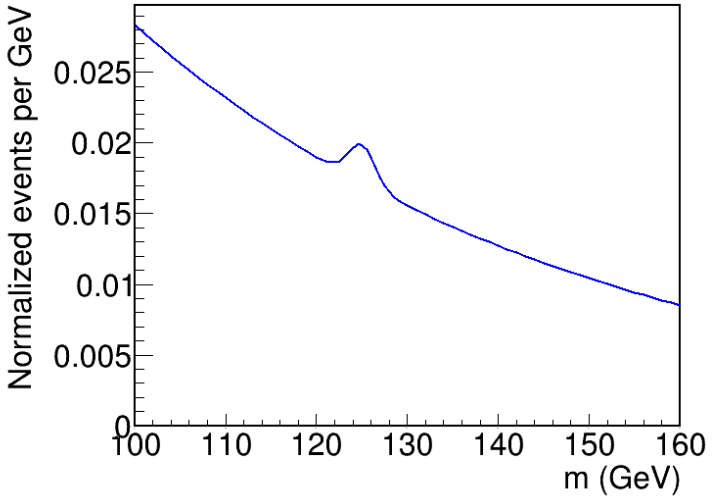
Model describes the distribution of the observable: **P(data; parameters)**

⇒ Possible outcomes of the experiment, for given parameter values

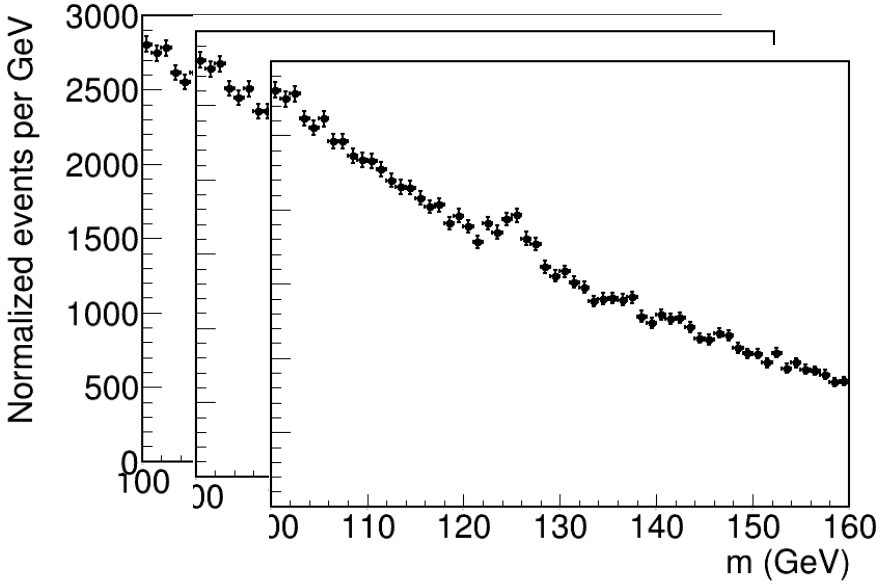
Can draw random events according to PDF : **generate *pseudo-data***

$P(\lambda = 5)$ → **2, 5, 3, 7, 4, 9, ….**

Each entry = separate “experiment”



Generate



P(data; parameters)

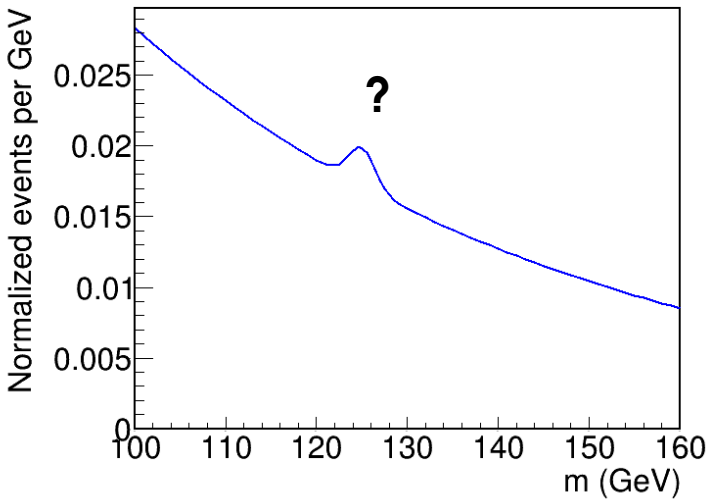
What a PDF is also for: Likelihood

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

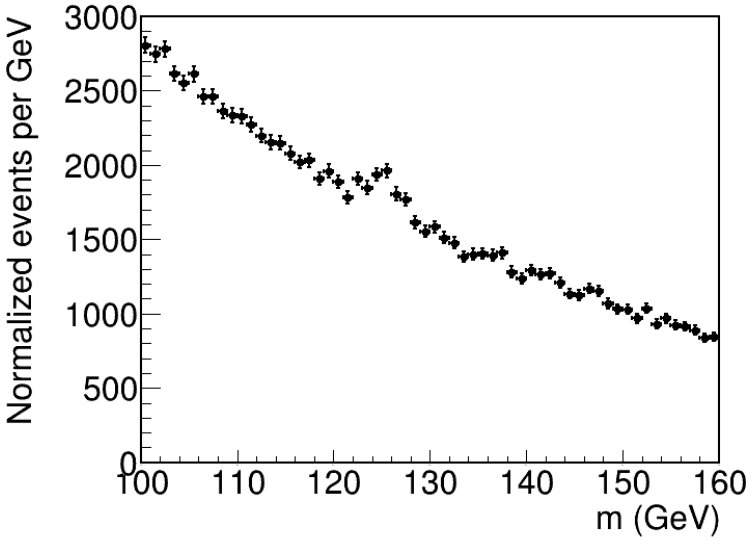
⇒ Possible outcomes of the experiment, for given parameter values

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

$P(\lambda = ?)$ ⟸ **5**



Estimate ⟸



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

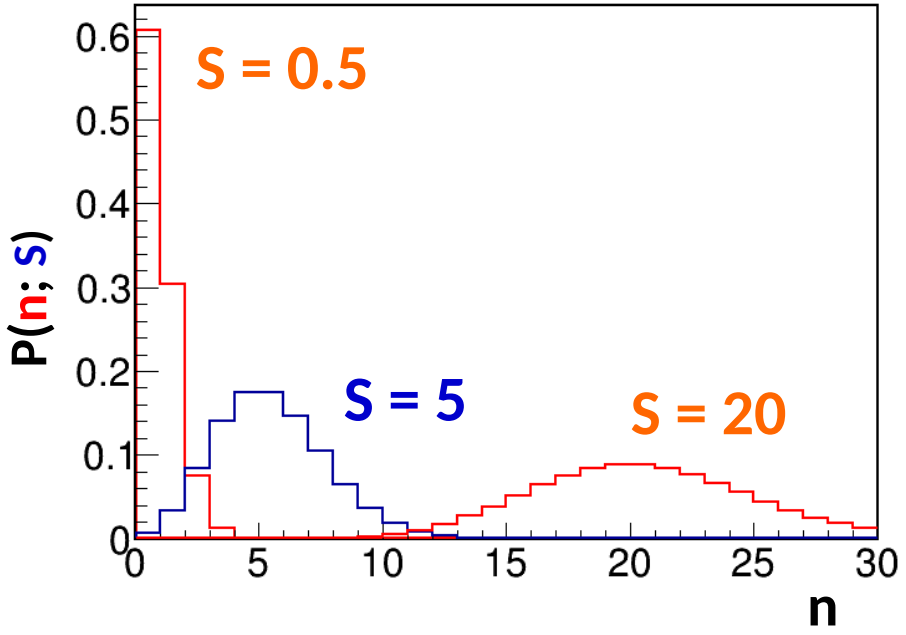
→ **Same as the PDF,** but seen as function of the parameters

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**

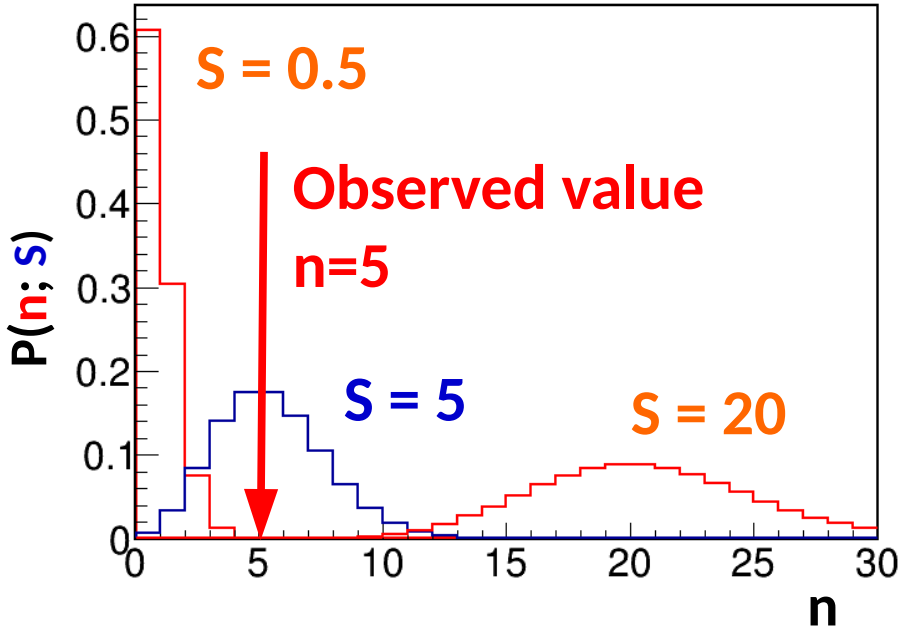
No guarantee it is the true value (data may be "unlikely") but sensible estimate

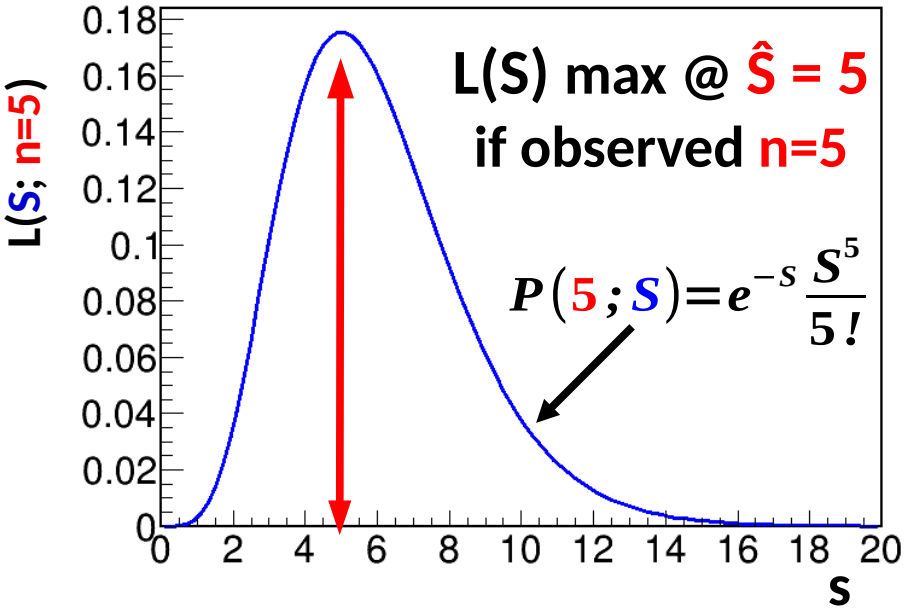
Maximum Likelihood Estimation

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

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

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



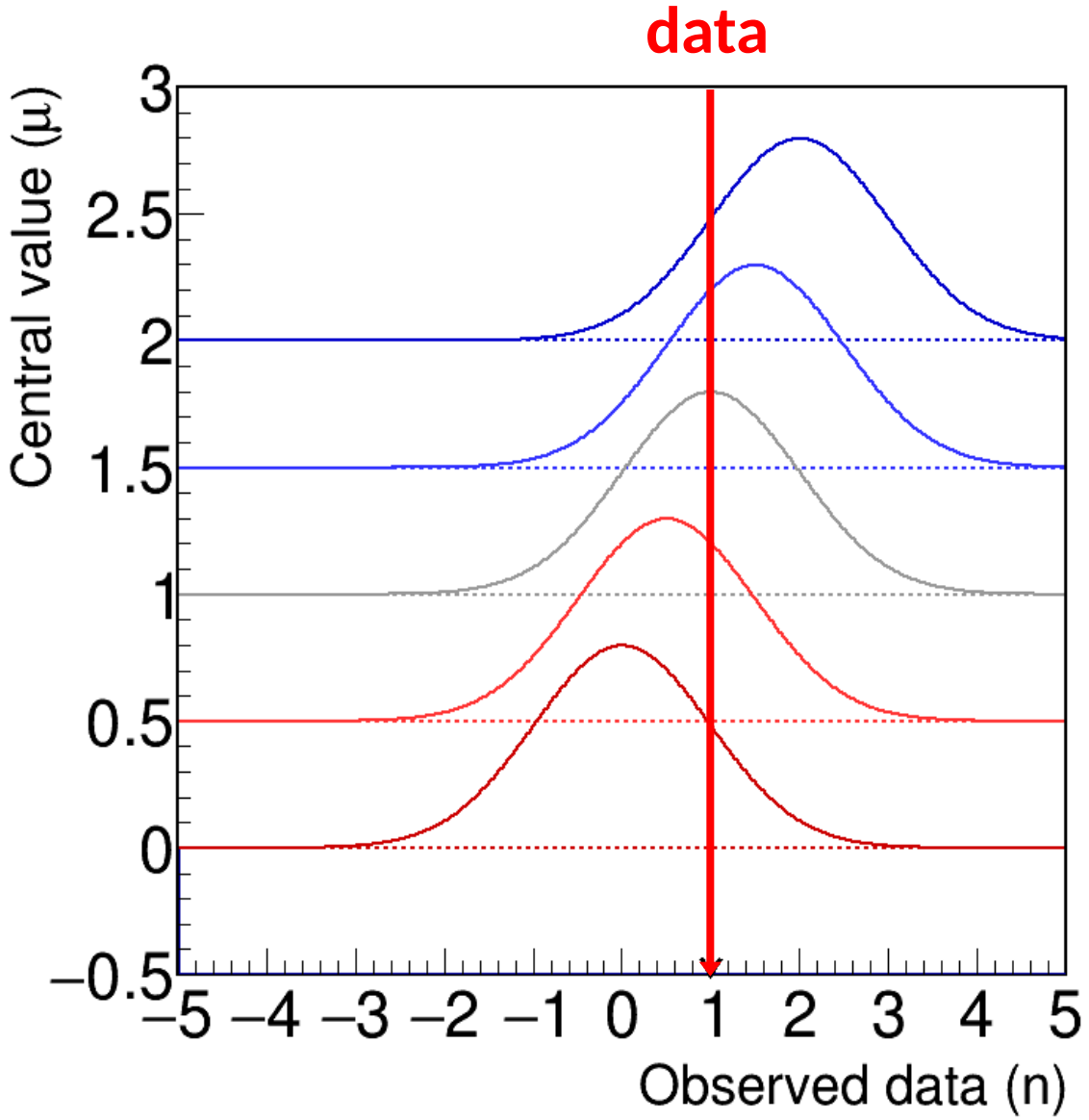


MLE: the value of μ for which **this data** was ***most likely to occur***

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

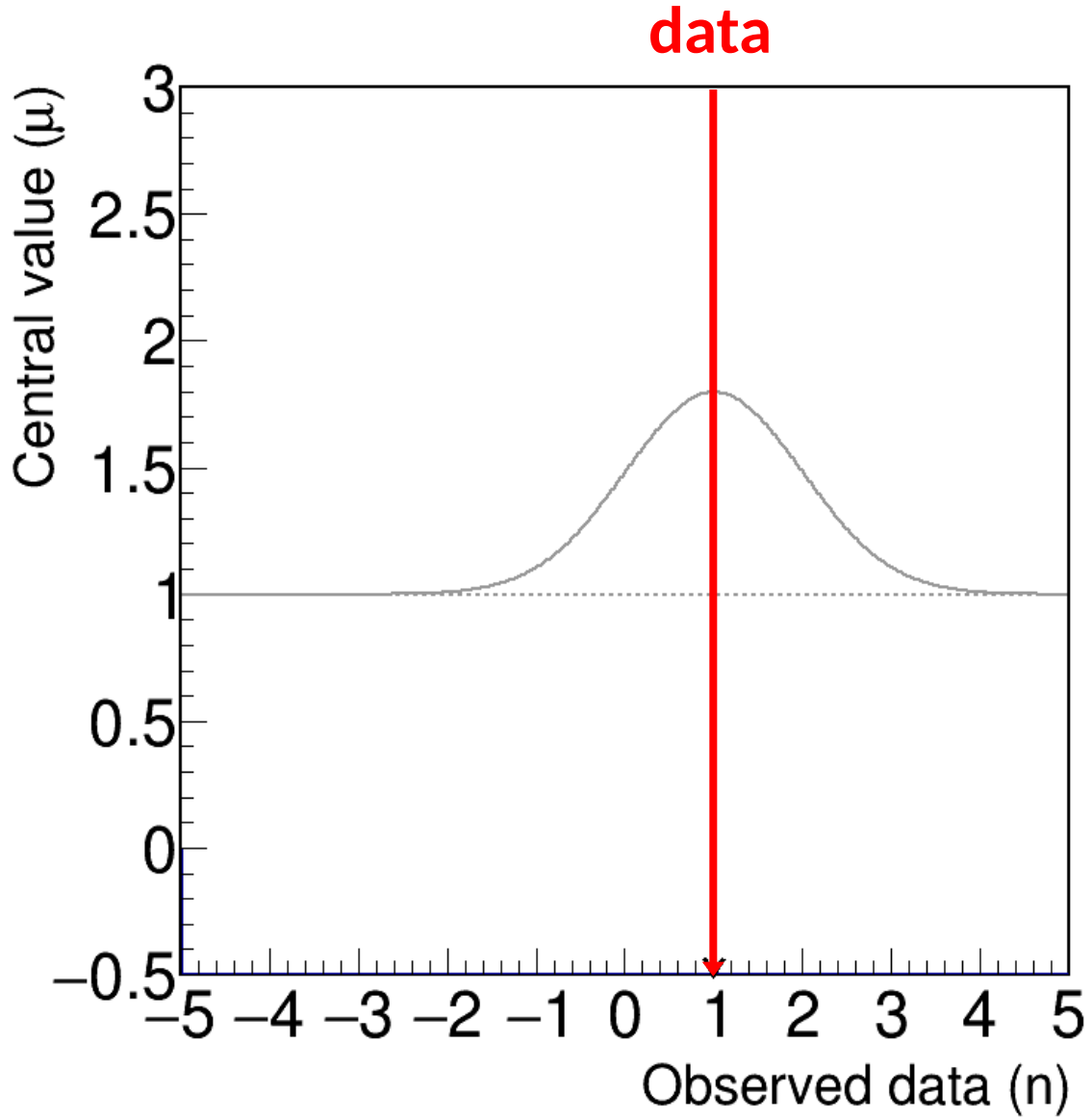
No guarantee it is the true value (data may be "unlikely") but sensible estimate

Gaussian case



Best-fit of Gaussian PDF mean to observed data

Gaussian case



Best-fit of Gaussian PDF mean to observed data

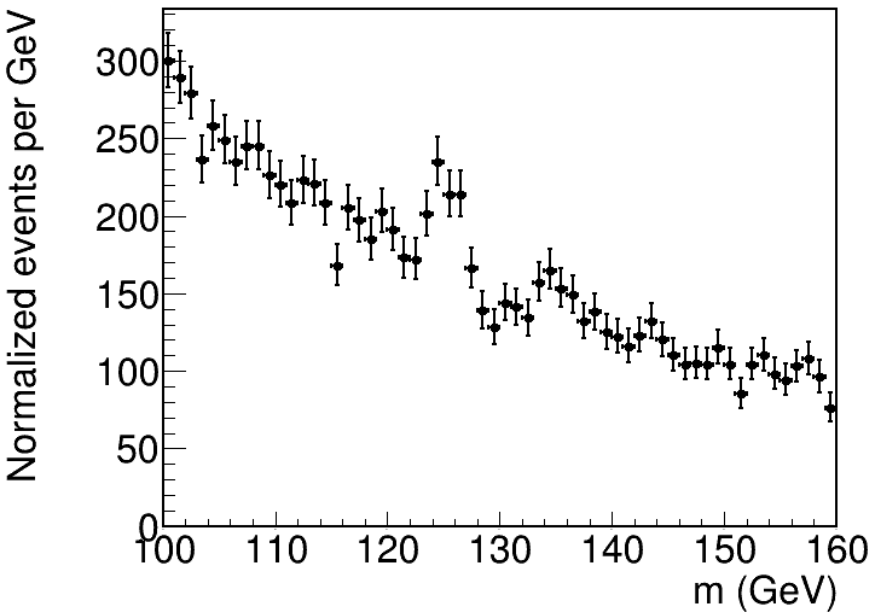
Multiple Gaussian bins

PDF for independent Gaussian bins:

$$P(n_i;\mu)=\prod_{i=1}^{N_{\text{bins}}} G(n_i;y_i(\mu),\sigma_i)$$

Likelihood

$$L(\mu)=\prod_{i=1}^{N_{\text{bins}}}\frac{1}{\sqrt{2\pi\sigma_i}}\exp\left(-\frac{(n_i-y_i(\mu))^2}{2\sigma_i^2}\right)$$



-2 log Likelihood

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

Doesn't depend on μ

Gaussian case: MLE = Minimum χ^2 = Least-squares minimization.

General case: Typically need non-linear minimization.

HEP practice: **MINUIT** (C++ library within ROOT, numerical gradient descent)

scipy.minimize: many algorithms, NumPy/TF/PyTorch/… backends.

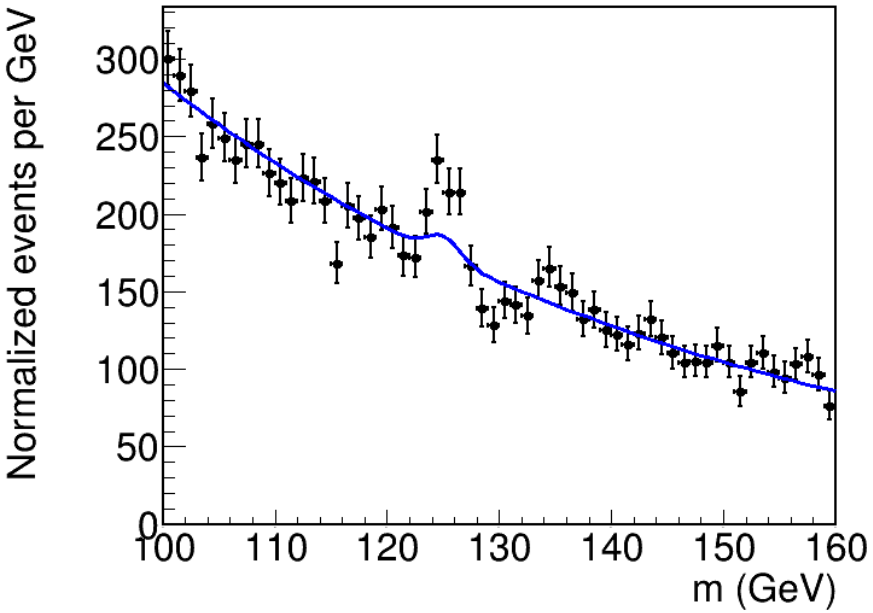
Multiple Gaussian bins

PDF for independent Gaussian bins:

$$P(n_i;\mu)=\prod_{i=1}^{N_{\text{bins}}} G(n_i;y_i(\mu),\sigma_i)$$

Likelihood

$$L(\mu)=\prod_{i=1}^{N_{\text{bins}}}\frac{1}{\sqrt{2\pi\sigma_i}}\exp\left(-\frac{(n_i-y_i(\mu))^2}{2\sigma_i^2}\right)$$



-2 log Likelihood

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

Doesn't depend on μ

Gaussian case: MLE = Minimum χ^2 = Least-squares minimization.

General case: Typically need non-linear minimization.

HEP practice: **MINUIT** (C++ library within ROOT, numerical gradient descent)

scipy.minimize: many algorithms, NumPy/TF/PyTorch/… backends.

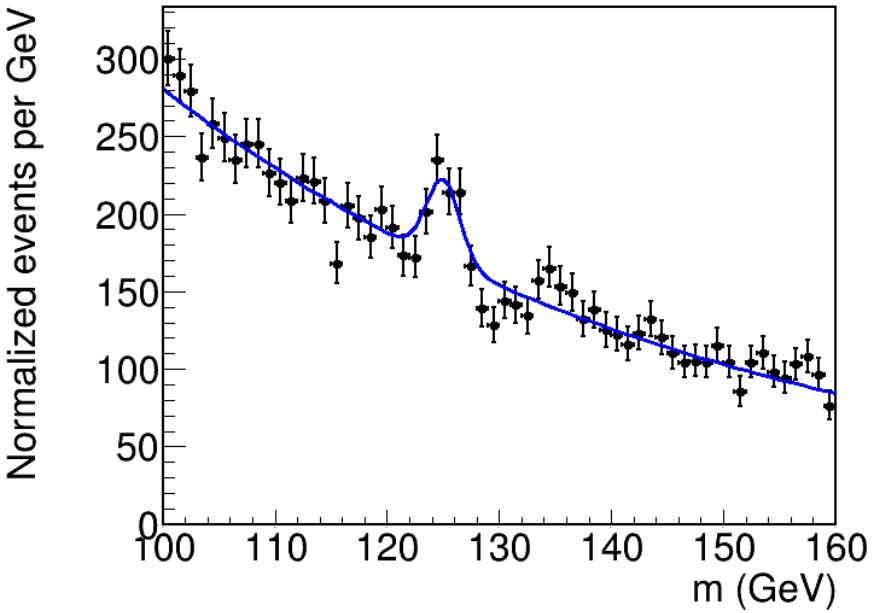
Multiple Gaussian bins

PDF for independent Gaussian bins:

$$P(n_i;\mu)=\prod_{i=1}^{N_{\text{bins}}} G(n_i;y_i(\mu),\sigma_i)$$

Likelihood

$$L(\mu)=\prod_{i=1}^{N_{\text{bins}}}\frac{1}{\sqrt{2\pi\sigma_i}}\exp\left(-\frac{(n_i-y_i(\mu))^2}{2\sigma_i^2}\right)$$



-2 log Likelihood

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

Doesn't depend on μ

Gaussian case: MLE = Minimum χ^2 = Least-squares minimization.

General case: Typically need non-linear minimization.

HEP practice: **MINUIT** (C++ library within ROOT, numerical gradient descent)

scipy.minimize: many algorithms, NumPy/TF/PyTorch/… backends.

Confidence Intervals

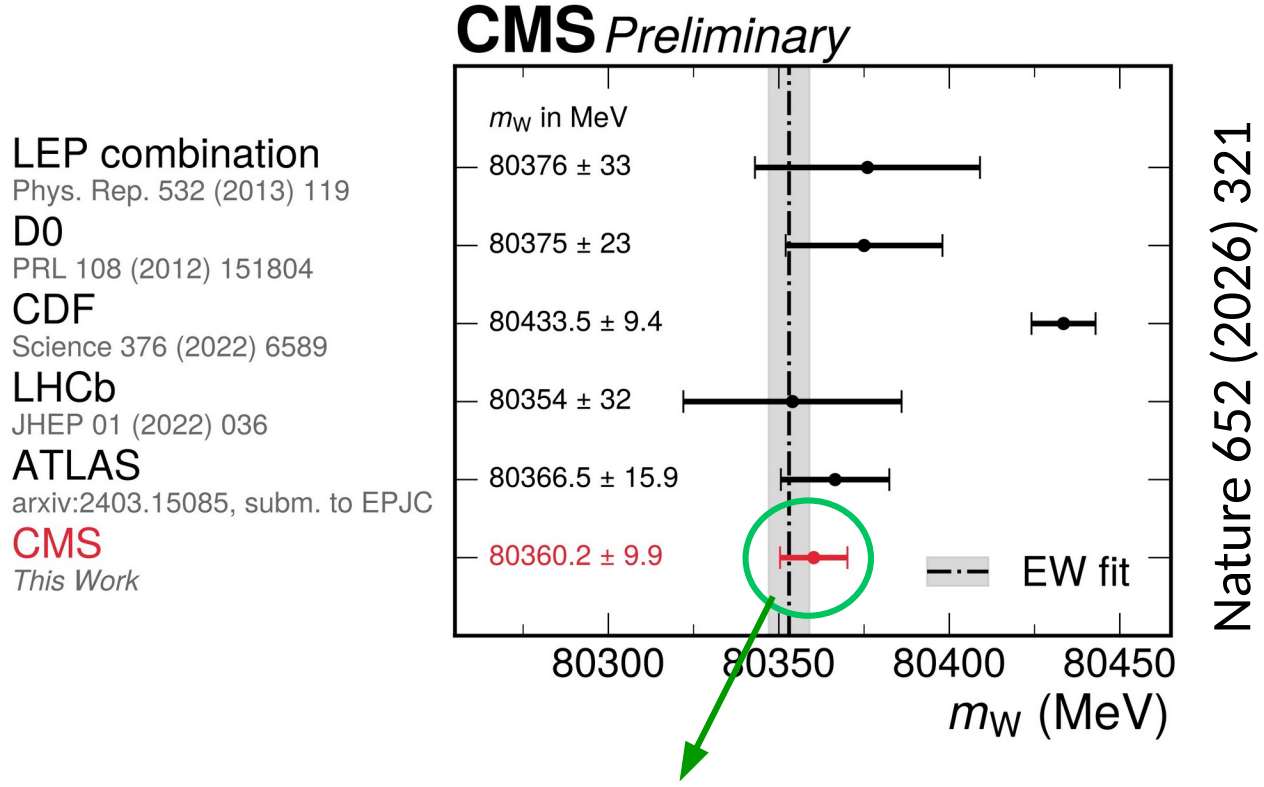
Uncertainties on best-fit values

"Best-fit" value from MLE : $\hat{\mu} = arg\,max\, L(\mu) = arg\,min\left[\lambda(\mu) = -2\log L(\mu)\right]$

However we also need to estimate the associated **uncertainties.**

Uncertainties carry important messages about measurements!

What is the meaning of the uncertainty bar ?



"We don't know the true value, but **there is a 68.3% chance that it is within the bar**"

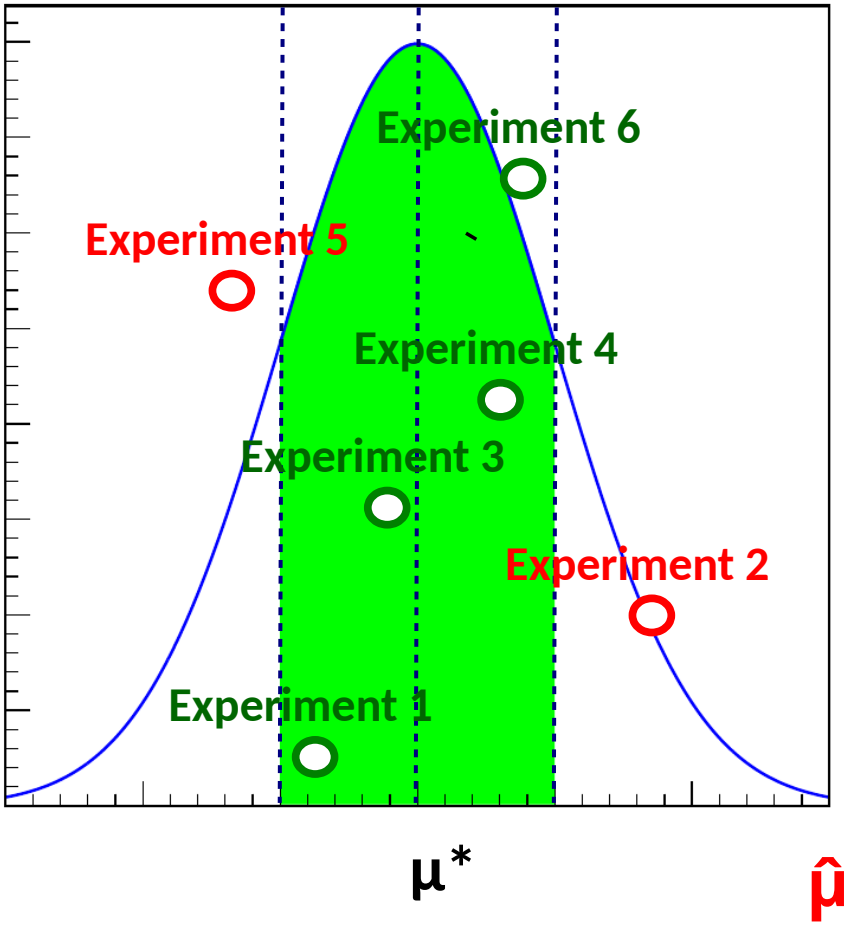
What we want

For a Gaussian, $P(\mu - \sigma < n < \mu + \sigma) = 68.3\,\%$ **68.3% confidence = "1σ"**

We repeat the same experiment multiple times, want to report a 1σ interval (68.3% CL) in each case.

Goal: we want 68.3% of these intervals to contain the true value μ*.

Crucially, must be able to do so without knowing μ* !



"We don't know the true value, but **there is a 68.3% chance that it is within the bar**"

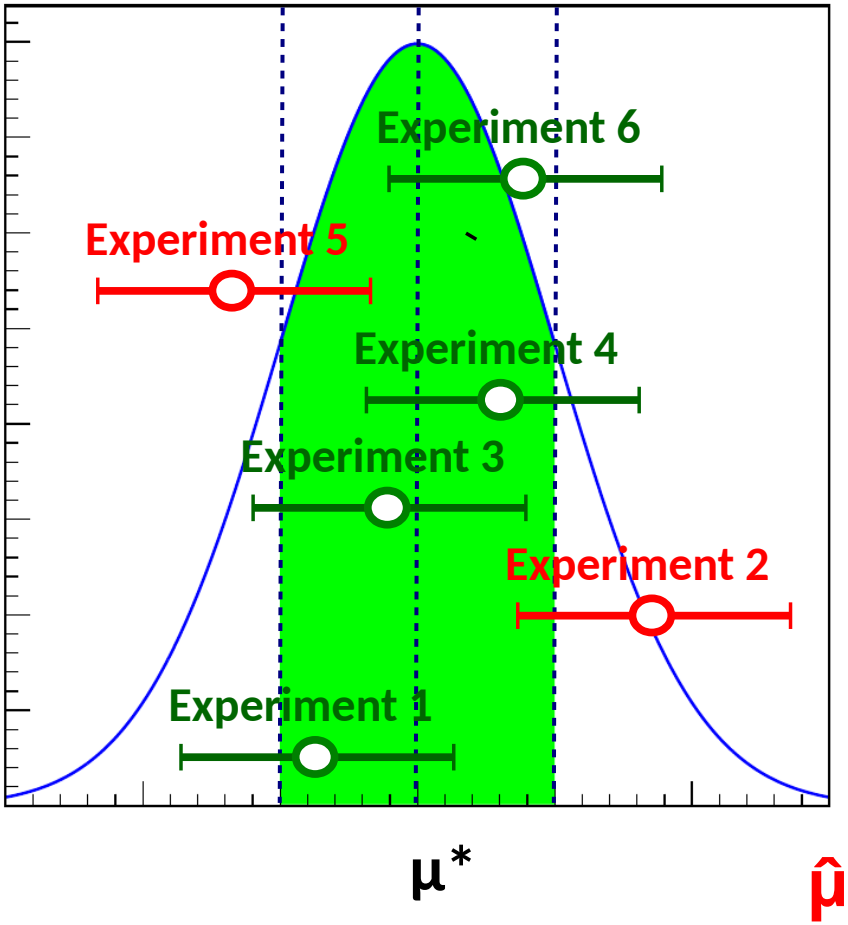
What we want

For a Gaussian, $P(\mu - \sigma < n < \mu + \sigma) = 68.3\,\%$ **68.3% confidence = "1σ"**

We repeat the same experiment multiple times, want to report a 1σ interval (68.3% CL) in each case.

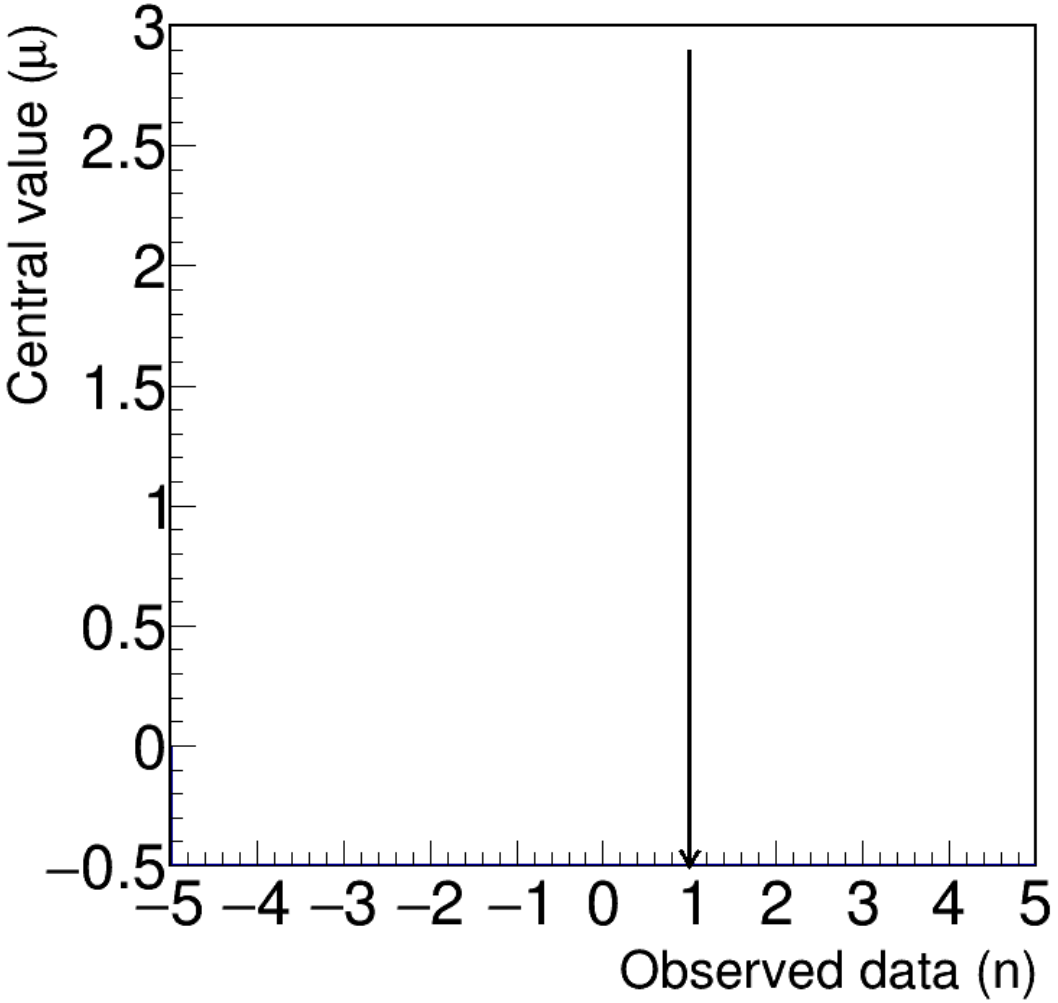
Goal: we want 68.3% of these intervals to contain the true value μ*.

Crucially, must be able to do so without knowing μ* !

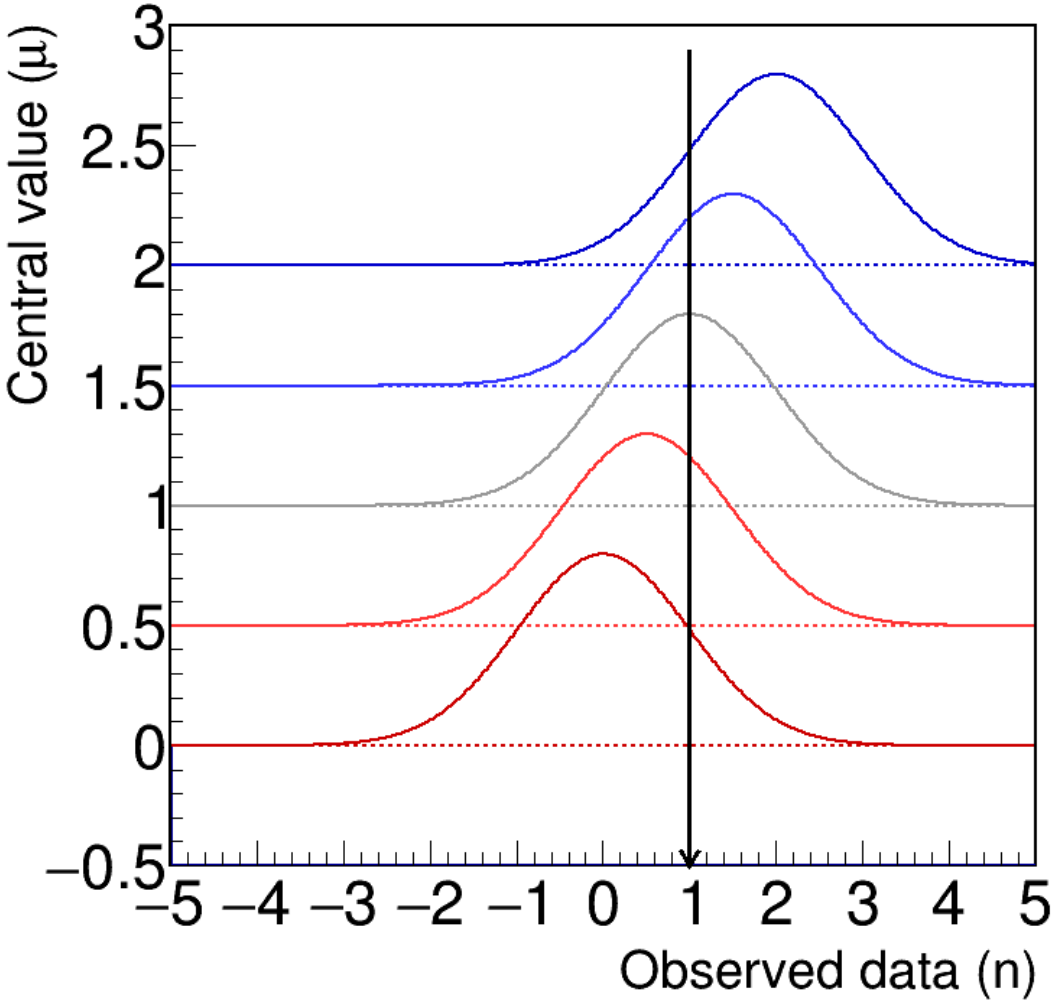


"We don't know the true value, but **there is a 68.3% chance that it is within the bar**"

Gaussian confidence intervals

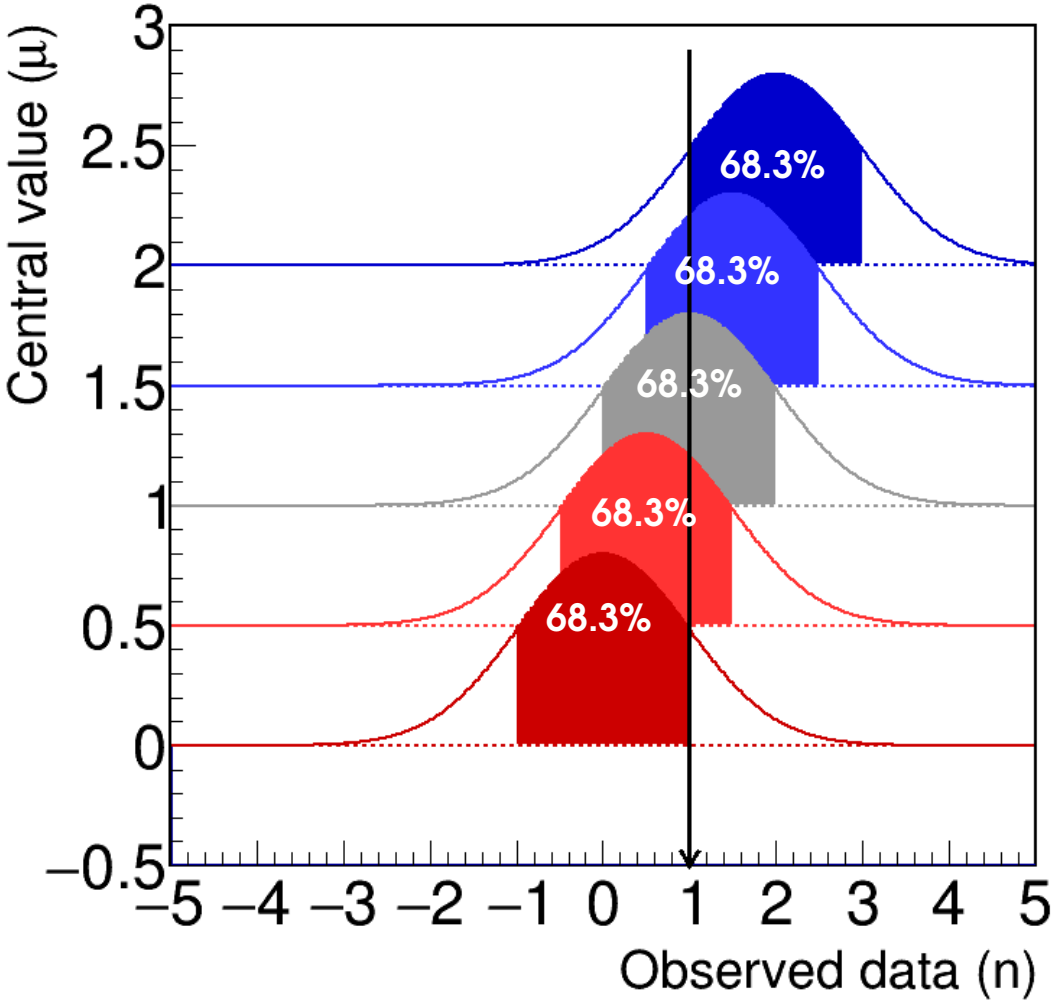


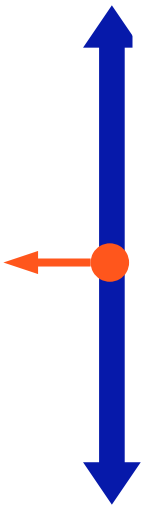
Gaussian confidence intervals



Central value
$\hat{\mu} = n$

Gaussian confidence intervals



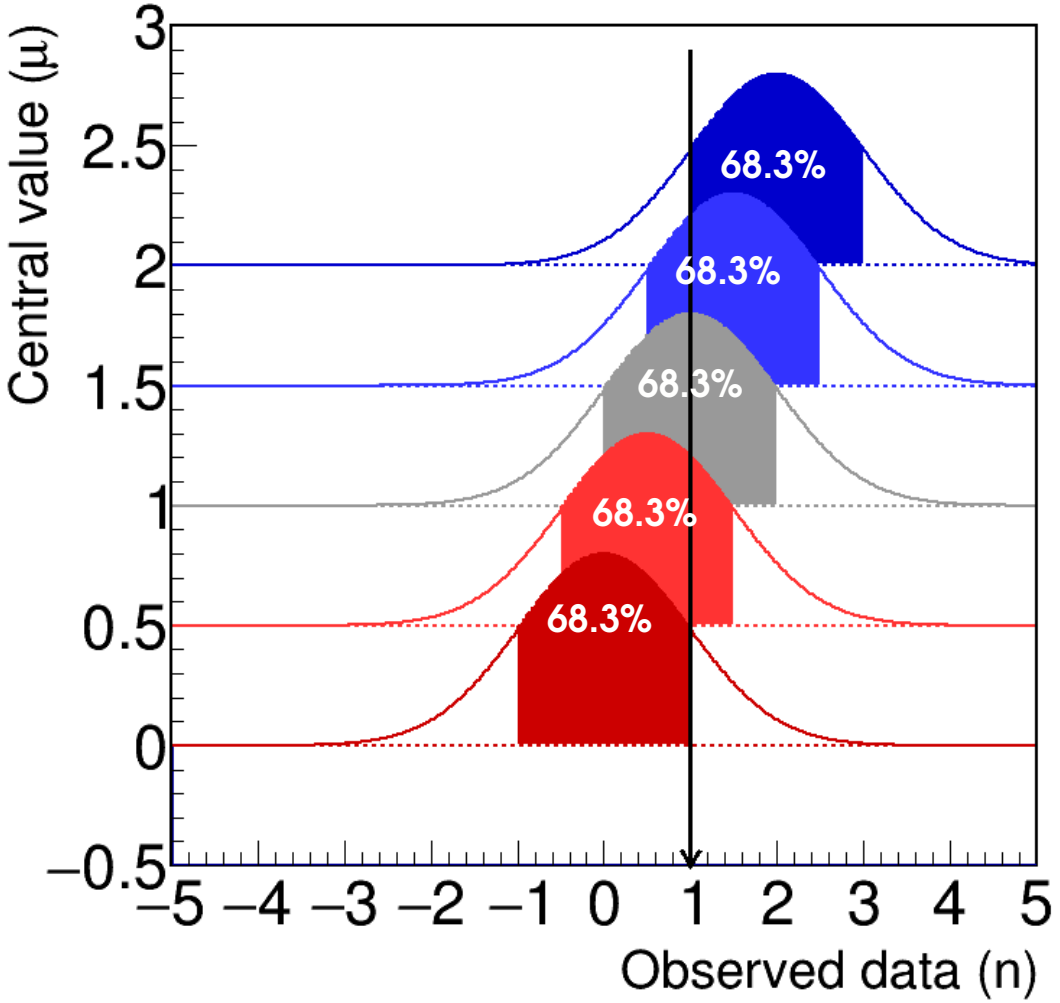


Other likely values

Central value
$\hat{\mu} = n$

How big an interval do we need around $\hat{\mu} = n$ to include the true value μ^* **68.3% of the time** ?

Gaussian confidence intervals



$$P\left(\mu^* - \sigma < n < \mu^* + \sigma\right) = 68.3\,\%$$

$$P\left(n - \sigma < \mu^* < n + \sigma\right) = 68.3\,\%$$

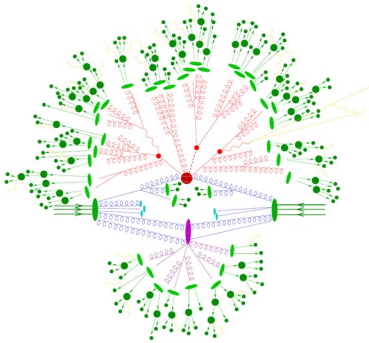
μ = n ± σ at 68.3% CL ("1σ")

Still a statement on n!

The reported interval **n ± σ contains the true value μ* 68.3% of the time**

Works even if we don't know μ* !

Takeaways (1)



Random data must be described using a **statistical model**. Usual cases:

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

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

How do we obtain the values of the POIs ?

Takeaways (2)

How do we obtain the values of the POIs ?

- **Central value** ("best-fit value") from Maximum likelihood estimation (MLE)

$$\hat{\mu} = arg\,max\, L(\mu) = arg\,min\left[\lambda(\mu) = -2\log L(\mu)\right]$$

→ **Gaussian case**: MLE identical to least squares minimization:

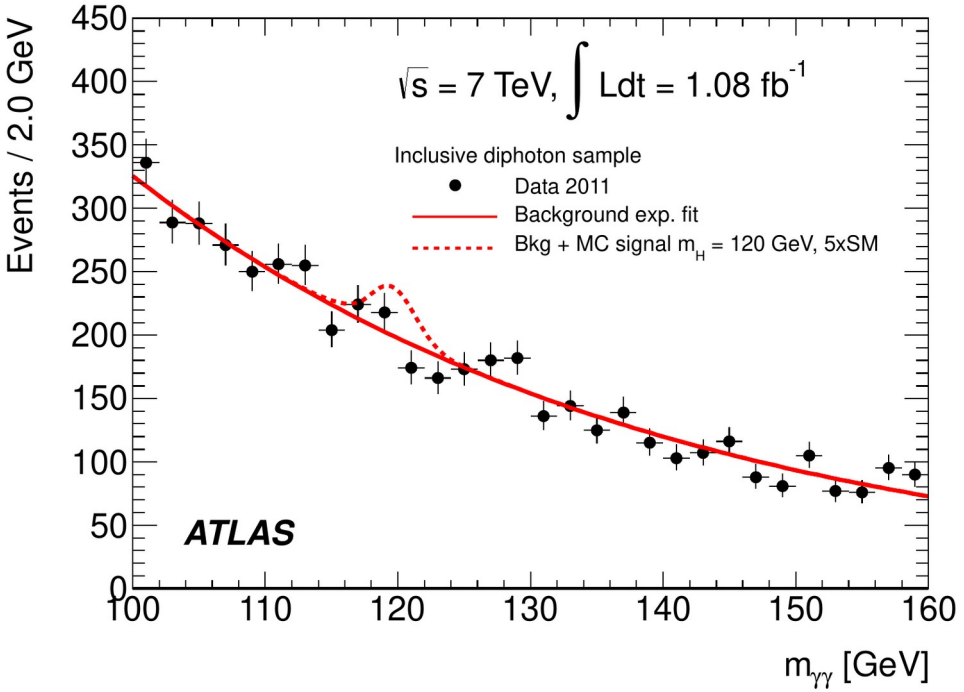
$$\lambda(\mu) = -2\log L(\mu) = \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - y_i(\mu)}{\sigma_i}\right)^2$$

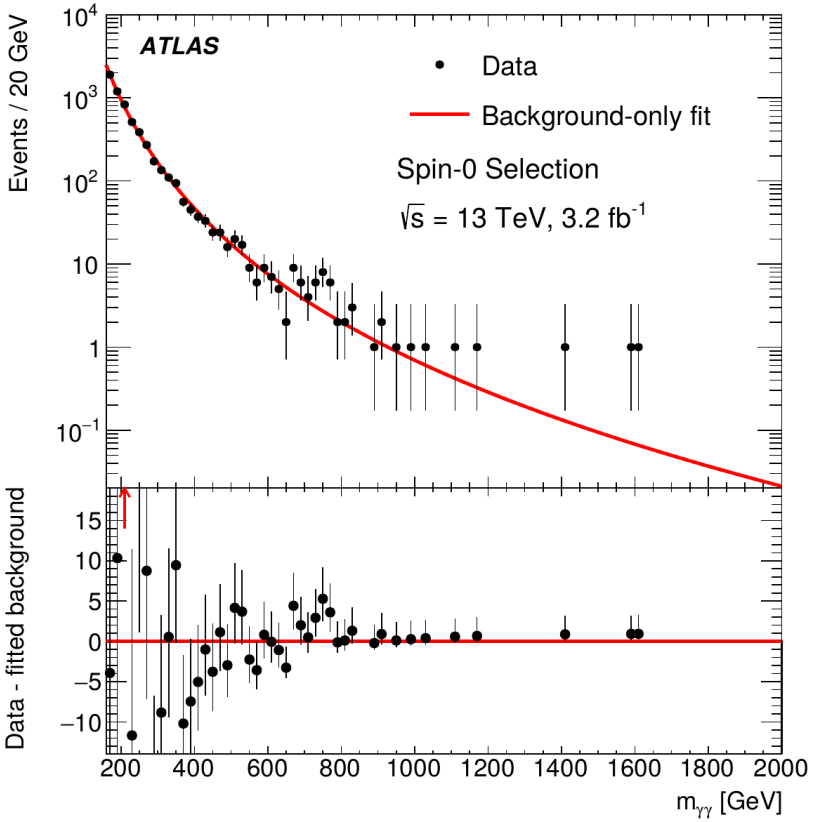
- **Uncertainties** ("error bar") : interval that is guaranteed to contain the true value 68.3% of the time ("1σ" case -- also 95% CL, etc.) → **Confidence interval**

→ **Gaussian case**: 68.3% CL interval is $\hat{\mu} \pm \sigma$ where σ is the Gaussian width.

Next steps: methods to obtain intervals, etc. for realistic models

Hypothesis Testing

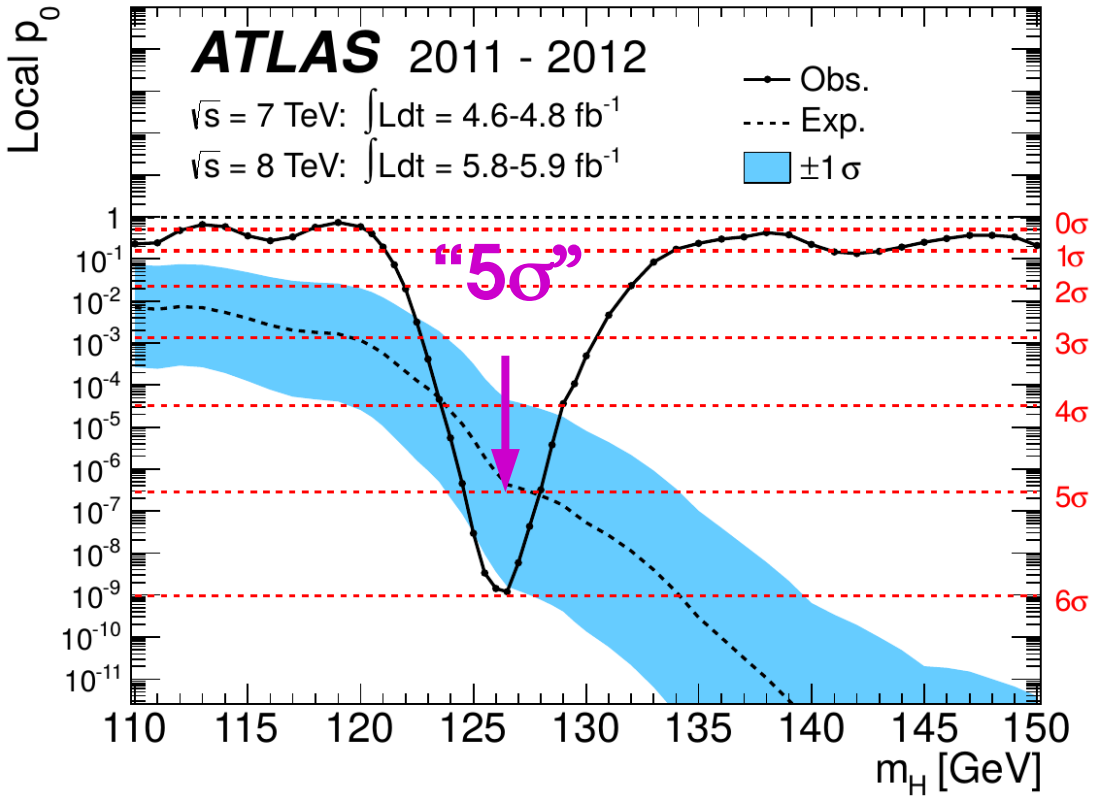




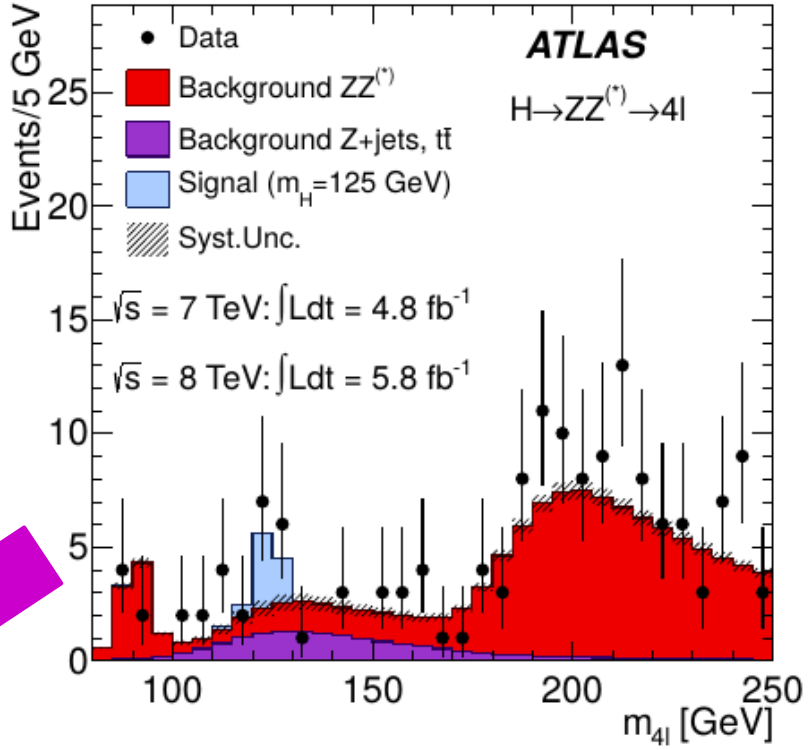
Discovery Testing

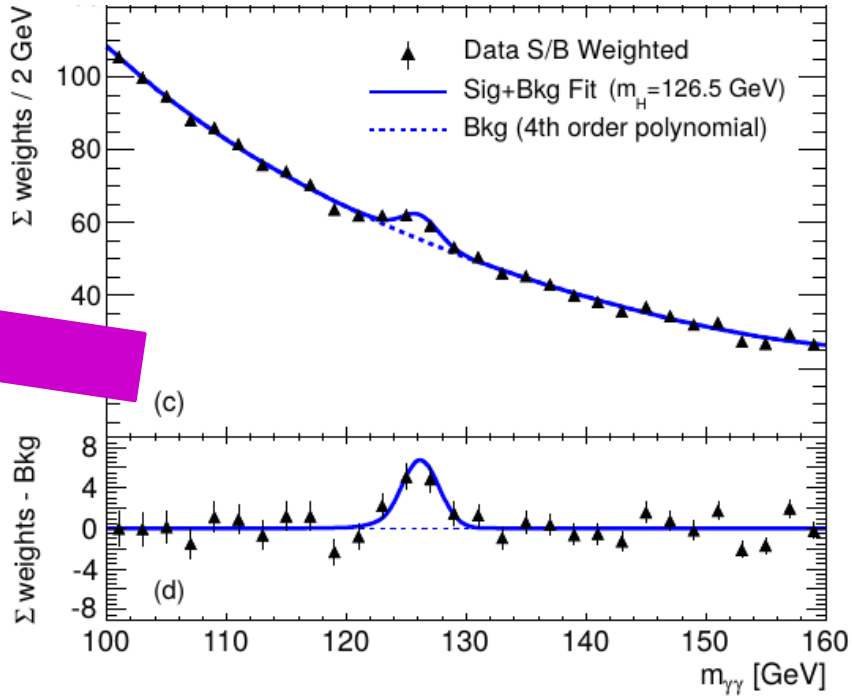
We see an unexpected feature in our data, is it a signal for new physics or a fluctuation ?

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

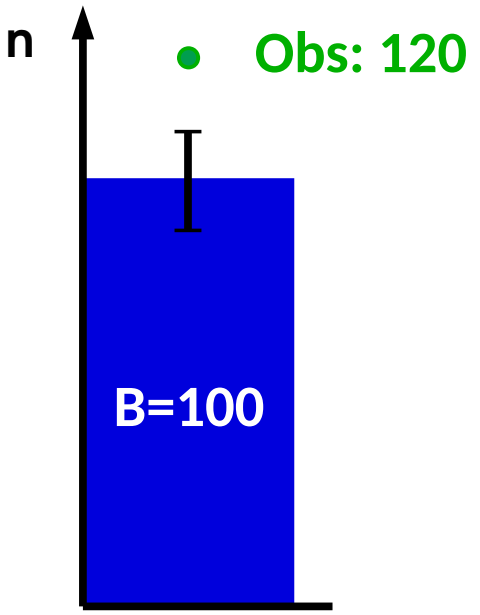


Phys. Lett. B 716 (2012) 1-29





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**.

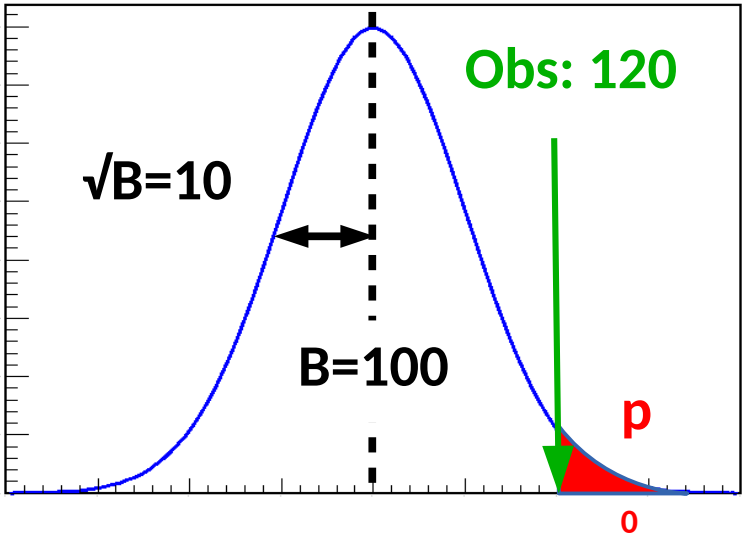
$$S = n_{obs} - B$$

Uncertainty on B is √B = 10

⇒ **Significance** Z = 2

⇒ we are **~2σ** away from S=0.

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



P-value p_0: fraction of outcomes that is *at least as extreme as the observation.*

Gaussian quantiles : $p_0 = 1 - \Phi(Z)$ $\quad \Phi(Z) = \int_{-\infty}^{Z} G(u;0,1)\,du$

In our case, get Z=2 about **p_0 ~ 2.3%** of the time if S = 0 ⇒ **Rare, but not exceptional**

Discovery Testing

Z	$P(\lvert x - X_0 \rvert > Z\sigma)$
1	0.317
2	0.046
3	0.003
4	6×10^{-5}
5	6×10^{-7}

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**.

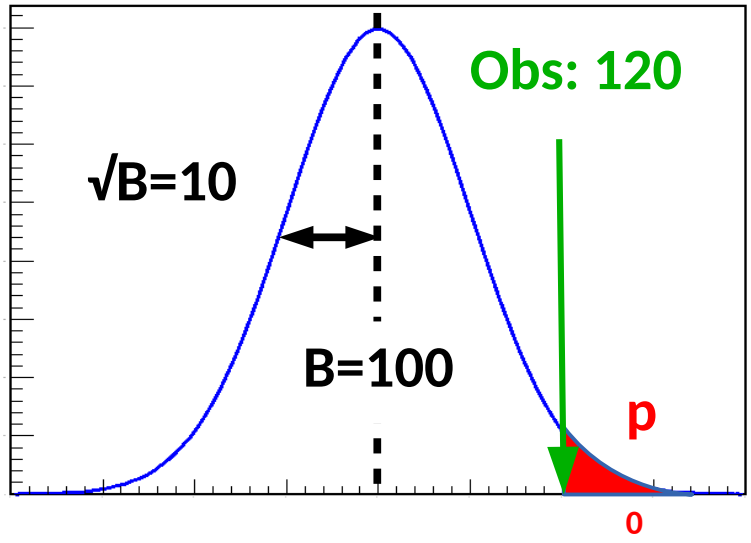
$$S = n_{obs} - B$$

Uncertainty on B is √B = 10

⇒ **Significance** Z = 2

⇒ we are **~2σ** away from S=0.

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

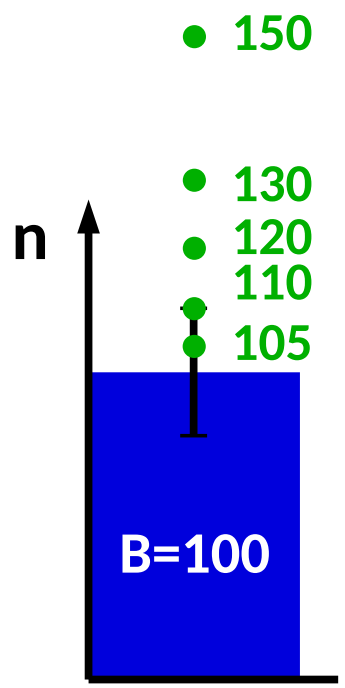


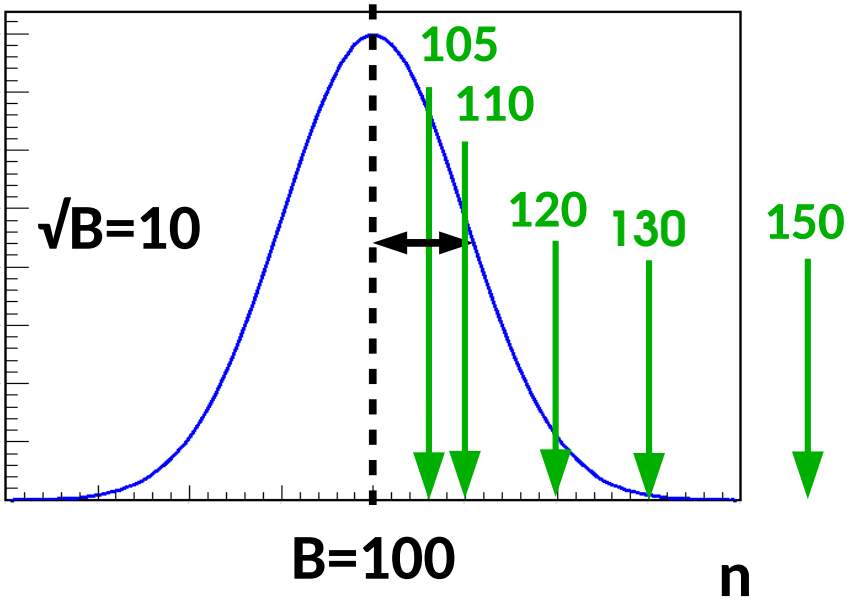
P-value p_0: fraction of outcomes that is *at least as extreme as the observation.*

Gaussian quantiles : $p_0 = 1 - \Phi(Z)$ $\quad \Phi(Z) = \int_{-\infty}^{Z} G(u;0,1)\, du$

In our case, get Z=2 about p_0 **~ 2.3%** of the time if S = 0 ⇒ **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

Now need to be able to do the same in realistic cases:
- **Determine S**
- **Compute Z and p_0**

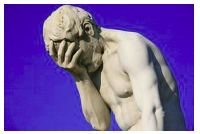
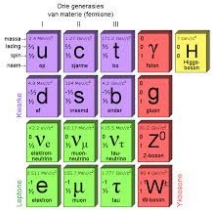
Evidence

Discovery

General Hypothesis Testing

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

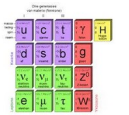
→ **Goal** : decide if H_0 is favored or disfavored using a test based on the data.

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

General Hypothesis Testing

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

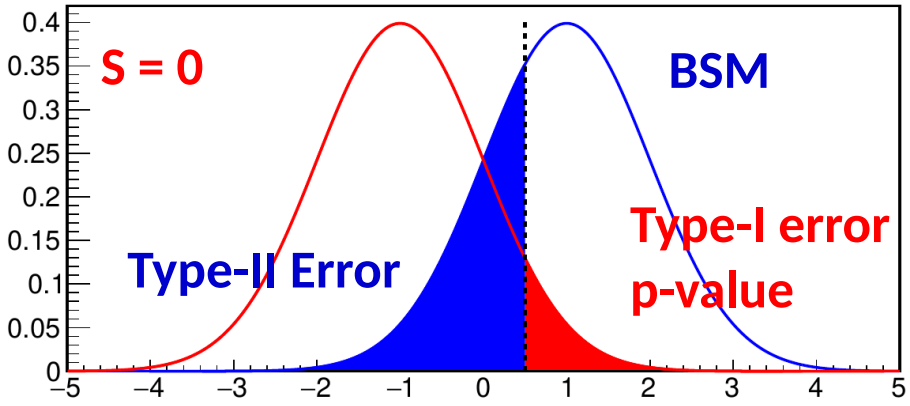
	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	**Discovery!** 	**Type-II error (Missed discovery)** 
H_0 is true (Nothing new)	**Type-I error (False discovery)** 	**No new physics, none found** 

a.k.a. p-value, significance

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

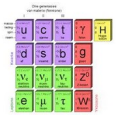
→ **Goal**: test that minimizes Type-II errors **for a given level of Type-I error**.

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



General Hypothesis Testing

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

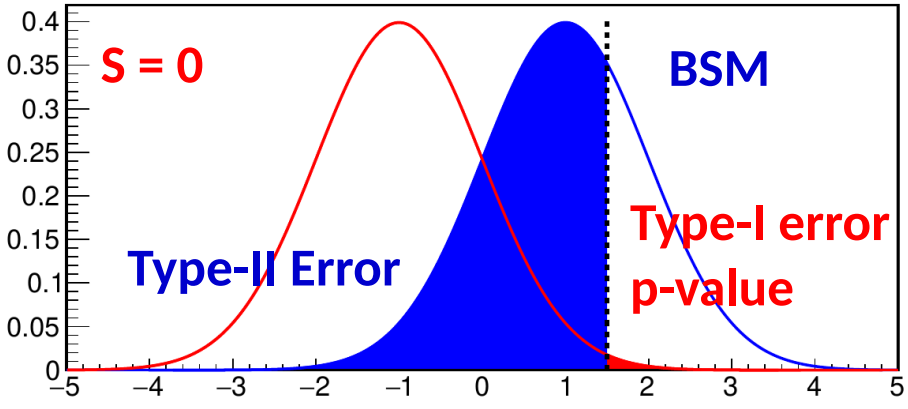
	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	**Discovery!** 	**Type-II error (Missed discovery)** 
H_0 is true (Nothing new)	**Type-I error (False discovery)** 	**No new physics, none found** 

a.k.a. 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**.

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



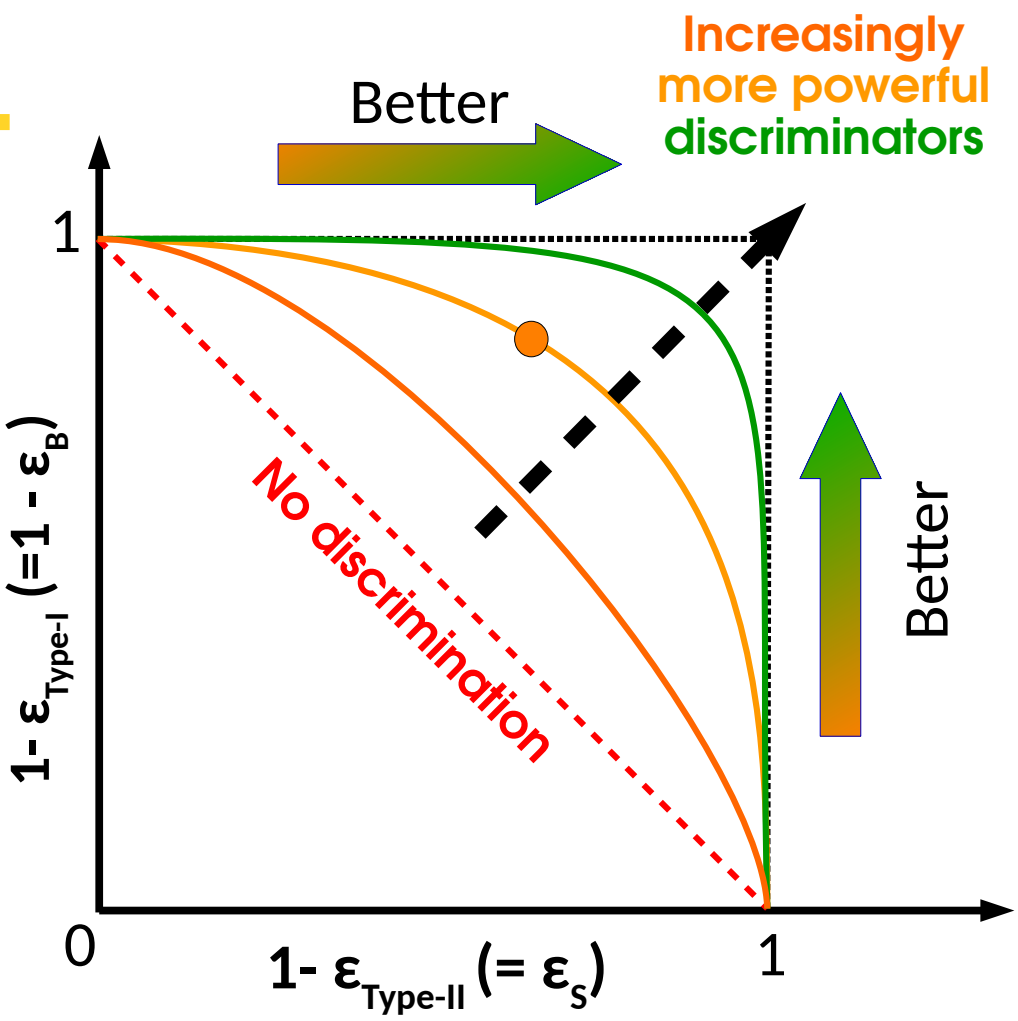
ROC Curves

"Receiver operating characteristic" (ROC) Curve:

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

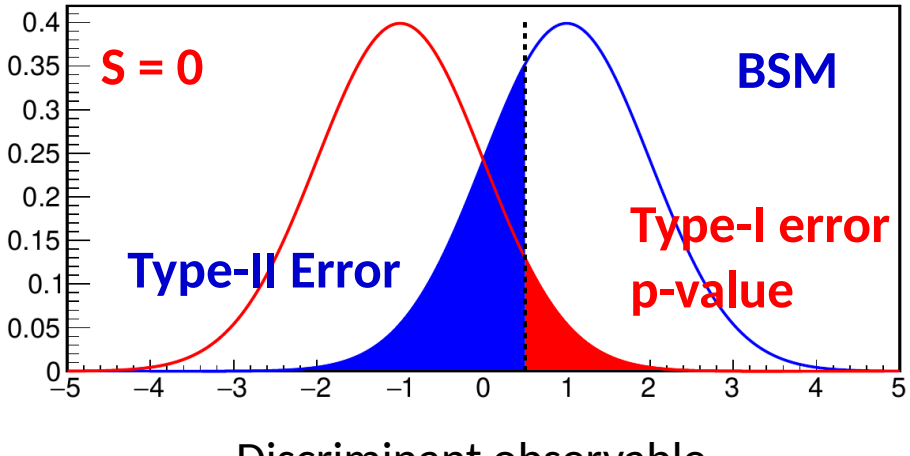
→ All curves monotonically decrease from (0,1) to (1,0)

→ Better discriminators more "bent" towards (1,1)



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

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



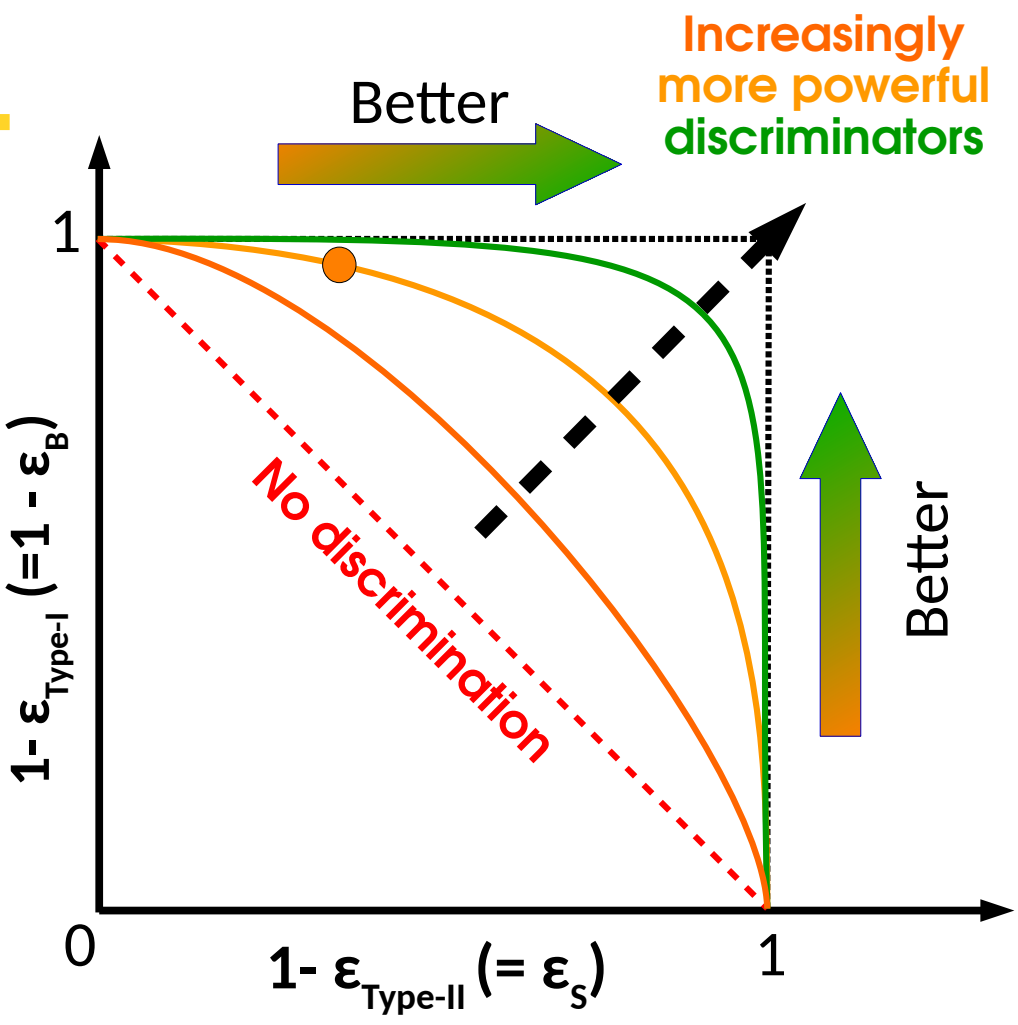
ROC Curves

"Receiver operating characteristic" (ROC) Curve:

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

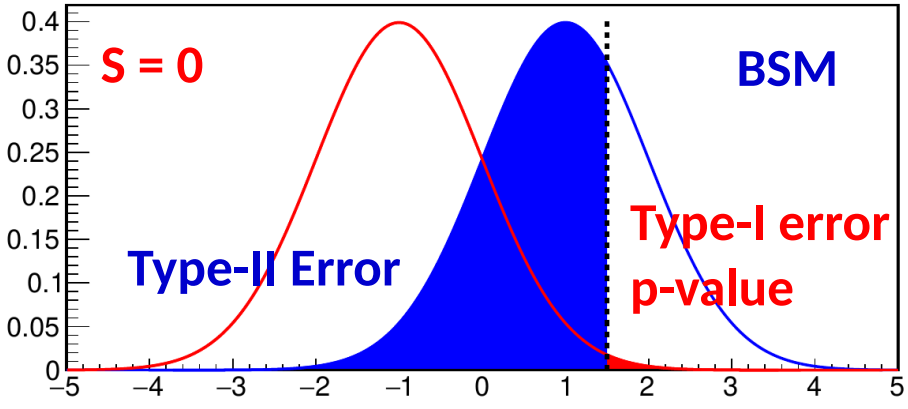
→ All curves monotonically decrease from (0,1) to (1,0)

→ Better discriminators more "bent" towards (1,1)



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

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



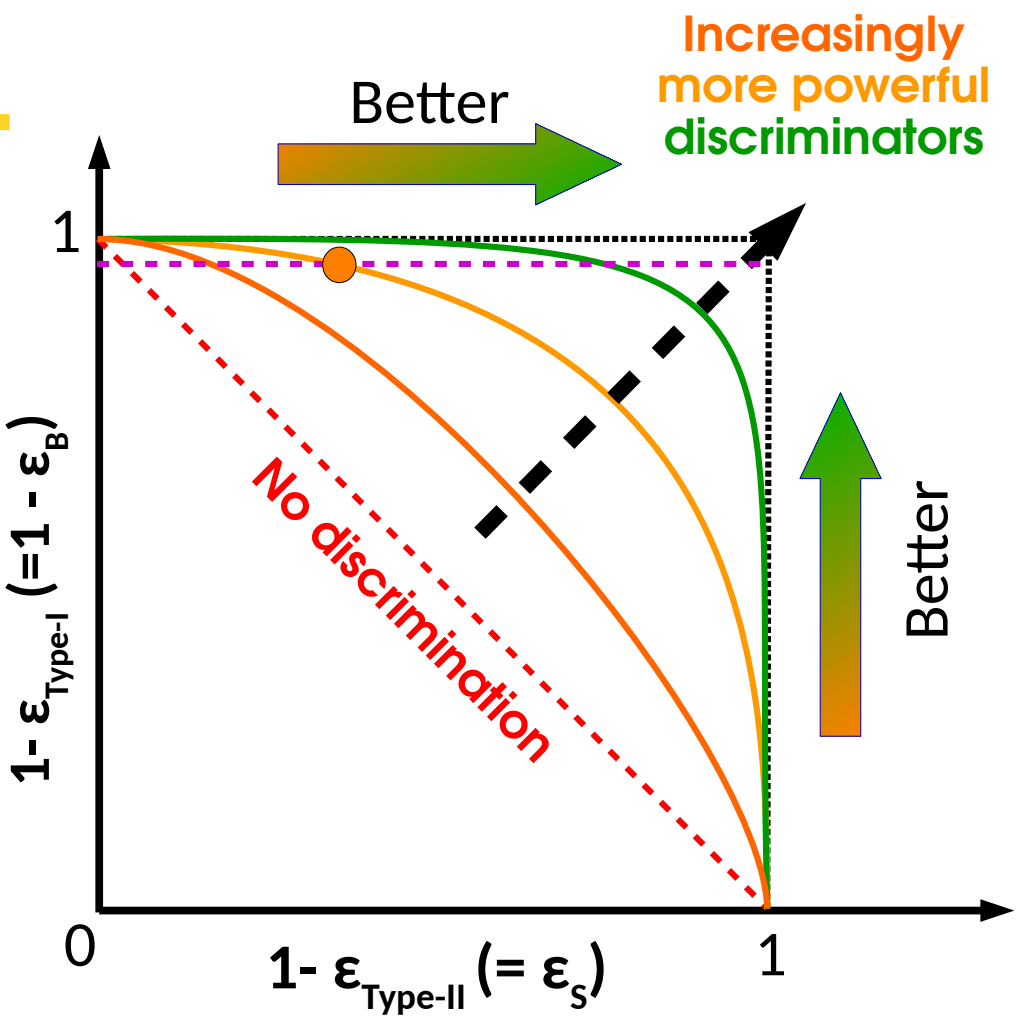
ROC Curves

"Receiver operating characteristic" (ROC) Curve:

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

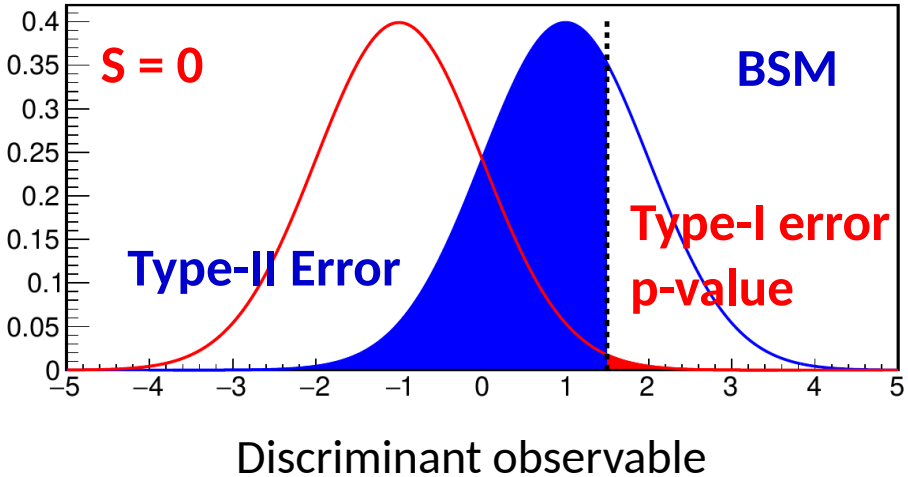
→ All curves monotonically decrease from (0,1) to (1,0)

→ Better discriminators more "bent" towards (1,1)



→ **Goal**: test that minimizes Type-II errors **for a 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)*

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

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

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

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

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

Optimal → minimizes Type-II uncertainties for given level of Type-I uncertainties.

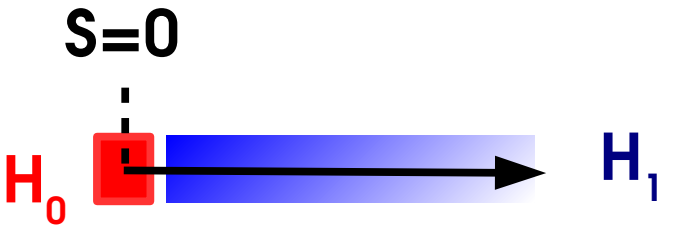
In the following: all tests based on LR, will focus on p-values (Type-I errors), trusting that Type-II errors are anyway as small as they can be…

Discovery: Test Statistic

Cowan, Cranmer, Gross & Vitells, E
ur.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 Likelihood ratio: $\dfrac{L(S=0)}{L(\hat{S})}$

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

Note: for $\hat{S} < 0$, set $q_0=0$ to reject negative signals (“*one-sided* test statistic”)

Discovery p-value

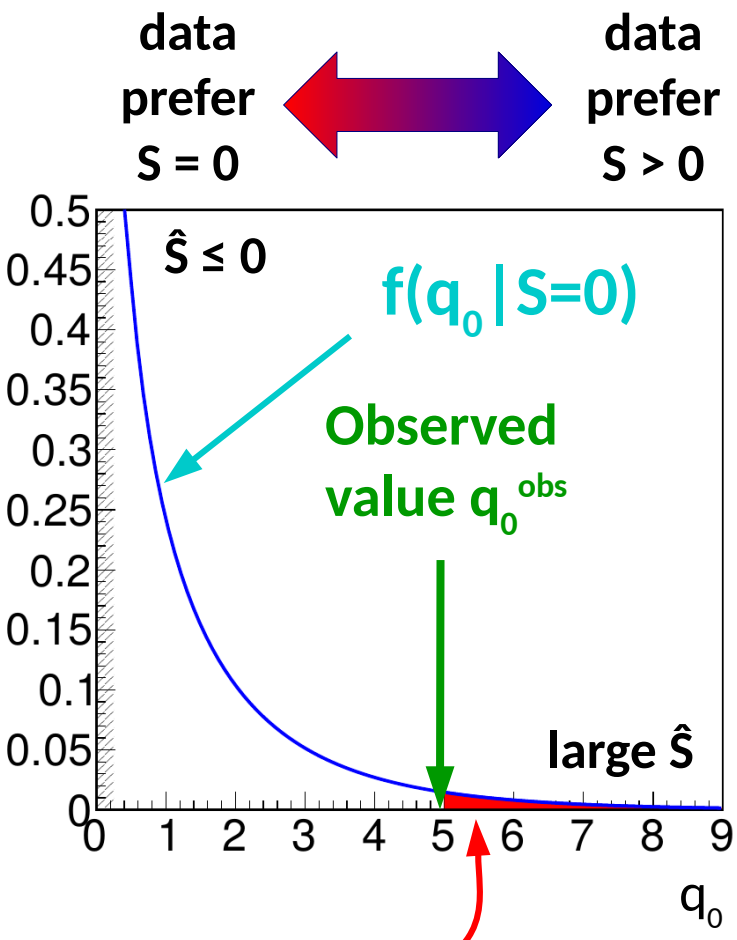
Large values of $-2\log\frac{L(S=0)}{L(\hat{S})}$ means that:

⇒ observed $\hat{S}$ is far from 0

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

Is it still compatible with $H_0(S=0)$?

Compute *p-value* in the tail of the distribution to exclude H_0 (… and **claim a discovery!**)



$$p_0 = \int_{q_0^{obs}}^{\infty} f(q_0|S=0)\, dq_0$$

Need to know $f(q_0 \mid S=0)$, the distribution of q_0 in the S=0 scenario…

Hard problem in general!

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) :

2-line "proof" : asymptotically L and S are Gaussian, so

$$L(S) = \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}}{\sigma}\right)^2\right] \Rightarrow q_0 = \left(\frac{\hat{S}}{\sigma}\right)^2$$

$$\Rightarrow \sqrt{q_0} = \frac{\hat{S}}{\sigma} \sim G(0,1) \Rightarrow q_0 \sim \chi^2(n_{dof}=1)$$

Wilks' Theorem:

q_0 is distributed as χ^2 (n_{POI}) for S = 0

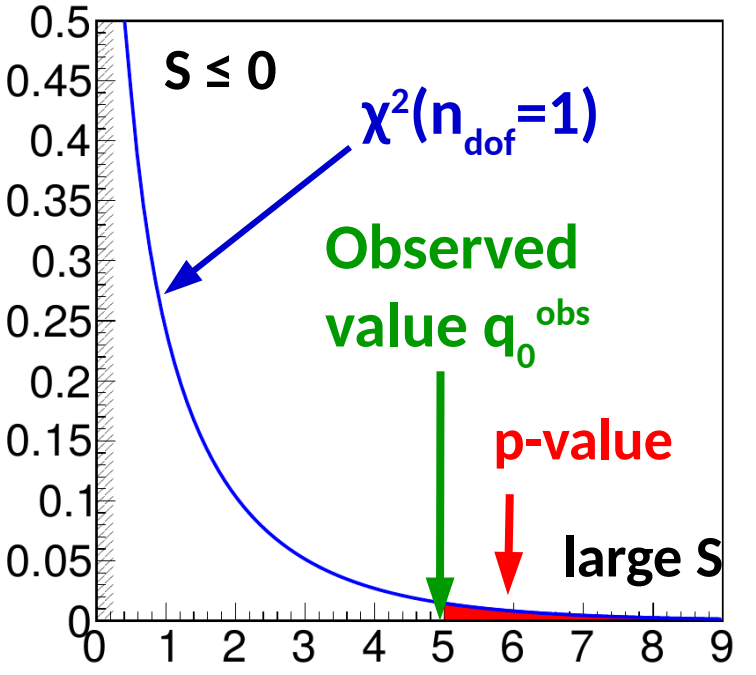
⇒ **1 POI : $\sqrt{q_0}$ follows a normal distribution G(0,1)**

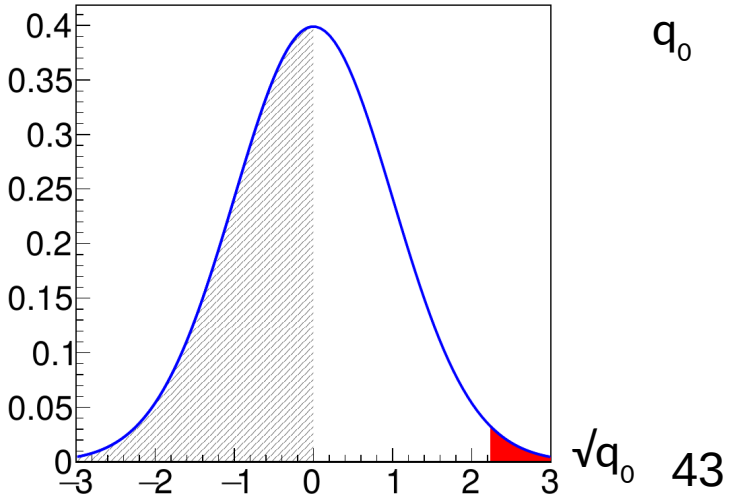
⇒ Can compute p-values from Gaussian quantiles

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

⇒ Even more simply, the significance is:

$$Z = \sqrt{q_0}$$





"**Weak" Gaussian assumption**:

- Compute q_0 using the exact (non-Gaussian) model
- Assume Gaussianity only for the **distribution** of q_0.

Typically works well already for O(5) events

Testing for discovery using asymptotics

1. Build the statistical model of the measurement, **P(data; μ)**

2. Define the likelihood, **L(μ) = P(data; μ)**

3. Compute the test statistic $q_0 = -2\log\frac{L(S=0)}{L(\hat{S})} \quad \hat{S} \geq 0$

4. Compute the significance $Z = \sqrt{q_0}$

5. Or the discovery p-value $p_0 = 1 - \Phi(\sqrt{q_0})$

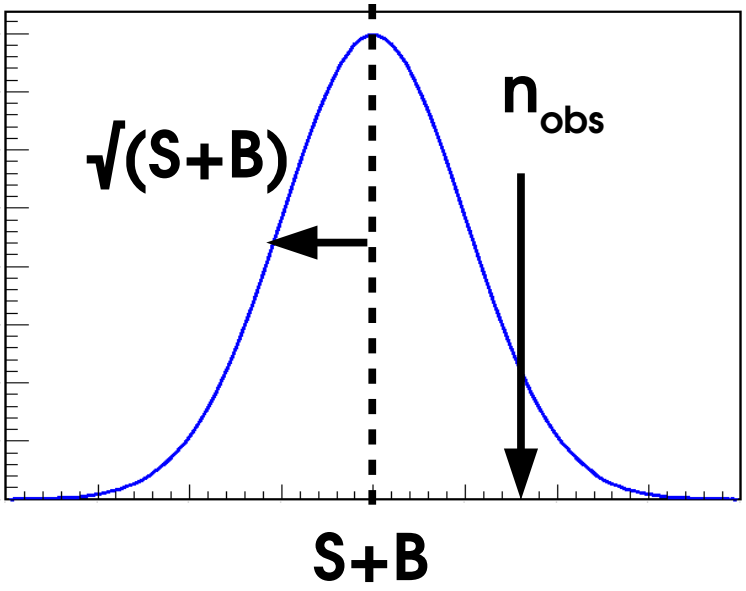
Valid in the **asymptotic limit**, i.e. large enough datasets.

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 :

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

→ Find the best-fit value (MLE) $\hat{S}$ for the signal
(can use $\lambda = -2 \log L$ instead of L for simplicity)

→ Find the expression of q_0 for $\hat{S} > 0$.

→ Find the expression for the significance

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

Homework 2: Poisson Counting

Same problem as Homework 1, 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

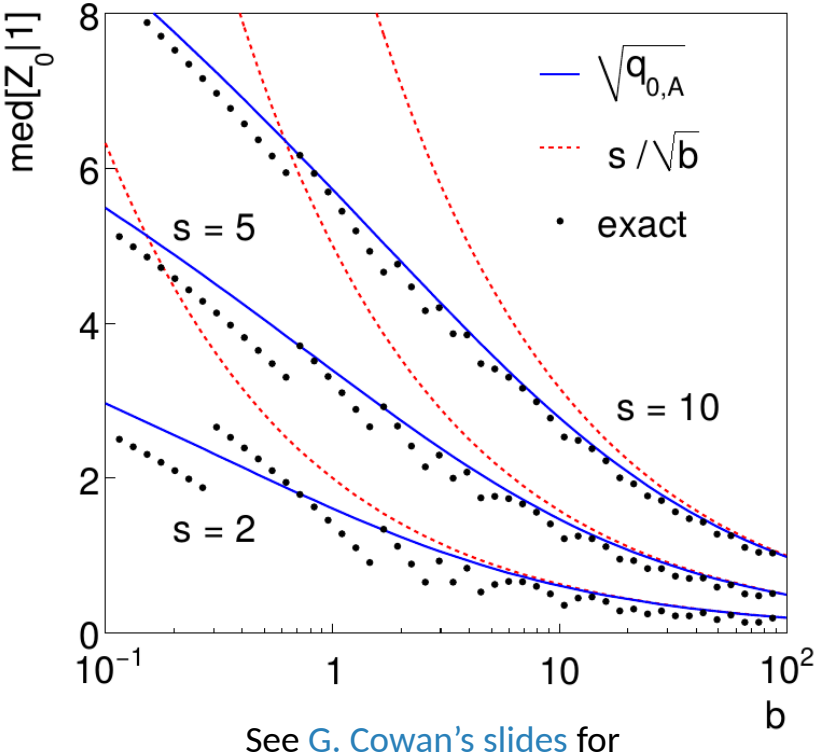
Solution:

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

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)

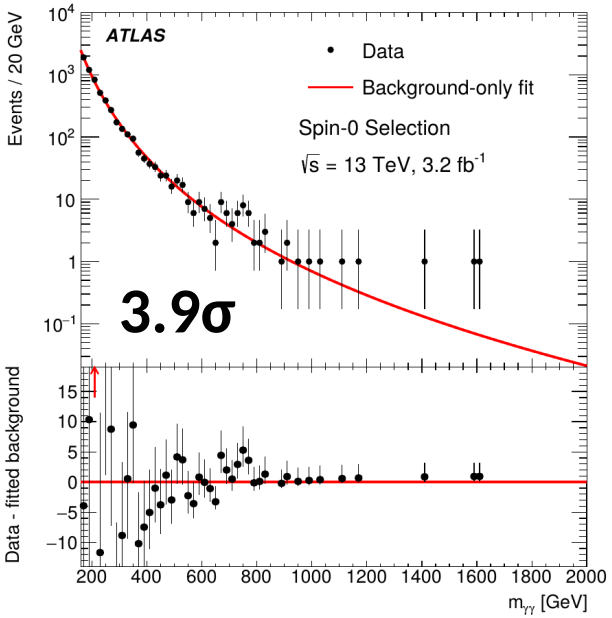
Discovery: 5σ ($p_0 = 3\ 10^{-7}$ ⇔ 1 chance in 3.5M)

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 : Hypothesis Tests

Given a PDF **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** $\frac{L(H_0)}{L(H_1)}$

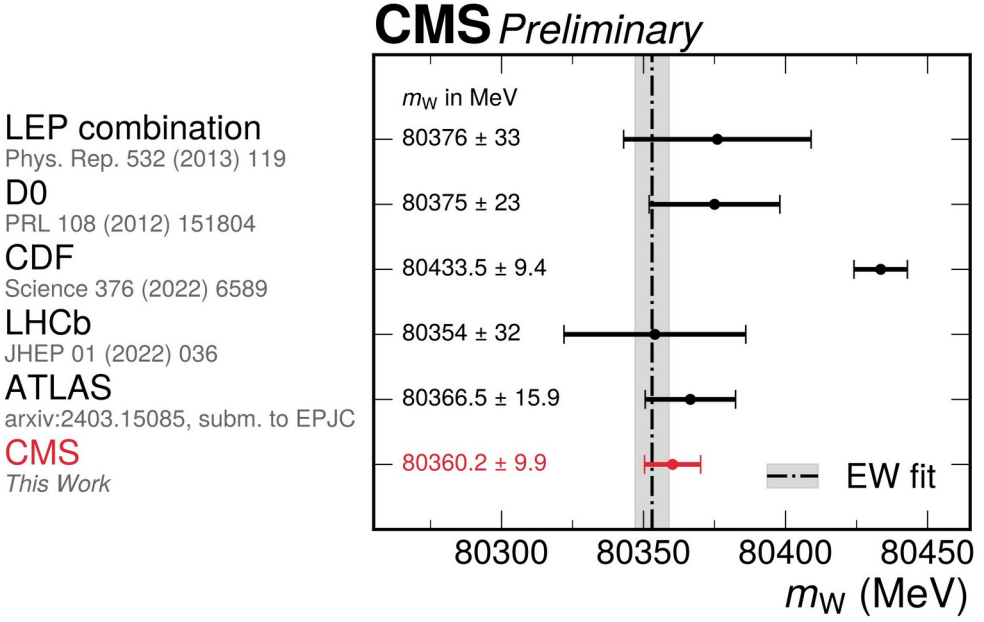
To test for **discovery**, use $q_0 = -2\log\frac{L(S=0)}{L(\hat{S})} \quad \hat{S} \geq 0$

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

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

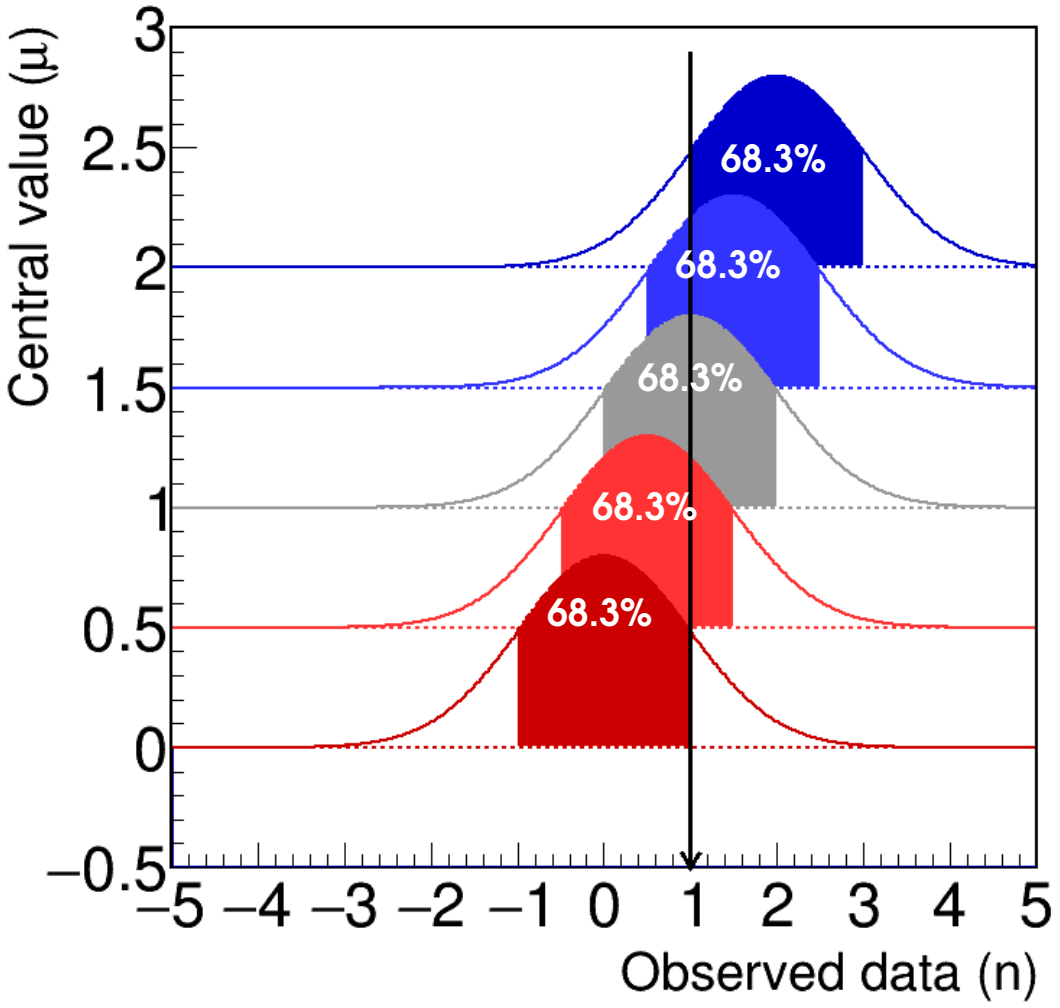
For a single **Poisson** measurement, $Z = \sqrt{2\left[(\hat{S}+B)\log\left(1+\frac{\hat{S}}{B}\right) - \hat{S}\right]}$

Confidence Intervals using Likelihoods



Reminder: Gaussian confidence intervals

Confidence interval at xx% CL : guaranteed to contain the true value xx% of the time



$$P\left(\mu^* - \sigma < n < \mu^* + \sigma\right) = 68.3\,\%$$

$$P\left(n - \sigma < \mu^* < n + \sigma\right) = 68.3\,\%$$

The interval **n ± σ** contains μ* 68.3% of the time ⇒ **68.3% ("1σ") CL interval**

What about the general (non-Gaussian) case ?

Likelihood Intervals

Probability to observe the data **for a given μ**.

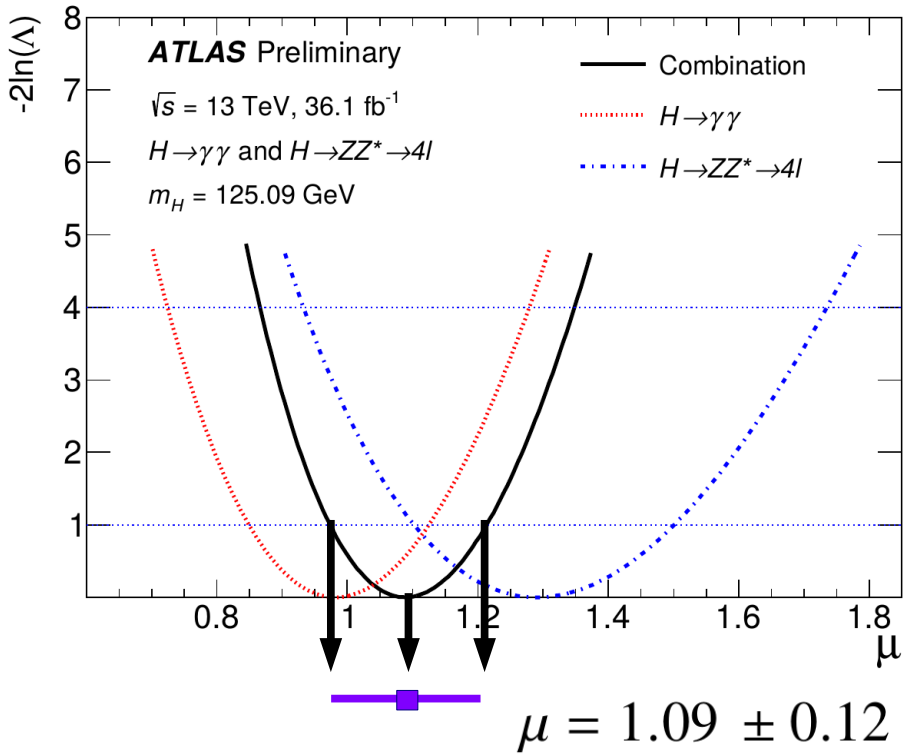
$$t(\mu) = -2\log\frac{L(\mu)}{L(\hat{\mu})}$$

Probability to observe the data **for best-fit $\hat{\mu}$.**

Confidence intervals from L(μ):

- Test various values μ against $\hat{\mu}$, using the **Profile Likelihood Ratio t(μ).**
- Minimum (=0) for μ=$\hat{\mu}$, rises away from $\hat{\mu}$.
- Optimal properties thanks to the Neyman-Pearson lemma.

ATLAS-CONF-2017-047



Case of a 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**
- Minimum occurs at μ = $\hat{\mu}$ (=n)

$t(\mu_\pm) = 1 \Rightarrow \mu = n \pm \sigma$ gives 1σ interval

Likelihood Intervals

Confidence intervals from L(μ):

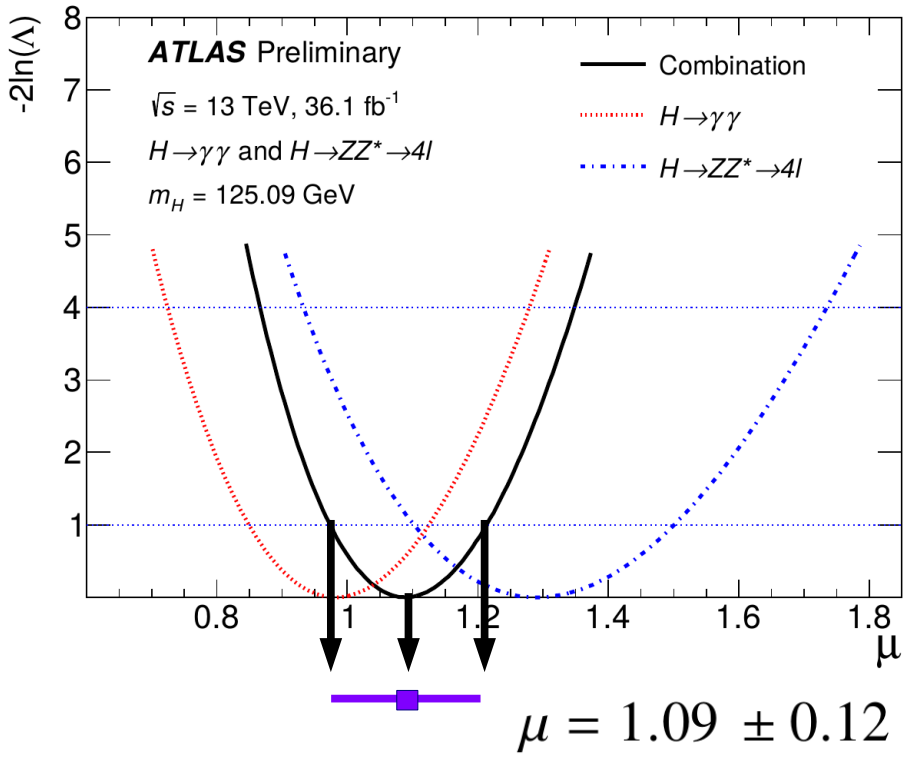
- Test various values μ against $\hat{\mu}$, using the **Profile Likelihood Ratio t(μ).**
- Minimum (=0) for $\mu=\hat{\mu}$, rises away from $\hat{\mu}$.
- Optimal properties thanks to the Neyman-Pearson lemma.

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 $\hat{\mu}$.**

ATLAS-CONF-2017-047



General case:

- Generally not a perfect parabola
- Minimum still at $\mu = \hat{\mu}$

Asymptotic approximation

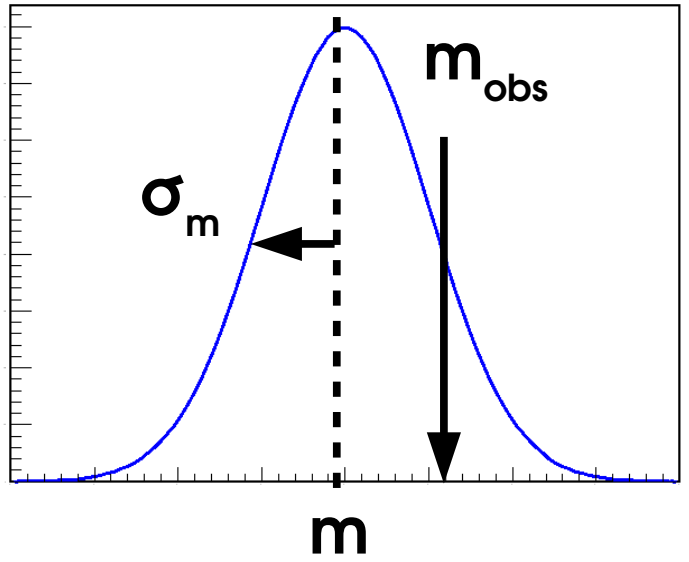
→ Compute t(μ) using the exact L(μ)

→ **1σ interval given by t(μ) = 1**

(other thresholds for other intervals)

→ Again, "weak" Gaussian assumptions

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 1σ (68.3% CL) interval on m

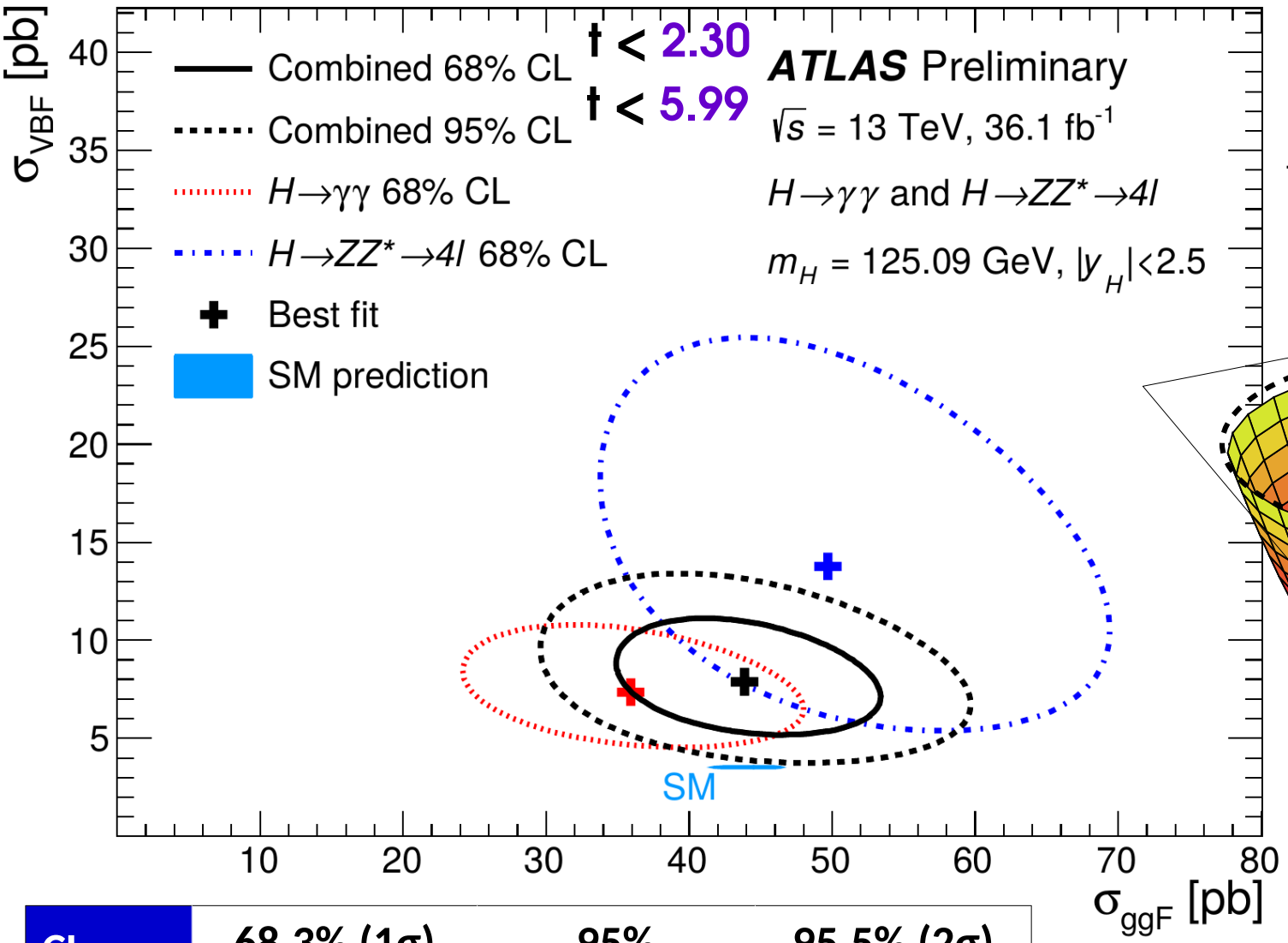
Solution: $m = m_{obs} \pm \sigma_m$

→ As expected!

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

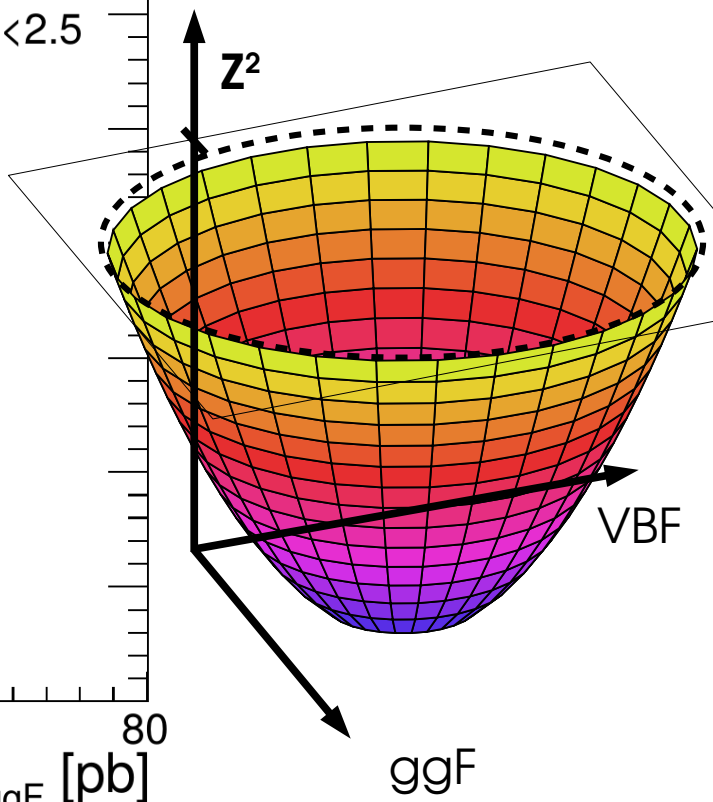
2D Example: Higgs σ_{VBF} vs. σ_{ggF}

ATLAS-CONF-2017-047



$$t = -2\log\frac{L(X_0, Y_0)}{L(\hat{X}, \hat{Y})} \sim \chi^2(n_{dof}=2)$$

$t_{ggF,VBF}$



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

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

Gaussian case: elliptic paraboloid surface

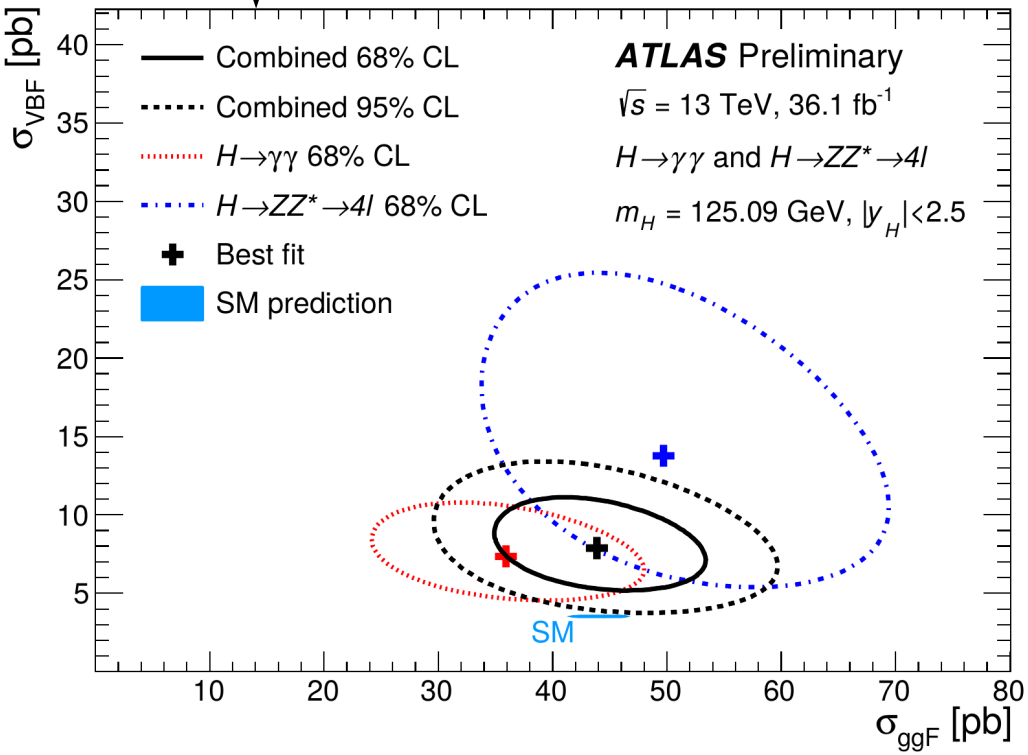
Reparameterization

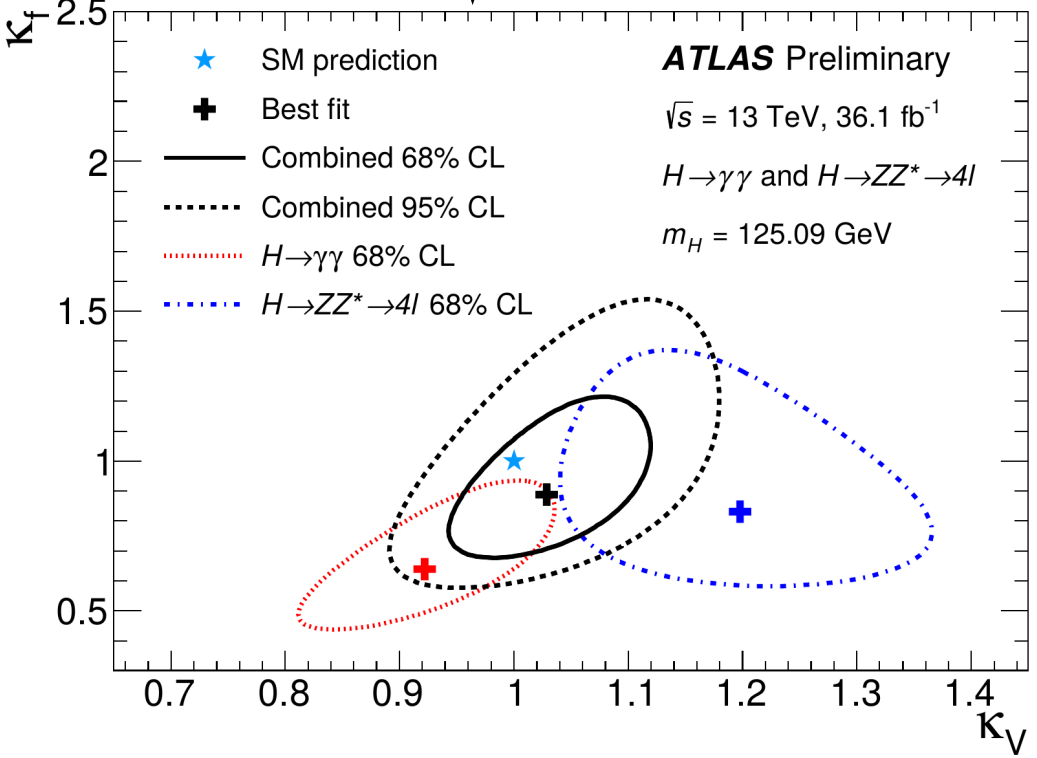
Start with basic measurement in terms of e.g. **($\sigma \times B$)**

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

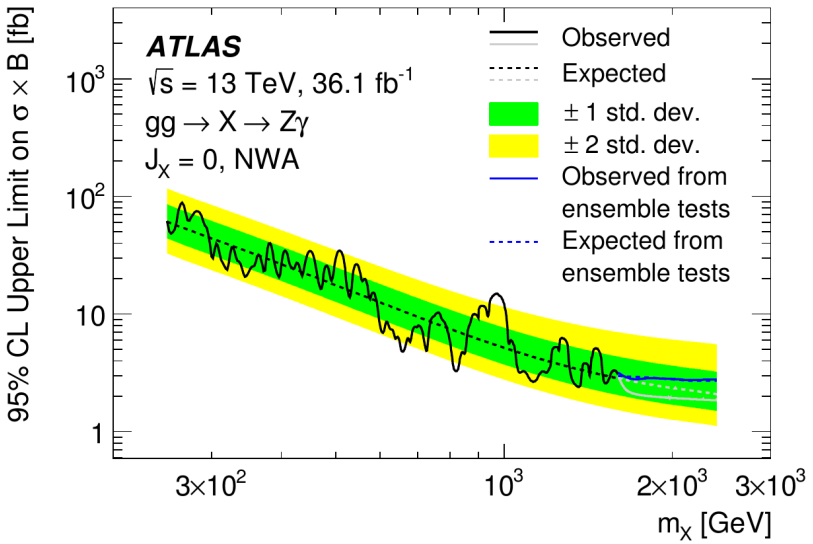
e.g. Higgs couplings: σ_{ggF}, σ_{VBF} sensitive to Higgs coupling modifiers κ_V, κ_F.

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





Upper Limits



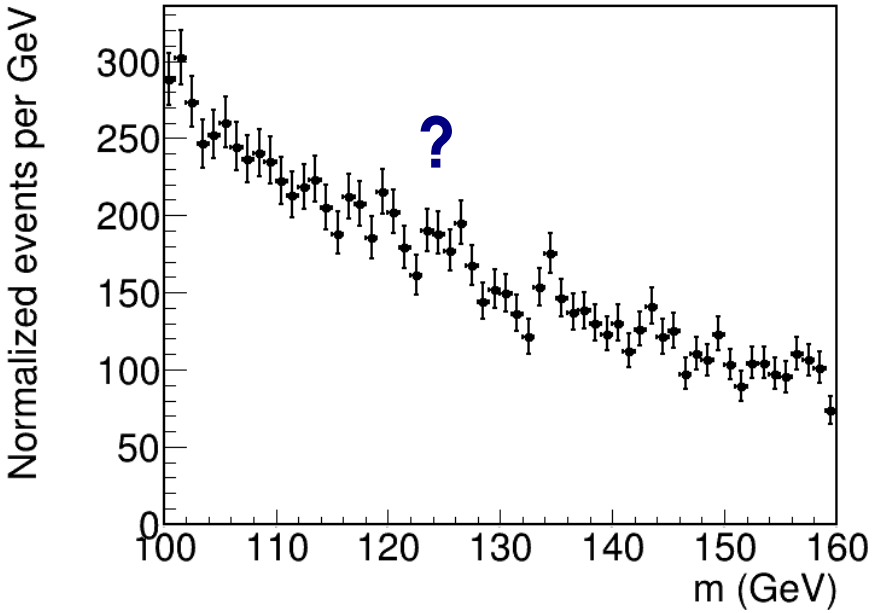
Upper limits

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"



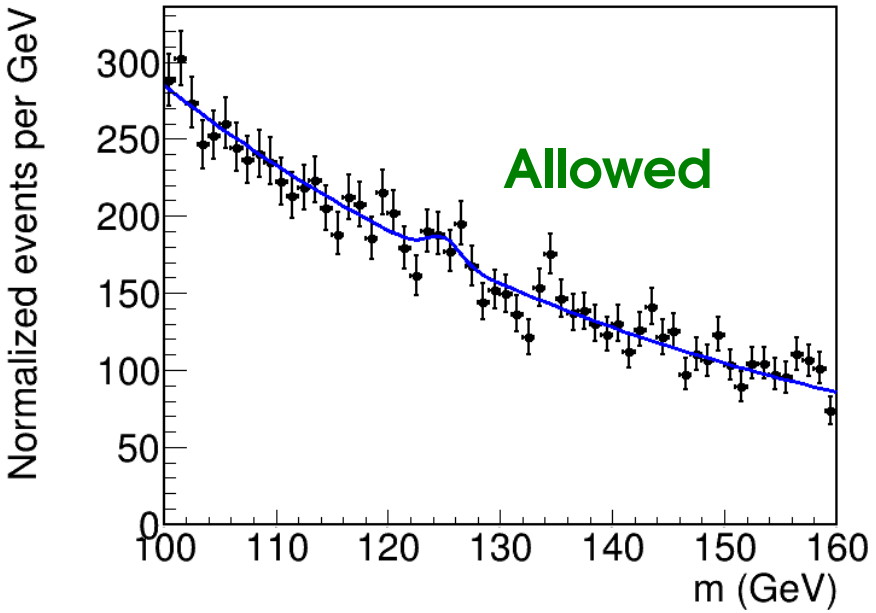
Upper limits

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"



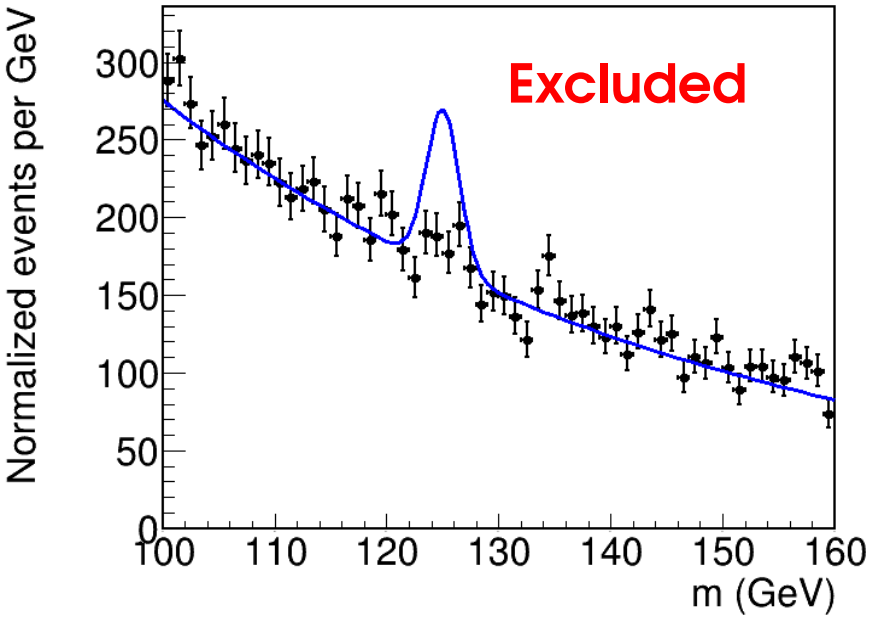
Upper limits

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"



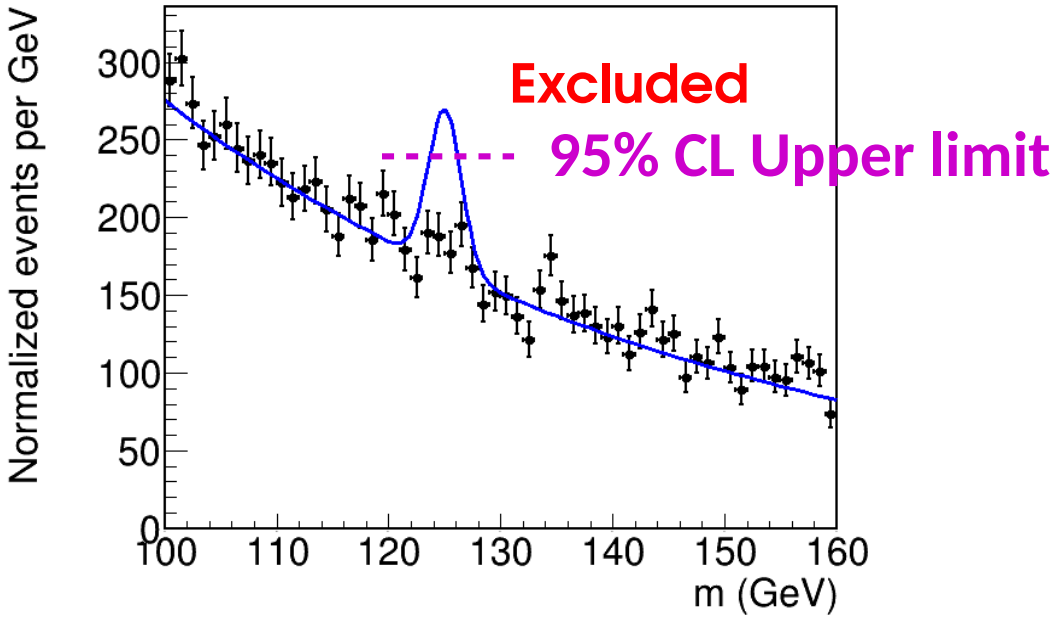
Upper limits

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

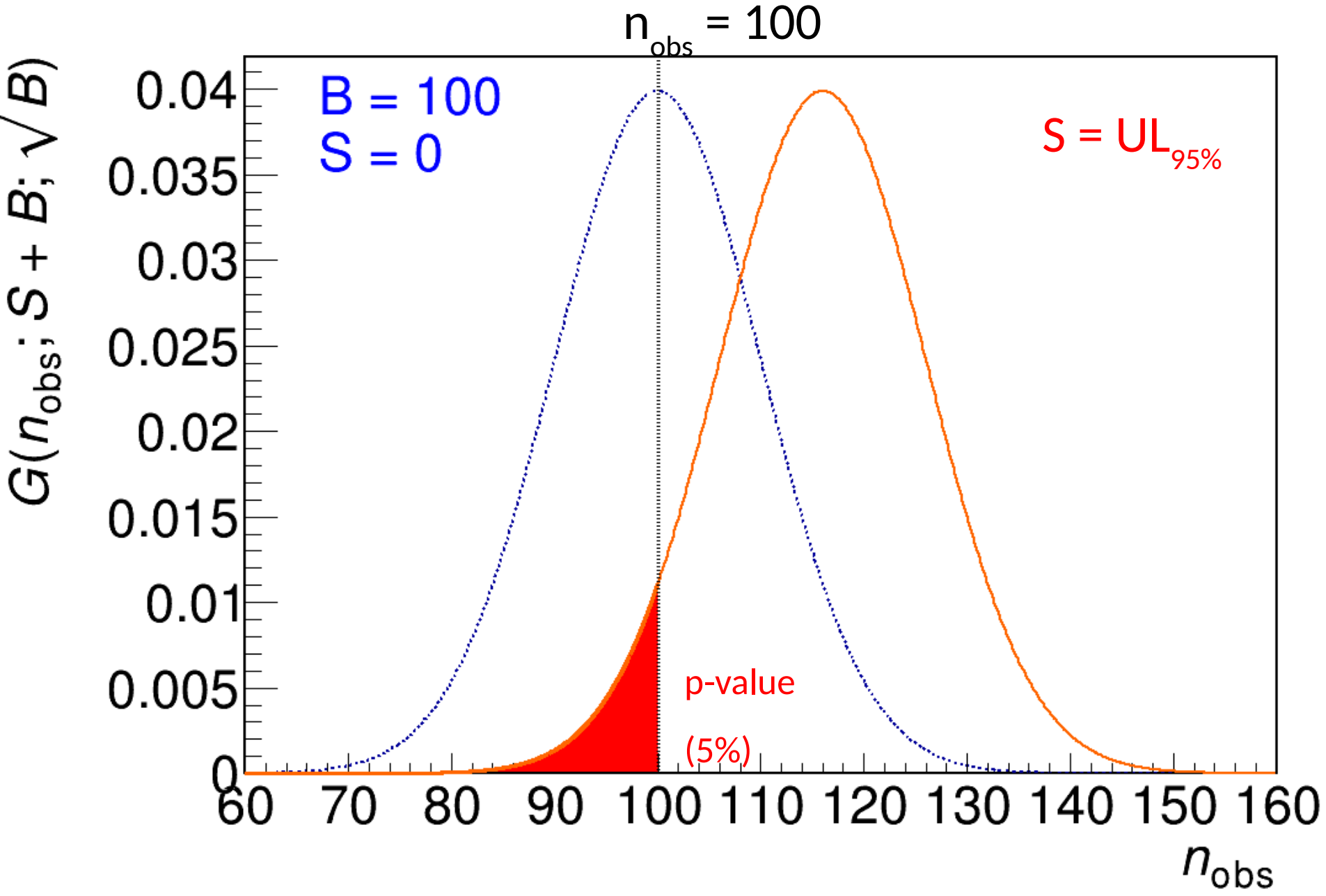
→ More interesting to **exclude large signals**

⇒ **Upper limits on signal yield**

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

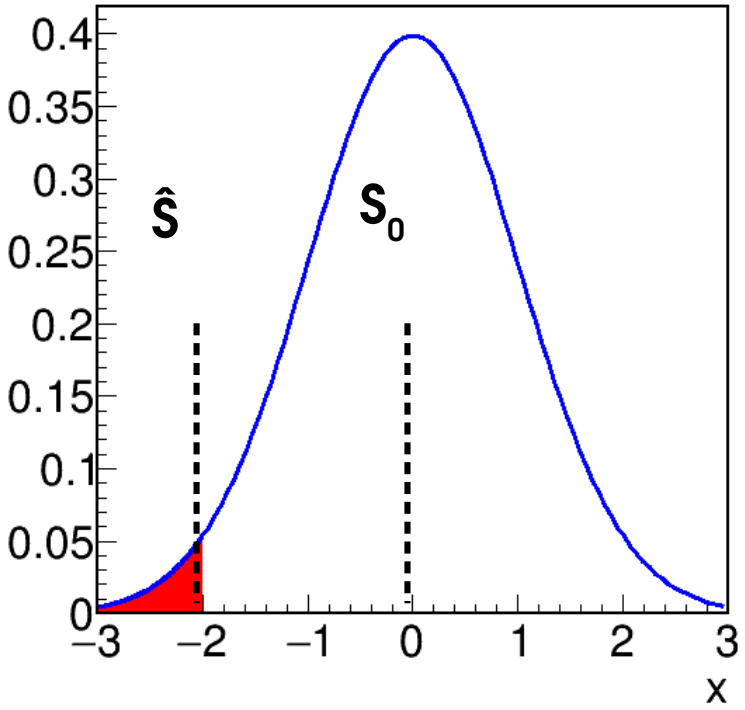


Upper Limit @ 95% CL in Gaussian counting (S=0)



Test Statistics for Limit-Setting

Limit-setting: try to exclude values of S that are above $\hat{S}$.

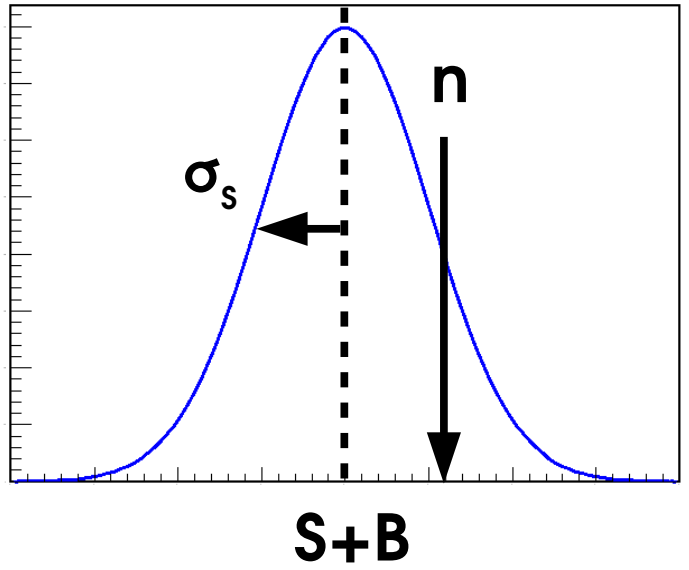


$$q(S_0) = \begin{cases} -2\log\dfrac{L(S=S_0)}{L(\hat{S})} & S_0 > \hat{S} \\ 0 & S_0 \le \hat{S} \end{cases}$$

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

Discovery was also one-sided, for S>0

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 $q(S_0)$

→ 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 !) even when S ≥ 0.

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

Usual p-value for $S=S_0$

p-value for S=0

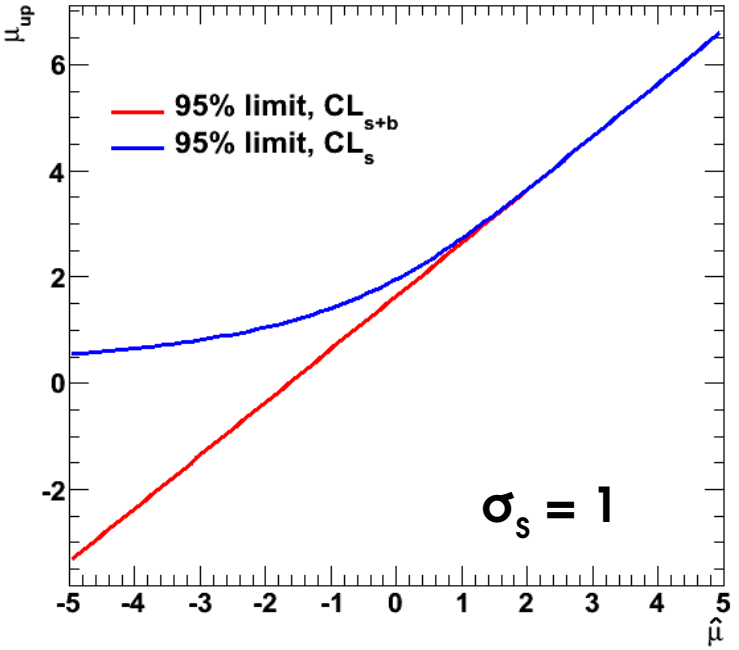
Known feature – to avoid, usual solution in HEP is to use CL_s "modified p-value"

⇒ Compute exclusion relative to that of S=0

→ Somewhat ad-hoc, but good properties…

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

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



Drawback: ***overcoverage***

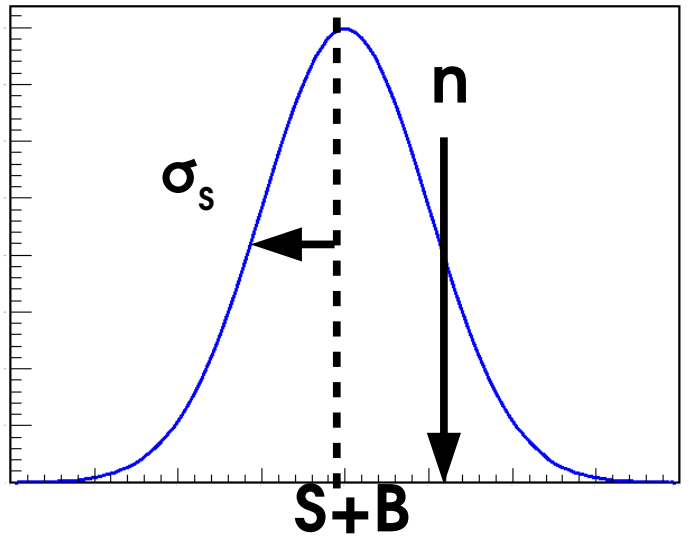
→ limit is claimed to be 95% CL, but actually ≥95% CL for small p_B.

Homework 5: CL_s in the 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: $S_{up} = \hat{S} + 1.64\,\sigma_S$ at 95 % CL

CL_s upper limit :

→ Compute p_{S0} (same as for CLs+b)

→ Compute 1-p_B (hard!)

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

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

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

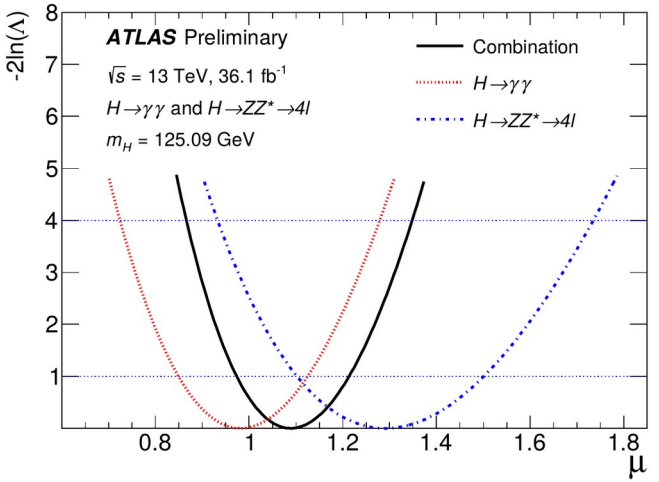
$$S_{up}(n_{obs}=0) = \log(20) = 2.996 \approx 3$$

Takeaways: Confidence intervals and Upper Limits

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

→ Crossings with $t(\mu_0) = 1$ for 1σ intervals (in 1D)

Gaussian regime: $\mu = \hat{\mu} \pm \sigma_\mu$ at 68.3% CL (1σ interval)

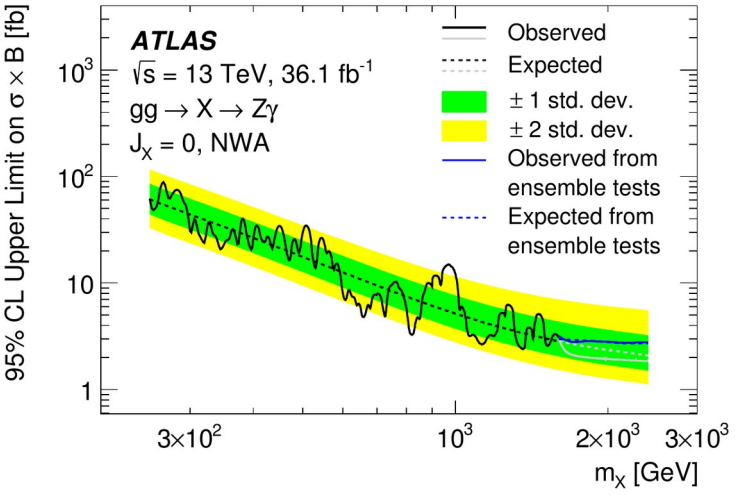


Limits : use LR-based test statistic: $q_{S_0} = -2\log \frac{L(S=S_0)}{L(\hat{S})} \quad S_0 \geq \hat{S}$

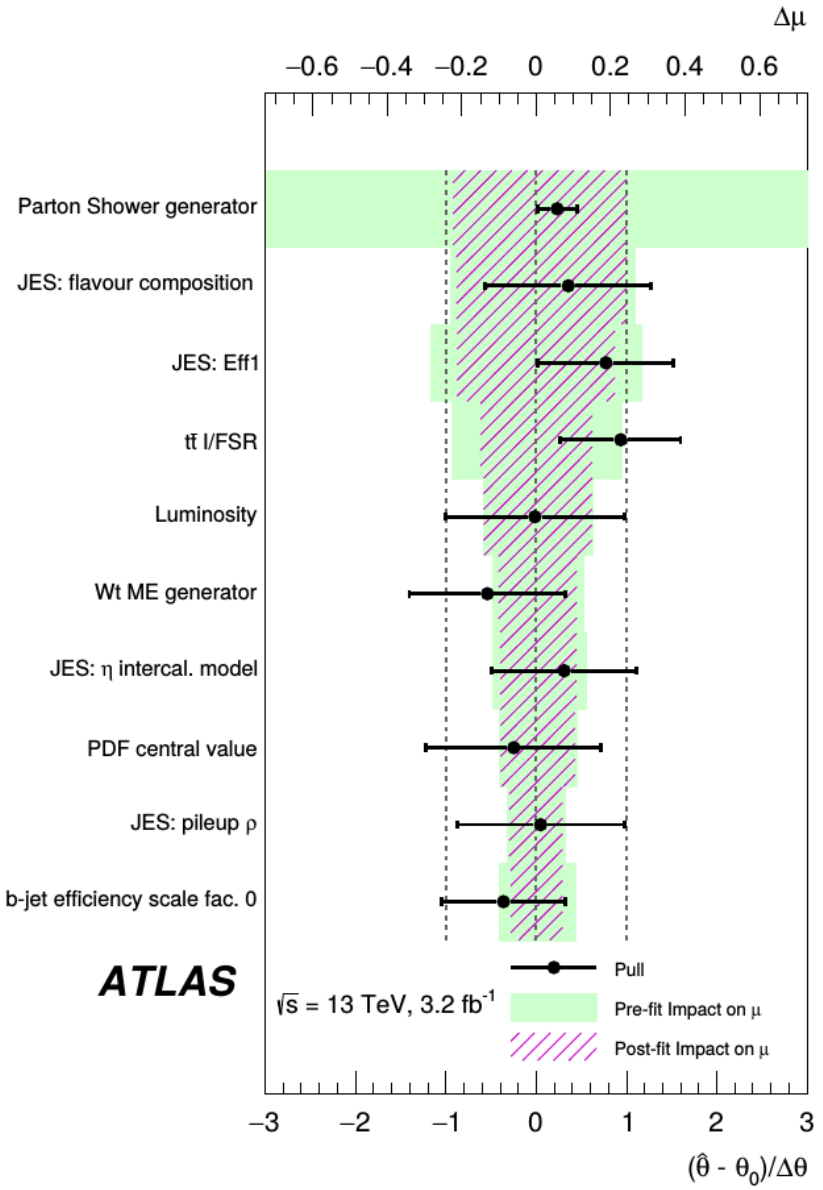
→ Use **CL_s procedure** to avoid negative limits

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

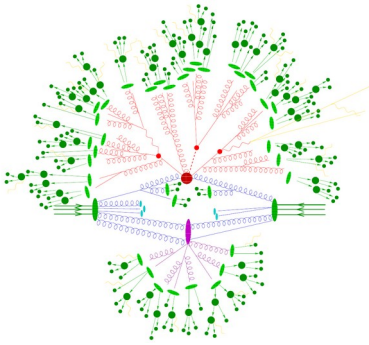
Poisson regime, n=0 : **S < 3 events at 95% CL**



Systematic Errors



Reminder on statistical modeling



Random data must be described using a **statistical model**. Usual cases:

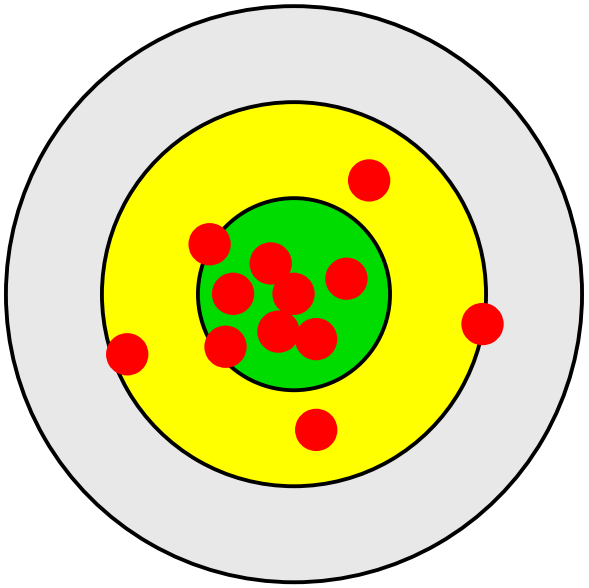
Description	Observable	Likelihood
Counting	**n**	**Poisson** $L(S,B)=e^{-(S+B)}\frac{(S+B)^{n}}{n!}$
Binned shape analysis	n_i, i = 1 .. N_{bins}	**Poisson product** $L(S,B)=\prod_{i=1}^{n_{bins}} e^{-(S f_i^{sig}+B f_i^{bkg})}\frac{(S f_i^{sig}+B f_i^{bkg})^{n_i}}{n_i!}$
Unbinned shape analysis	m_i,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

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

How do we obtain the values of the POIs ?

Systematic Errors

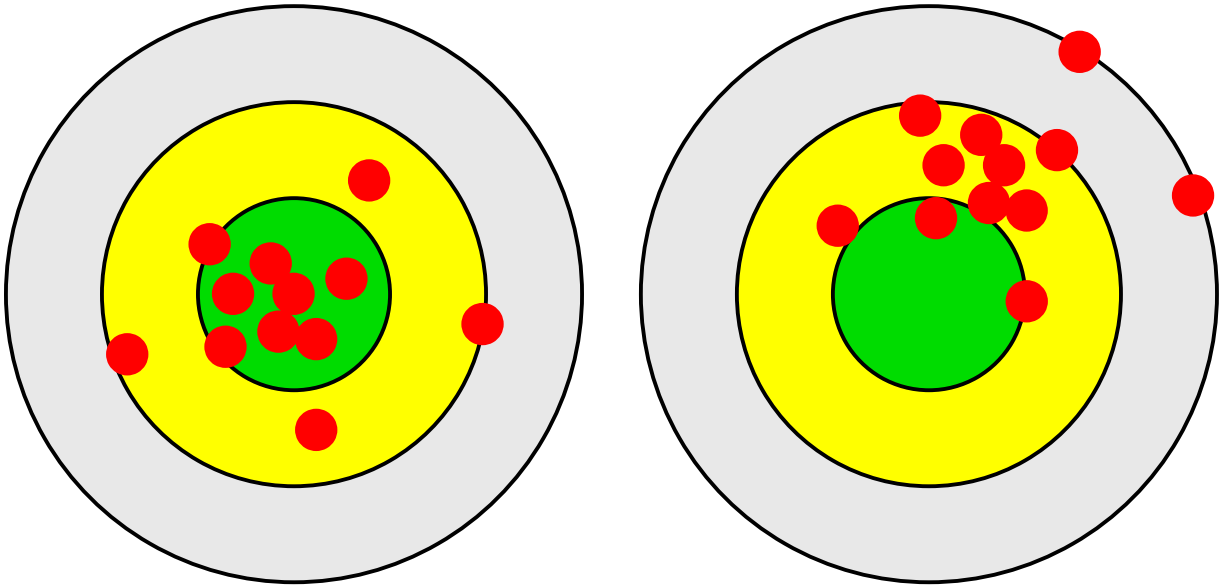
The PDF expresses *uncertainties on the outcome of a measurement*



These uncertainties are called **Statistical Uncertainties.**

Systematic Errors

The PDF expresses *uncertainties on the outcome of a measurement*



These uncertainties are called **Statistical Uncertainties.**

However **the model itself may not be known exactly**

→ We also need **systematic uncertainties,** i.e. *on the form of the PDF itself*.

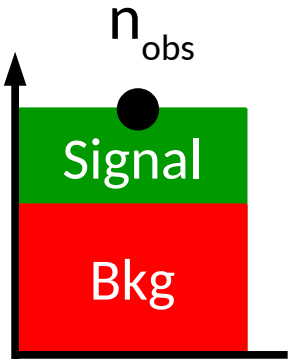
Systematics

"Systematic uncertainty is, in any statistical inference procedure, the uncertainty due to the incomplete knowledge of the probability distribution of the observables. G. Punzi, *What is systematics ?*

Systematics = what we don't know about the random process.

How to describe them in practice ?

→ **Parameterize using additional nuisance parameters (NPs)**



But: completely free NPs can spoil the measurement (e.g. free B ?...)

⇒ **Add constraints in the statistical model**

$$P(\mu, \theta; \text{data}) = P_{\text{measurement}}(\mu, \theta; \text{data})\ C(\theta)$$

POI — Systematics NP — Measurement PDF — NP Constraint term

C(θ) represents **external knowledge** about the NPs that we inject into the statistical model – e.g. to say that "**B = 100 ± 5**"

Frequentist Systematics

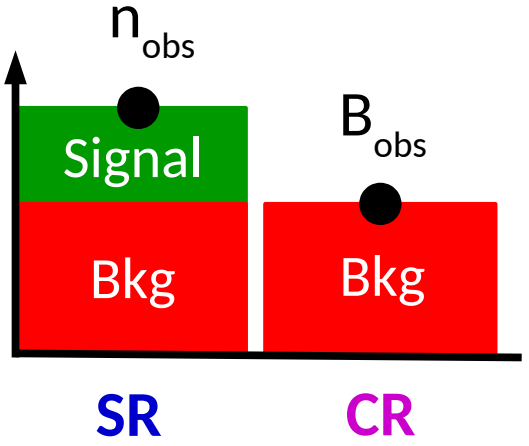
Idea: Systematics NP = NPs measured in a separate ***auxiliary* measurements** (*e.g.* B)

→ Build the **combined PDF** of the main+auxiliary measurements

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

Independent measurements ⇒ Just a product

In the combined likelihood, **systematic NPs are constrained**
⇒ Measured simultaneously with the POIs. in a fit to data.



→ Often no clear setup for auxiliary measurements
(e.g. theory simulation uncertainties)
⇒ **Define constraints "by hand"** ("pseudo-measurement")

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

Profiling nuisance parameters

How to deal with nuisance parameters in likelihood ratios ?

→ **Let the data choose ⇒ use the best-fit values (*Profiling*)**

Profile Likelihood Ratio (PLR)

$\hat{\hat{\theta}}(S_0)$ best-fit value for $S = S_0$ (**conditional** MLE)

$$t(S_0) = -2\log\frac{L(S=S_0, \hat{\hat{\theta}}(S_0))}{L(\hat{S}, \hat{\theta})}$$

$\hat{\theta}$ overall best-fit value (**unconditional** MLE)

Wilks' Theorem* : *same benefits as plain likelihood ratio without NPs:

$t(S_0)$ is distributed as a $\chi^2(n_{dof}=1)$ *also with NPs present*

→ Profiling "builds in" the effect of the NPs

⇒ **Can use $t(S_0)$ to compute uncertainties, significance, etc. same as before**

Homework 7: Gaussian Profiling

Counting experiment with background uncertainty: **n = S + B** :

→ **Signal region (SR)**: $n_{obs} \sim G(S + B, \sigma_{stat})$

→ **Control region (CR)**: $B_{obs} \sim G(B, \sigma_{bkg})$

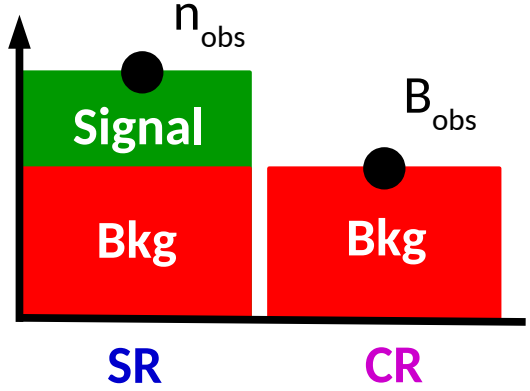
$$L(S,B) = G(n_{obs}; S+B, \sigma_{stat})\, G(B_{obs}; B, \sigma_{bkg})$$

Recall: Signal region only (fixed B): $t(S) = \left(\frac{S - n_{obs}}{\sigma_{stat}}\right)^2$ $S = (n_{obs} - B) \pm \sigma_{stat}$

→ Compute the best-fit (MLEs) for S and B

→ Show that the conditional MLE for B is

$$\hat{\hat{B}}(S) = B_{obs} + \frac{\sigma_{bkg}^2}{\sigma_{stat}^2 + \sigma_{bkg}^2}(\hat{S} - S)$$



→ Compute the profile likelihood t(S)

→ Compute the 1σ confidence interval on S

Answer: $S = (n_{obs} - B_{obs}) \pm \sqrt{\sigma_{stat}^2 + \sigma_{bkg}^2}$ $\sigma_S = \sqrt{\sigma_{stat}^2 + \sigma_{bkg}^2}$

Stat uncertainty (on n) and systematic (on B) add in quadrature

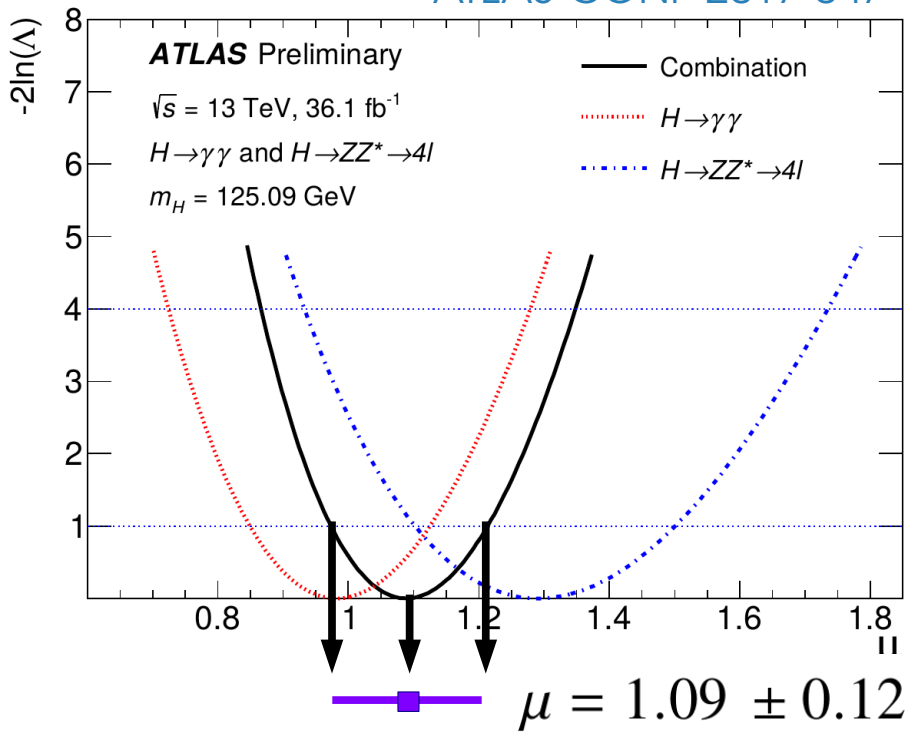
Example: Higgs cross-section measurement

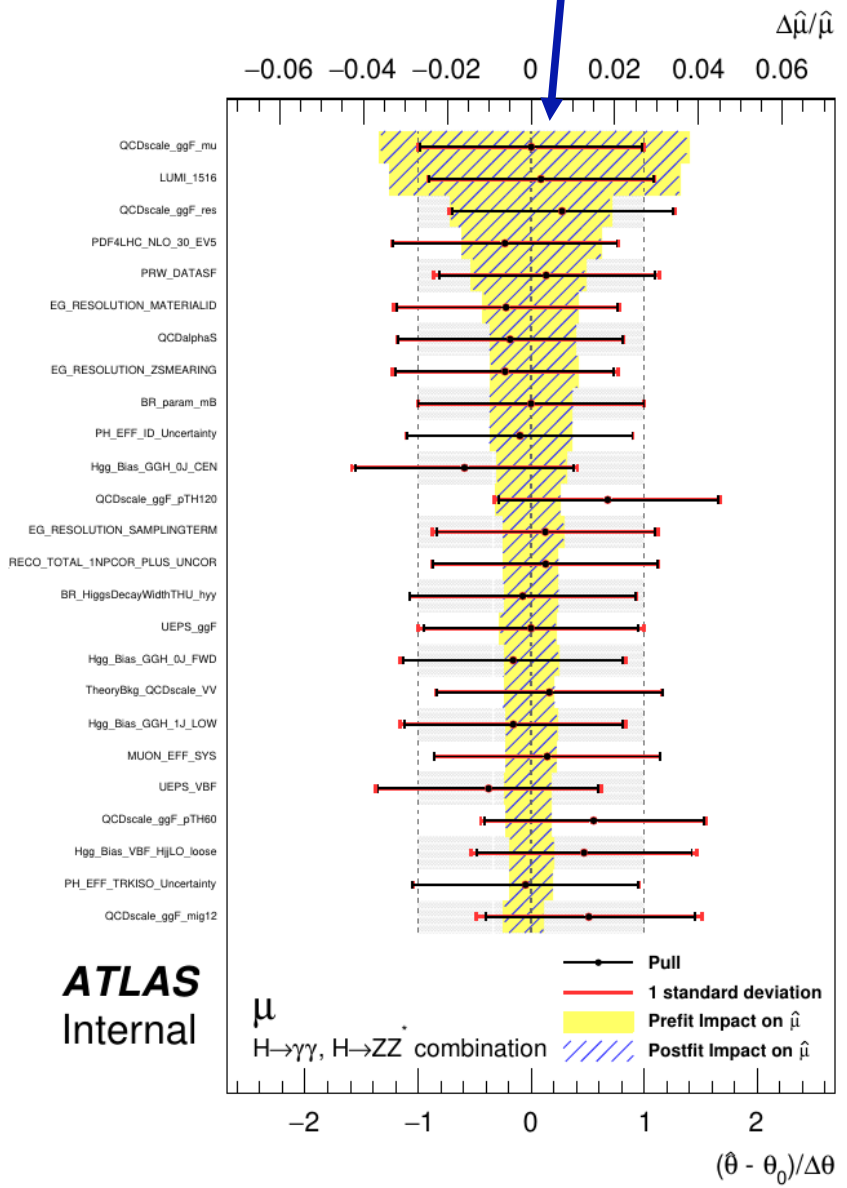
Confidence intervals from Profile Likelihood

- O(100) nuisance parameters also included.
- Best-fit values of the main ones usually provided as part of measurement results.

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

ATLAS-CONF-2017-047



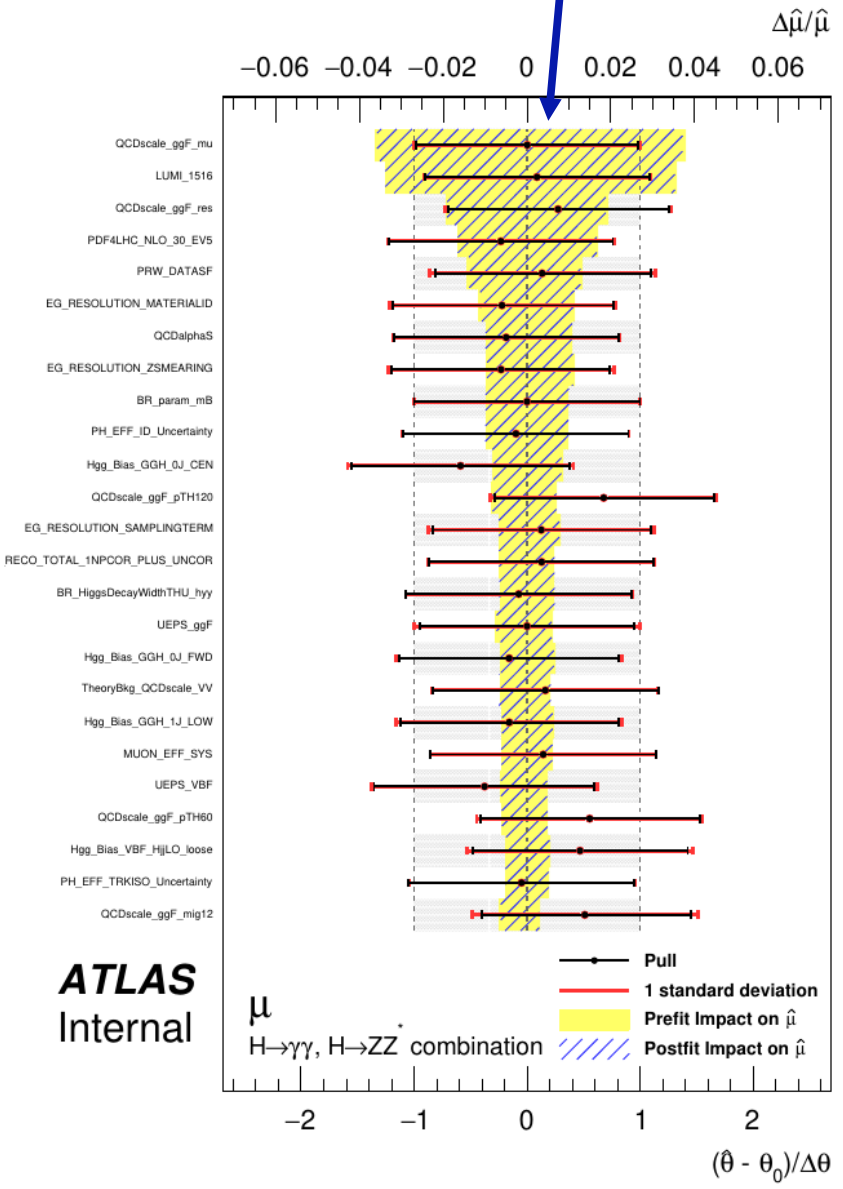


Pull-Impact plots

Confidence intervals from Profile Likelihood

- O(100) nuisance parameters also included.
- Standard presentation: **Pull-impact plot**
  - **Pull:** $(\hat{\theta}-\theta_0) / \sigma_\theta$
  - **Impact:** change in μ from ±1σ shift in θ.

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



Nominally:

- **Pull = 0** : i.e. the pre-fit expectation
- **Pull uncertainty = 1** : uncertainty on the POI is given by the Gaussian constrains

In practice, can have

- **Pull ≠ 0**: if data differs from prefit model
  ⇒ Needs investigation if large
- **Pull Uncertainty < 1** : effect of systematic is *constrained* by the data
  ⇒ Needs to check if this legitimate.

Pull-Impact plots

13 TeV single-t XS (arXiv:1612.07231)

Confidence intervals from Profile Likelihood

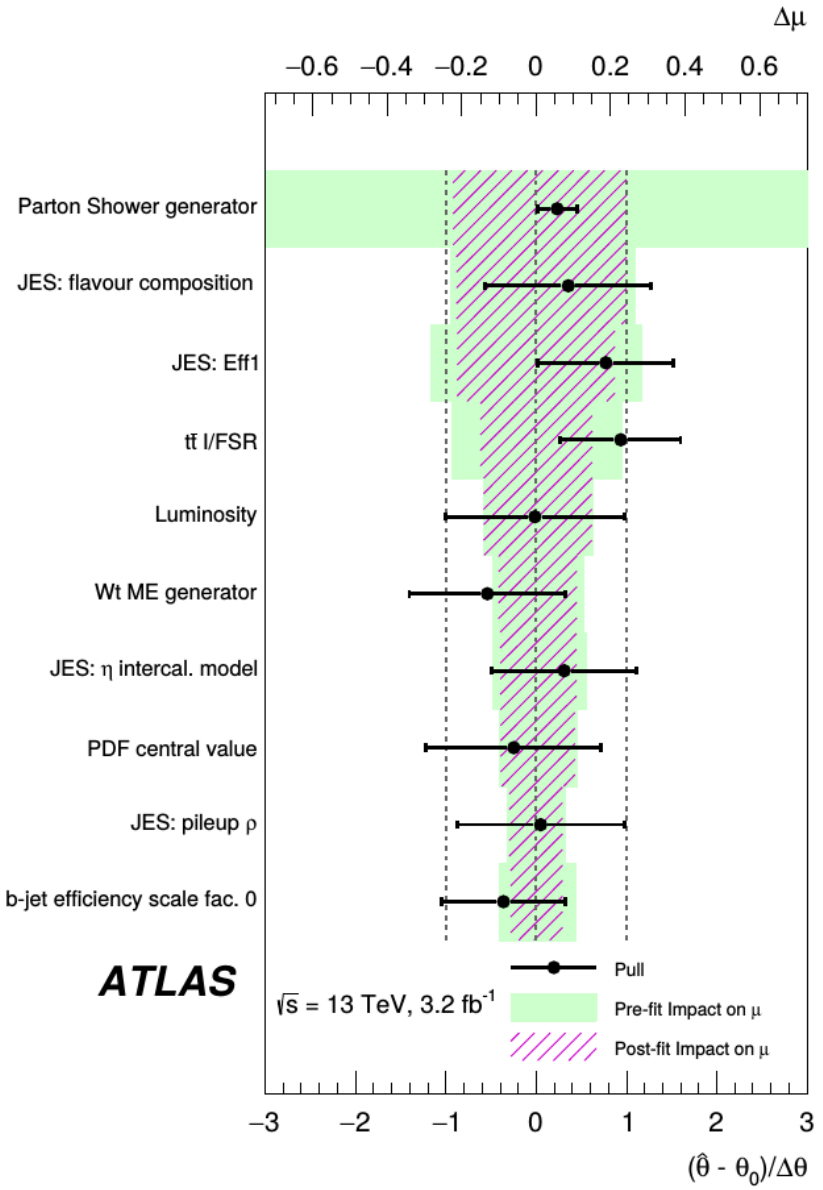
- O(100) nuisance parameters also included.
- Standard presentation: **Pull-impact plot**
  - **Pull:** $(\hat{\theta}-\theta_0) / \sigma_\theta$
  - **Impact:** change in μ from ±1σ shift in θ.

Nominally:

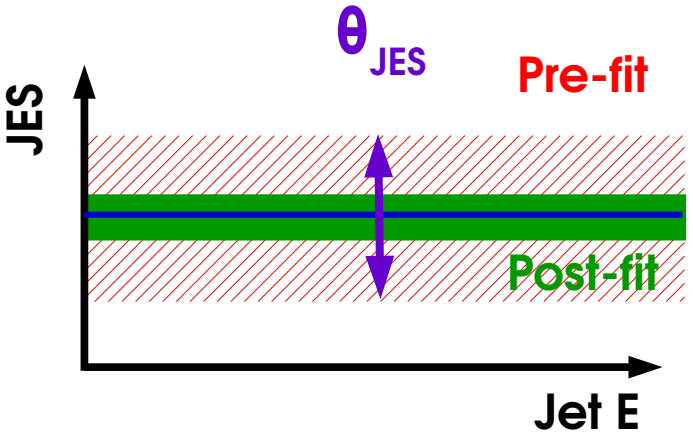
- **Pull = 0** : i.e. the pre-fit expectation
- **Pull uncertainty = 1** : uncertainty on the POI is given by the Gaussian constrains

In practice, can have

- **Pull ≠ 0**: if data differs from prefit model
  ⇒ Needs investigation if large
- **Pull Uncertainty < 1** : effect of systematic is *constrained* by the data
  ⇒ Needs to check if this legitimate.



Profiling Issues



Too simple modeling can have unintended effects

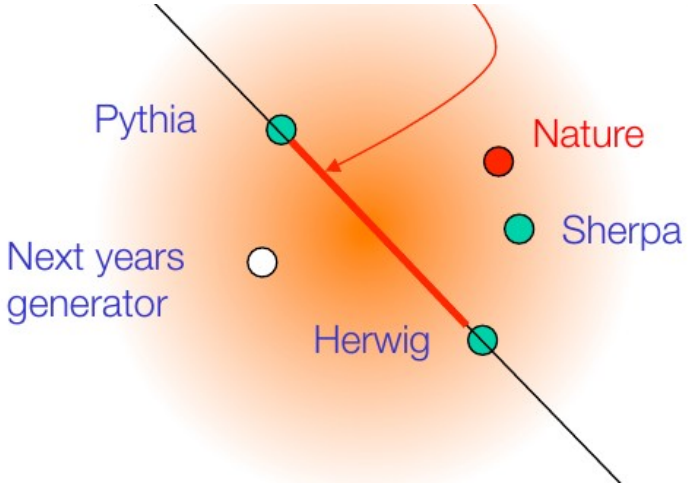
→ e.g. single Jet E scale parameter:

⇒ Low-E jets calibrate high-E jets – intended ?

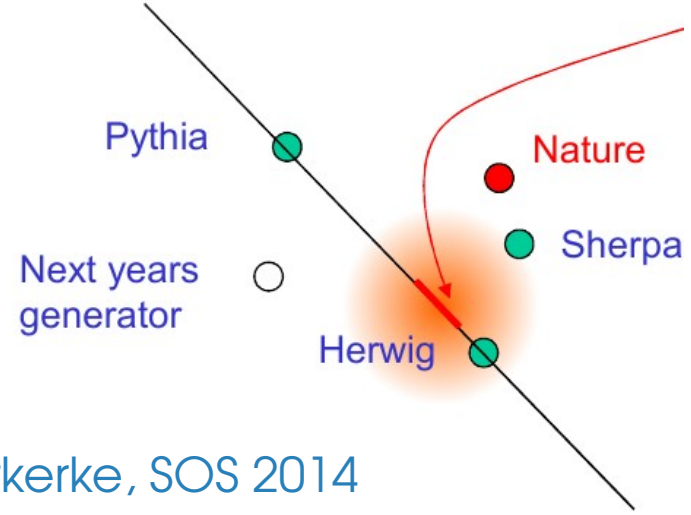
***Two-point* uncertainties**:

→ Interpolation may not cover full configuration space, can lead to too-strong constraints

Pre-fit constraint

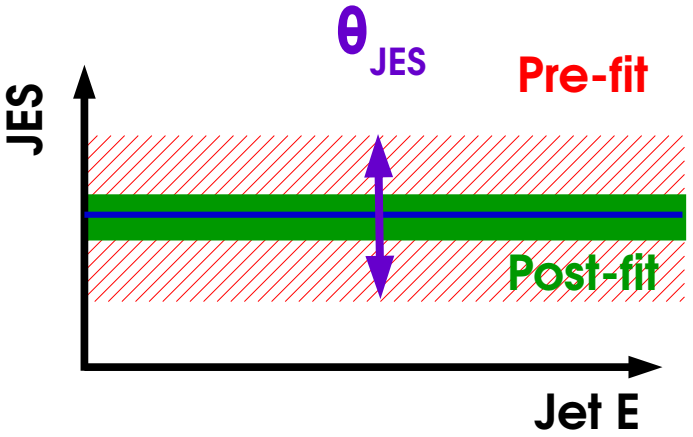


Post -fit constraint



W. Verkerke, SOS 2014

Profiling Issues



Too simple modeling can have unintended effects

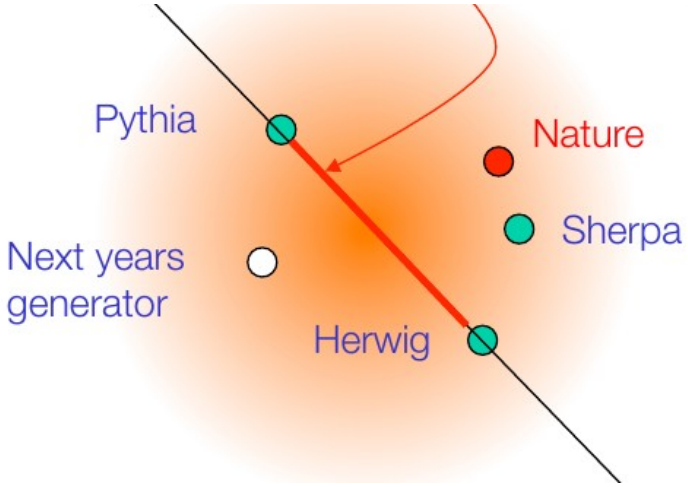
→ e.g. single Jet E scale parameter:

⇒ Low-E jets calibrate high-E jets – intended ?

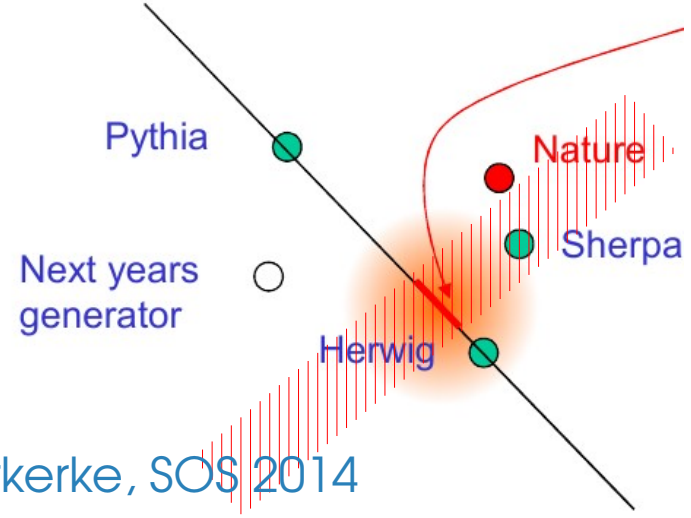
***Two-point* uncertainties**:

→ Interpolation may not cover full configuration space, can lead to too-strong constraints

Pre-fit constraint



Post -fit constraint



W. Verkerke, SOS 2014

Takeaways: Profiling

When testing a hypothesis, use the best-fit values of the nuisance parameters: ***Profile Likelihood Ratio***.

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

Allows to include systematics as nuisance parameters subject to constraints

Profiling systematics includes their effect into the total uncertainty.

Gaussian case: $\sigma_{\text{total}} = \sqrt{\sigma^2_{\text{stat}} + \sigma^2_{\text{syst}}}$

Guaranteed to work well as long as everything is Gaussian, but typically also robust against non-Gaussian behavior.

Profiling can have unintended effects : need to carefully check behavior

Bayesian Analysis

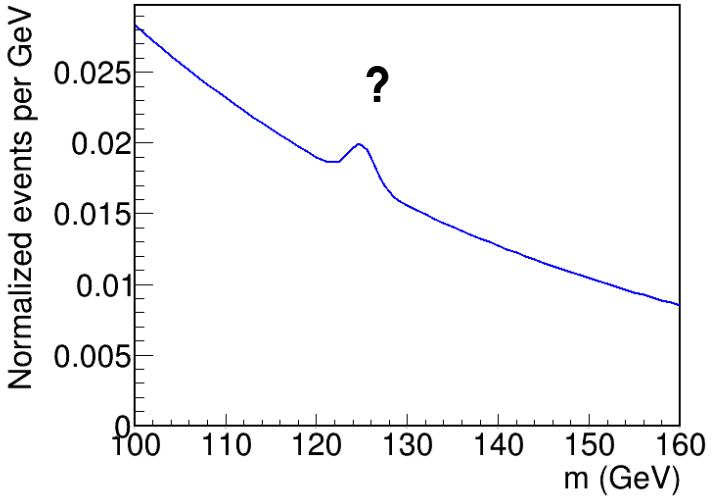
Bayesian methods

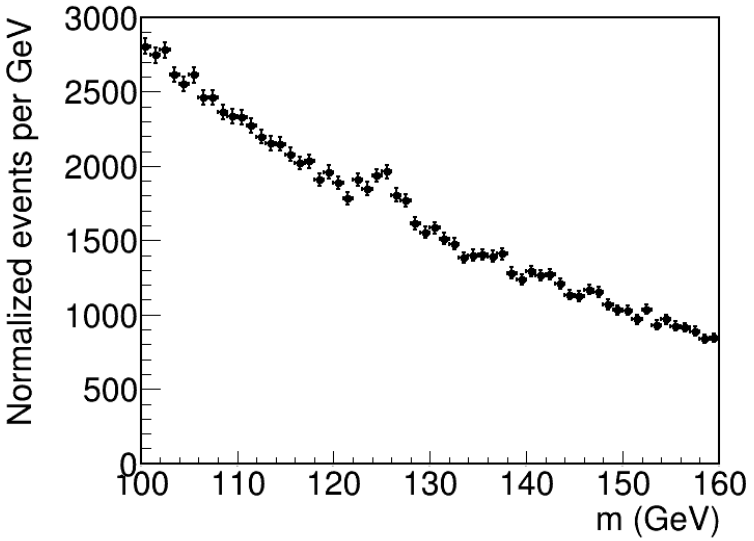
Remember the problem discussed this morning:

- PDFs give possible outcomes for known parameters
- We already know the outcome, and want information on the parameters

$P(\lambda = ?)$ ⟵ **2**

Estimate ⟵





Solution: maximum likelihood estimation of the parameters, given the data

This is a (good) solution ("classical/frequentist") **but there is another way.**

Bayesian methods

Bayesian methods: promote parameters (POIs and NPs) to random variables

→ Represent our best knowledge of their value, not the true values.

Can use **Bayes' Theorem** to obtain *a PDF for the parameters*

Bayes' Theorem $P(\mu|n) = P(n|\mu)\,\frac{P(\mu)}{P(n)}$

Prior PDF on μ: represents our knowledge before the measurement

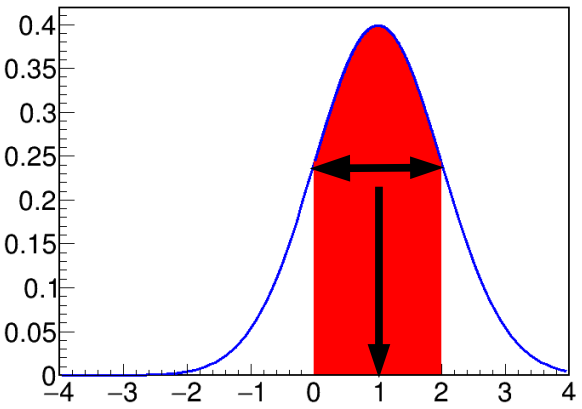
Posterior PDF: represents our total knowledge from prior + measurement

Measurement PDF, same as for the frequentist P(n;μ)

Norm. factor: adjusted so P(μ|n) is normalized to 1

Immediately useful to get intervals on μ:

- Peak of P(μ|n) gives the central value : *Maximum a posteriori* (MAP).
- 68.3% interquantile gives the 1σ interval



Problem: what to use for the prior ?...

Bayesian methods

Systematics and nuisance parameters:

Each NP is considered a random variable: Bayes theorem gives **P(μ, θ | n)**

Define a prior **π(θ)** for each nuisance parameter.

⇒ Obtain **P(μ|n)** for μ alone by integrating out the θ:

$$P(\mu|n) = \int P(\mu, \theta|n)\, \pi(\theta)\, d\theta$$

Use probability distribution P(μ) to compute intervals and limits as before.

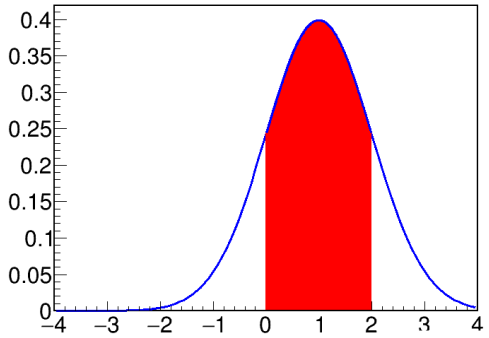
68.3% CL interval:

$$\int_A^B P(\mu|n)\, d\mu = 68.3\,\%$$

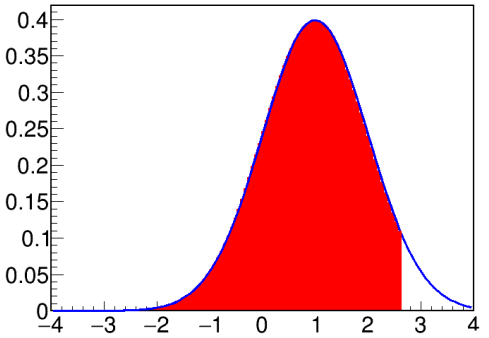
Here CL means "Credibility Level")

95% CL upper limit

$$\int_{-\infty}^{L} P(\mu|n)\, d\mu = 95\,\%$$



μ



μ

Bayesian vs. frequentist

> *"Bayesians address the question everyone is interested in, by using assumptions no-one believes.*
> *Frequentists use impeccable logic to deal with an issue of no interest to anyone."*
>
> – Louis Lyons

Many points of commonality, but some important differences!

Bayesian analysis typically:

⊕ Conceptually simpler – frequentist results often difficult to interpret

⊖ No simple way to test for discovery

⊕ Hybrid methods sometimes used (frequentist discovery + Bayesian systs)

⊖ No support for NPs constrained in data

⊖ Integration over NPs can be CPU-intensive (but can use MCMC methods)

⊕ Minimization over many NPs also not a simple problem for frequentist case…

⊖ Need to specify priors, which often contains some arbitrariness – e.g. a prior flat in one parameterization is usually not flat in another.

⊕ Can use Jeffreys' or reference priors to avoid this, although difficult in practice.

⊕ Frequentist and Bayesian results often agree, so not a big issue in practice!

Homework 8: Bayesian methods and CL_s

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

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

→ **What is the 95% CL upper limit on S, given a measurement n_{obs} ?**

1. CL_s computation:

- Use the result of Homework 7 to compute the PLR for S
- Use the result of Homework 6 to compute the CLs upper limit

2. Bayesian computation:

- Integrate P(n; S, θ) over θ to get the marginalized P(n| S)
- Use Bayes' theorem to compute P(S|n) ∝ P(n|S) P(S), with P(S) a flat prior over S>0.
- Find the 95% CL limit by solving $\int_{S_{up}}^{\infty} P(S|n)\,dS = 5\,\%$

Solution:

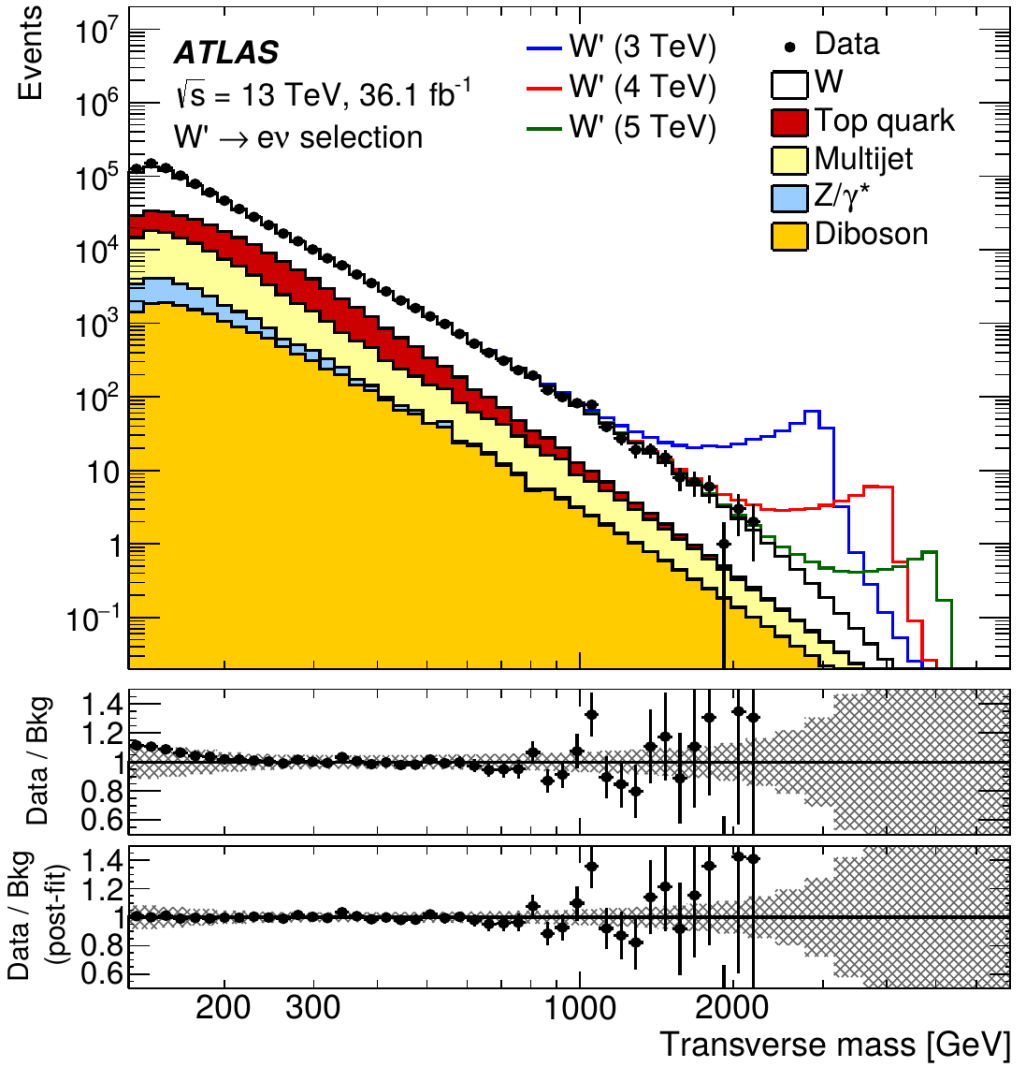
(in both cases!)

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

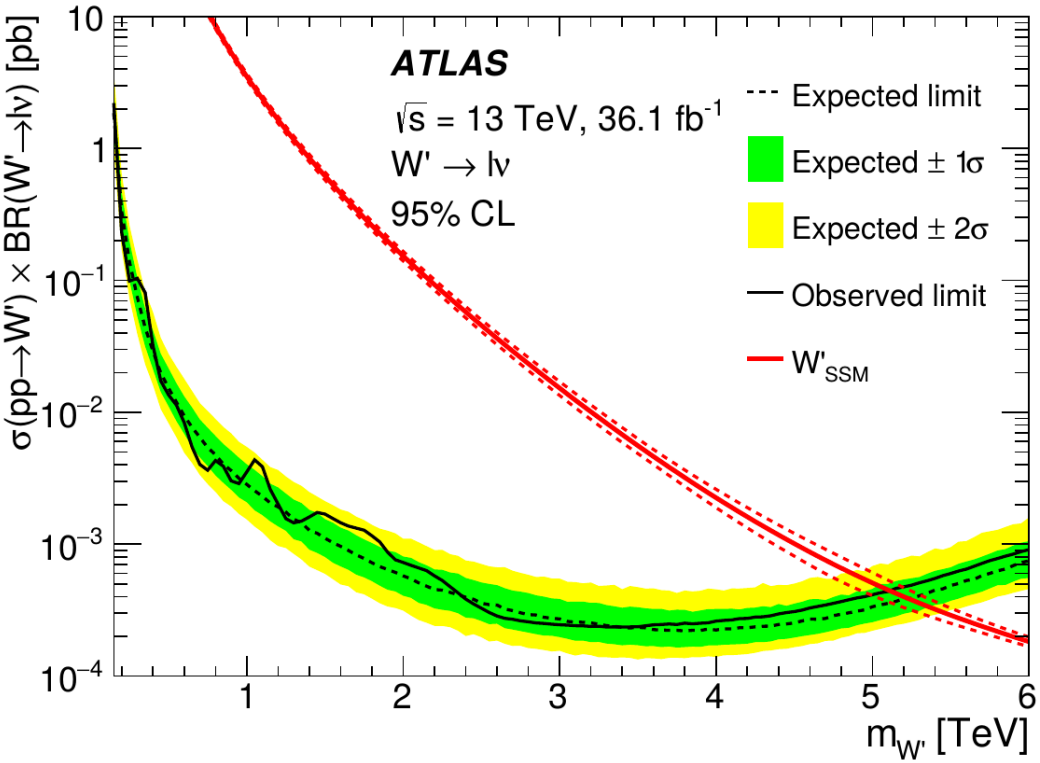
Example: W'→lν Search

arXiv:1706.04786

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



Trigger	Multijet background
Lepton reconstruction and identification	Top extrapolation
Lepton momentum scale and resolution	Diboson extrapolation
$E_{\mathrm{T}}^{\mathrm{miss}}$ resolution and scale	PDF choice for DY
Jet energy resolution	PDF variation for DY
Pile-up	EW corrections for DY
	Luminosity



Other Topics

Generating Pseudo-data

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

Þ Possible outcomes of the experiment, for given parameter values

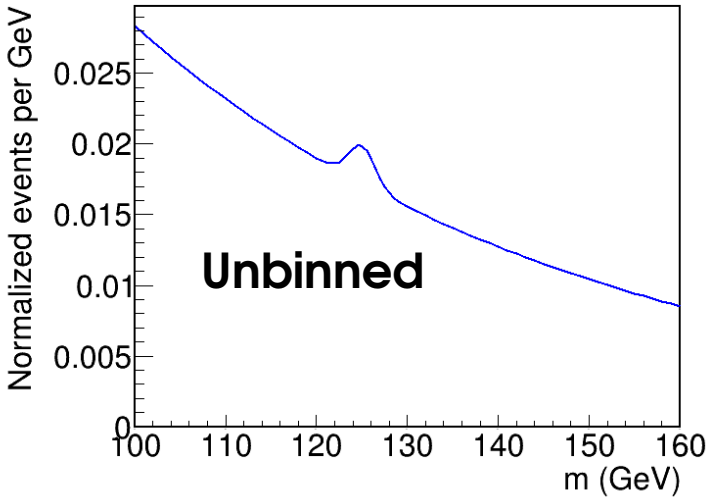
Can draw random events according to PDF : **generate *pseudo-data***

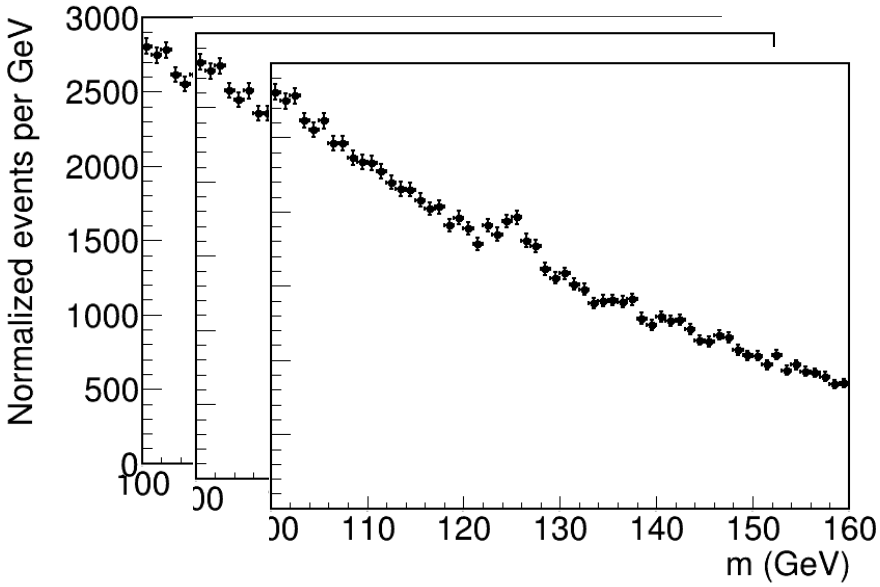
$$P(\lambda = 5)$$

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

Each entry = separate "experiment"

Generate





Expected Results: Toys

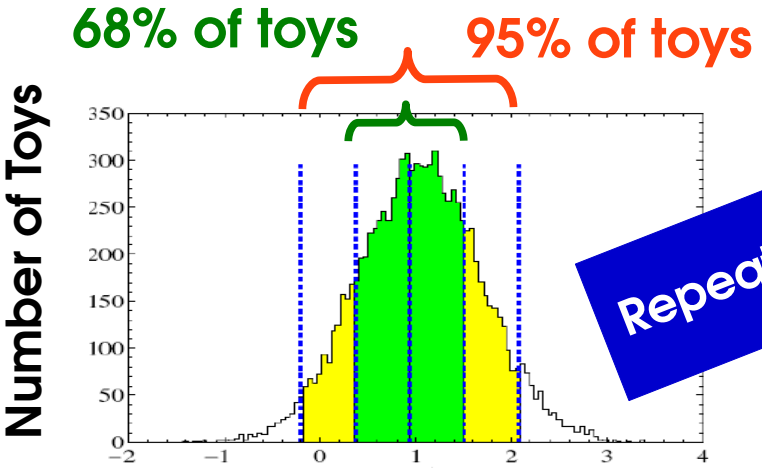
***Expected* results**: median outcome under a given hypothesis

→ usually B-only for searches, but other choices possible.

Two main ways to compute:

→ **Pseudo-experiments (*toys*):**

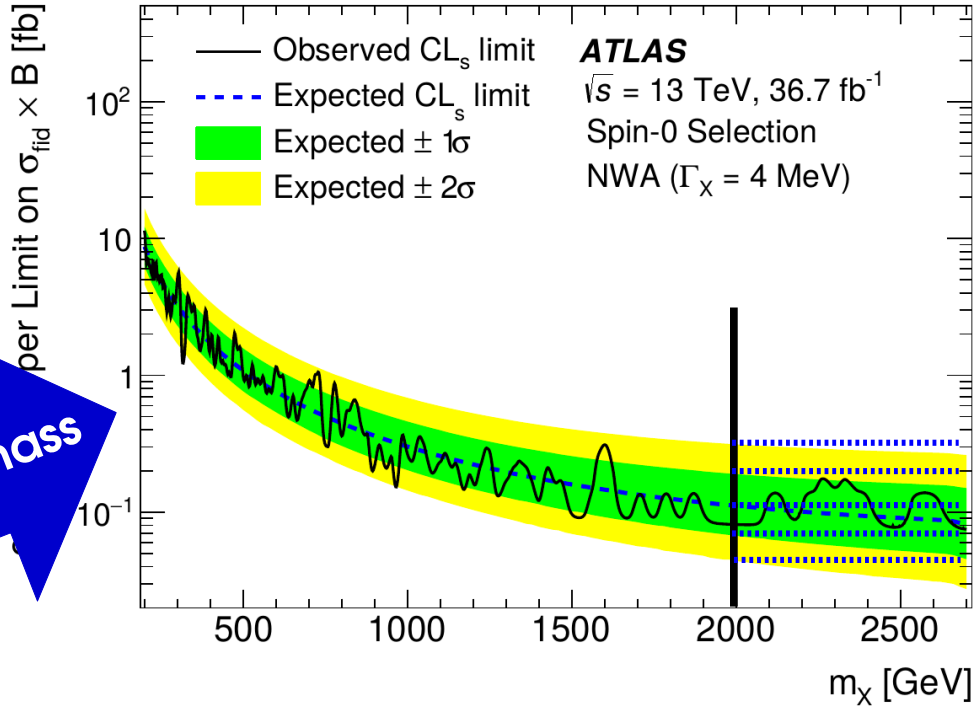
- Generate a pseudo-dataset in B-only hypothesis
- Compute limit
- Repeat and histogram the results
- Central value = median, bands based on quantiles



Eur.Phys.J.C71:1554,2011

Repeat for each mass

Phys. Lett. B 775 (2017) 105



Expected Results: 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

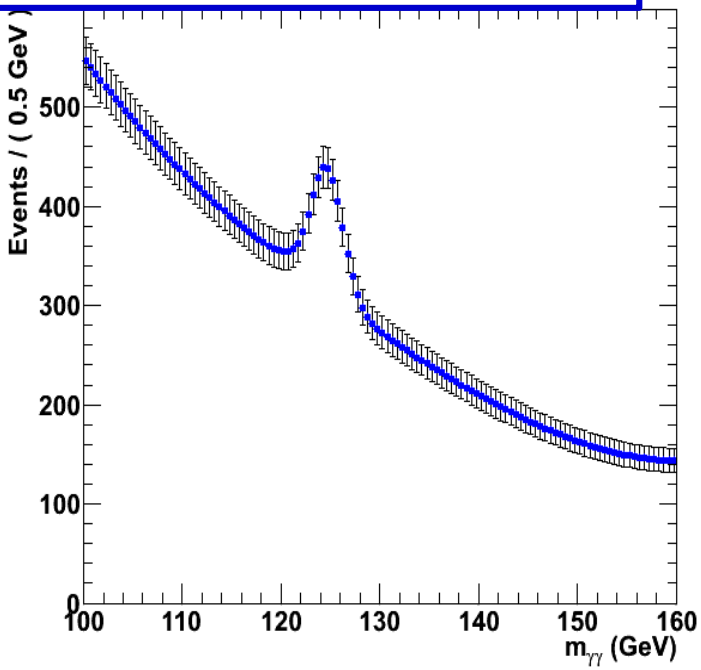
→ ***Asimov Datasets***

- 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**

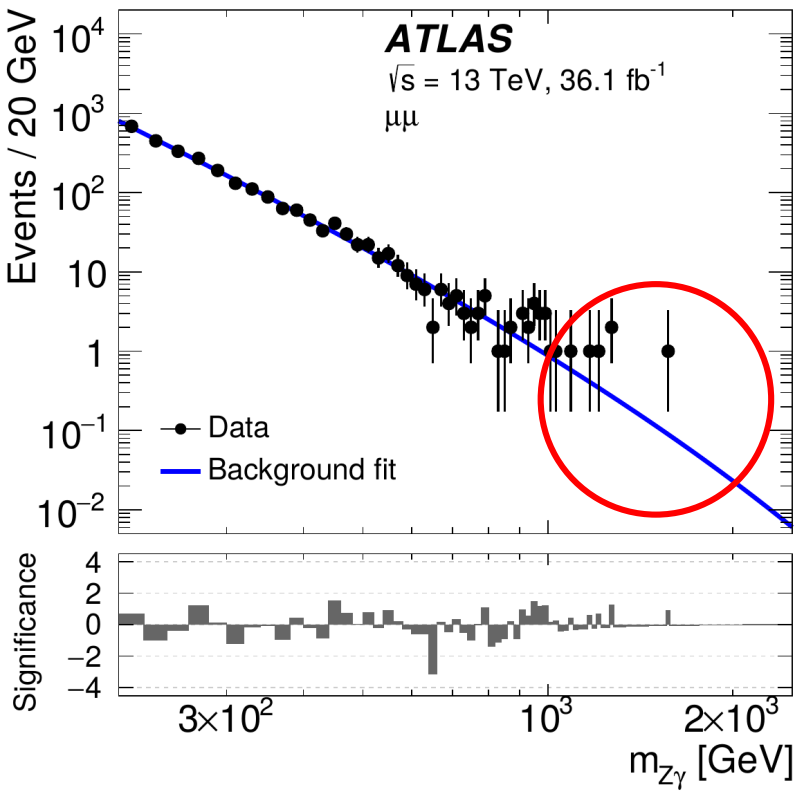


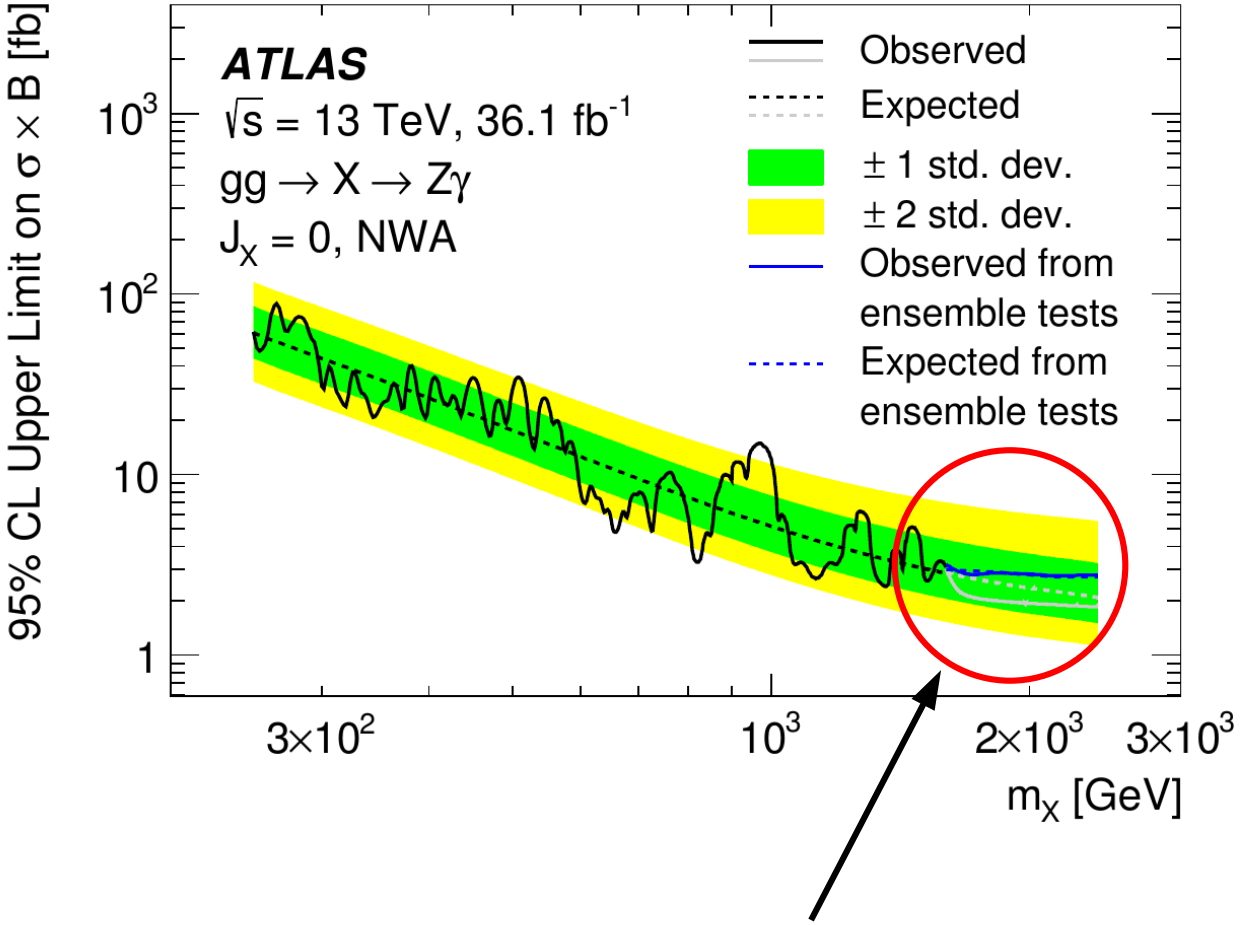
Toys: Example

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

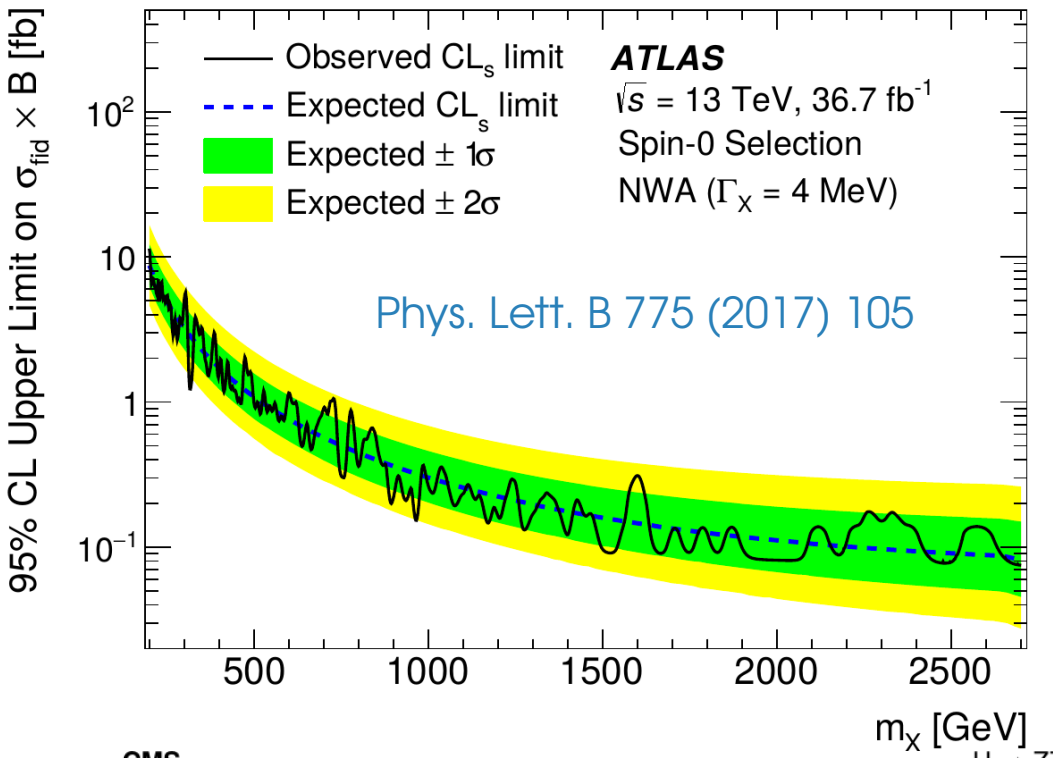




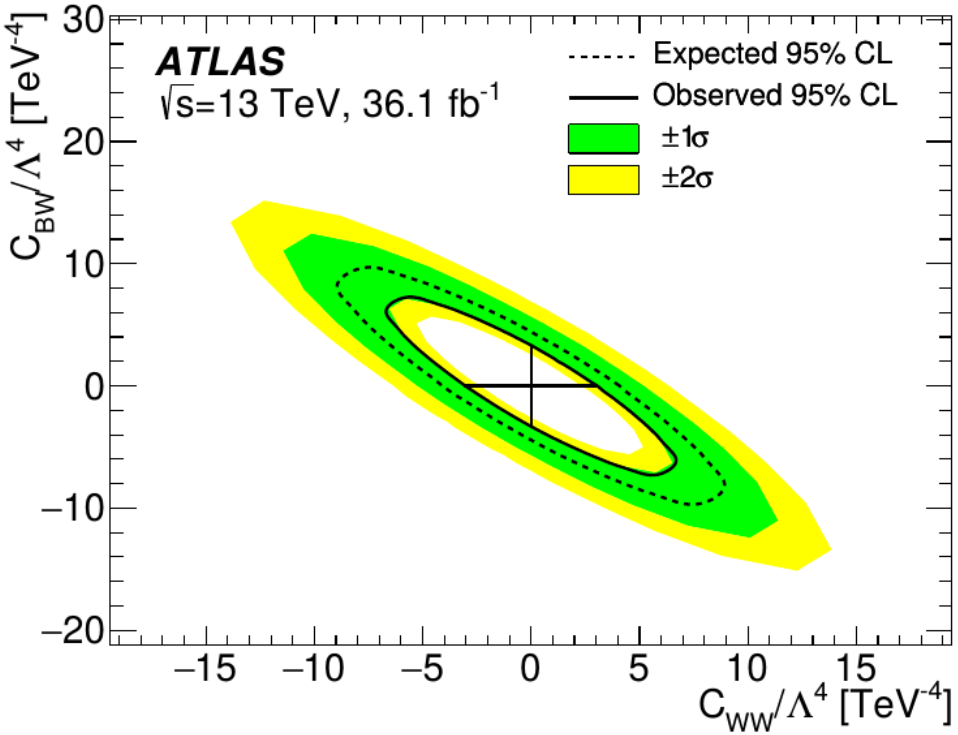
Asimov results (in gray) give optimistic result compared to toys (**in blue**)

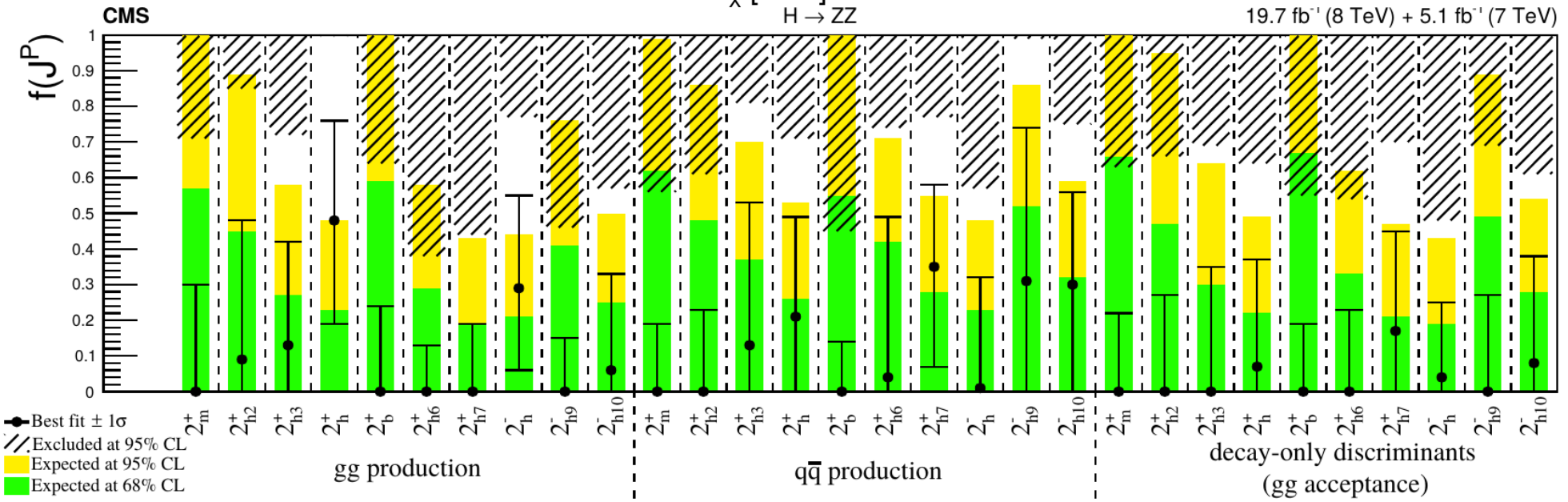
Upper Limit Examples

ATLAS 2015-2016 4l aTGC Search



Phys. Lett. B 775 (2017) 105





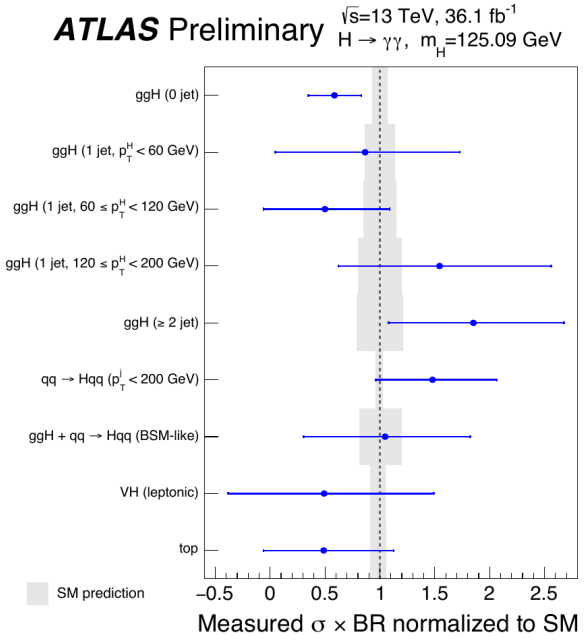
Phys. Rev. D 92 (2015) 012004

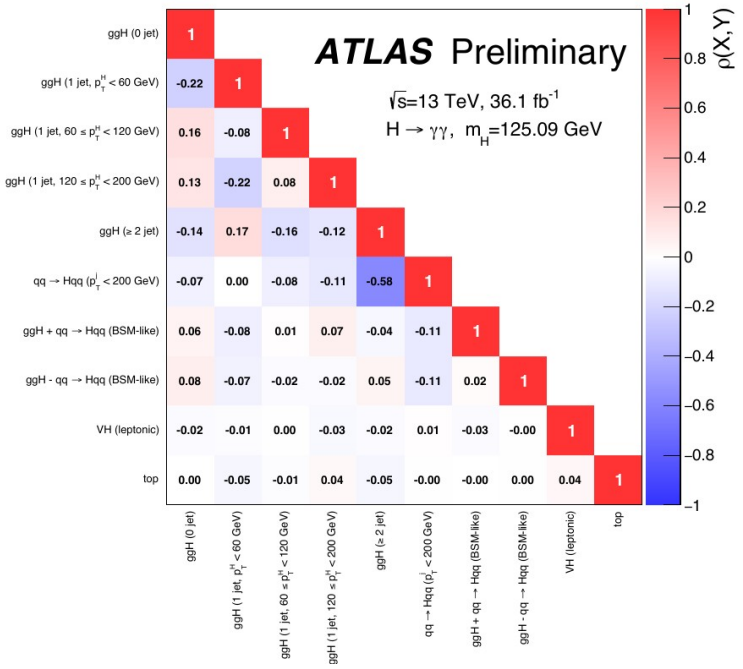
Presentation of Results

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

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

→ For example using the HEPData repository.





Non-Gaussian case: need publish full likelihood (e.g. here)

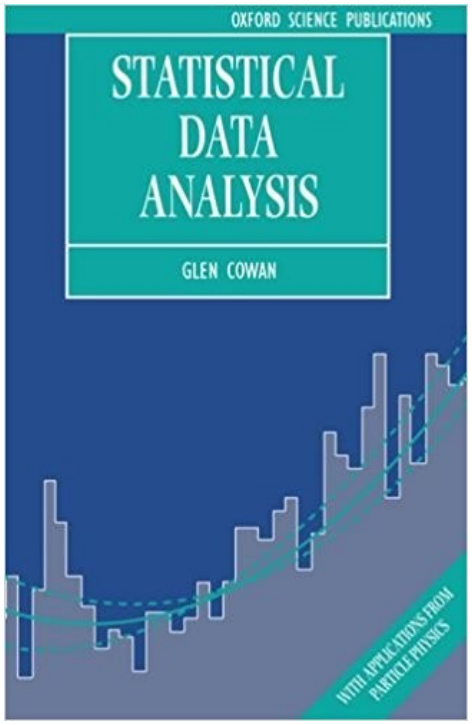
Not always done but increasingly common

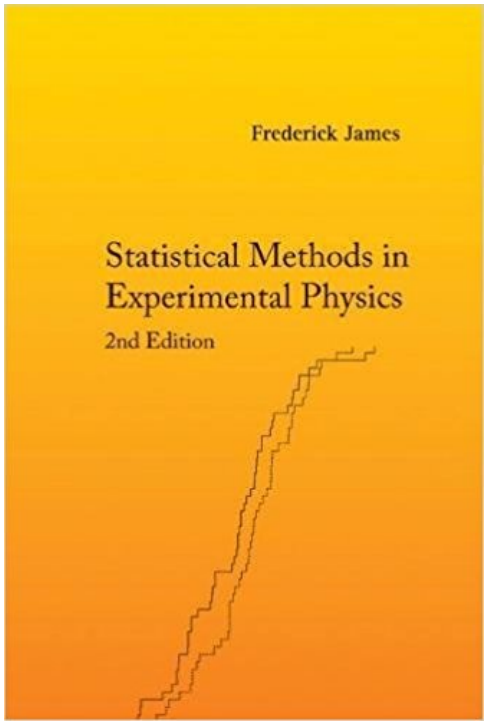
Conclusion

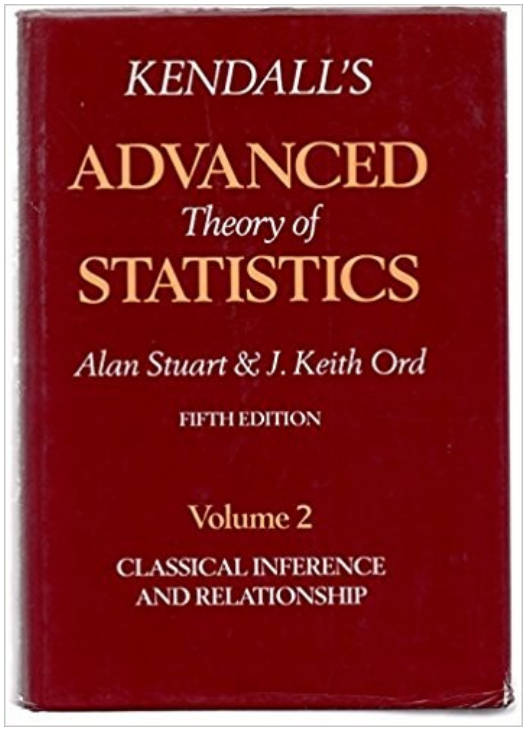
- Significant evolution in the statistical methods used in HEP
- Range of techniques, adapted to various situations and results.
- Benefits:
  - **Modeling the statistical process with high precision in difficult situations (large systematics, small signals)**
  - **Making optimal use of available information**
- Implemented in standard RooFit/RooStat toolkits within the ROOT framework, as well as other tools.
- Still many open questions and areas that could use improvement, e.g. on how to **make best use of the data** and **sharing our results**

Books and Courses

See A. Höcker's CERN Summer student program statistics course







Some other courses available online:

Glen Cowan's Cours d'Hiver and 2010 CERN Academic Training lectures

Kyle Cranmer's CERN Academic Training lectures

Louis Lyons'and Lorenzo Moneta's CERN Academic Training Lectures

Hands-on exercises

The Statistics course will include Hands-on exercises on **jupyter notebooks** (built using the **`numpy/scipy/pyplot`** stack) are part of the course.

If you have a computer with you, **please install anaconda** as this provides a consistent installation of python, JupyterLab, etc.

→ Alternatively, you can also install JupyterLab as a standalone package.

→ Another solution is to run the notebooks on the public jupyter servers at mybinder.org. This will probably be slower but avoids a local install.

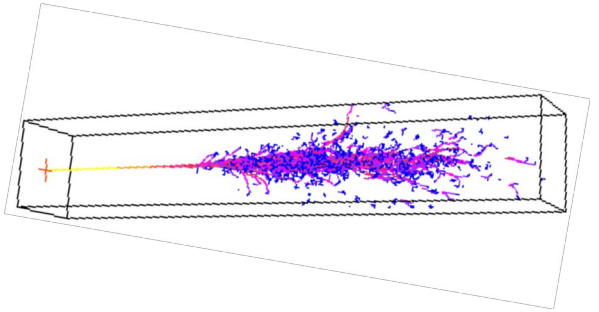
Lecture 1	Lecture Notes	notebook [solutions]	binder [solutions]
Lecture 2	Lecture Notes	notebook [solutions]	binder [solutions]

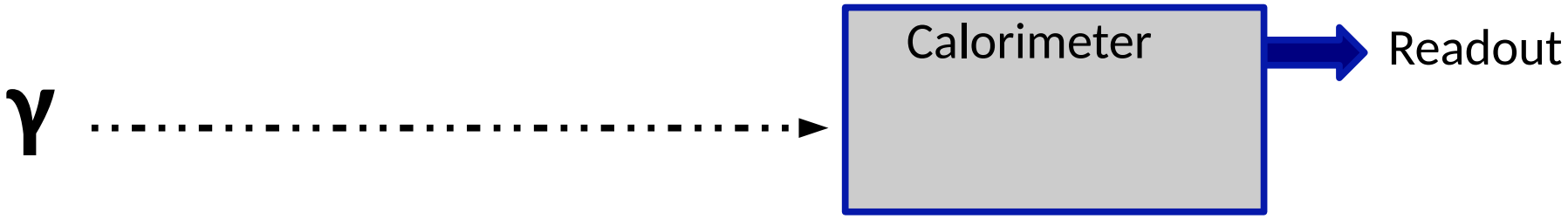
Extra Slides

Probability basics

Measurement Errors

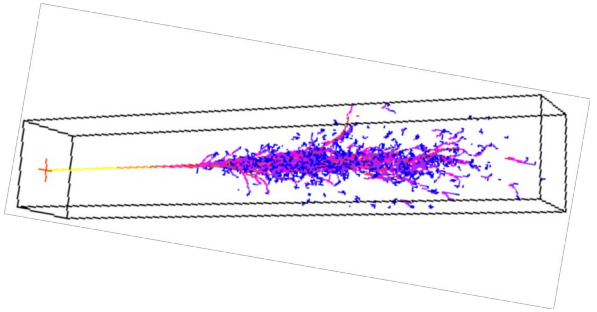
Example: measuring the energy of a photon in a calorimeter

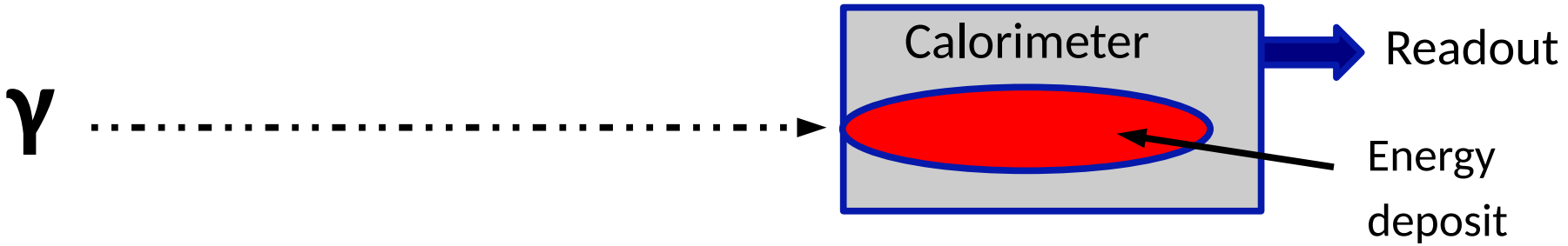


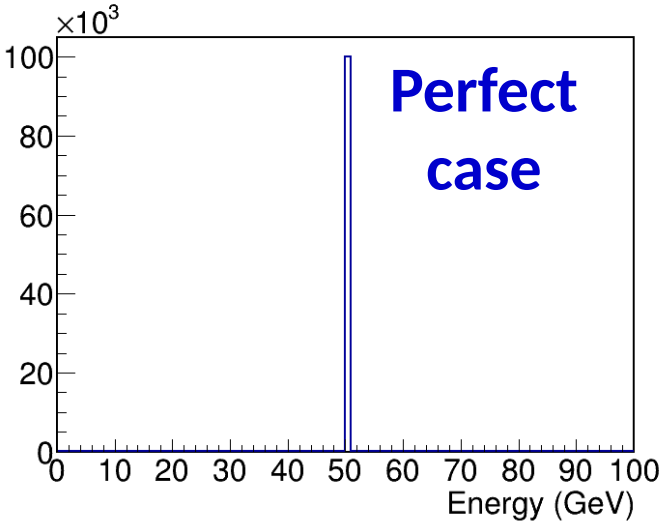


Measurement Errors

Example: measuring the energy of a photon in a calorimeter

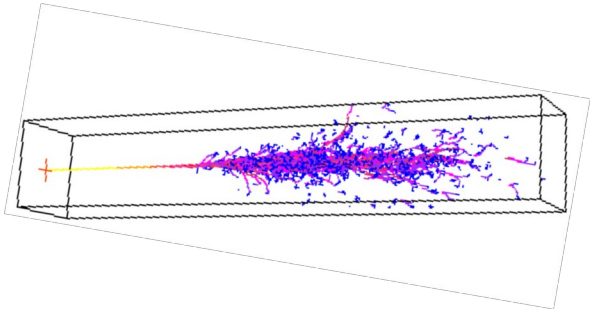


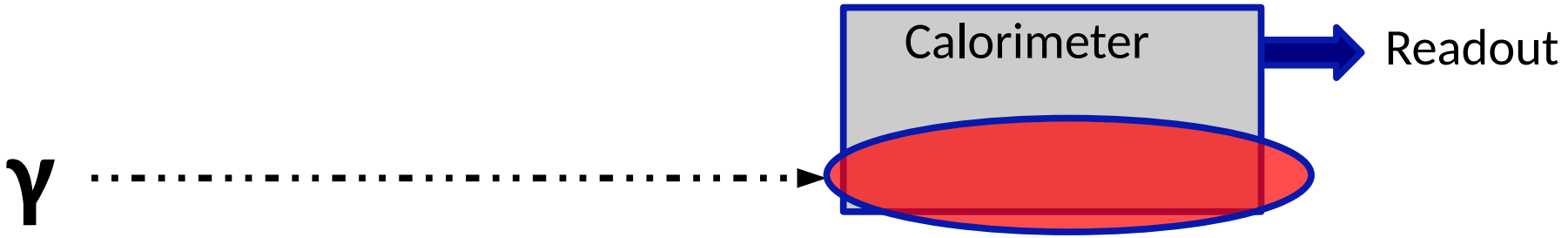


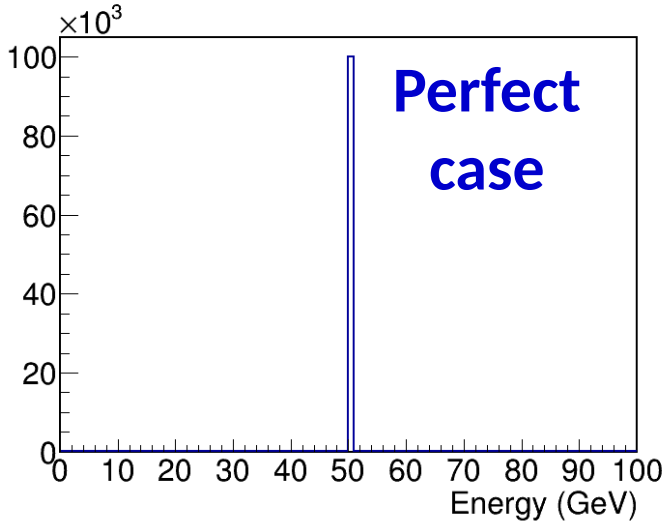


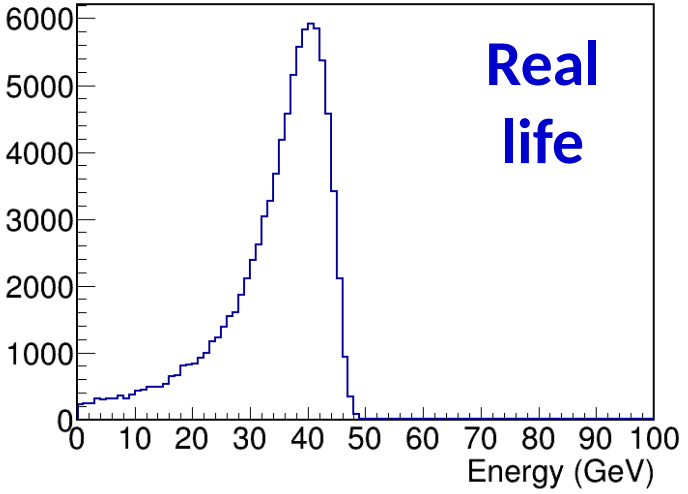
Measurement Errors

Example: measuring the energy of a photon in a calorimeter



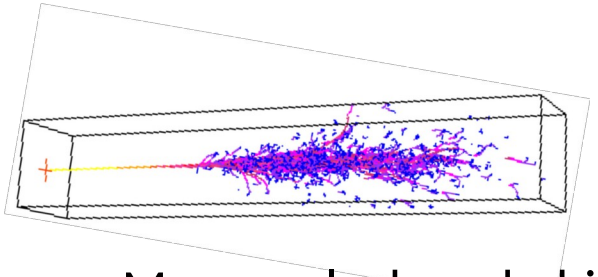


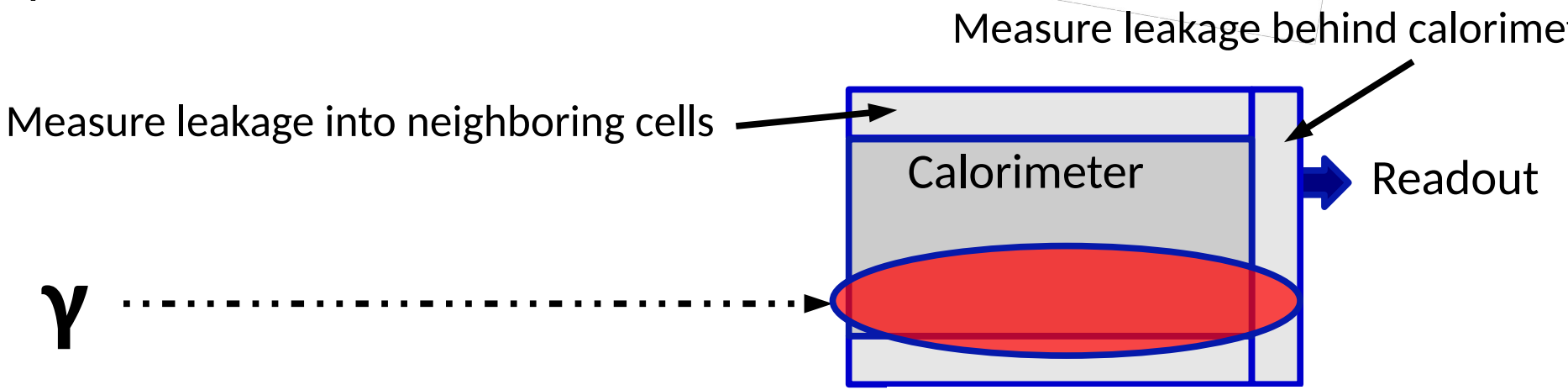


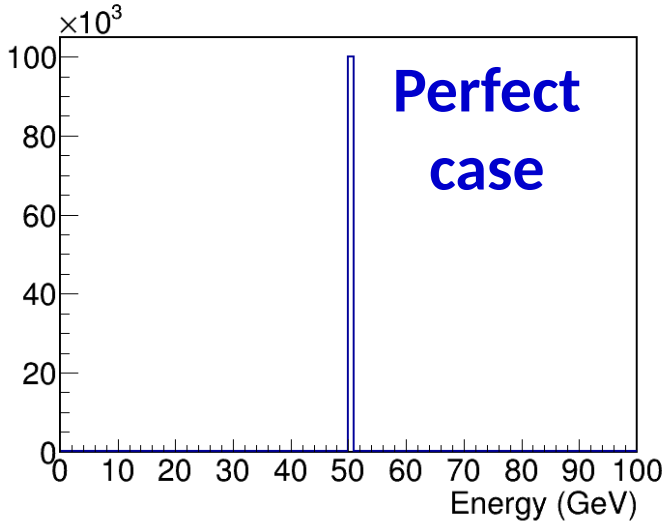


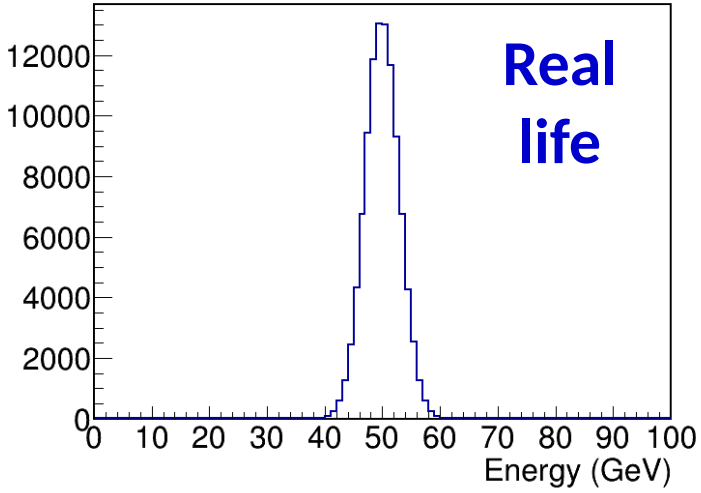
Measurement Errors

Example: measuring the energy of a photon in a calorimeter



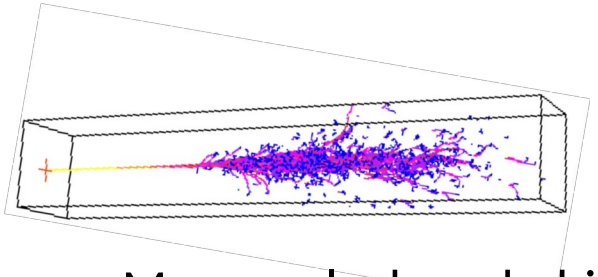


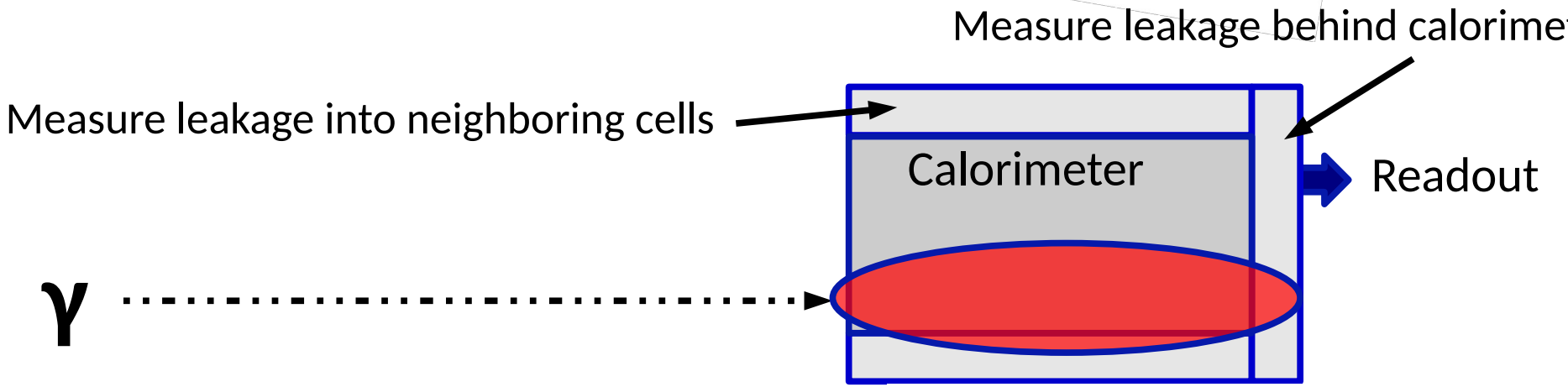


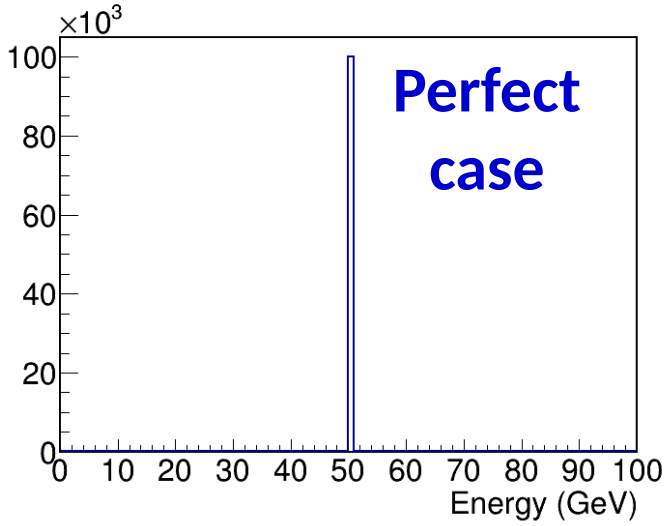


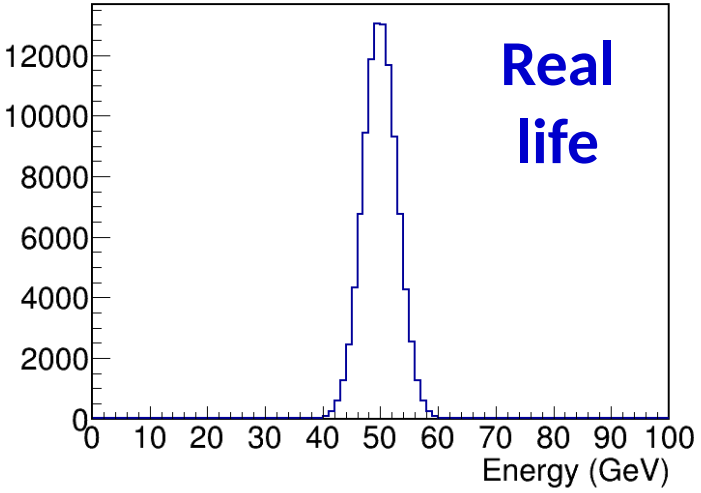
Measurement Errors

Example: measuring the energy of a photon in a calorimeter









Cannot predict the measured value for a given event

⇒ **Random process** ⇒ **Need a probabilistic description**

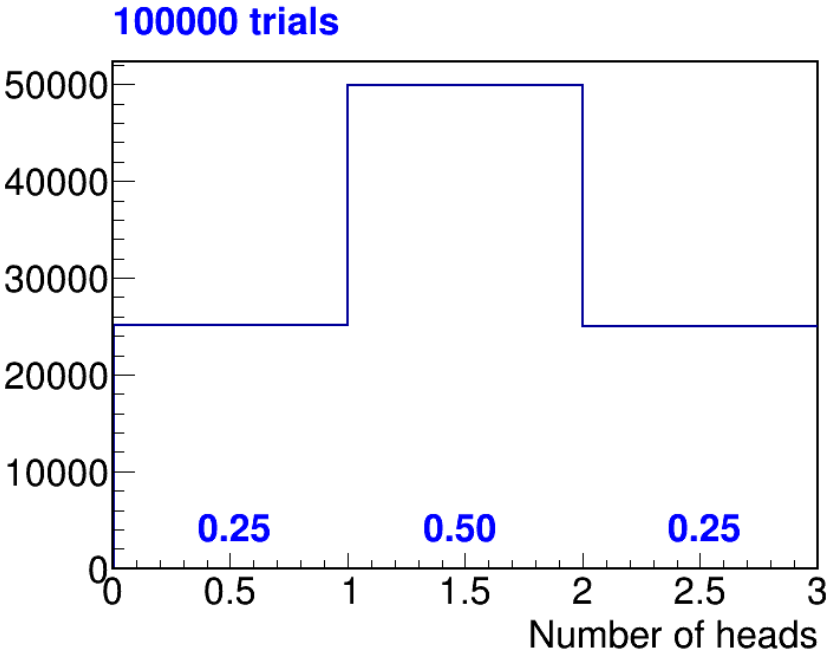
Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

→ Fractions of events in each bin i converge to a limit p_i



Probability distribution :

{ P_i } for i = 0, 1, 2

Properties

- $P_i > 0$
- $\Sigma P_i = 1$

Continuous Variables: PDFs

Continuous variable: can consider **per-bin** probabilities p_i, i=1.. n_{bins}

Bin size → 0 : **Probability distribution function P(x)**

High PDF value

⇒ High chance to get a measurement here

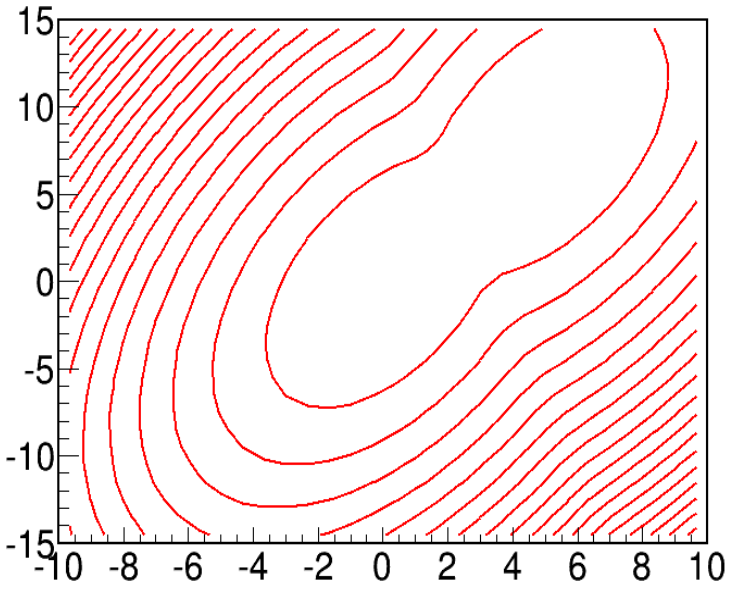
$$P(x) > 0, \quad \int P(x)\, dx = 1$$

x

Generalizes to **multiple variables** :

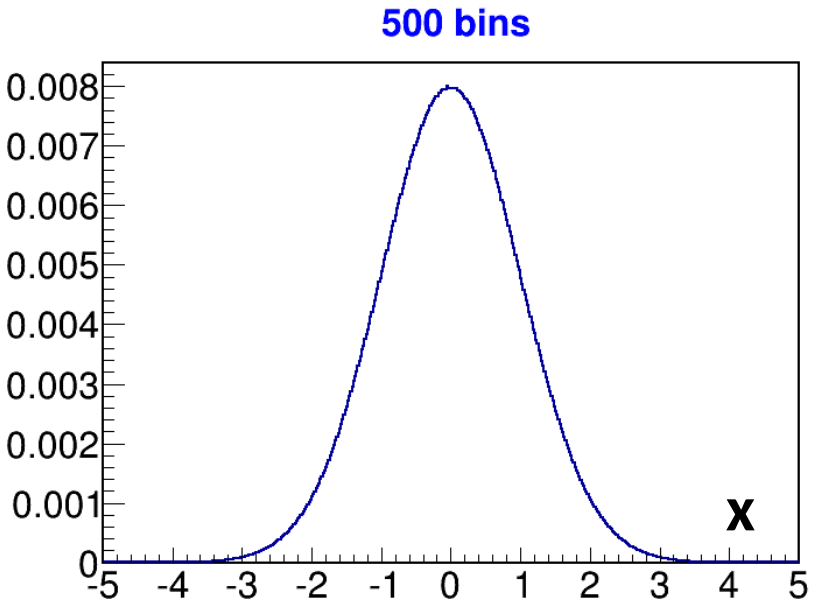
$P(x,y) > 0, \int P(x,y)\, dx\, dy = 1$

Contours: P(x,y)



Continuous Variables: PDFs

Continuous variable: can consider **per-bin** probabilities p_i, i=1.. n_{bins}



Bin size → 0 : **Probability distribution function P(x)**

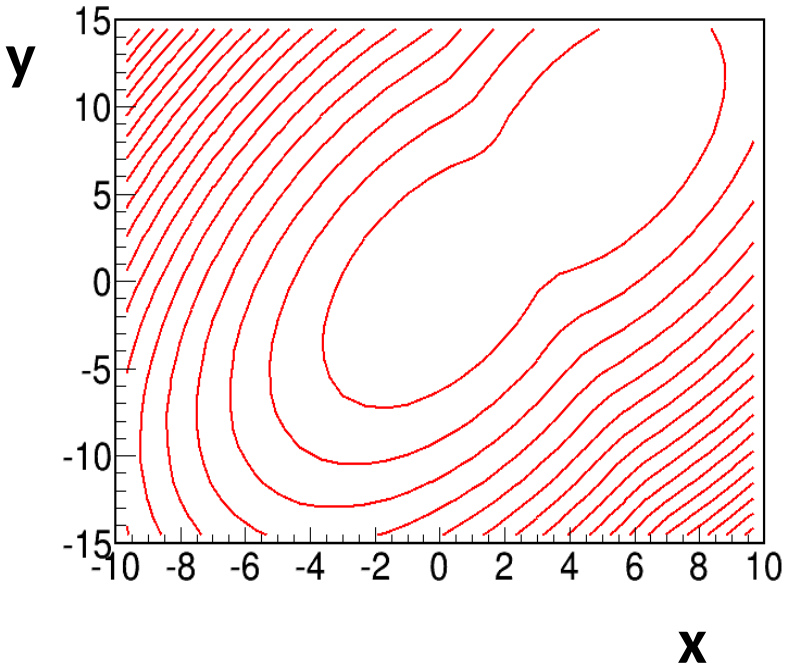
High PDF value

⇒ High chance to get a measurement here

$$P(x) > 0, \quad \int P(x)\, dx = 1$$

Generalizes to **multiple variables** :

$P(x,y) > 0, \int P(x,y)\, dx\, dy = 1$



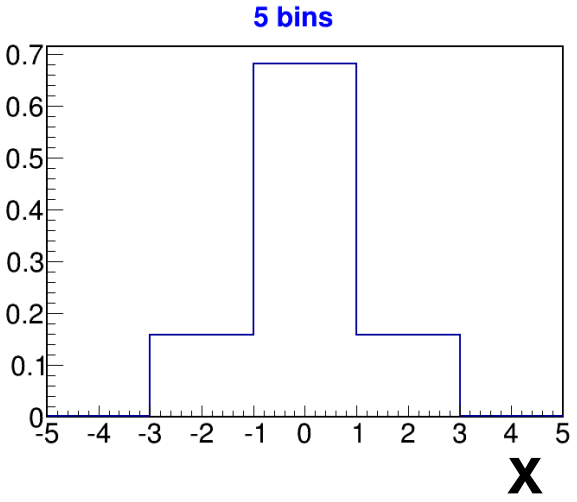
Contours: P(x,y)

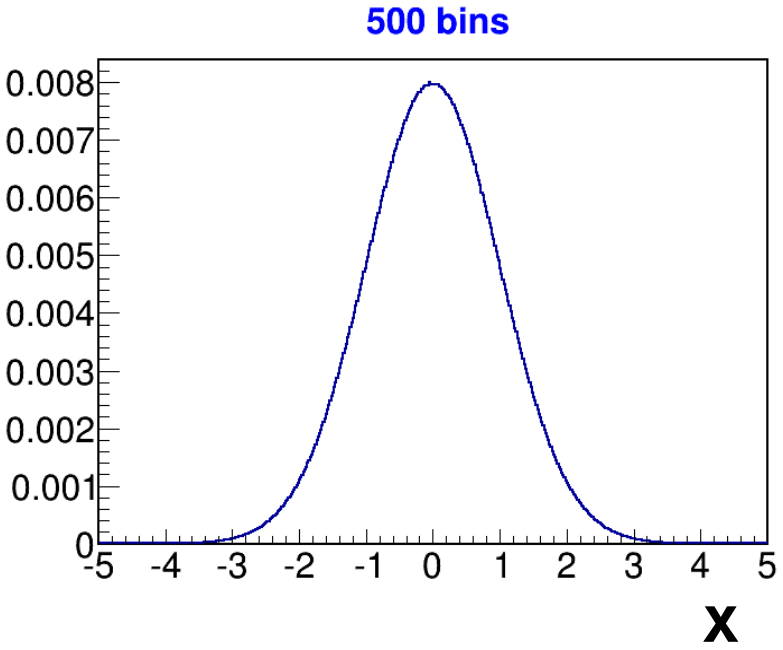
Some vocabulary…

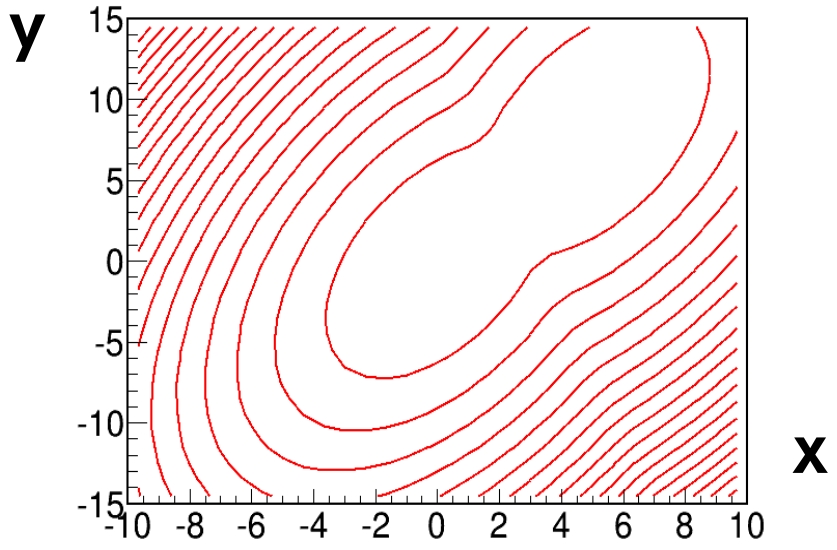
X, Y… are **Random Variables** (continuous or discrete), a.ka. **observables** :
→ X can take any value x, with probability **P(X=x).**

→ P(X=x) (also just P(x)) is the **PDF** of X,
a.k.a. the **Statistical Model.**

→ The **Observed data** is **one value** x_{obs} of X,
drawn from P(X=x).



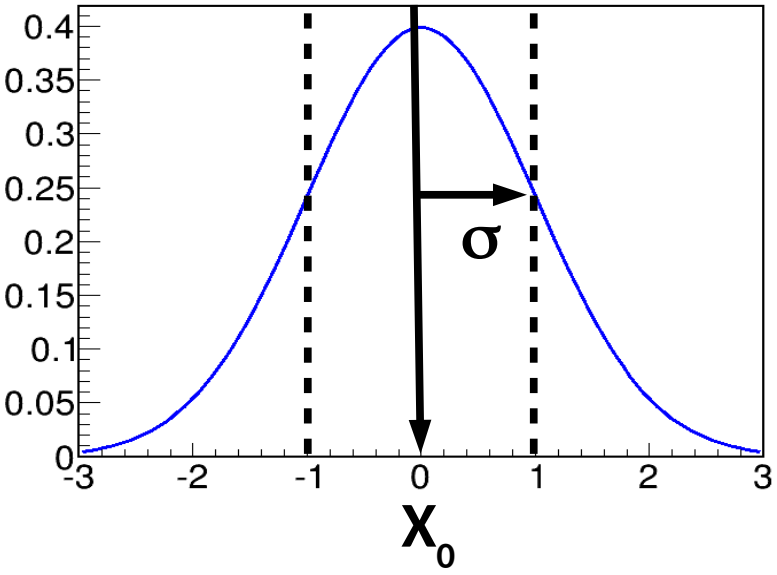




Gaussian PDF

Gaussian distribution:

$$G(x; X_0, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-X_0)^2}{2\sigma^2}}$$



→ **Mean** : X_0

→ **Variance** : σ^2 (⇒ RMS = σ)

Generalize to **N dimensions**:

$$G(x; X_0, C) = \frac{1}{[(2\pi)^N |C|]^{1/2}} e^{-\frac{1}{2}(x-X_0)^T C^{-1} (x-X_0)}$$

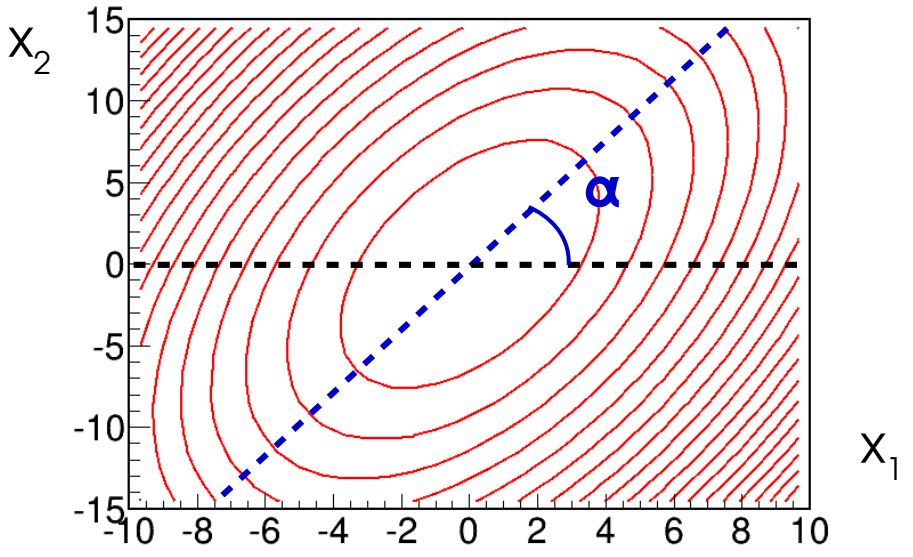
→ **Mean** : X_0

→ **Covariance matrix** :

$$C = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$\tan 2\alpha = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}$$



Central Limit Theorem

(*) Assuming $\sigma_X < \infty$ and other regularity conditions

For an observable X with **any**(*) **distribution**, one has

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \overset{n\to\infty}{\sim} G\left(\langle X\rangle, \frac{\sigma_X}{\sqrt{n}}\right)$$

What this means:

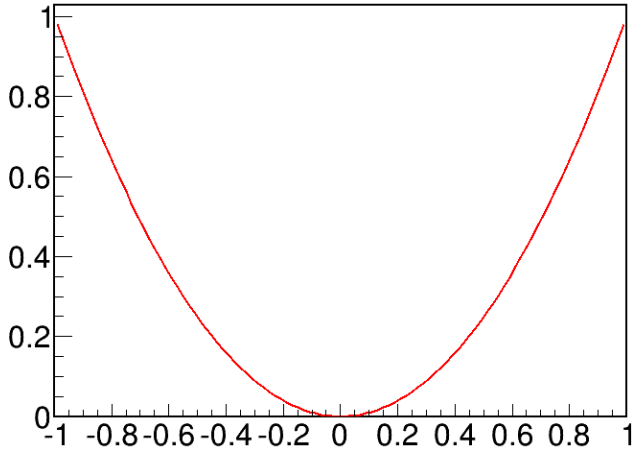
- **The average of many measurements is always Gaussian**, whatever the distribution for a single measurement
- The **mean** of the Gaussian is the **average of the single measurements**
- The **RMS** of the Gaussian **decreases as √n** : smaller fluctuations when averaging over many measurements

Another version: $$\sum_{i=1}^{n} x_i \overset{n\to\infty}{\sim} G\left(n\langle X\rangle, \sqrt{n}\,\sigma_X\right)$$

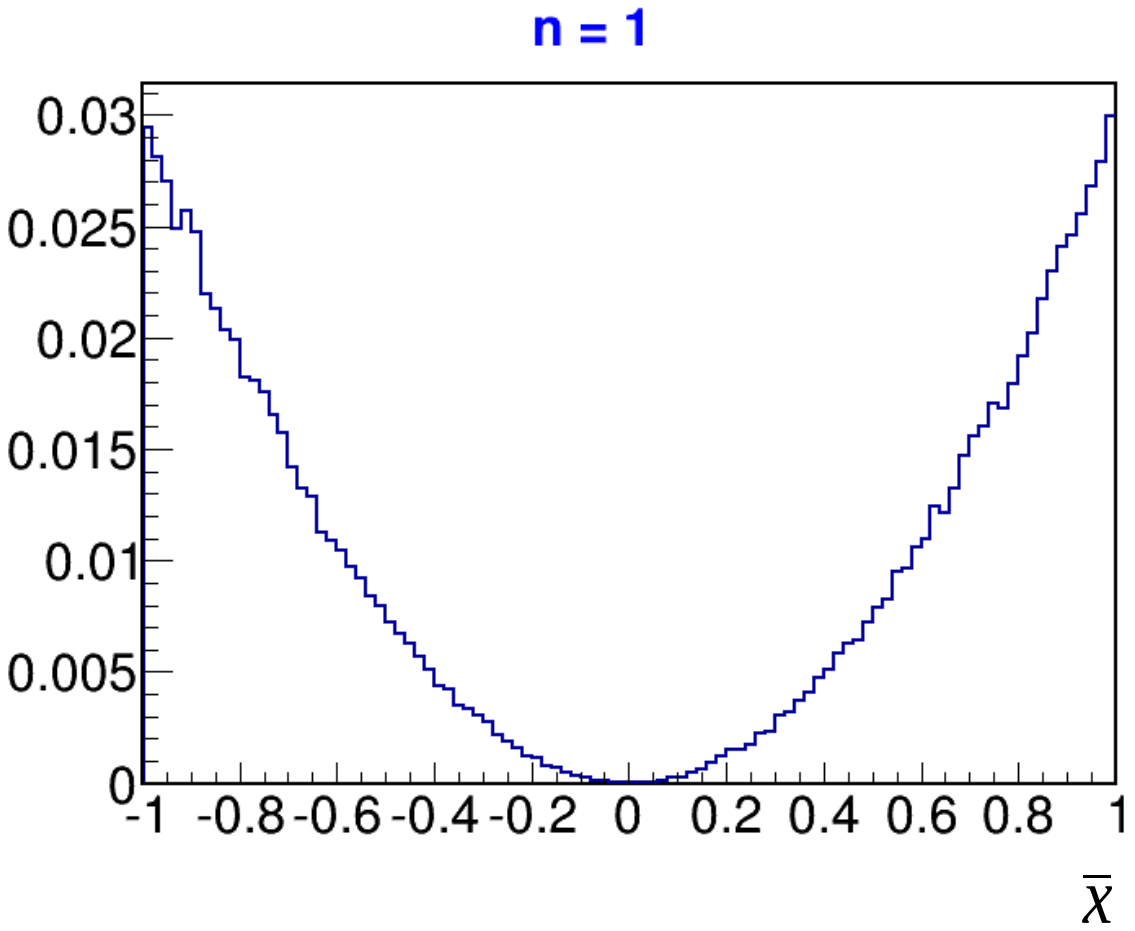
Mean scales like n, but RMS only like √n

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay cos θ*)



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

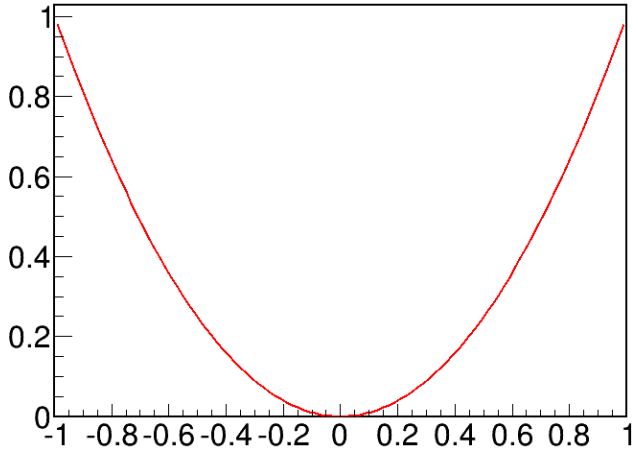


Distribution becomes Gaussian, although very non-Gaussian originally

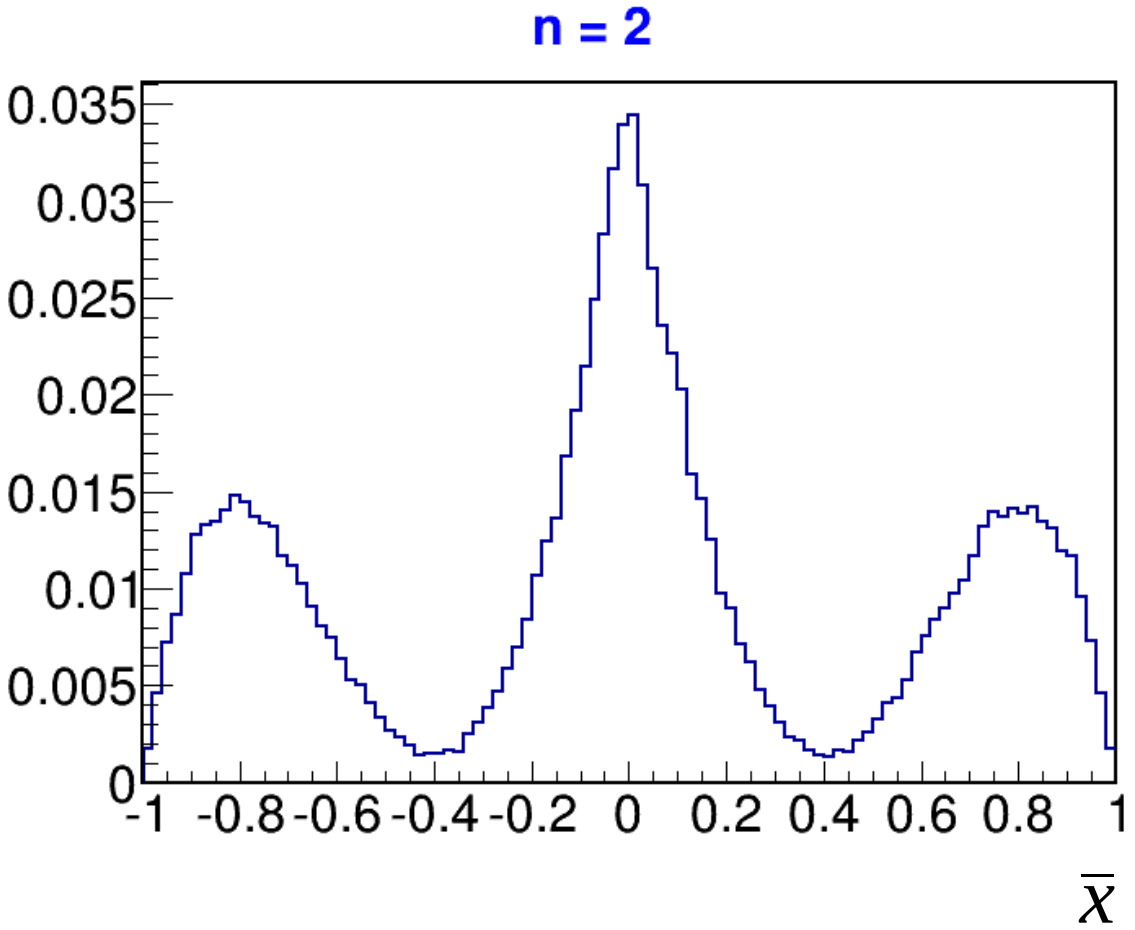
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay cos θ*)



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

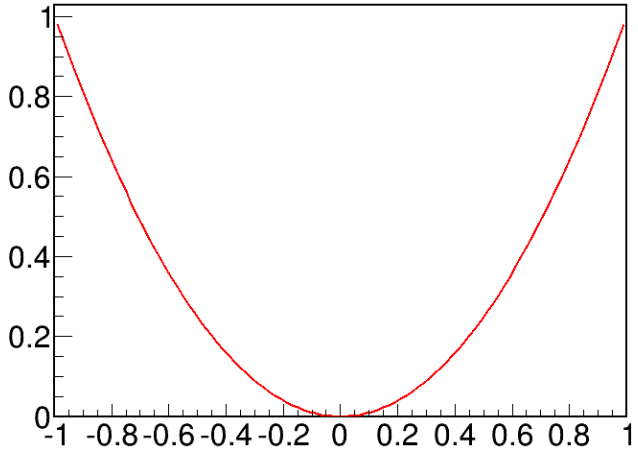


Distribution becomes Gaussian, although very non-Gaussian originally

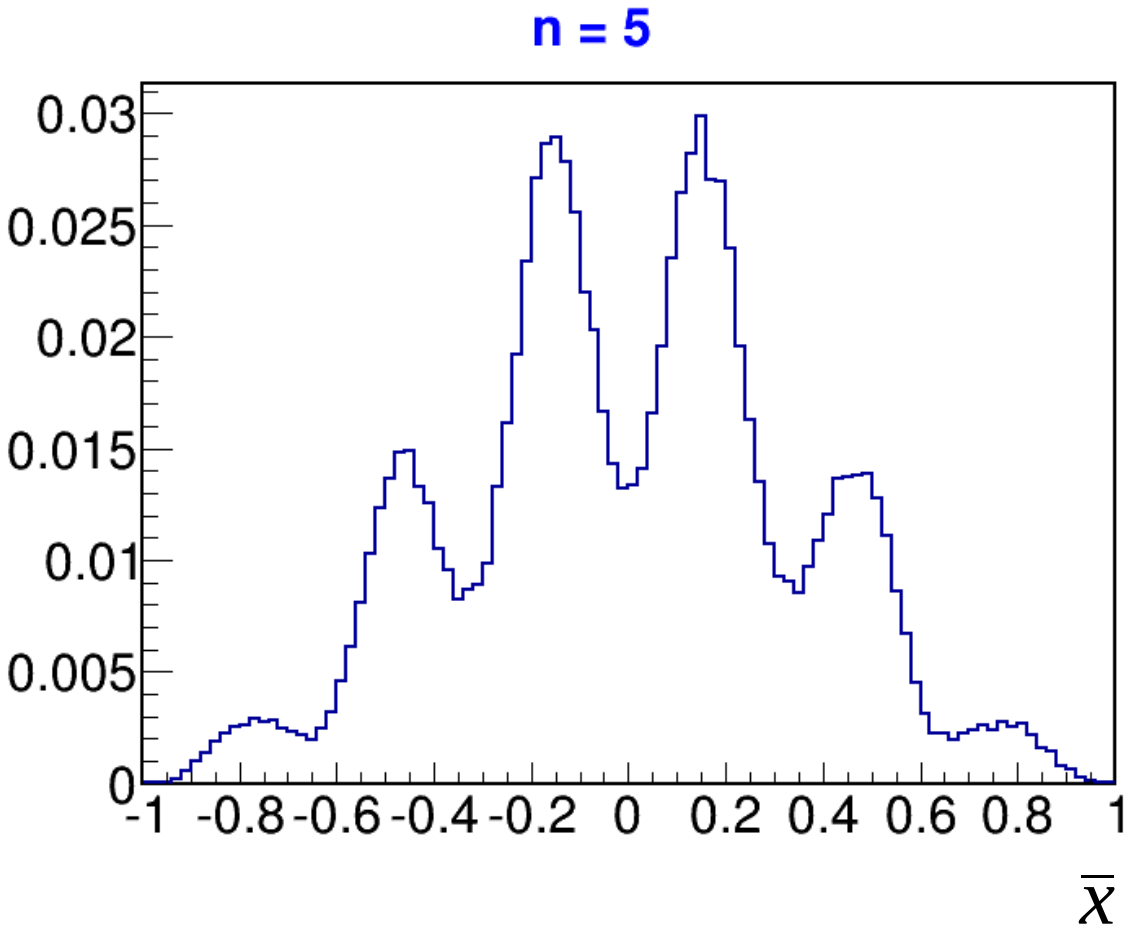
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay cos θ*)



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

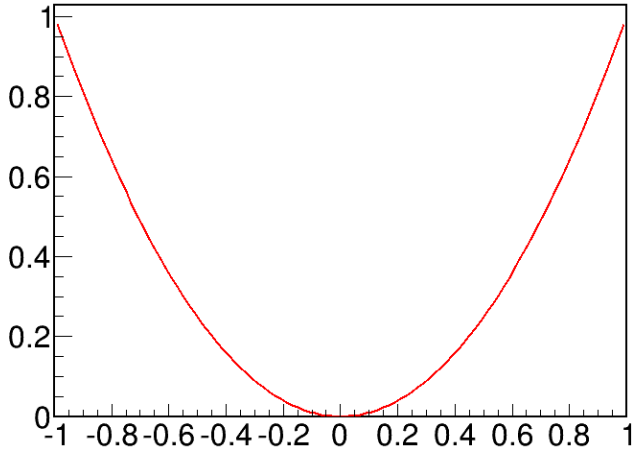


Distribution becomes Gaussian, although very non-Gaussian originally

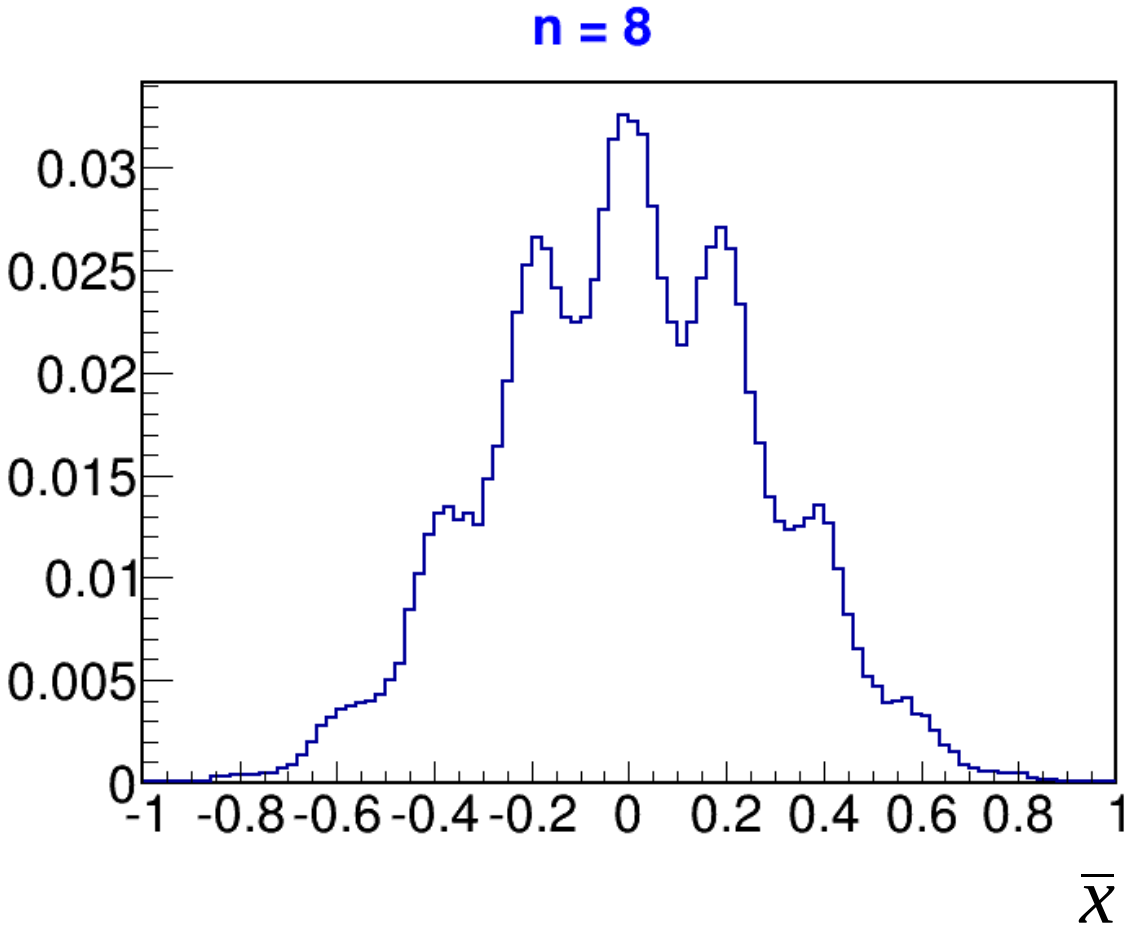
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay cos θ*)



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

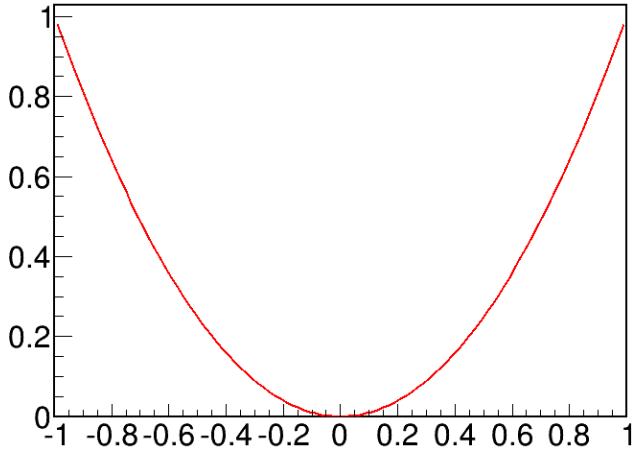


Distribution becomes Gaussian, although very non-Gaussian originally

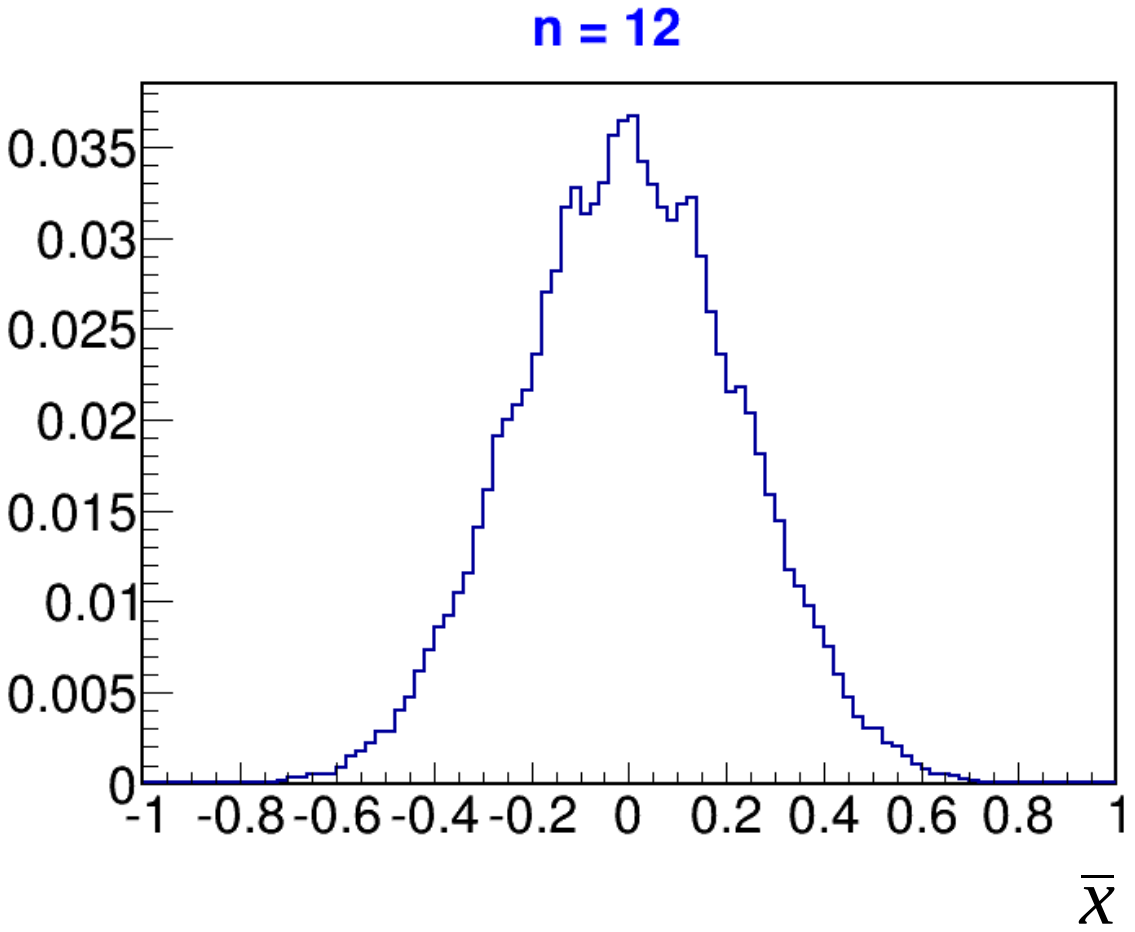
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay cos θ*)



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

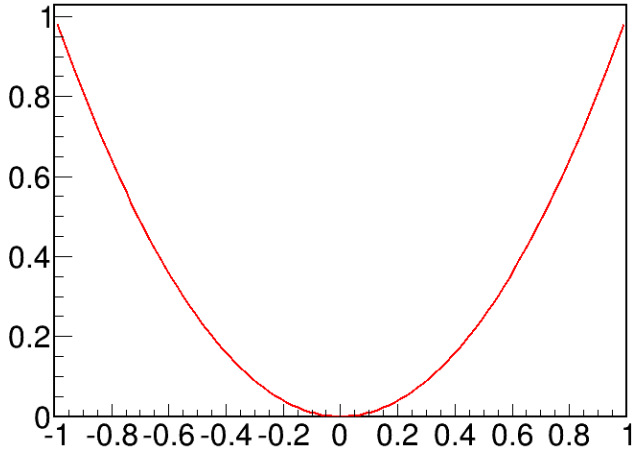


Distribution becomes Gaussian, although very non-Gaussian originally

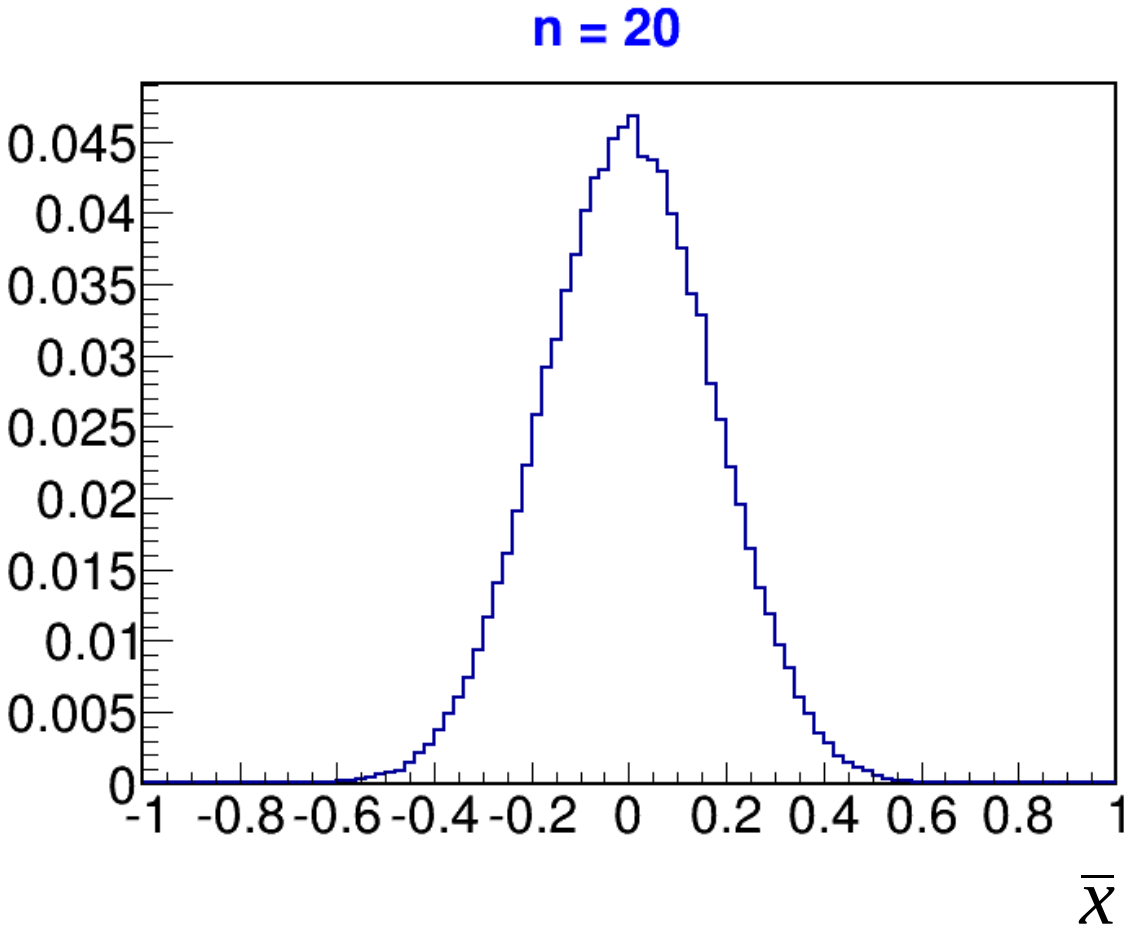
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay cos θ*)



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

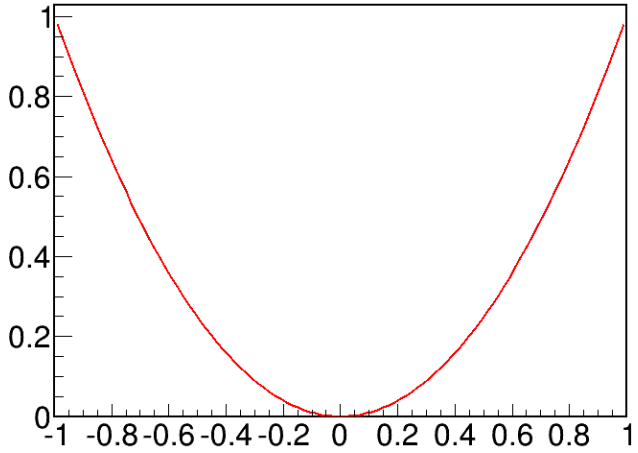


Distribution becomes Gaussian, although very non-Gaussian originally

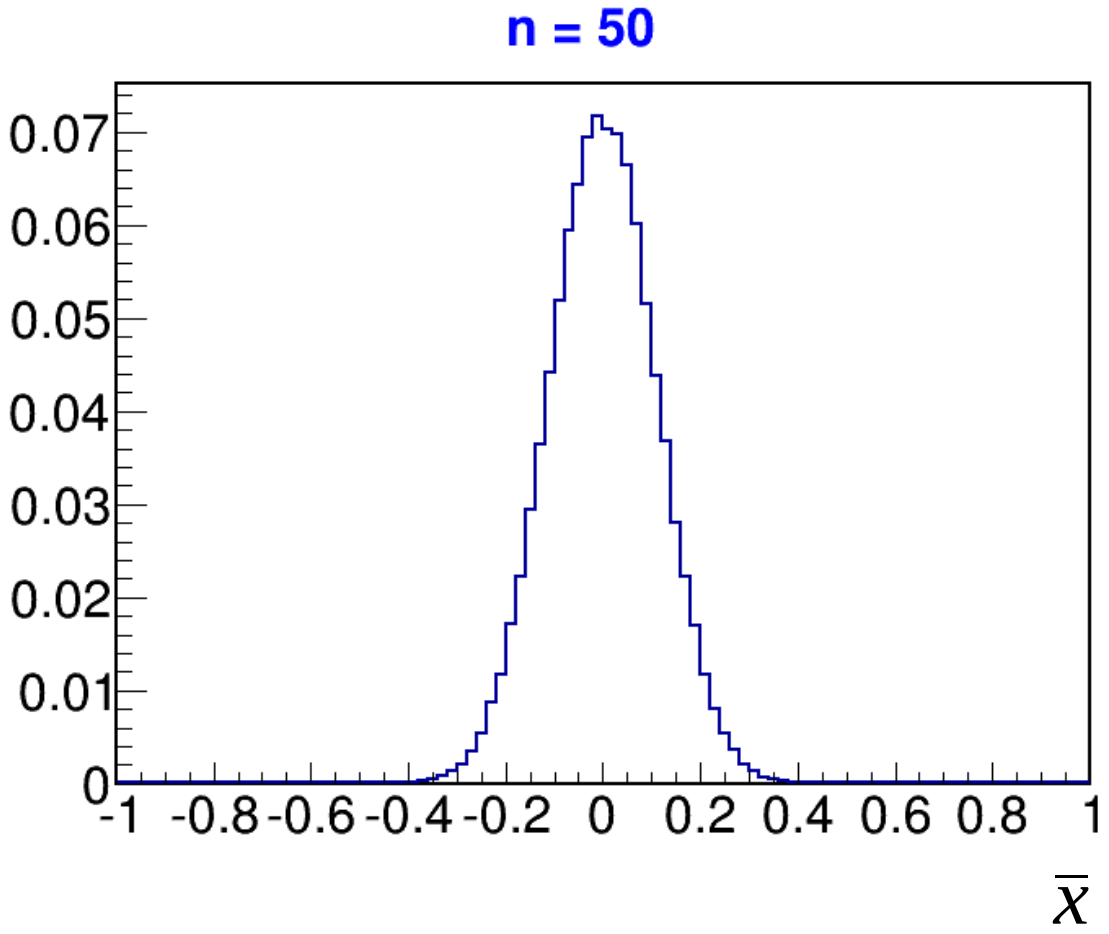
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay cos θ*)



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

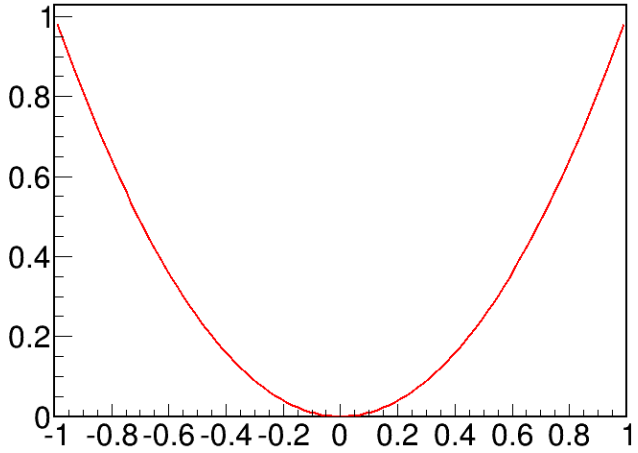


Distribution becomes Gaussian, although very non-Gaussian originally

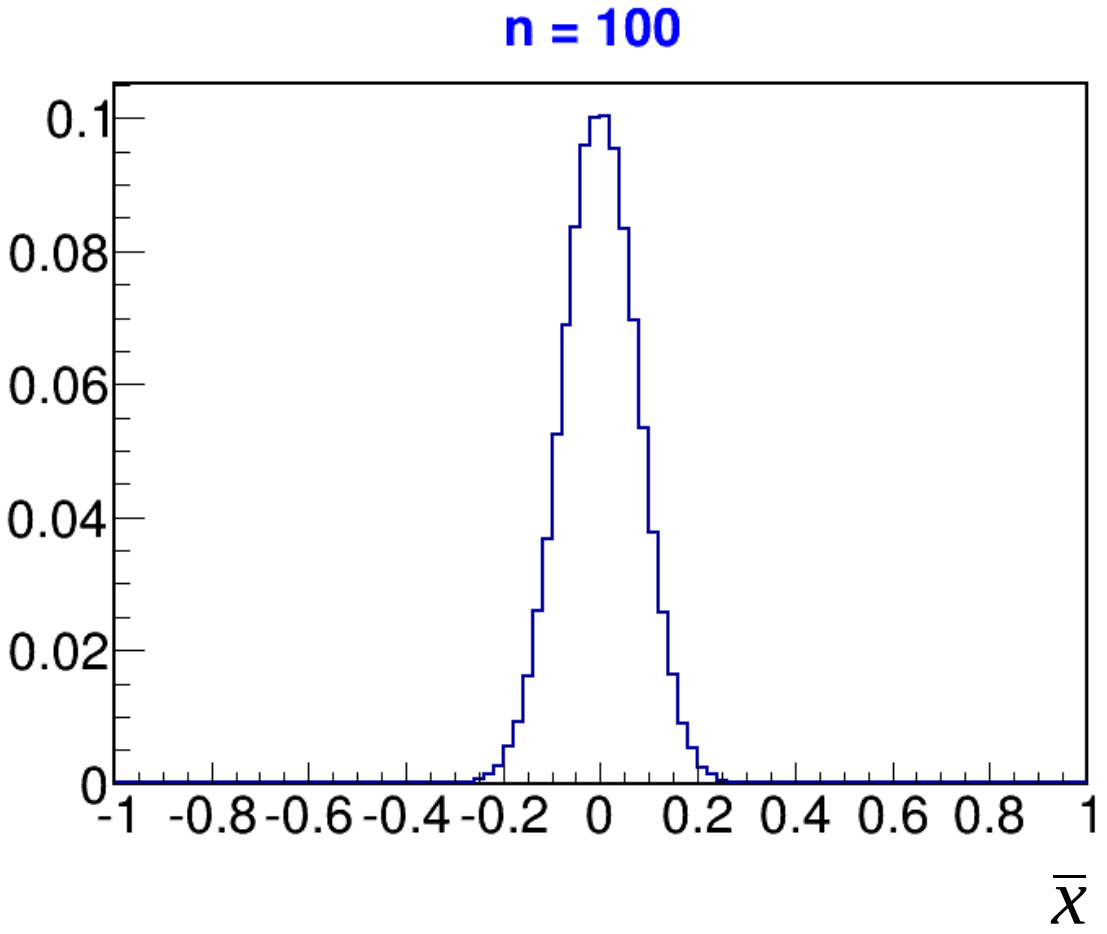
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay cos θ*)



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Distribution becomes Gaussian, although very non-Gaussian originally

Distribution becomes narrower as expected (as 1/√n)

Gaussian Quantiles

Consider $z = \left(\frac{x - X_0}{\sigma} \right)$ "*pull*" of x

For G(x; X_0,σ), we always have **z ~ G(z; 0,1) = N(z)**

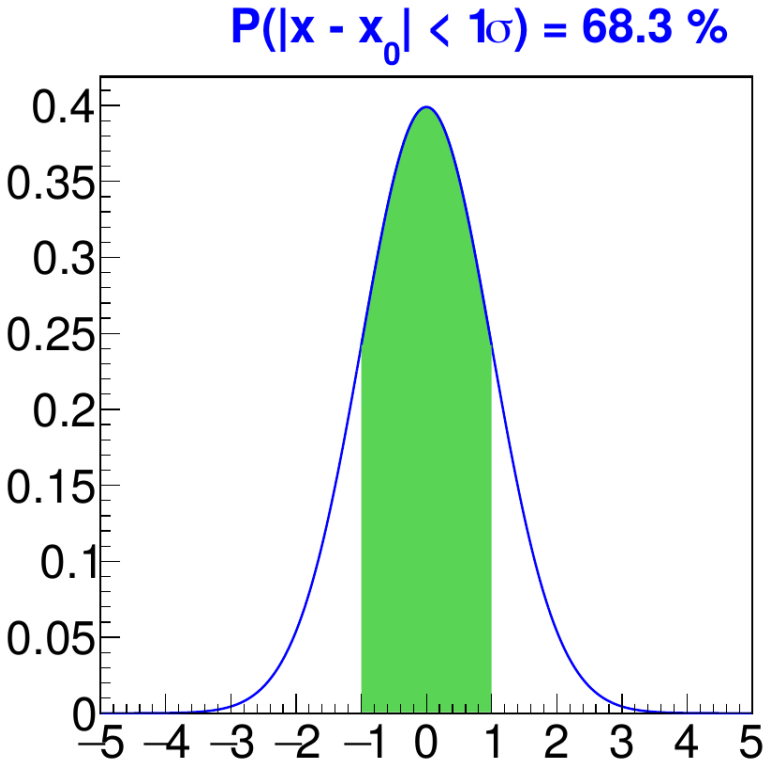
Probability P(|x – X_0| > Zσ) to be away from the mean:

Z	P(|x – X_0| > Zσ)
1	0.317
2	0.046
3	0.003
4	6×10^{-5}
5	6×10^{-7}

Cumulative Distribution Function (CDF)

of the Gaussian :

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



Gaussian Quantiles

Consider $z = \left(\frac{x - X_0}{\sigma} \right)$ "*pull*" of x

For G(x; X_0,σ), we always have **z ~ G(z; 0,1) = N(z)**

Probability P(|x – X_0| > Zσ) to be away from the mean:

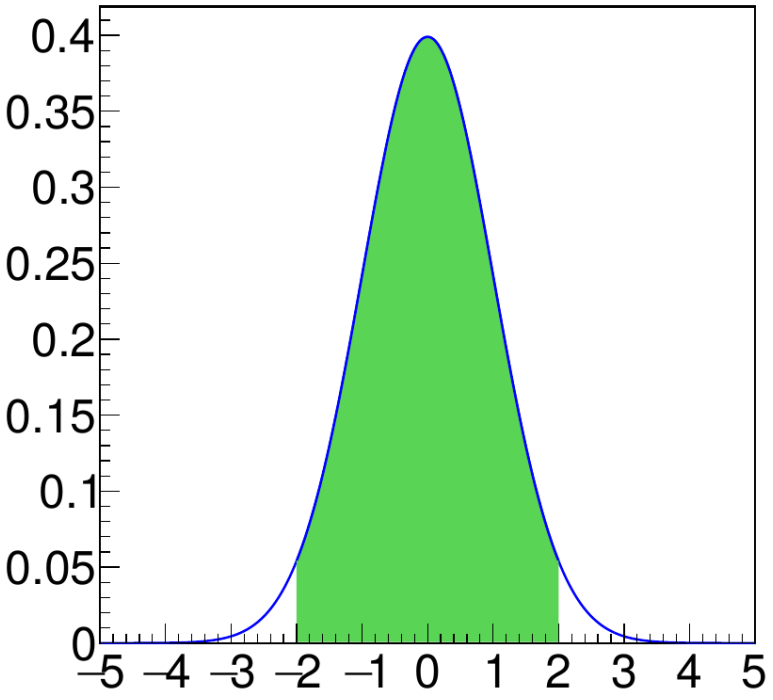
Z	P(|x – X_0| > Zσ)
1	0.317
2	0.046
3	0.003
4	6×10^{-5}
5	6×10^{-7}

Cumulative Distribution Function (CDF)

of the Gaussian :

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

P(|x - x_0| < 2σ) = 95.4 %



Gaussian Quantiles

Consider $z = \left| \frac{x - X_0}{\sigma} \right|$ "*pull*" of x

For G(x; X_0,σ), we always have **z ~ G(z; 0,1) = N(z)**

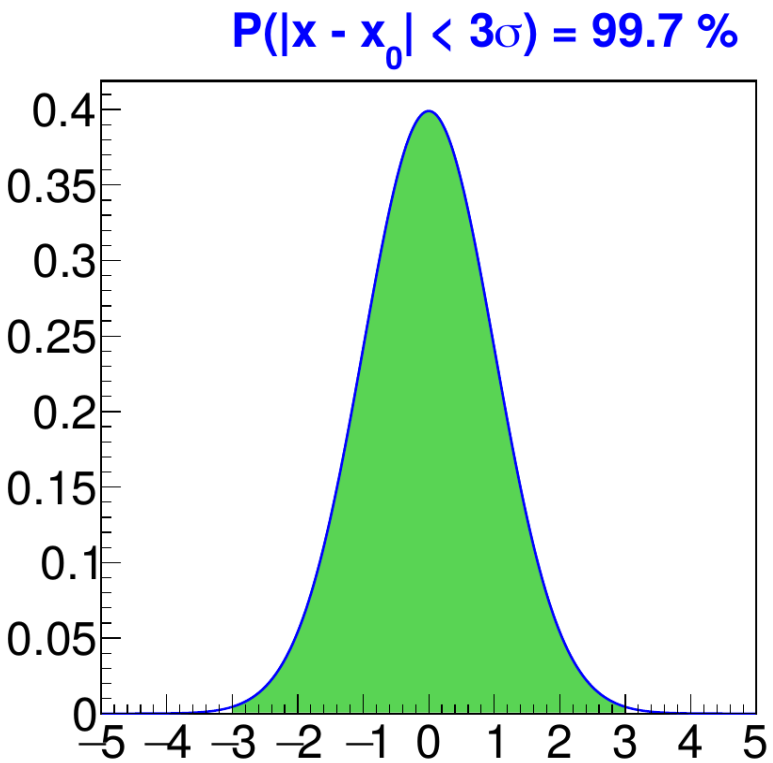
Probability P(|x – X_0| > Zσ) to be away from the mean:

Z	P(|x – X_0| > Zσ)
1	0.317
2	0.046
3	0.003
4	6×10^{-5}
5	6×10^{-7}

Cumulative Distribution Function (CDF)

of the Gaussian :

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



Chi-squared

Multiple Independent Gaussian variables x_i: Define

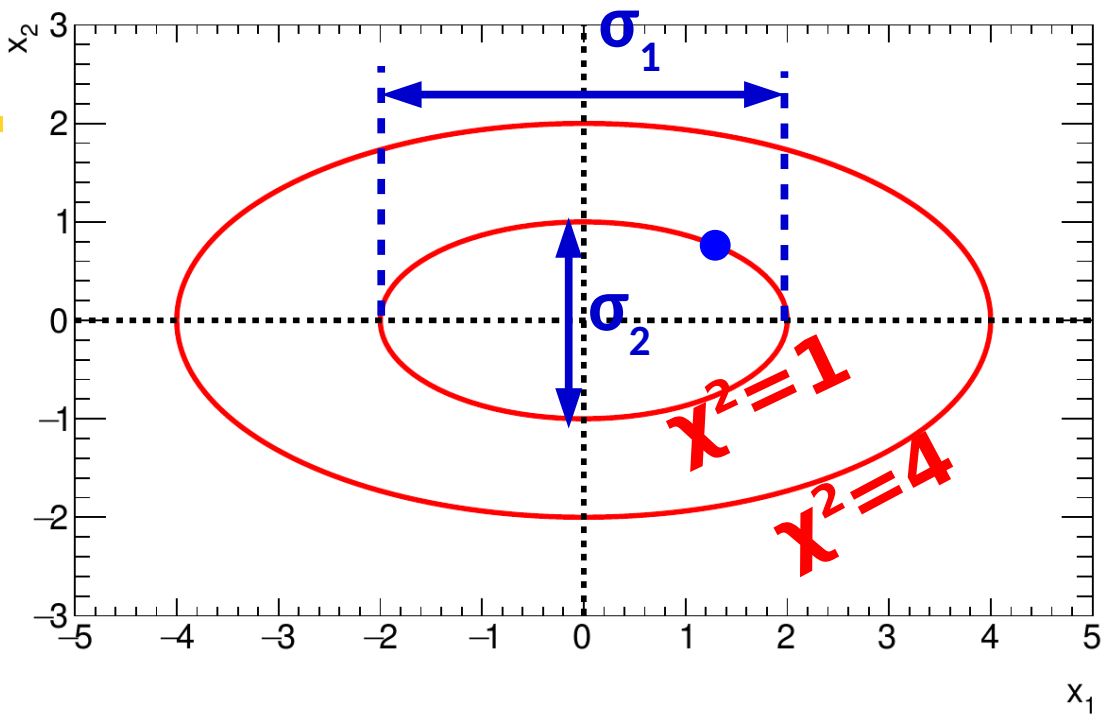
$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - x_i^0}{\sigma_i} \right)^2$$

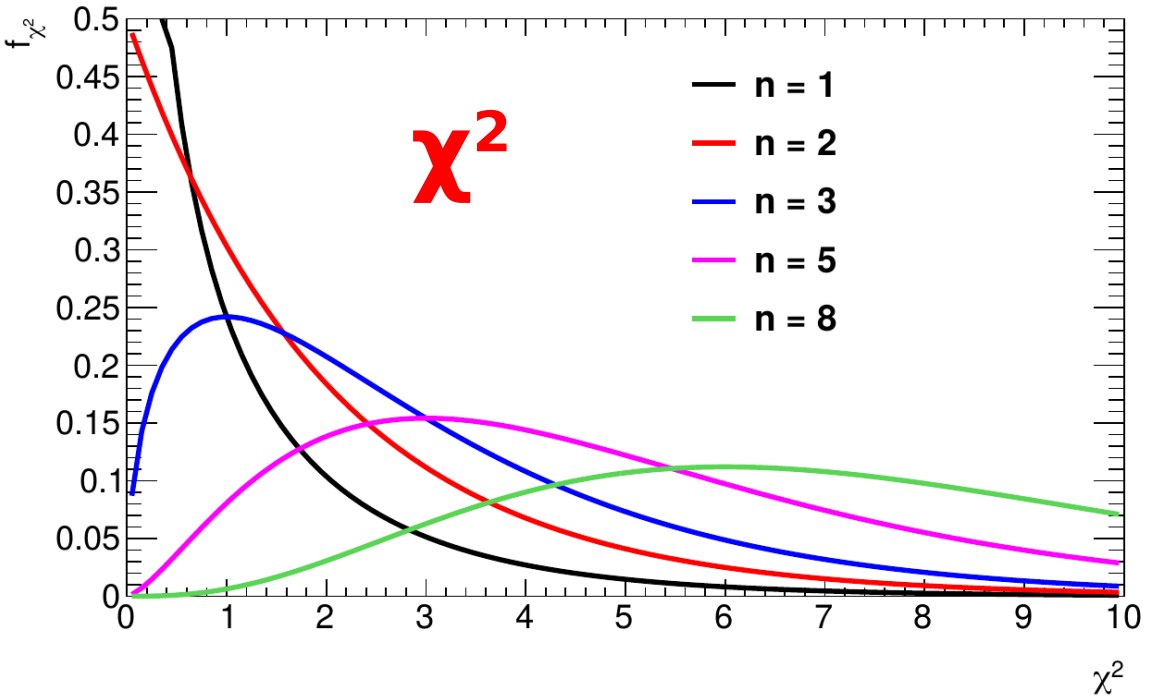
Measures global distance from reference point (x_1^0 .... x_n^0)

Distribution depends on n :

Rule of thumb:

$$\chi^2/n \text{ should be } \lesssim 1$$





Chi-squared

Multiple Independent Gaussian variables x_i: Define

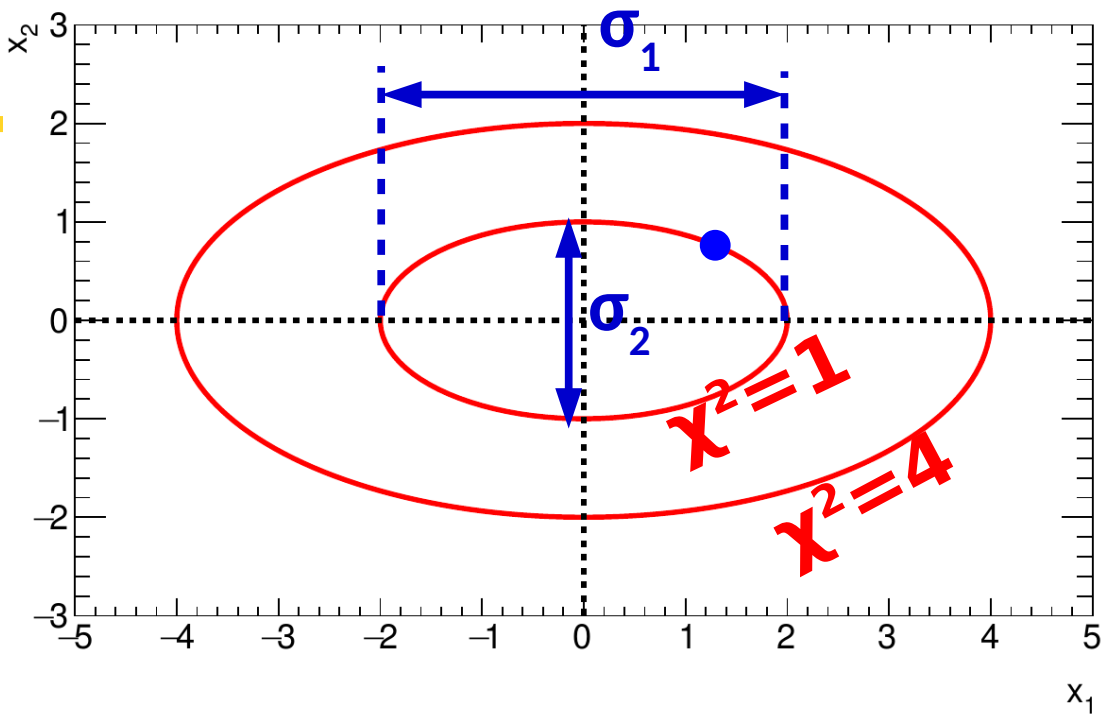
$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - x_i^0}{\sigma_i} \right)^2$$

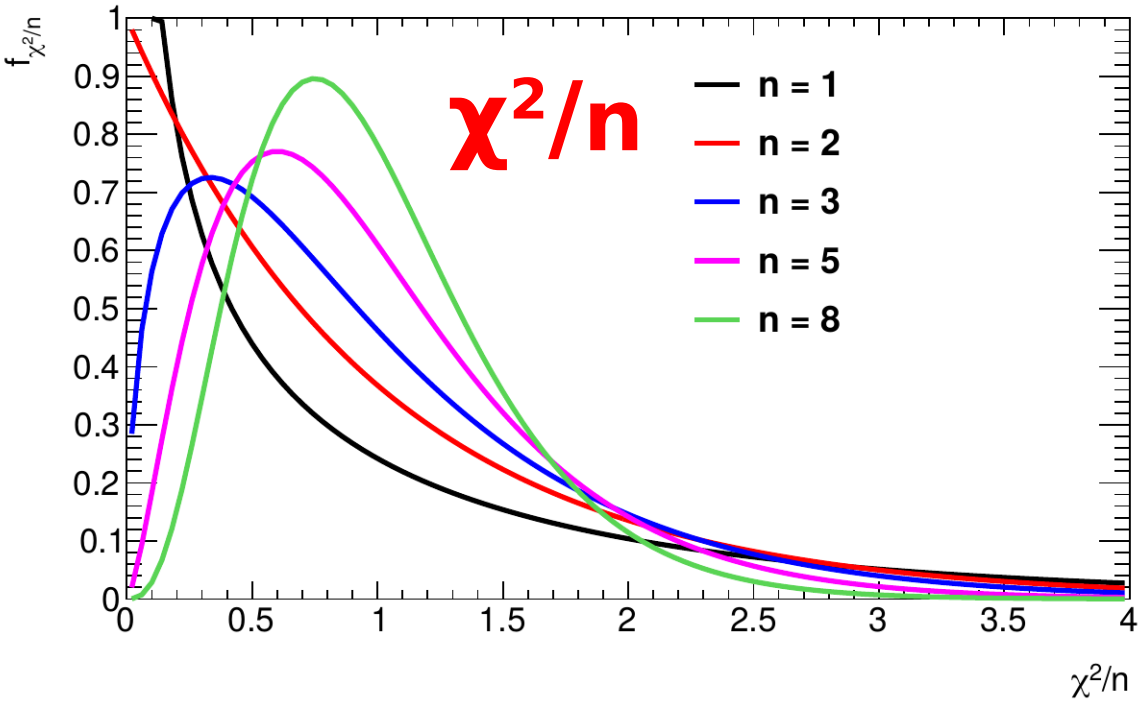
Measures global distance from reference point (x_1^0 .... x_n^0)

Distribution depends on n :

Rule of thumb:

χ^2/n **should be** $\lesssim 1$

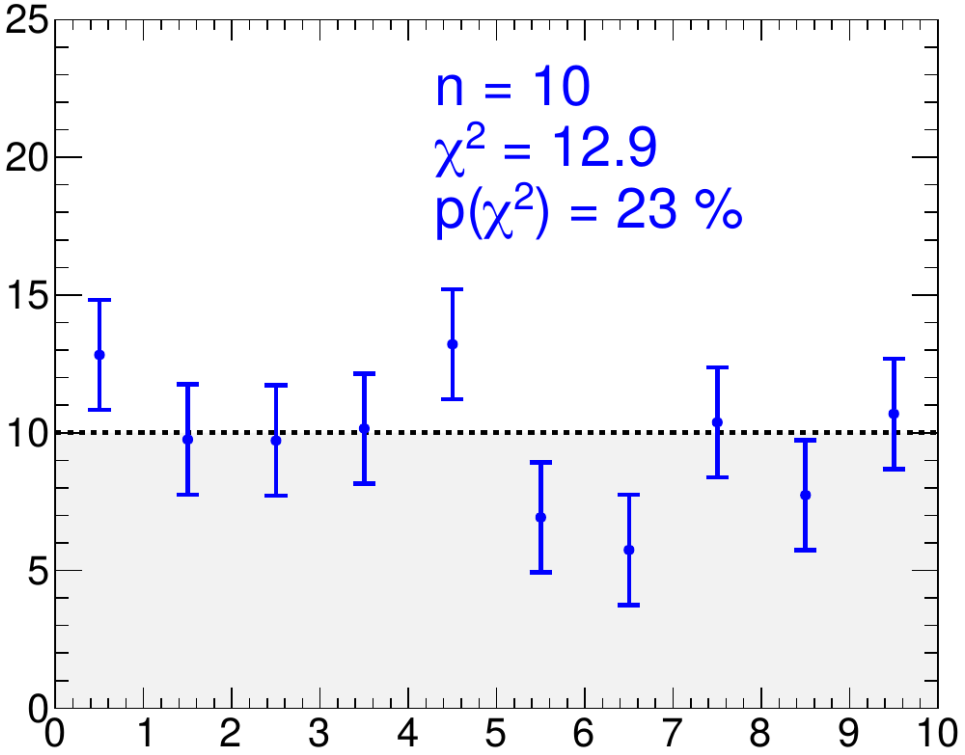




Histogram Chi-squared

Histogram χ^2 with respect to a reference shape:

- Assume an independent Gaussian distribution in each bin
- Degrees of freedom = (number of bins) – (number of fit parameters)



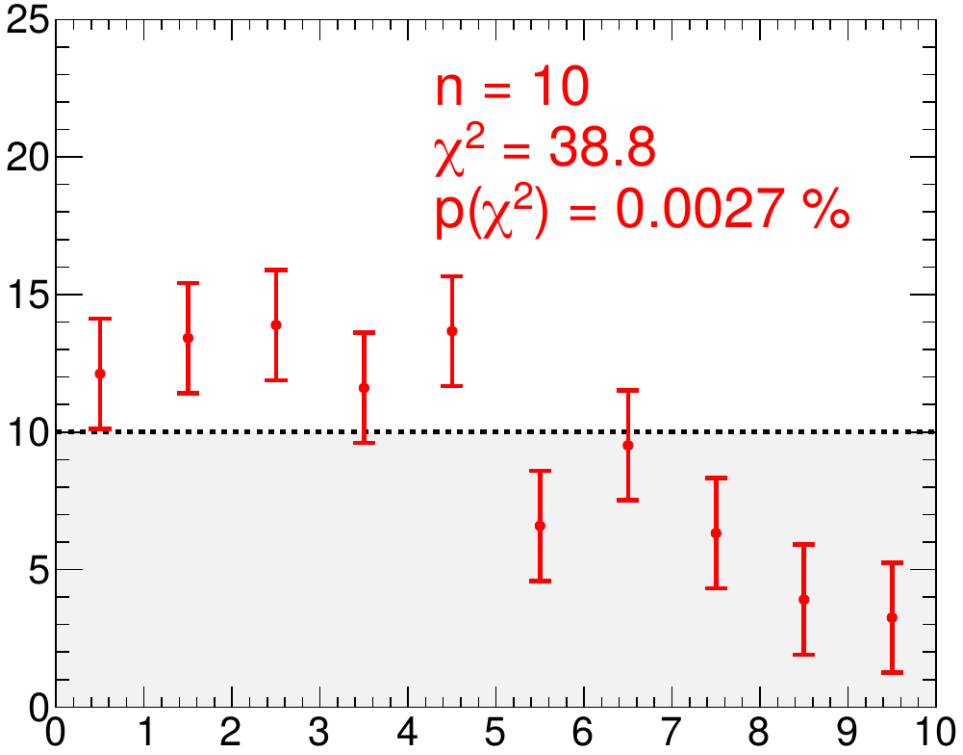
BLUE histogram vs. flat reference

$\chi^2 = 12.9$, $p(\chi^2=12.9, n=10) = 23\%$ ✔

Histogram Chi-squared

Histogram χ^2 with respect to a reference shape:

- Assume an independent Gaussian distribution in each bin
- Degrees of freedom = (number of bins) – (number of fit parameters)



BLUE histogram vs. flat reference

$\chi^2 = 12.9$, $p(\chi^2=12.9, n=10) = 23\%$ ✔

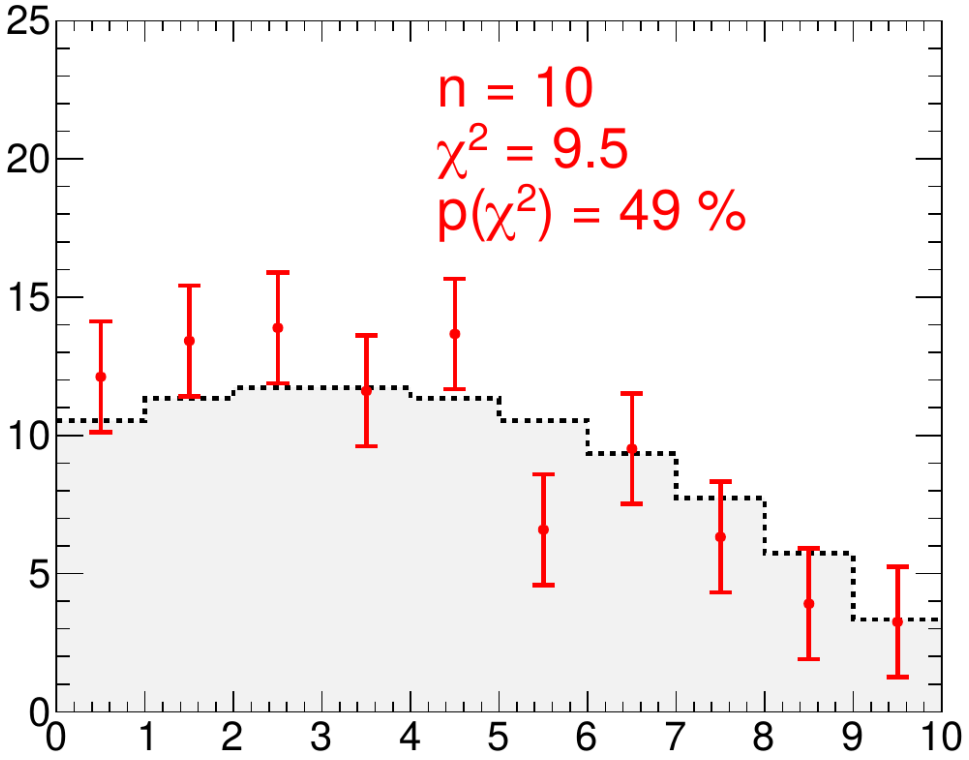
RED histogram vs. flat reference

$\chi^2 = 38.8$, $p(\chi^2=38.8, n=10) = 0.003\%$ ✘

Histogram Chi-squared

Histogram χ^2 with respect to a reference shape:

- Assume an independent Gaussian distribution in each bin
- Degrees of freedom = (number of bins) – (number of fit parameters)



BLUE histogram vs. flat reference

$\chi^2 = 12.9$, $p(\chi^2=12.9, n=10) = 23\%$ ✔

RED histogram vs. flat reference

$\chi^2 = 38.8$, $p(\chi^2=38.8, n=10) = 0.003\%$ ✘

RED histogram vs. correct reference

$\chi2 = 9.5$, $p(\chi2=9.5, n=10) = 49\%$ ✔

PDF Properties: Mean

E(X) = <X> : **Mean** of X – expected outcome **on average over many measurements**

$$\langle X \rangle = \sum_i x_i \, P_i \qquad \text{or}$$

$$\langle X \rangle = \int x \, P(x) \, dx$$

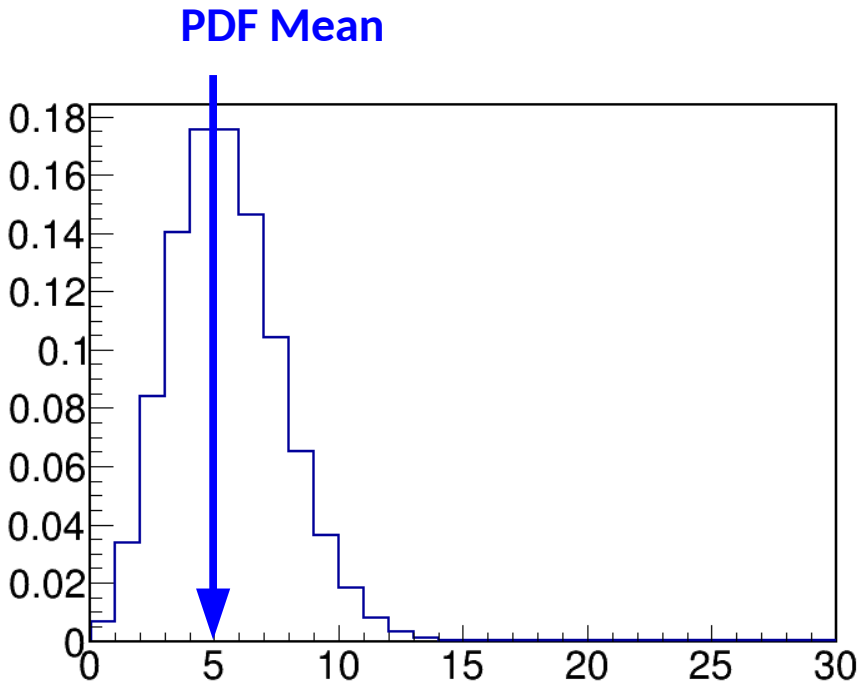
→ Property of the **PDF**

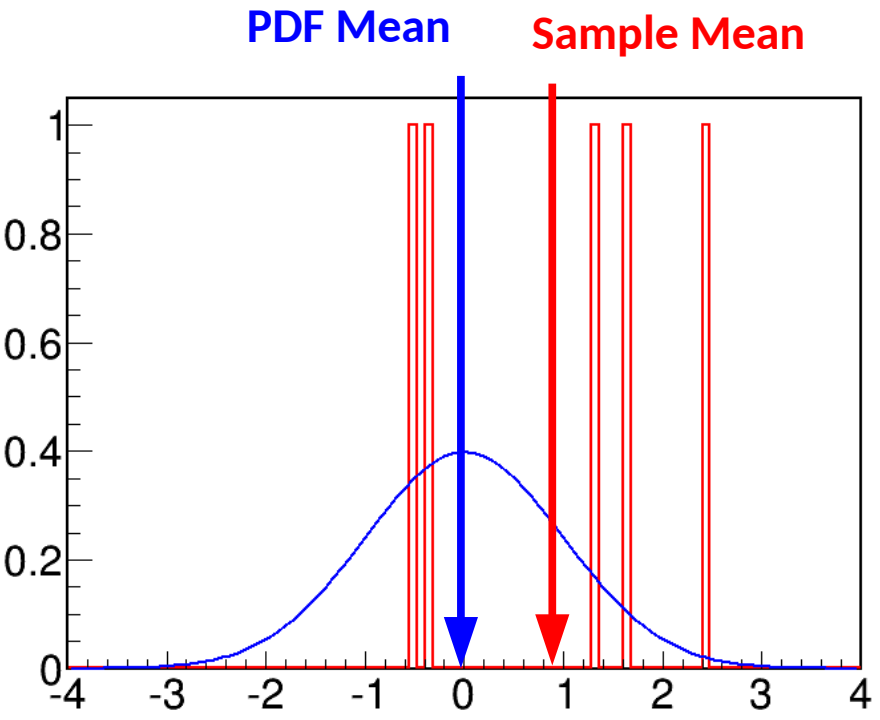
For measurements $x_1 \ldots x_n$,
then can compute the **Sample mean**:

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

→ Property of the **sample**

→ approximates the PDF mean.





PDF Properties: (Co)variance

Variance of X:

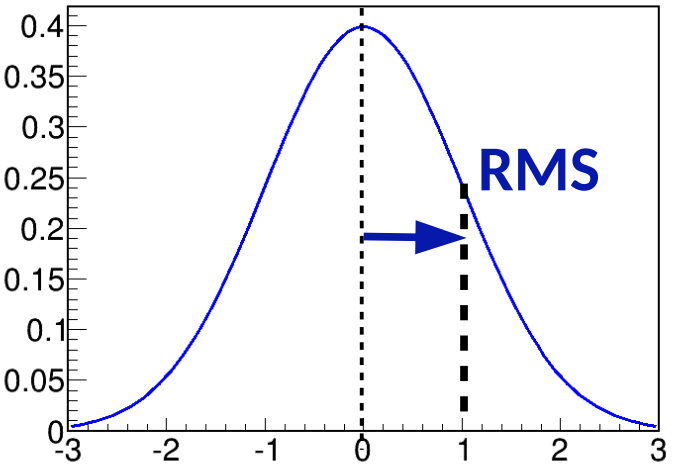
$$\mathrm{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$$

→ Average square of deviation from mean

→ **RMS(X) = √Var(X) = σ_x** ***standard deviation***

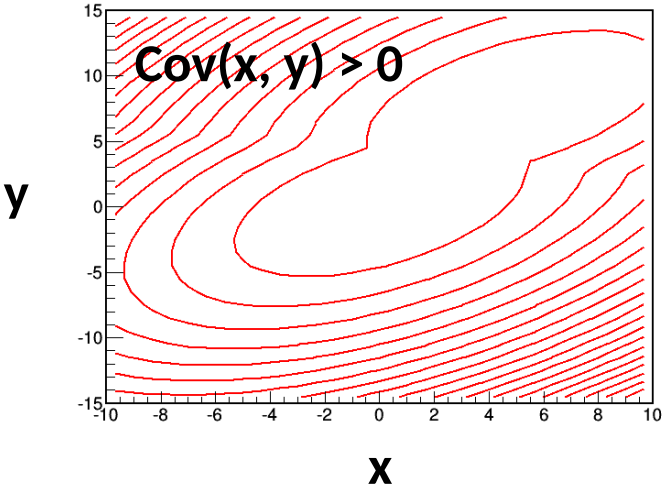
Can be approximated by **sample variance**:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$



Covariance of X and Y:

$$\mathrm{Cov}(X, Y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$$



→ Large if variations of X and Y are "synchronized"

Correlation coefficient $\rho = \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$ $-1 \le \rho \le 1$

PDF Properties: (Co)variance

Variance of X:

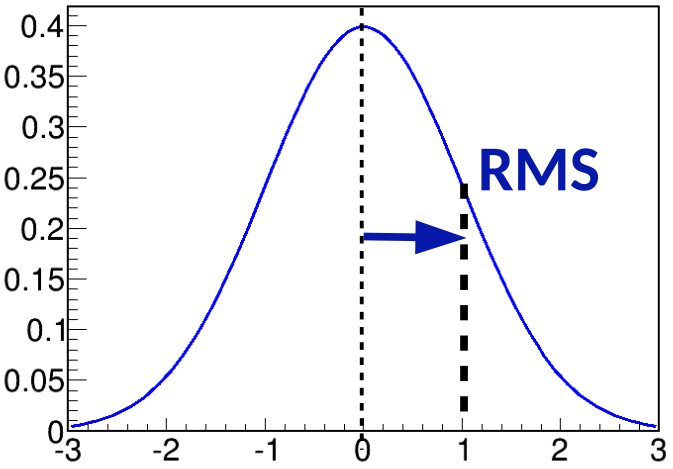
$$\mathrm{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$$

→ Average square of deviation from mean

→ **RMS(X) = √Var(X) = σ_x** ***standard deviation***

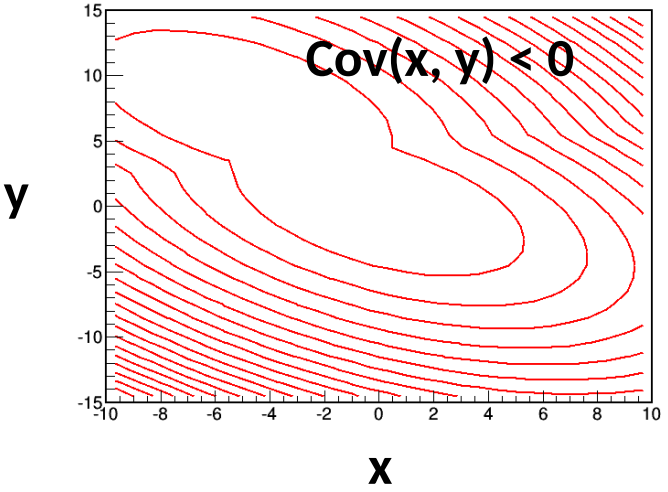
Can be approximated by **sample variance**:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$



Covariance of X and Y:

$$\mathrm{Cov}(X, Y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$$



→ Large if variations of X and Y are "synchronized"

Correlation coefficient $\rho = \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$ $-1 \le \rho \le 1$

PDF Properties: (Co)variance

Variance of X:

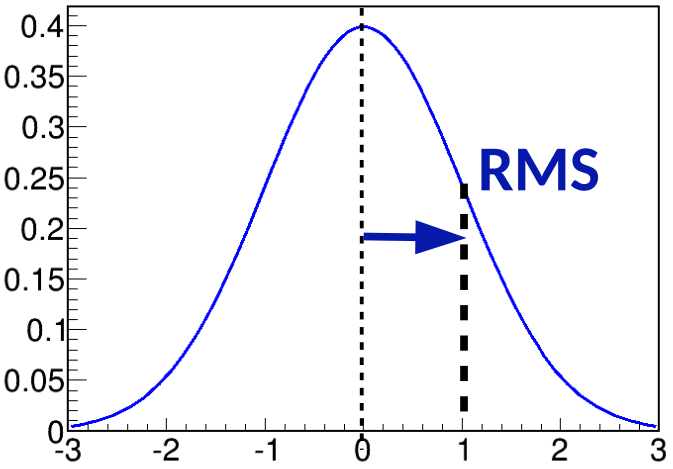
$$\mathbf{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$$

→ Average square of deviation from mean

→ **RMS(X) = √Var(X) = σ_x** ***standard deviation***

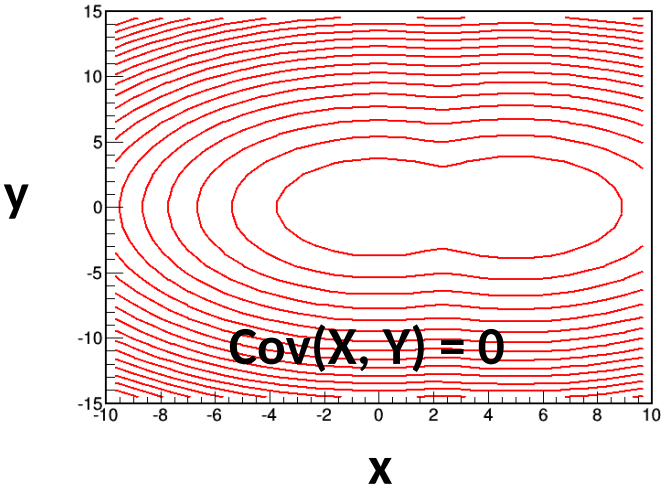
Can be approximated by **sample variance**:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$



Covariance of X and Y:

$$\mathbf{Cov}(X, Y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$$



→ Large if variations of X and Y are "synchronized"

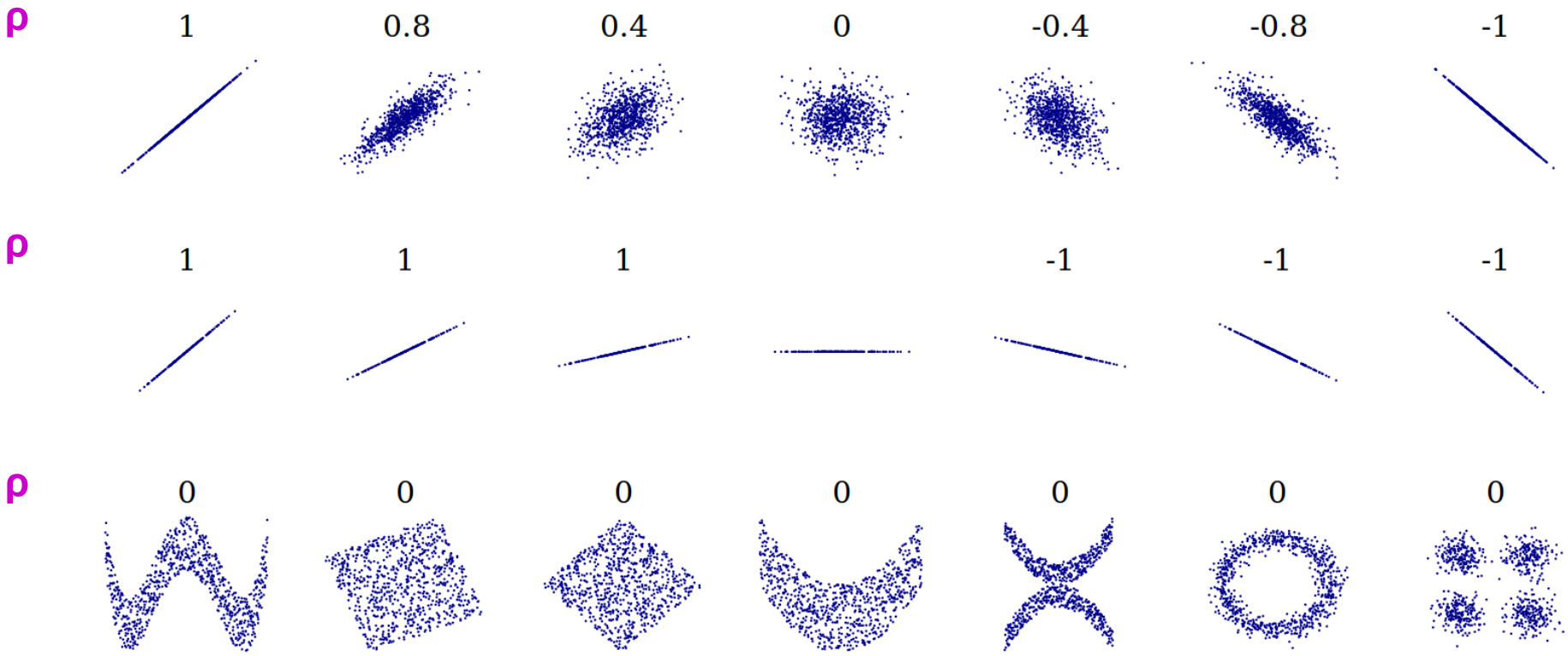
Correlation coefficient

$$\rho = \frac{\mathbf{Cov}(X, Y)}{\sqrt{\mathbf{Var}(X)\,\mathbf{Var}(Y)}}$$

$-1 \le \rho \le 1$

“Linear” vs. “non-linear” correlations

For non-Gaussian cases, the **Correlation coefficient ρ** is not the whole story:



$$\tan 2\alpha = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}$$

Source: Wikipedia

In particular, variables can still be correlated even when ρ=0 : *“Non-linear”* correlations.

Error Bars

Strictly speaking, ***the uncertainty is given by the model :***

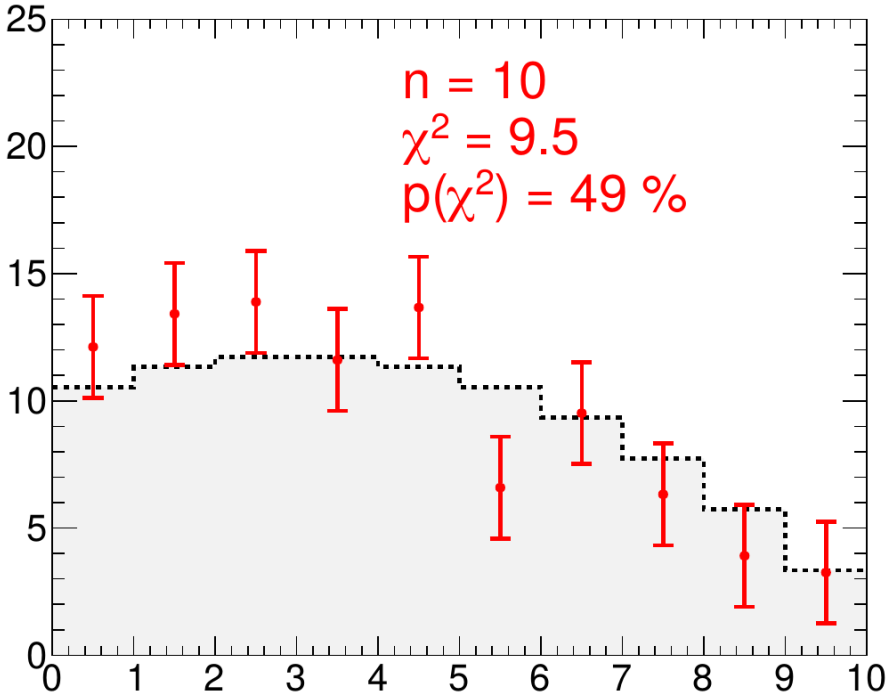
→ ***Bin central value*** *~ mean of the bin PDF*

→ ***Bin uncertainty*** *~ RMS of the bin PDF*

The data is just what it is, a simple observed point.

⇒ One should in principle **show the error bar on the prediction.**

→ In practice, the usual convention is to have **error bars on the data points**.



Error Bars

Strictly speaking, ***the uncertainty is given by the model :***

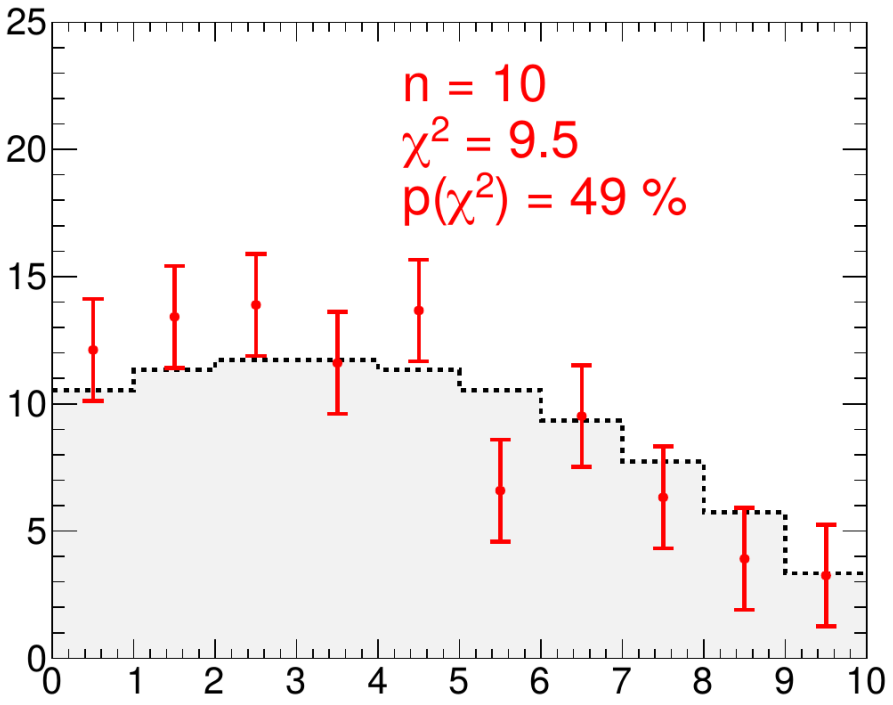
→ ***Bin central value*** *~ mean of the bin PDF*

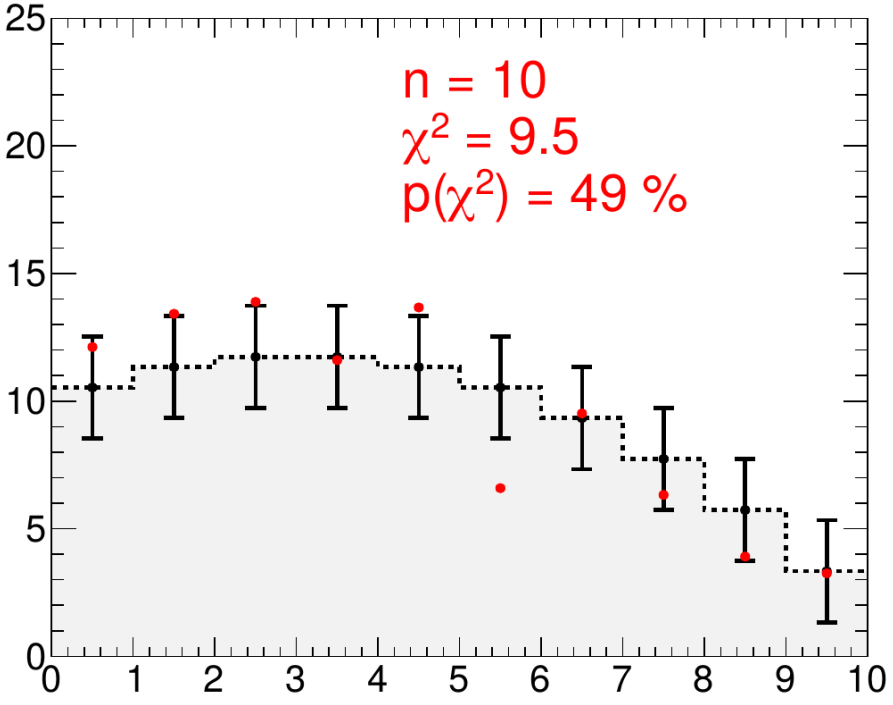
→ ***Bin uncertainty*** *~ RMS of the bin PDF*

The data is just what it is, a simple observed point.

⇒ One should in principle **show the error bar on the prediction.**

→ In practice, the usual convention is to have **error bars on the data points**.





Statistical Modeling

Rare Processes ?

HEP : almost always use Poisson distributions. Why ?

ATLAS :

- **Event rate ~ 1 GHz**

  (L~10^{34} $cm^{-2}s^{-1}$~10 nb^{-1}/s, σ_{tot}~10^{8} nb,)

- **Trigger rate ~ 1 kHz**

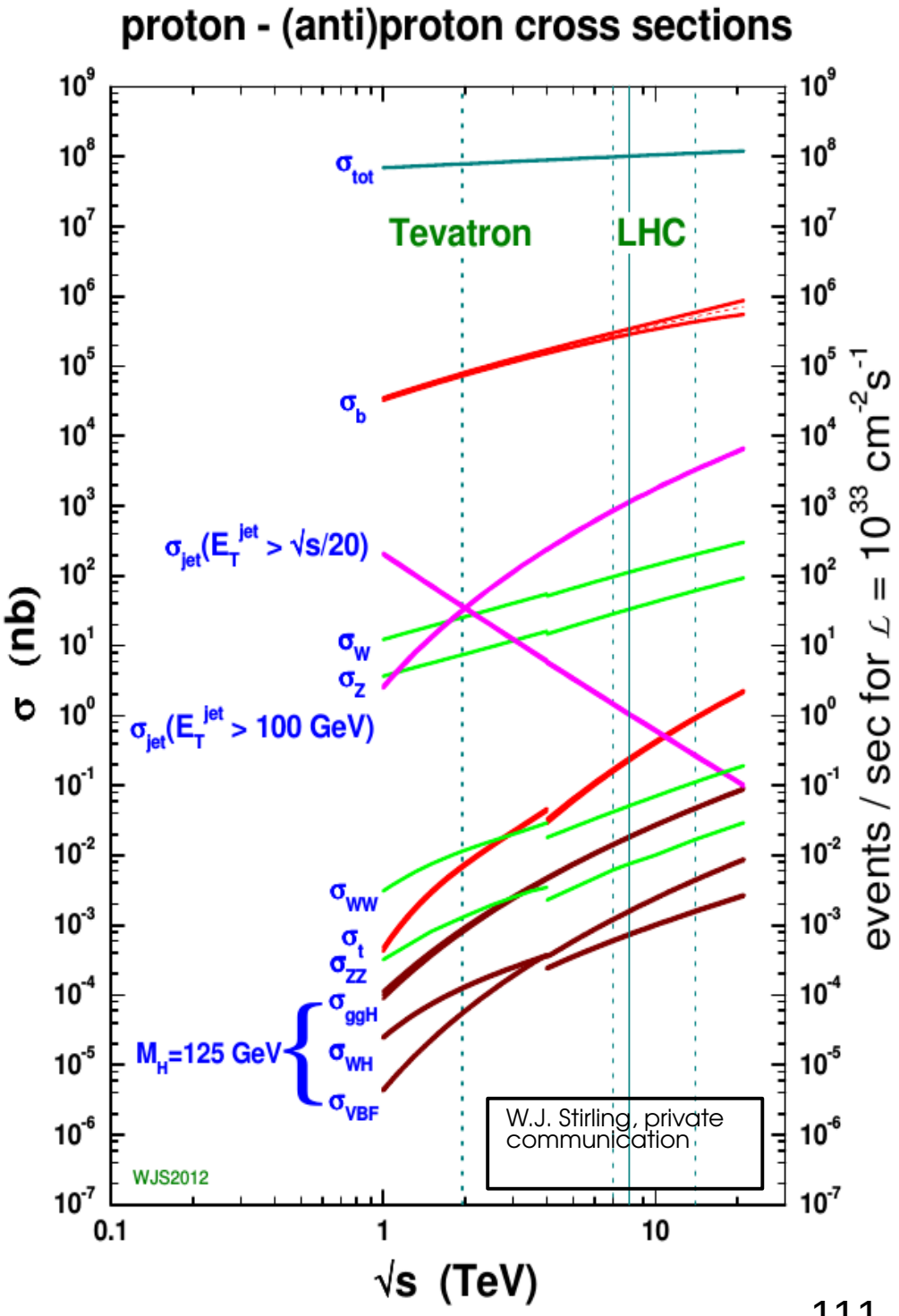
  (Higgs rate **~ 0.1 Hz**)

⇒ p ~ 10^{-6} ≪ 1 ($p_{H\to\gamma\gamma}$ ~ **10^{-13}**)

A day of data: **N ~ 10^{14} ≫ 1**

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

(Large N = design requirement, to get not-too-small l=Np…)

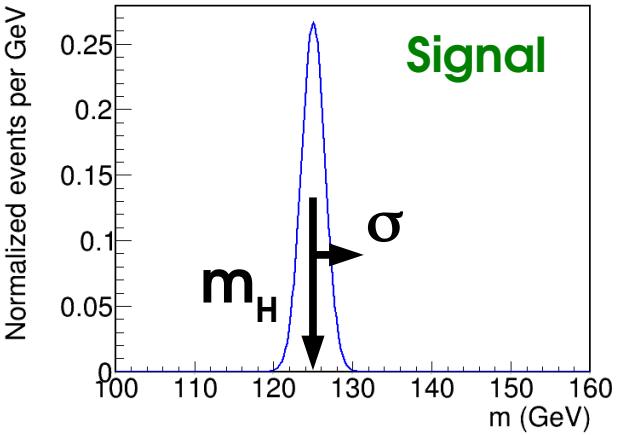


Unbinned Shape Analysis

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

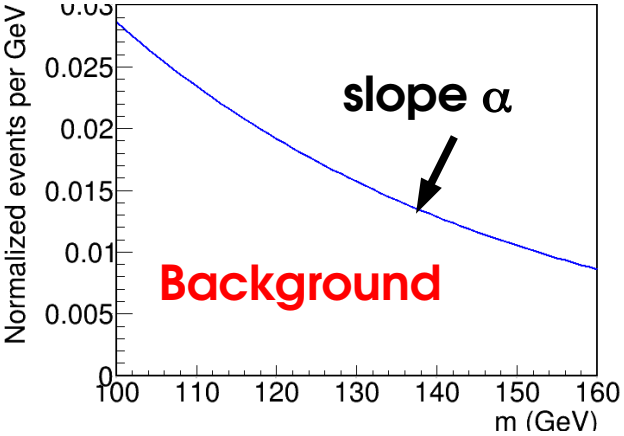
→ Describe shape of the **distribution of m**

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



H→γγ-inspired example:

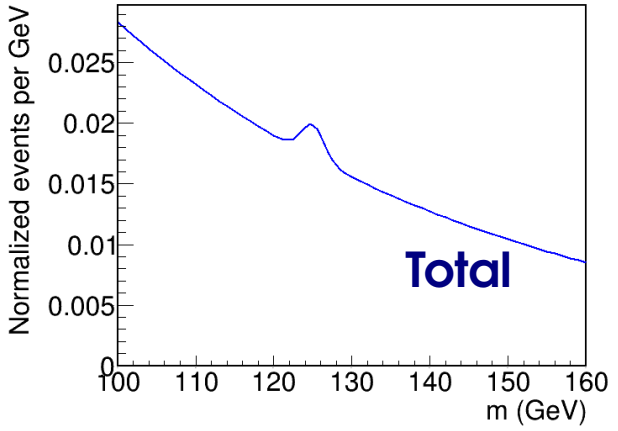
- Gaussian signal $P_{\text{signal}}(m) = G(m; m_H, \sigma)$
- Exponential bkg $P_{\text{bkg}}(m) = \alpha\, e^{-\alpha m}$



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

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

Poisson Example

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

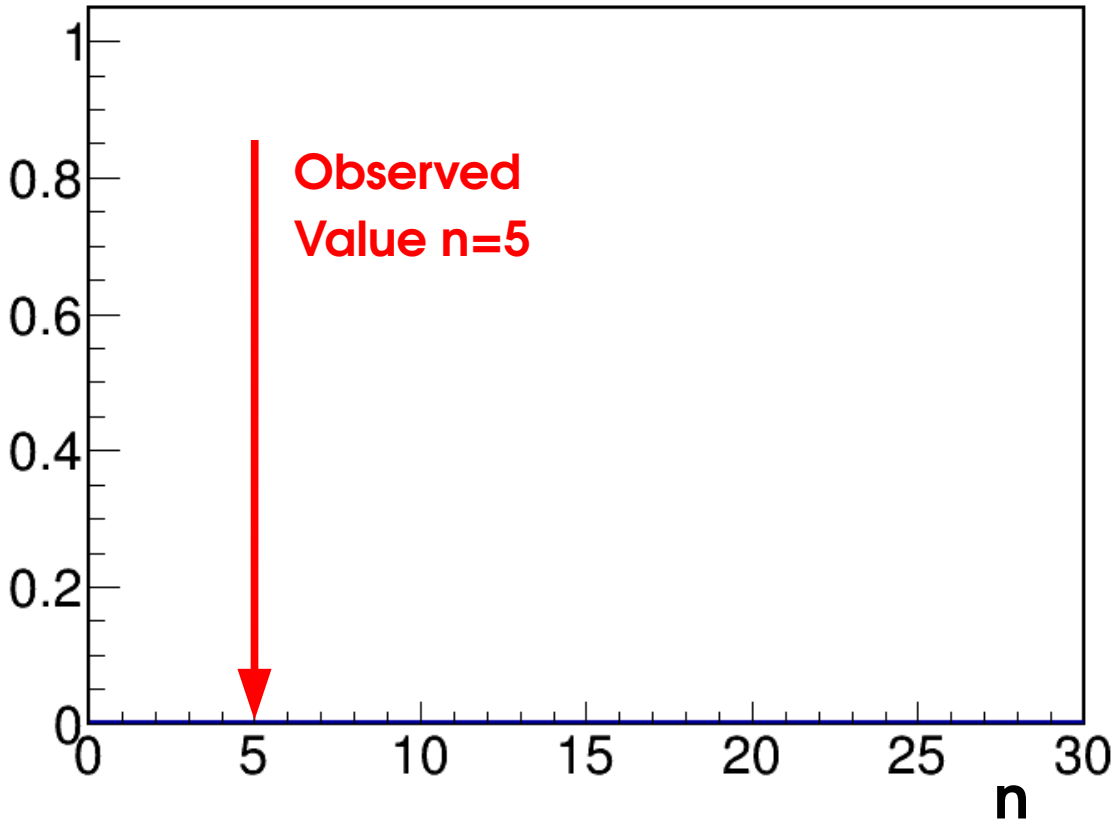
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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 :

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

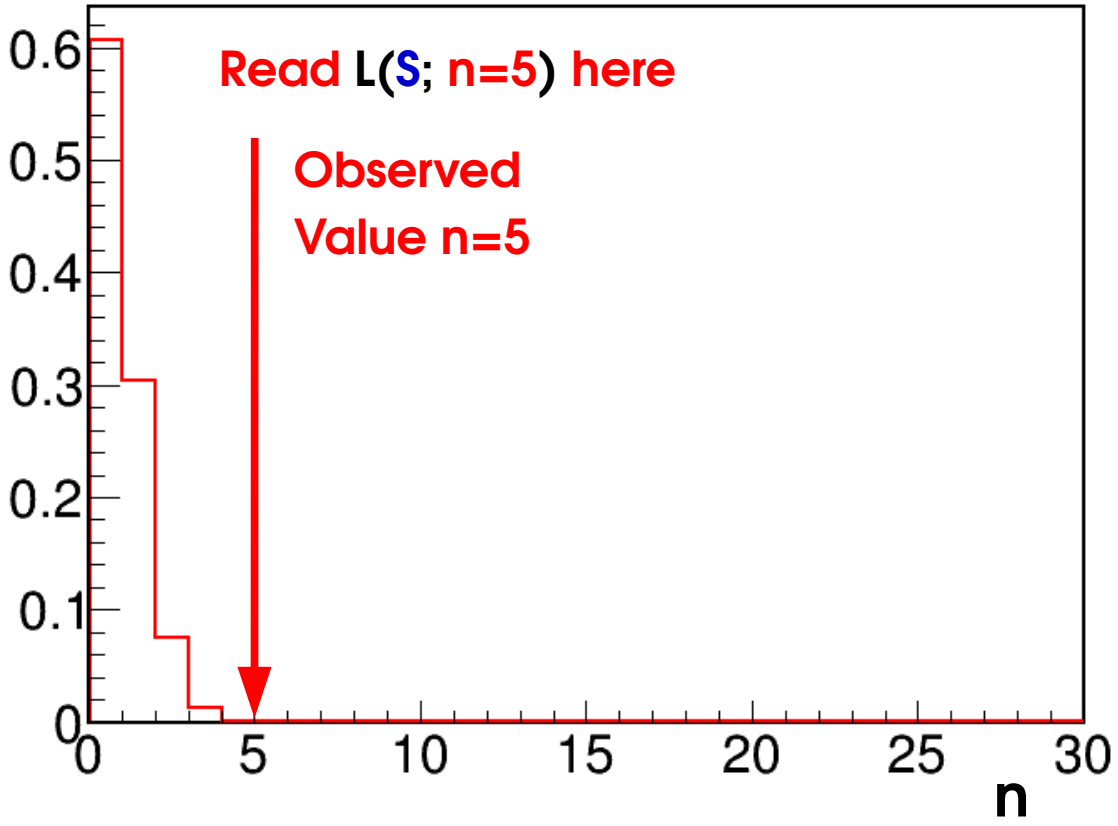
Say we **observe** n=5, want to infer information on the parameter S

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

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

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

P(S = 0.5)
Low
likelihood



Poisson Example

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

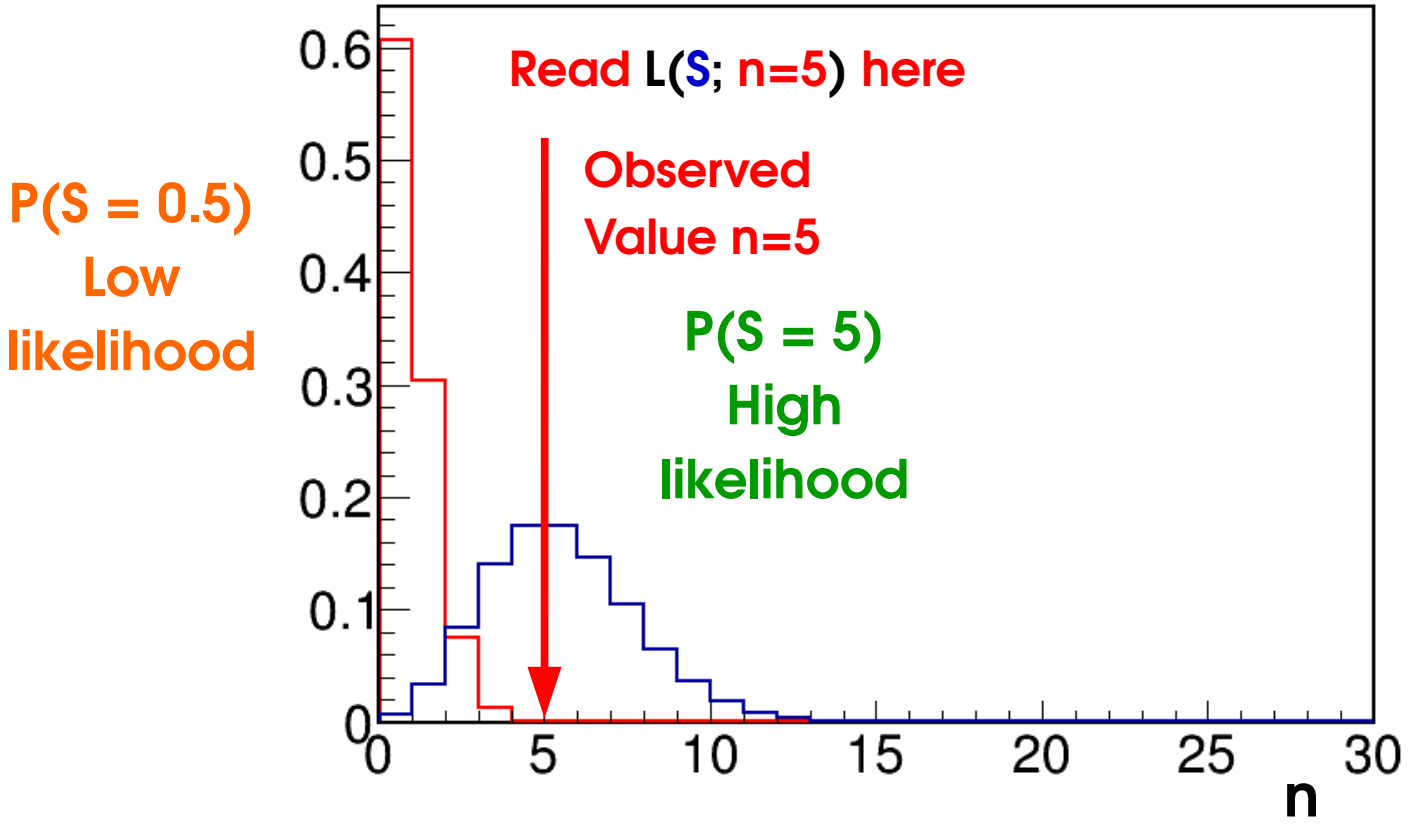
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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 :

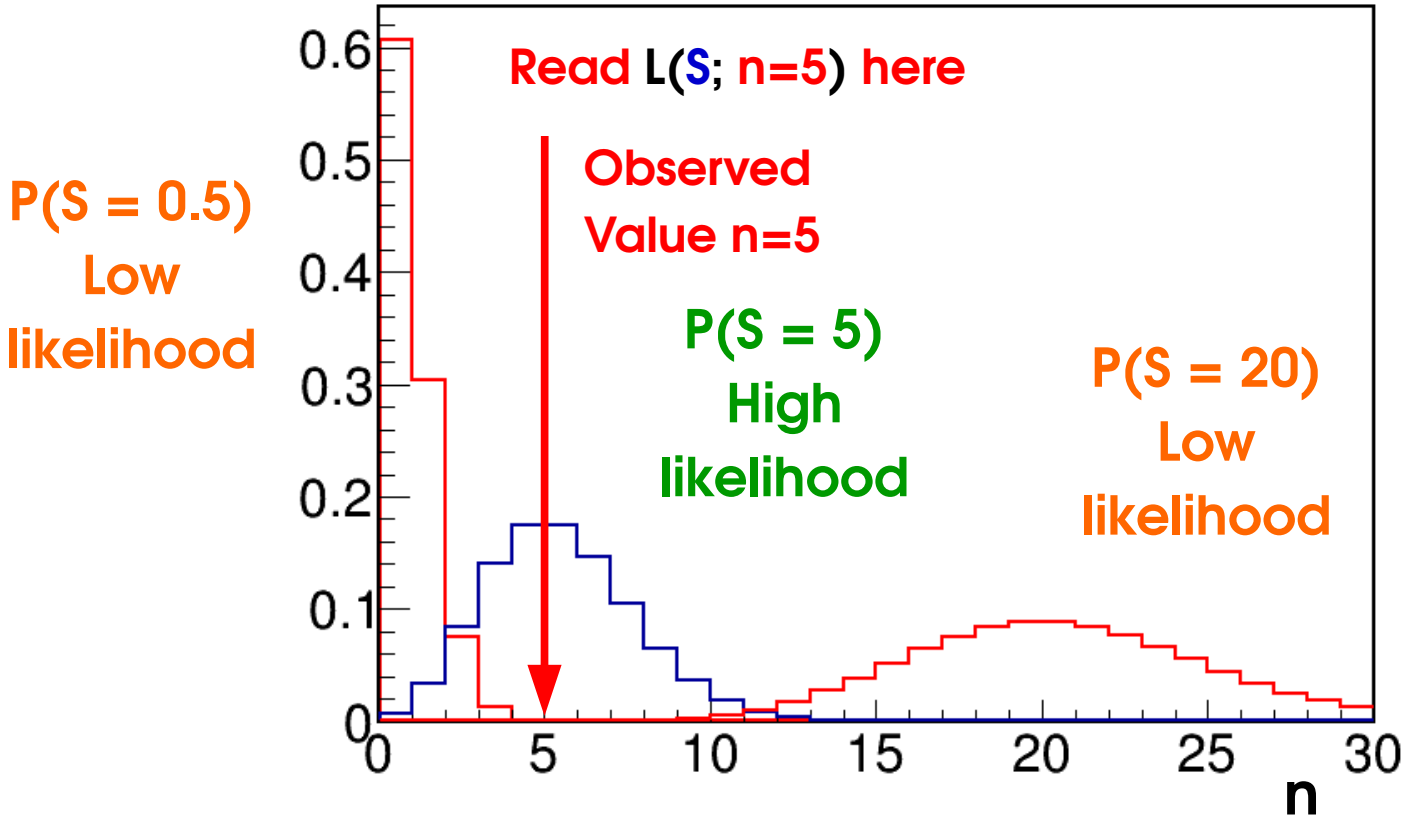
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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 :

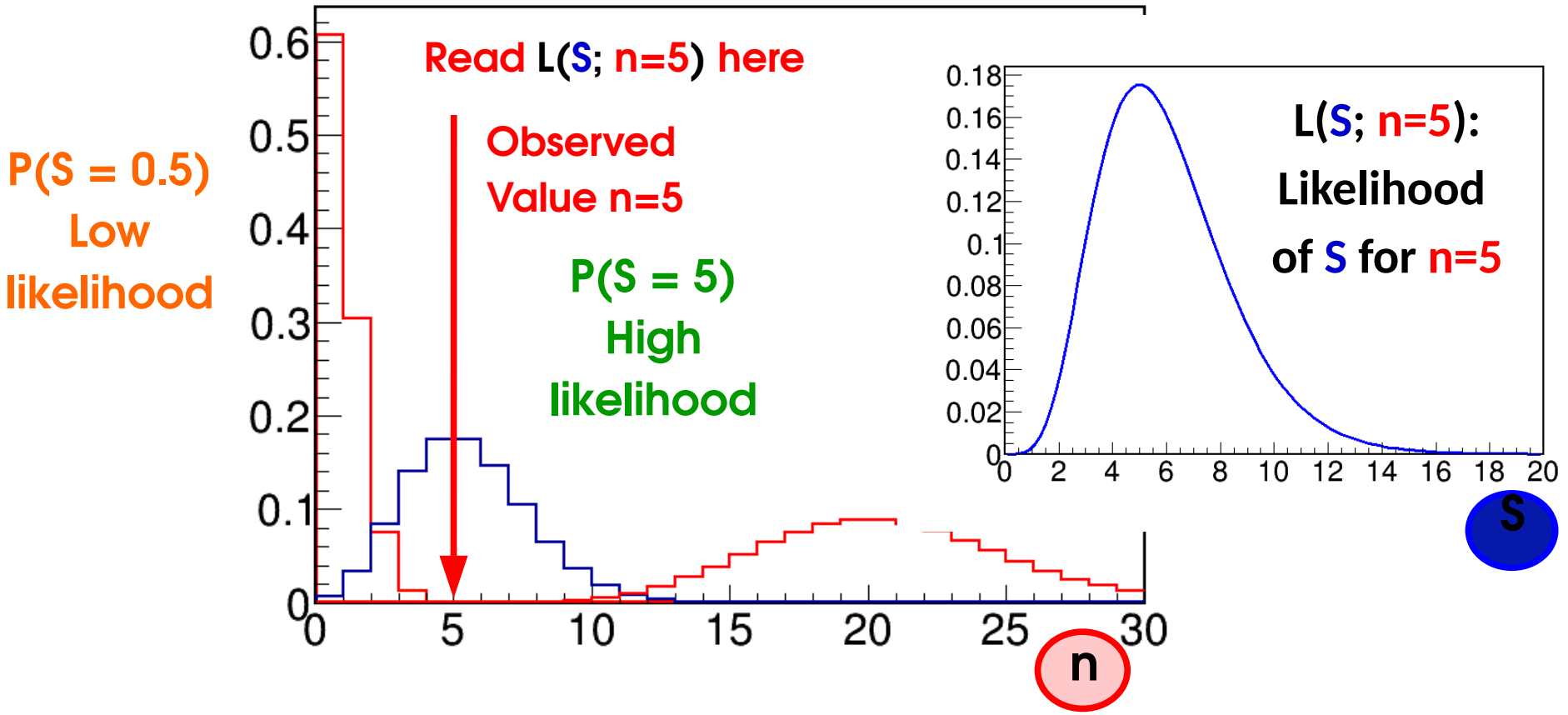
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

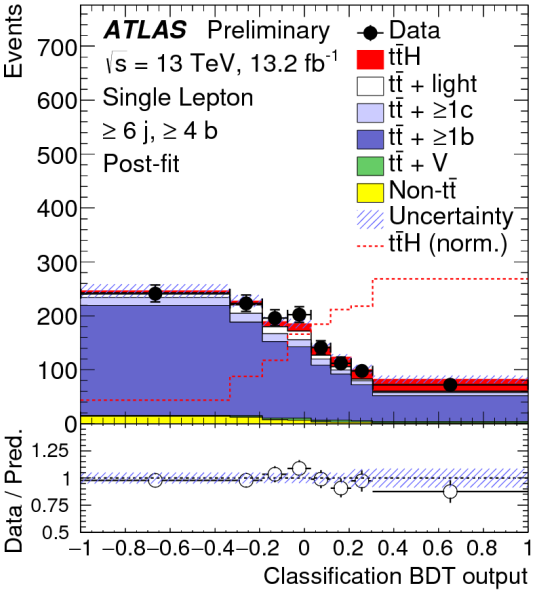
→ Try different values of S for a fixed data value n=5

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

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



MLEs in Shape Analyses



Binned shape analysis:

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

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

In practice easier to minimize

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

Needs a computer...

In the Gaussian limit

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

χ^2 formula!

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

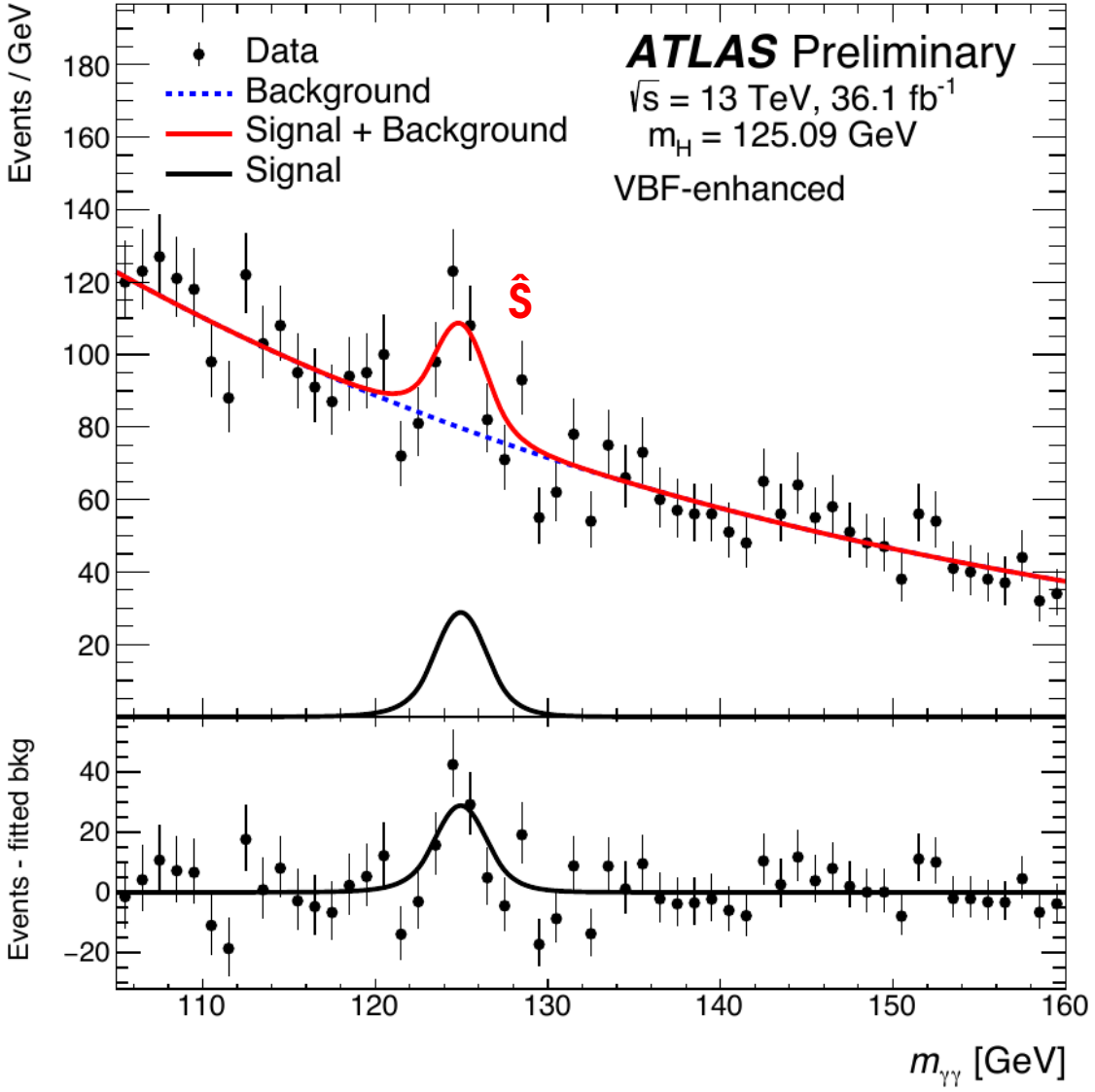
→ **Poisson MLE** (min λ_{Pois}) : ***Best fit value*** in a *likelihood* fit (in ROOT, fit option "L")

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

In both cases, MLE ⇔ *Best Fit*

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

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

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, $\hat{\mu} \overset{n\to\infty}{\to} \mu^*$

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

→ smaller σ_{Gauss} ⇒ more information.

Cramer-Rao bound: $$\text{Var}(\tilde{\mu}) \geq \frac{1}{I(\mu)}$$

For any estimator μ̃.

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

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

Gaussian case:

- For a Gaussian estimator μ̃

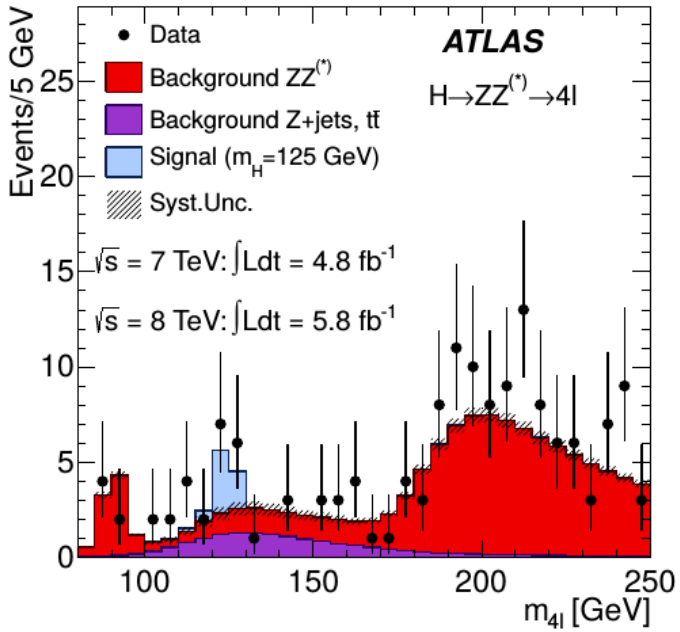
$$P(\tilde{\mu}) \propto \exp\left(-\frac{(\tilde{\mu} - \mu^*)^2}{2\sigma^2_{\tilde{\mu}}} \right)$$

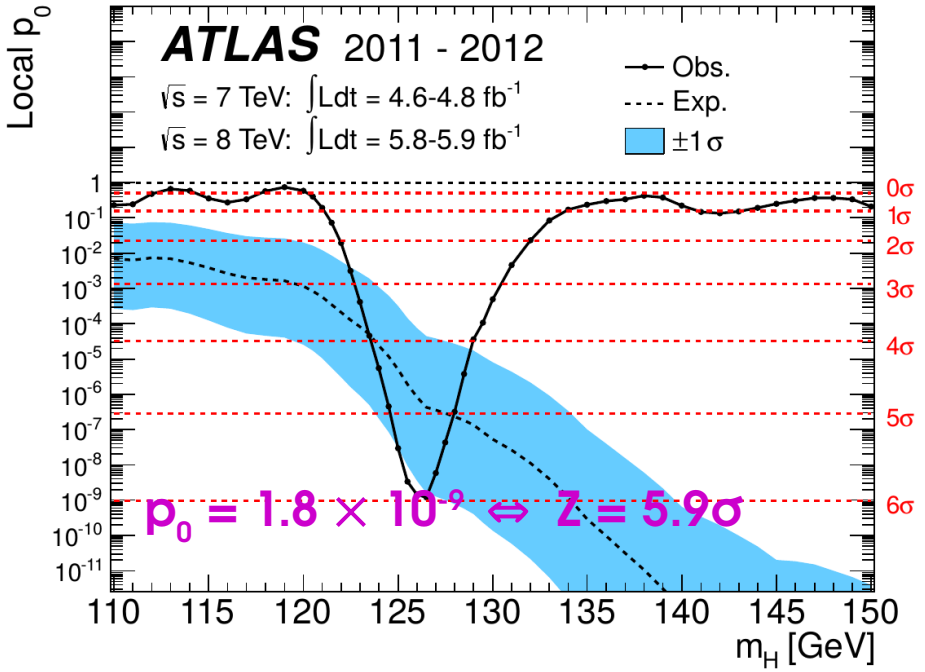
- MLE: $\text{Var}(\hat{\mu}) = \sigma_{\hat{\mu}}^2$

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

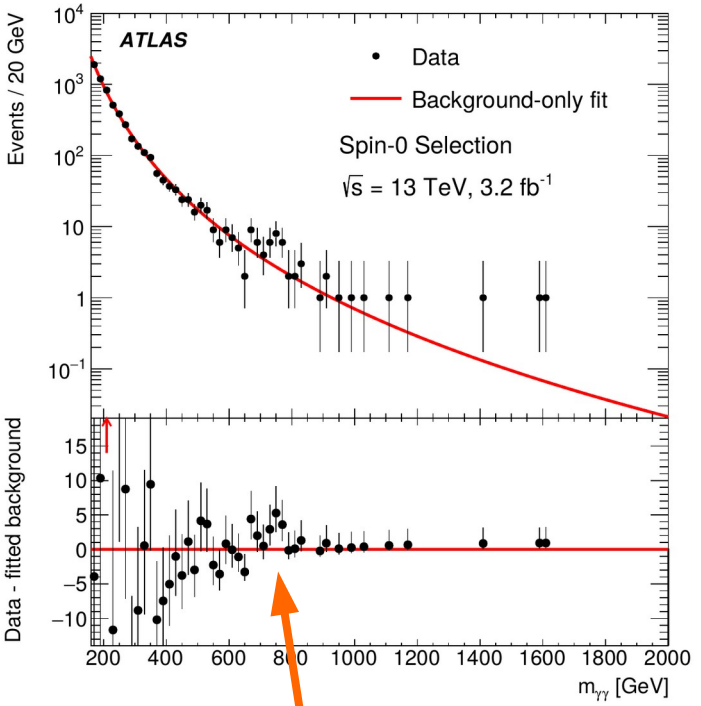
Some Examples

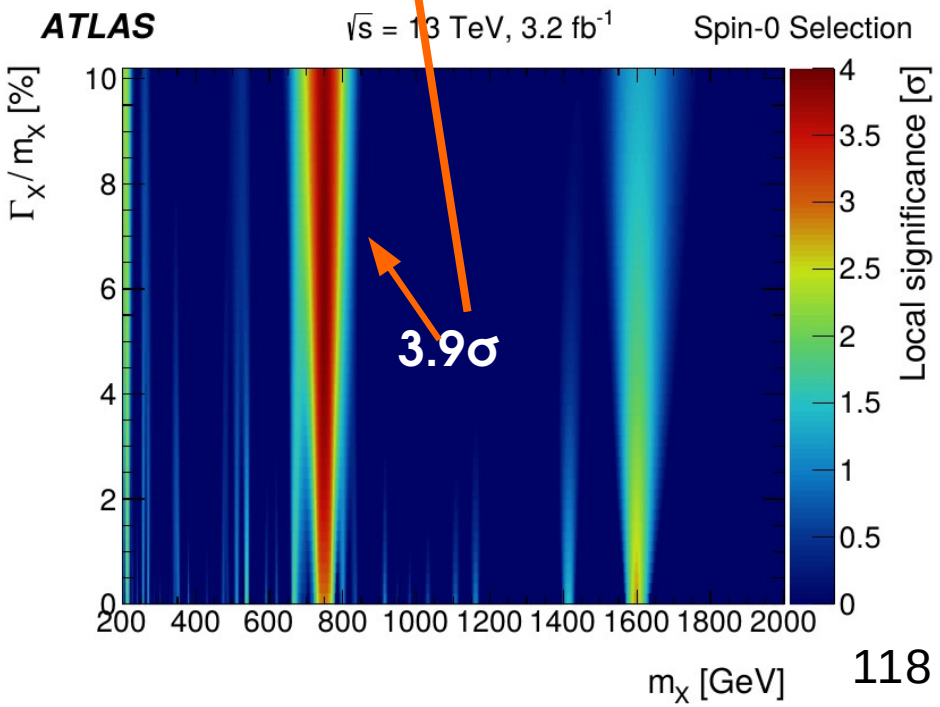
Higgs Discovery: Phys. Lett. B 716 (2012) 1-29





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



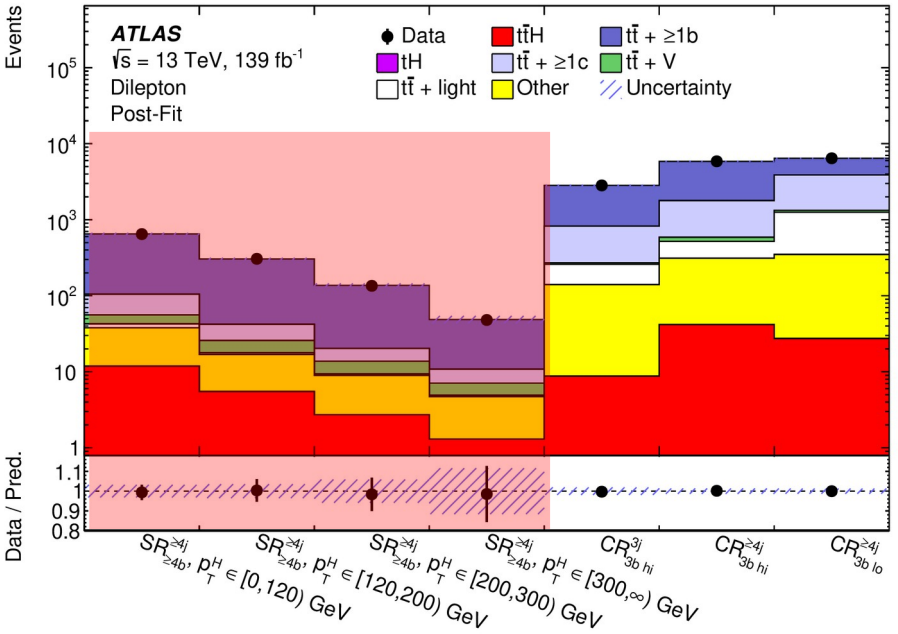


Categories

arXiv:2111.06712

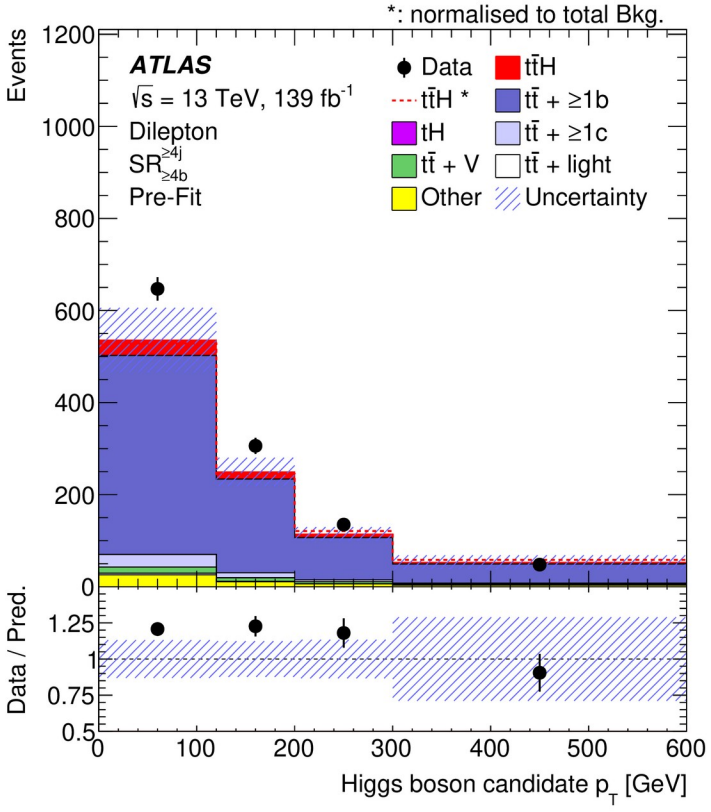
Multiple analysis regions often used.

→ Exploit better sensitivity in some regions



Here (ttH, H→bb analysis) 7 regions:

→ 4 *Signal Regions* (**SR**) split in p_T(Higgs)



Better sensitivity at high p_T

→ lower B backgrounds, higher S/B

Backgrounds levels from simulation here

→ Large systematic uncertainties!

Categories

arXiv:2111.06712

Multiple analysis regions often used.

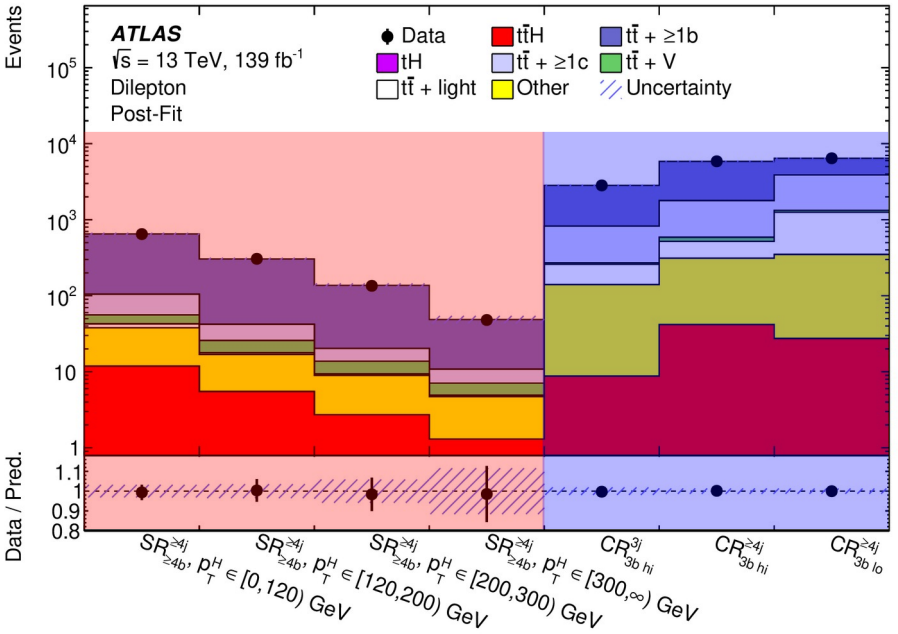
→ Exploit better sensitivity in some regions

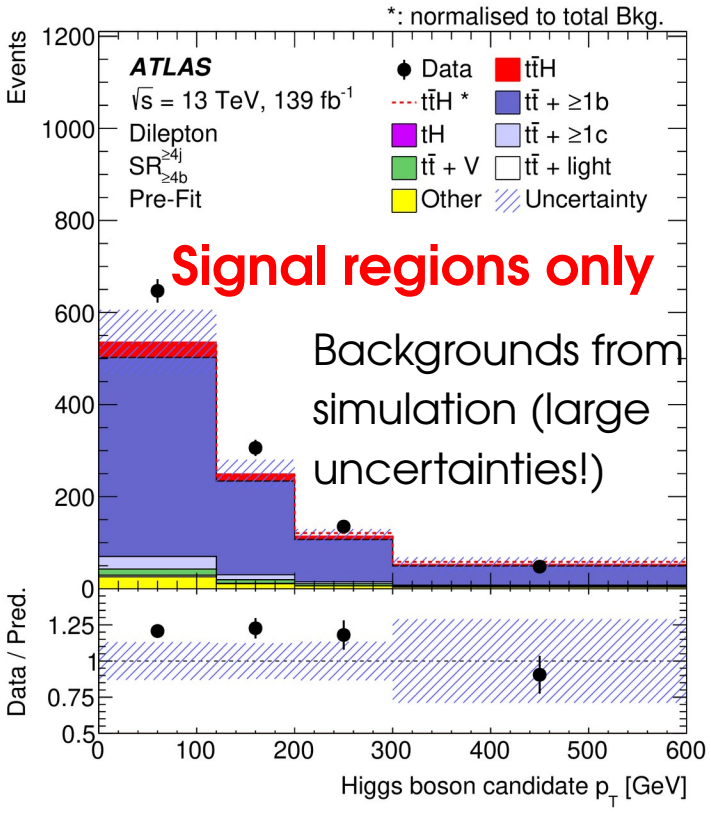
→ Constrain NPs: ***Control regions*** for bkgs

Here (ttH, H→bb analysis) 7 regions:

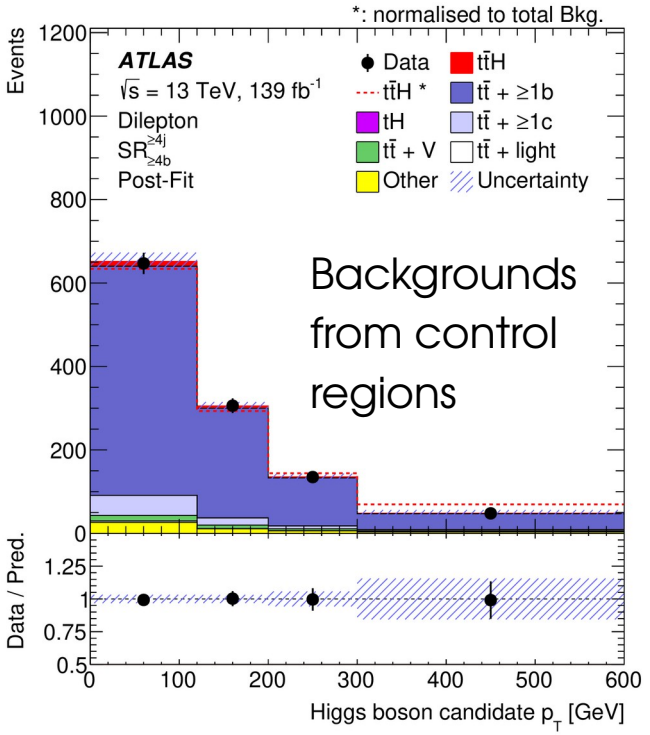
→ 4 *Signal Regions* (**SR**) split in p_T(Higgs)

→ 3 Background *Control Regions* (**CR**)





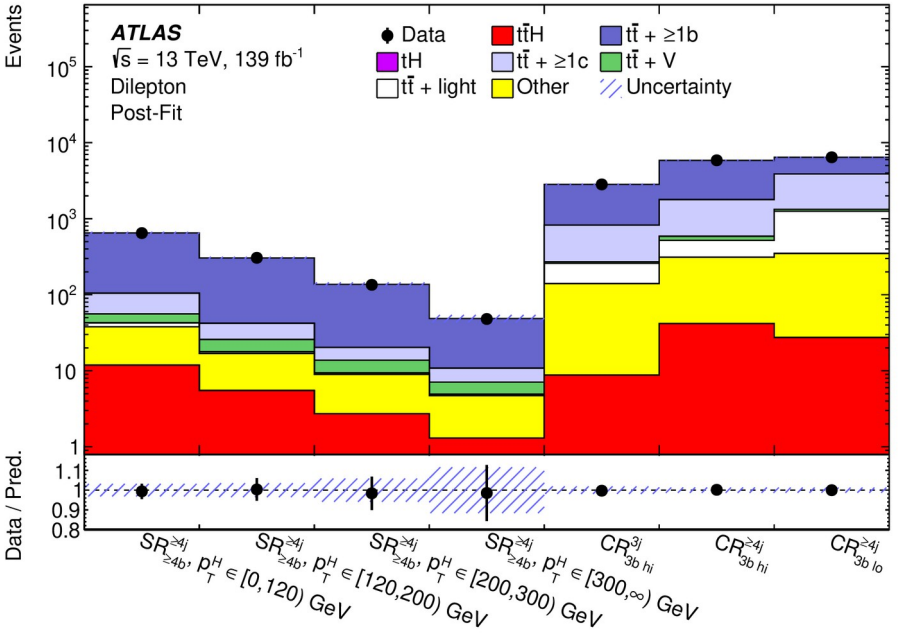
Include Background CRs



Signal + Bkg regions

Categories

arXiv:2111.06712



Multiple analysis regions often used.

→ Exploit better sensitivity in some regions

→ Constrain NPs: ***Control regions*** for bkgs

Here (ttH, H→bb analysis) **7** regions:

→ **4** *Signal Regions* (**SR**) split in p_T(Higgs)

→ **3** Background *Control Regions* (**CR**)

⇒ ***Combined PDF*** :

PDF for category k

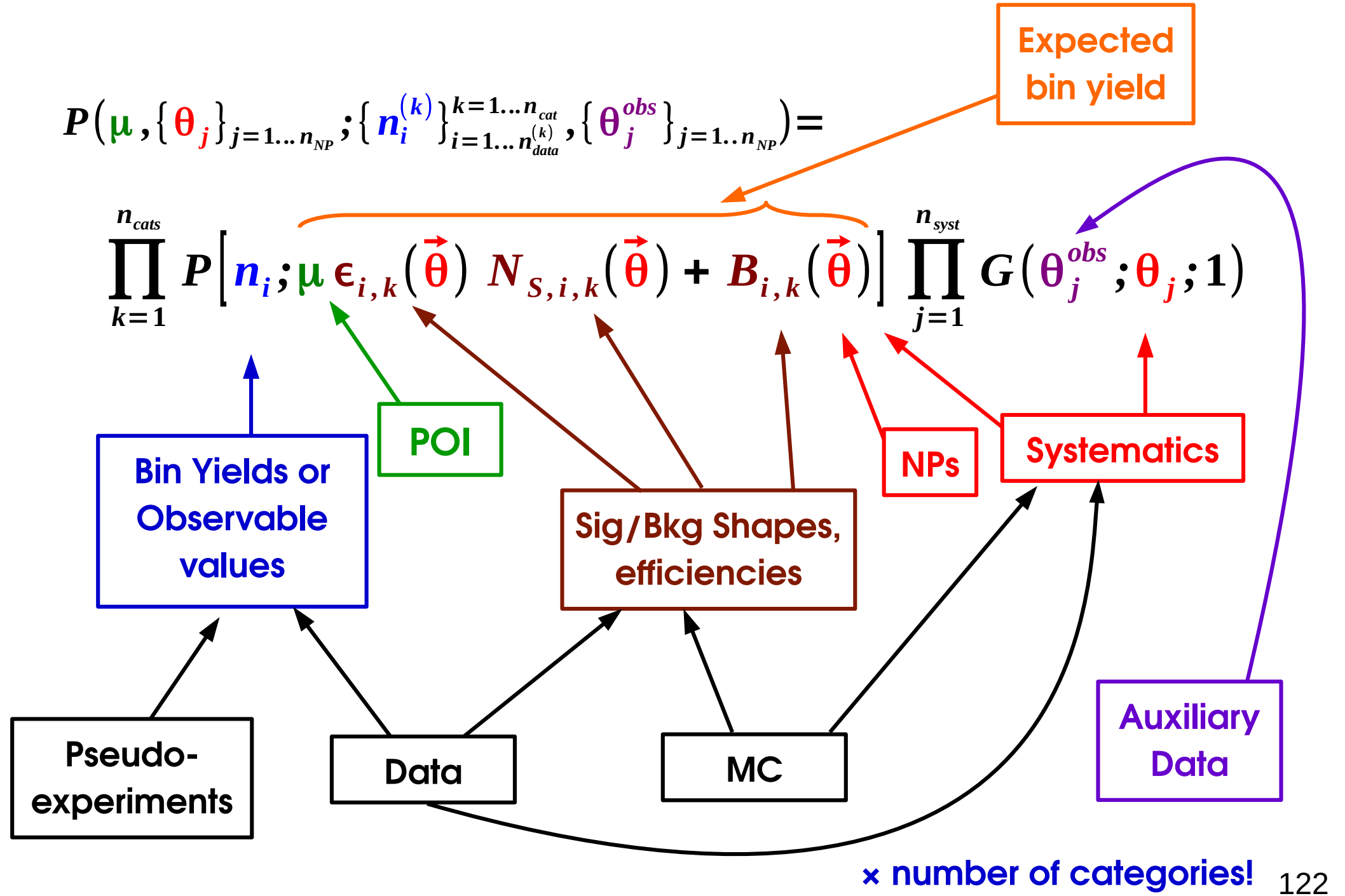
$$P\left(S, B; \left\{n_i^{(k)}\right\}_{i=1\ldots n_{evts}^{(k)}}^{k=1\ldots n_{cats}}\right) = \prod_{k=1}^{n_{cats}} P_k\left(S, B; \left\{n_i^{(k)}\right\}_{i=1\ldots n_{evts}^{(k)}}\right)$$

No overlaps between categories ⇒ No statistical correlations

⇒ can simply take product of individual PDFs.

Multiple categories allows to **constrain nuisance parameters** (e.g. **B**)

Counting model, the full version

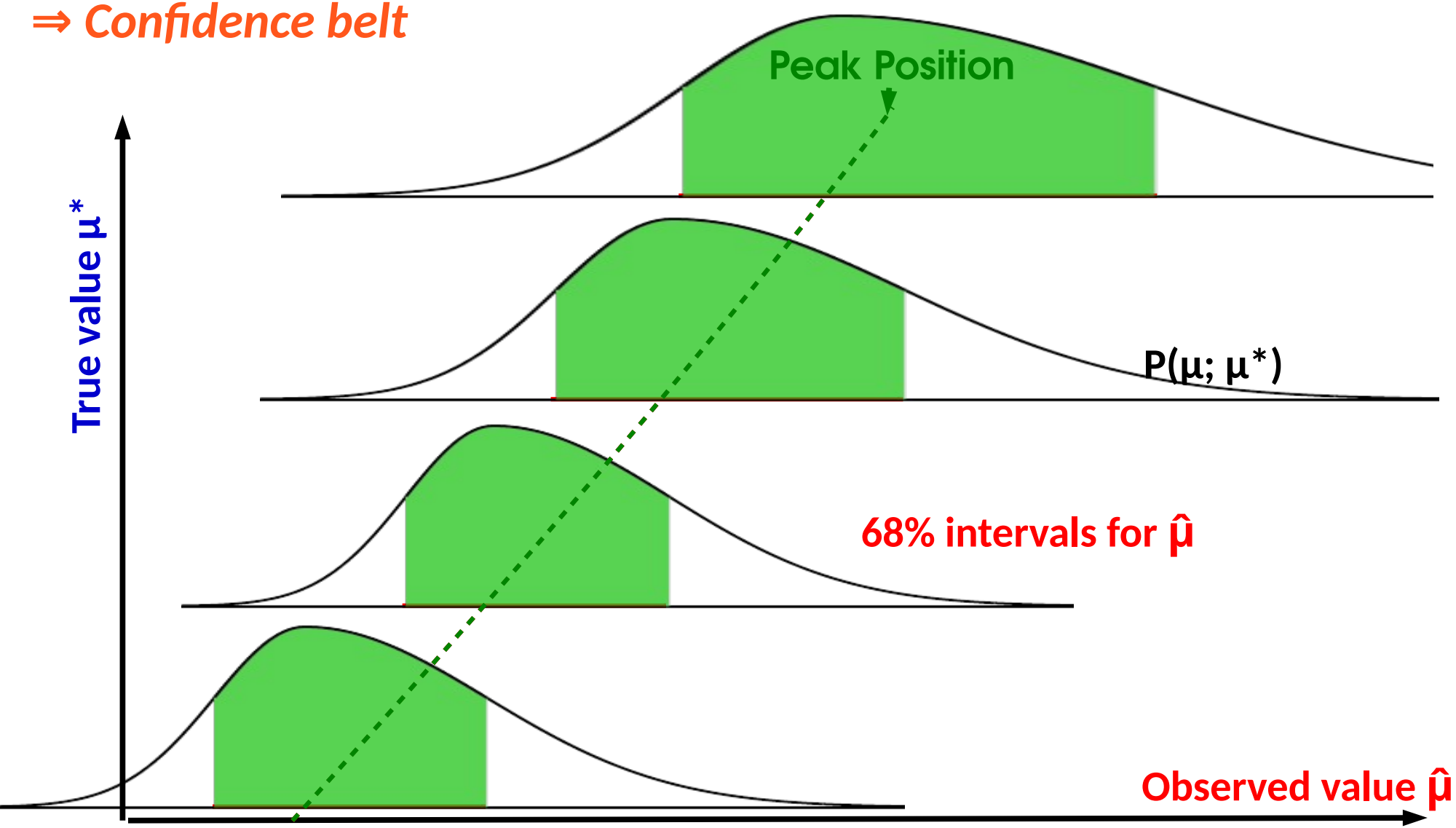


Confidence intervals

Neyman Construction

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

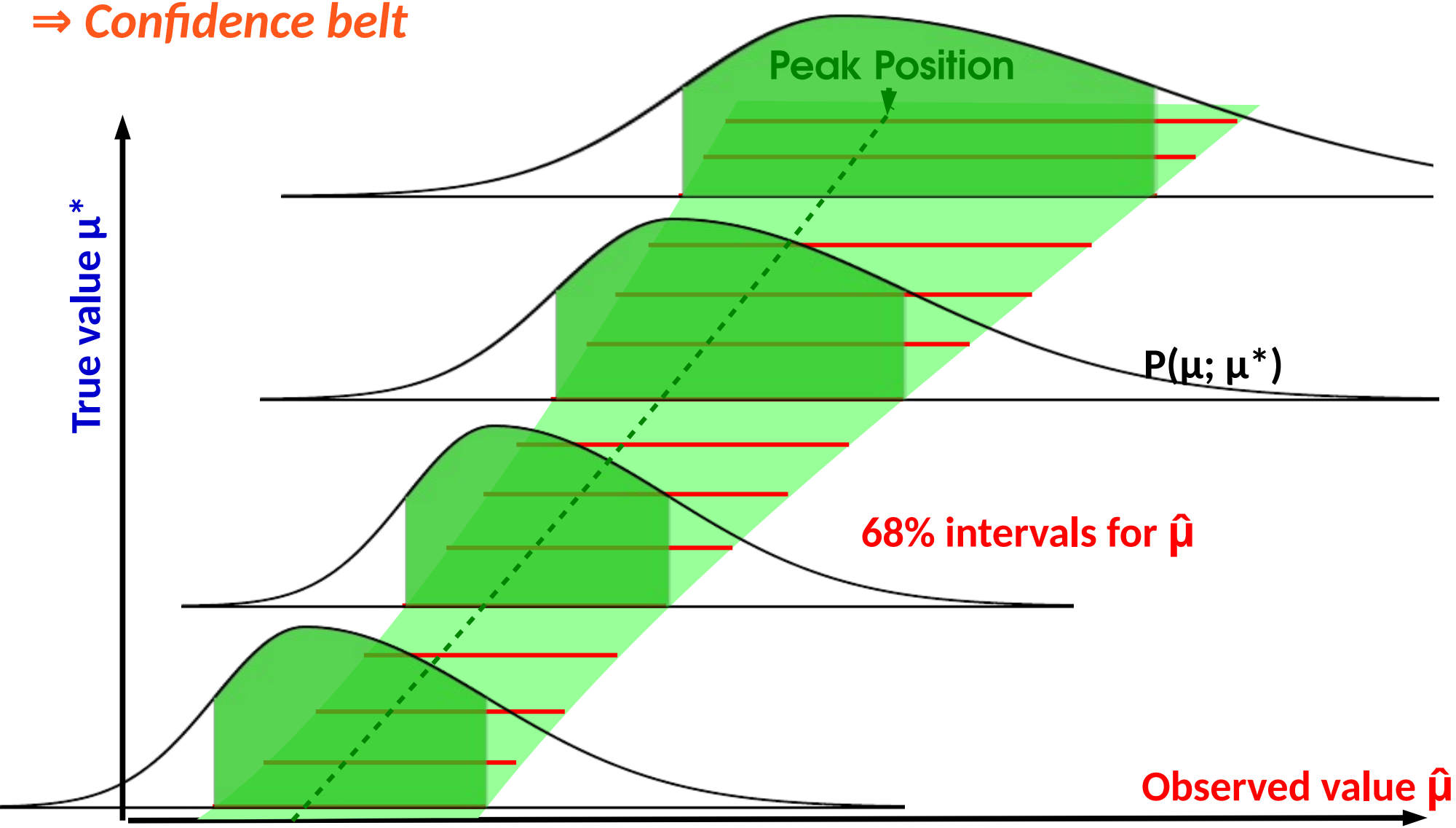
⇒ *Confidence belt*



Neyman Construction

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

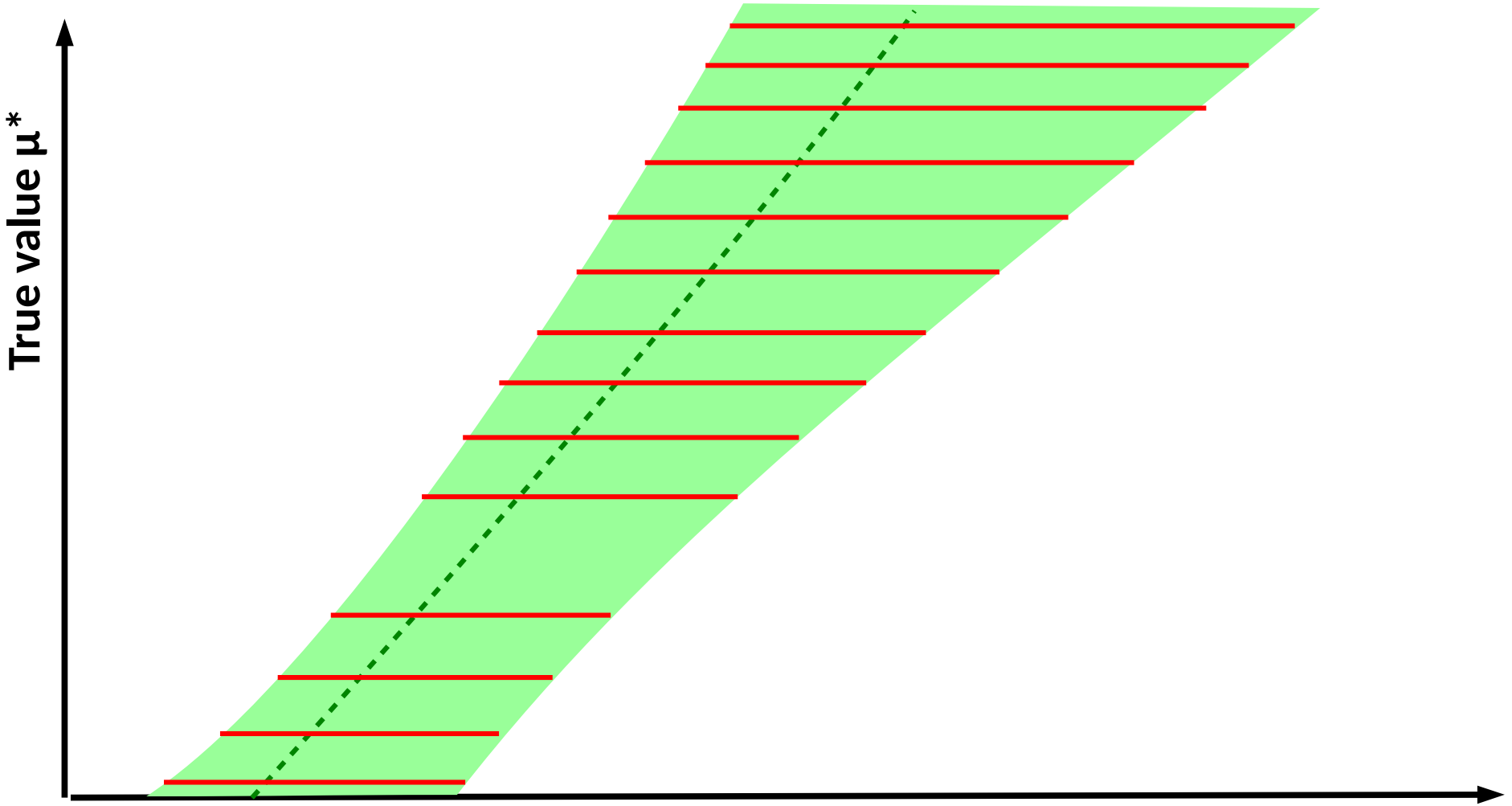
⇒ *Confidence belt*



Inversion using the Confidence Belt

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

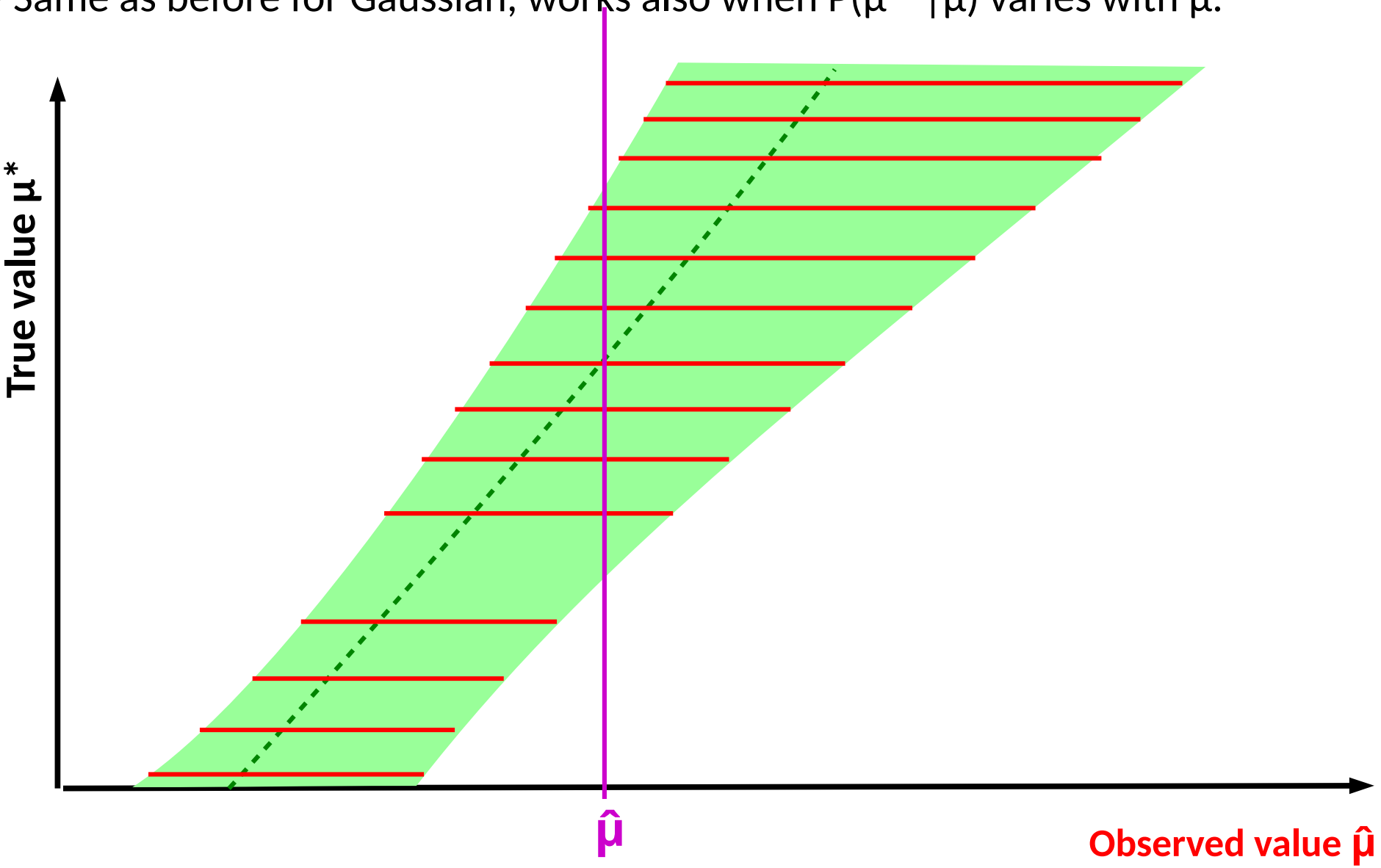
→ Same as before for Gaussian, works also when $P(\mu^{obs}|\mu)$ varies with μ.



Inversion using the Confidence Belt

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

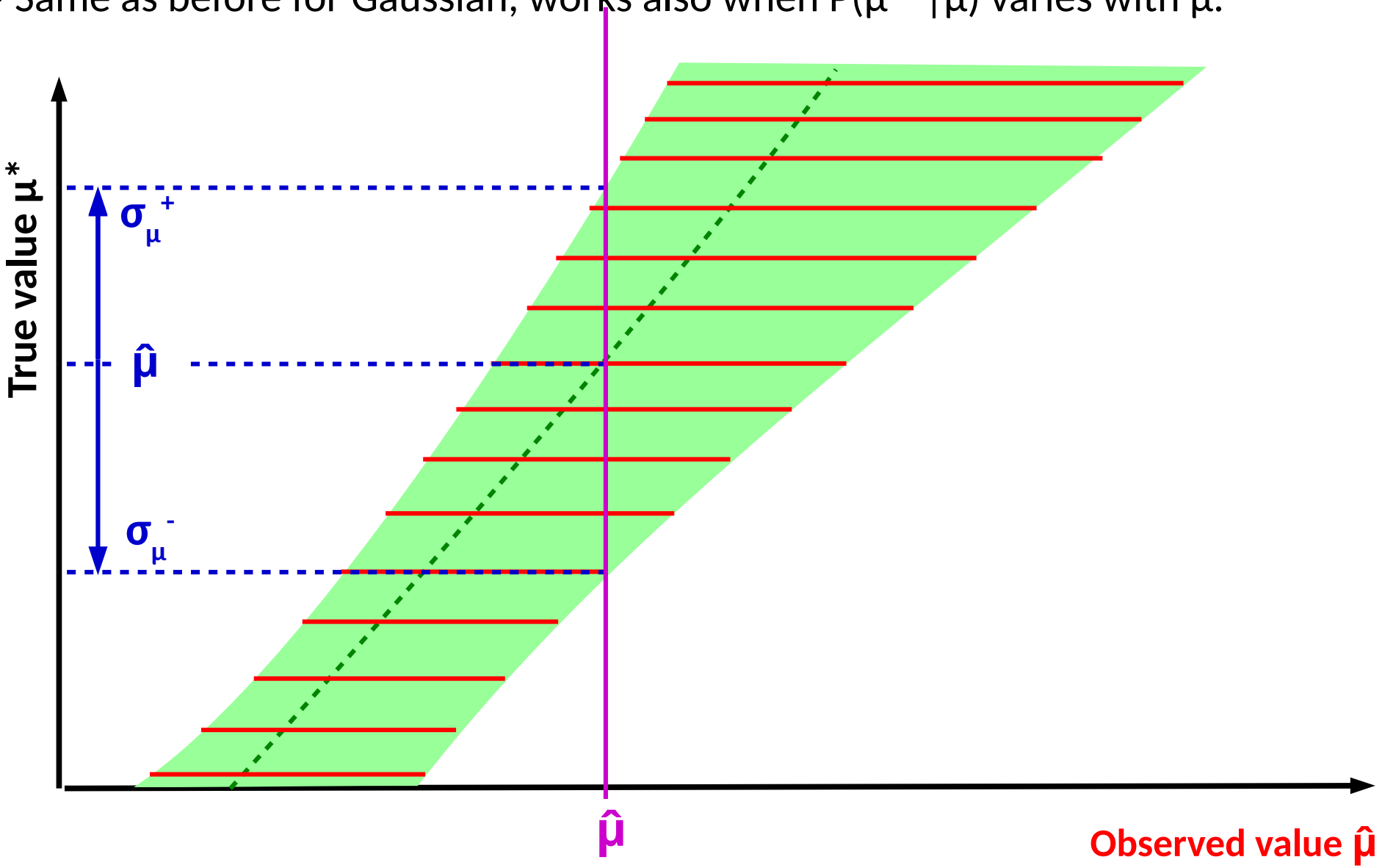
→ Same as before for Gaussian, works also when $P(\mu^{obs}|\mu)$ varies with μ.



Inversion using the Confidence Belt

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

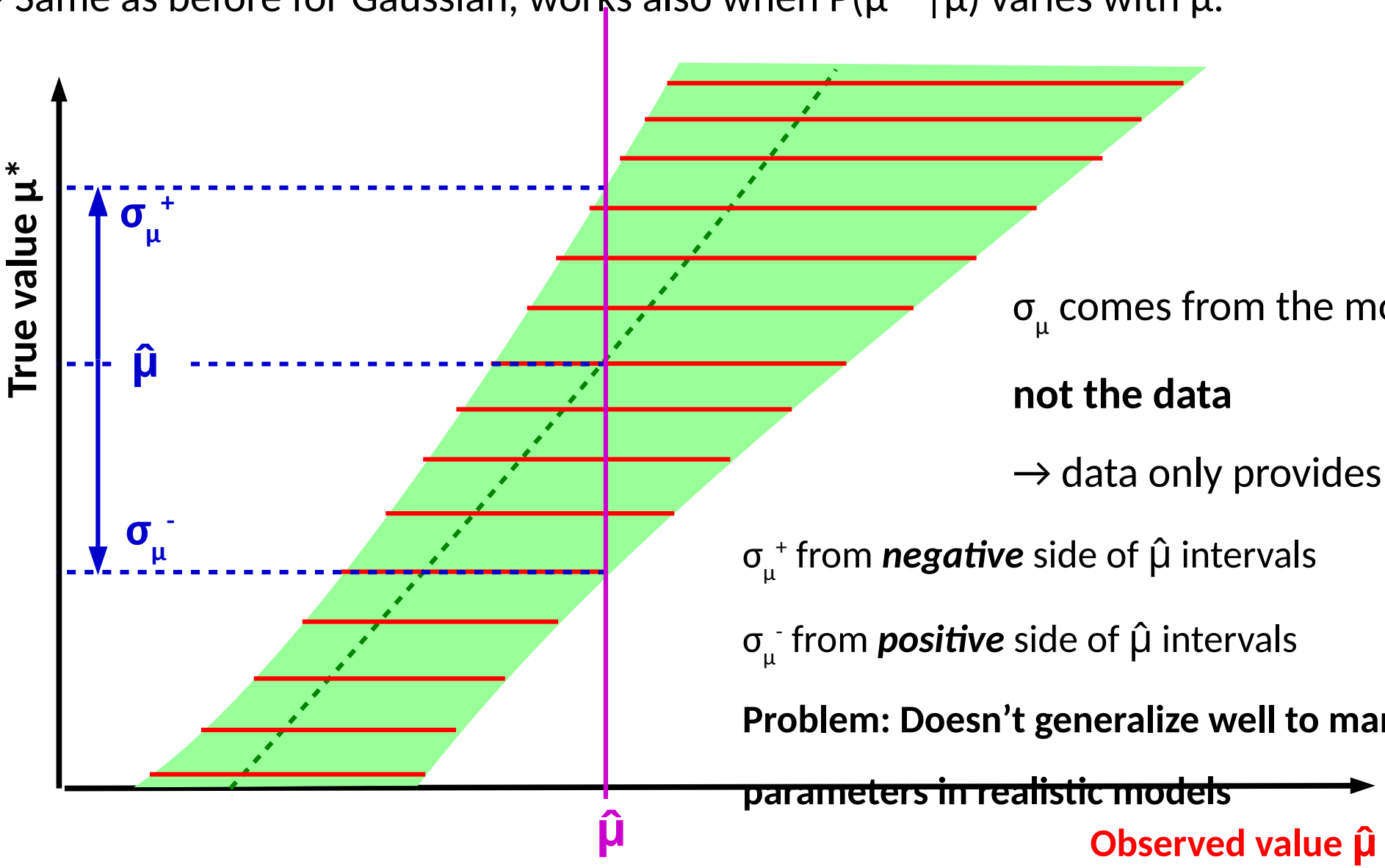
→ Same as before for Gaussian, works also when $P(\mu^{obs}|\mu)$ varies with μ.



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.



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

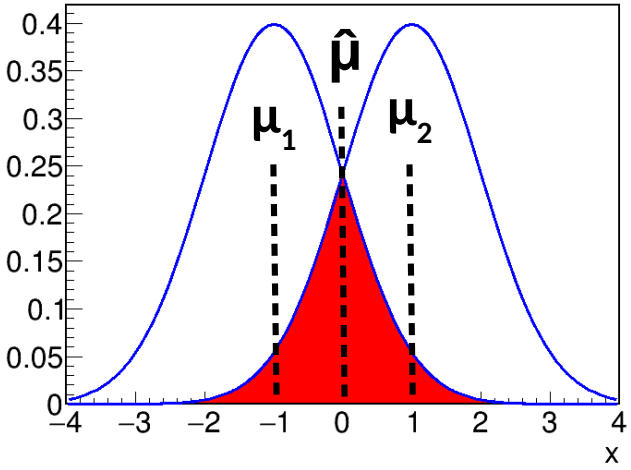
Problem: Doesn't generalize well to many parameters in realistic models

Upper Limits

Test Statistics for Limit-Setting

Confidence Interval :

Try to exclude μ values away from μ̂.

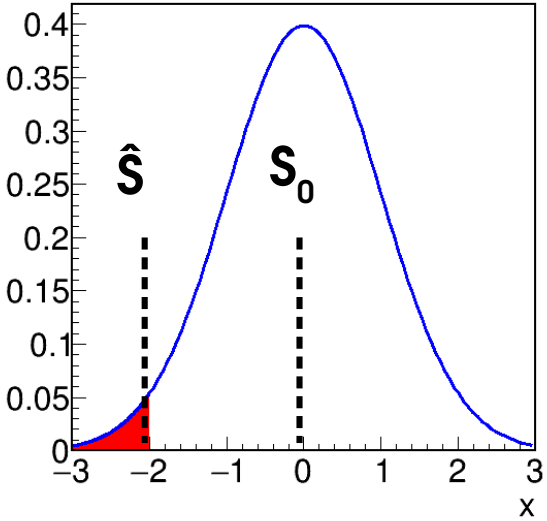


$$t(\mu_0) = -2\log\frac{L(\mu=\mu_0)}{L(\hat{\mu})}$$

"**Two-sided**" test

Limit-setting

Try to exclude values of S that are above Ŝ.



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

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

Discovery was also one-sided, for S>0

Getting the limit for a given CL

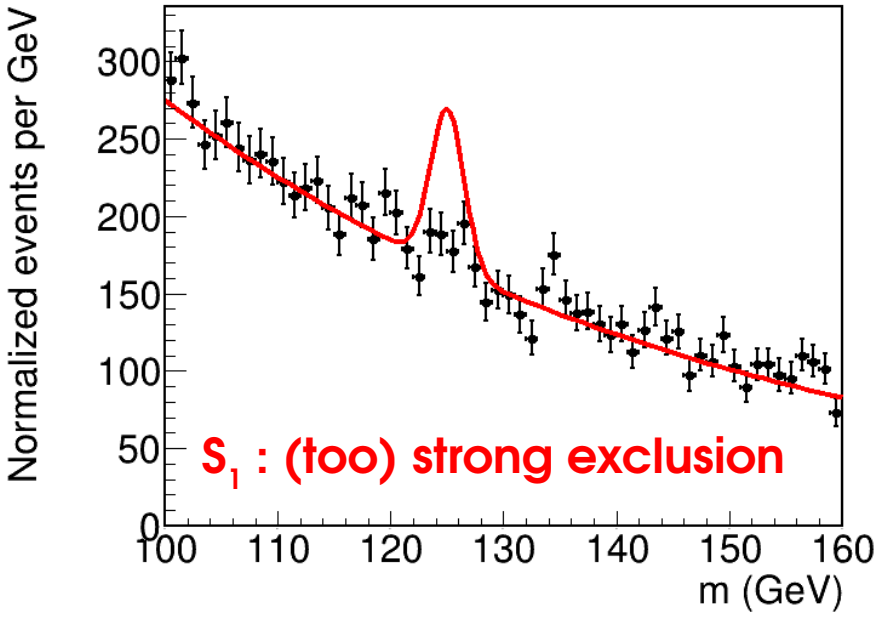
Procedure:

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

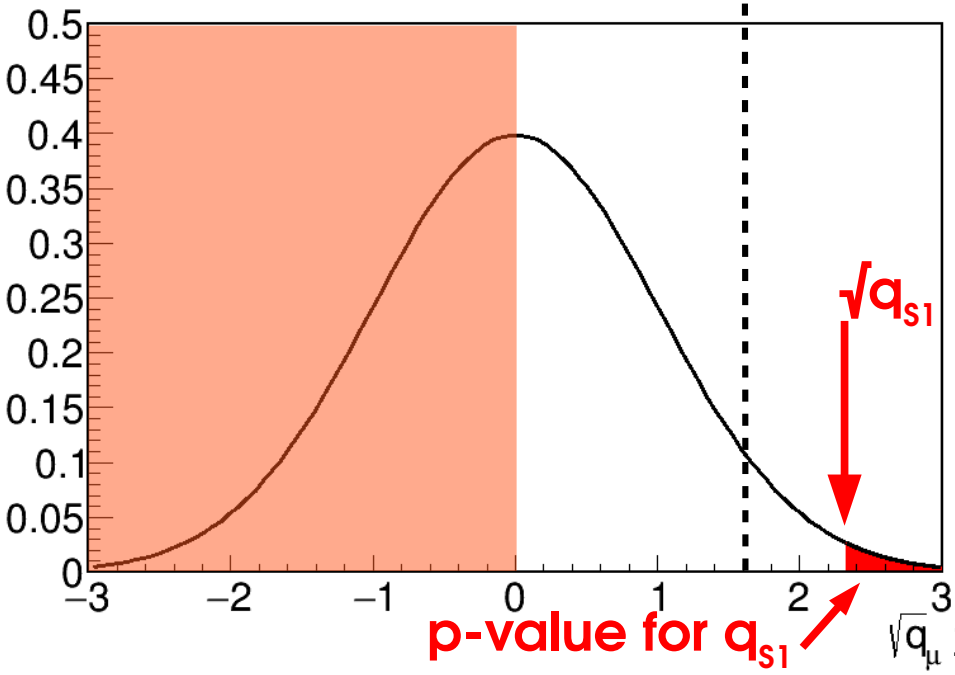
Asymptotics: $$p(S_0) = 1 - \Phi\left(\sqrt{q(S_0)}\right)$$

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**



$\sqrt{q(S)} = 1.64$
(p = 5%)



Getting the limit for a given CL

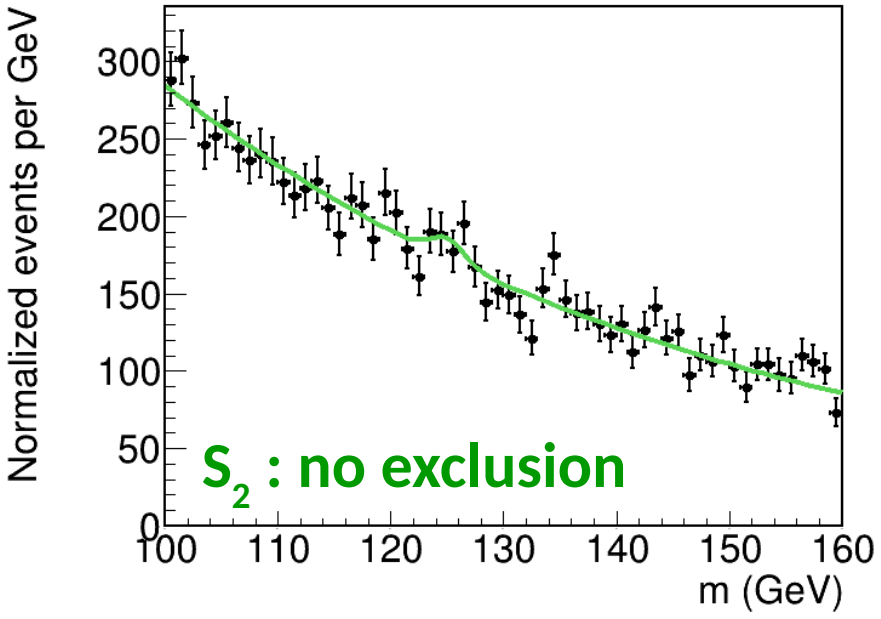
Procedure:

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

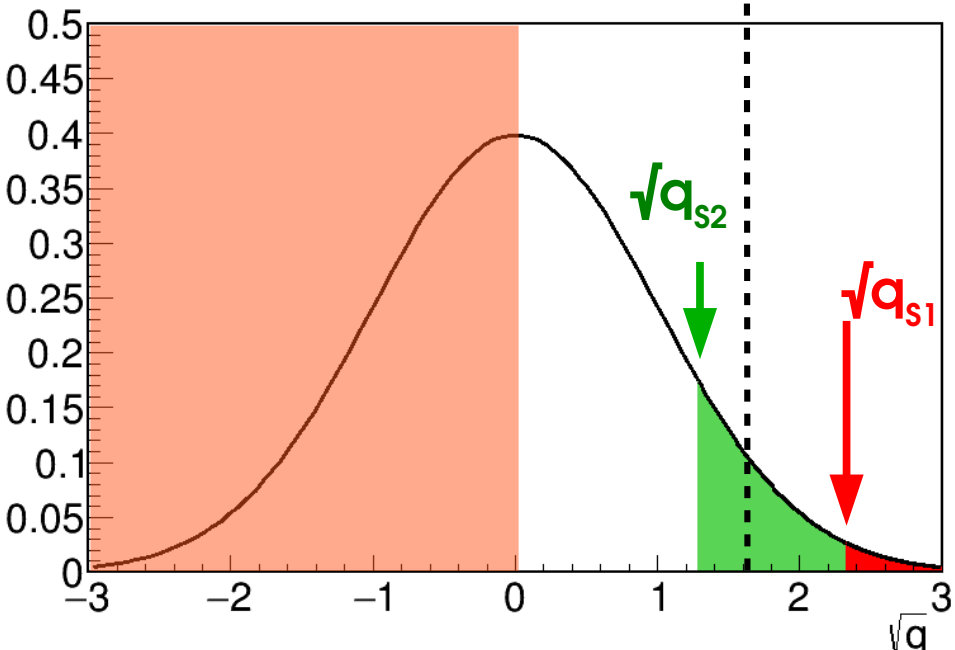
Asymptotics: $$p(S_0) = 1 - \Phi\left(\sqrt{q(S_0)}\right)$$

CL	p	Region
90%	10%	√q(S) > 1.28
95%	5%	√q(S) > 1.64
99%	1%	√q(S) > 2.33

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



√q(S) = 1.64
(p = 5%)



Getting the limit for a given CL

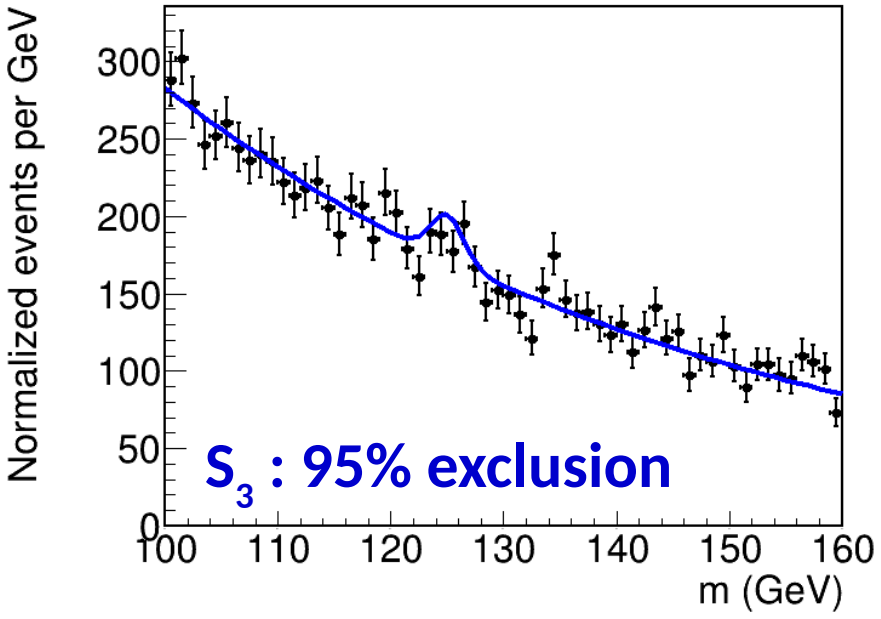
Procedure:

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

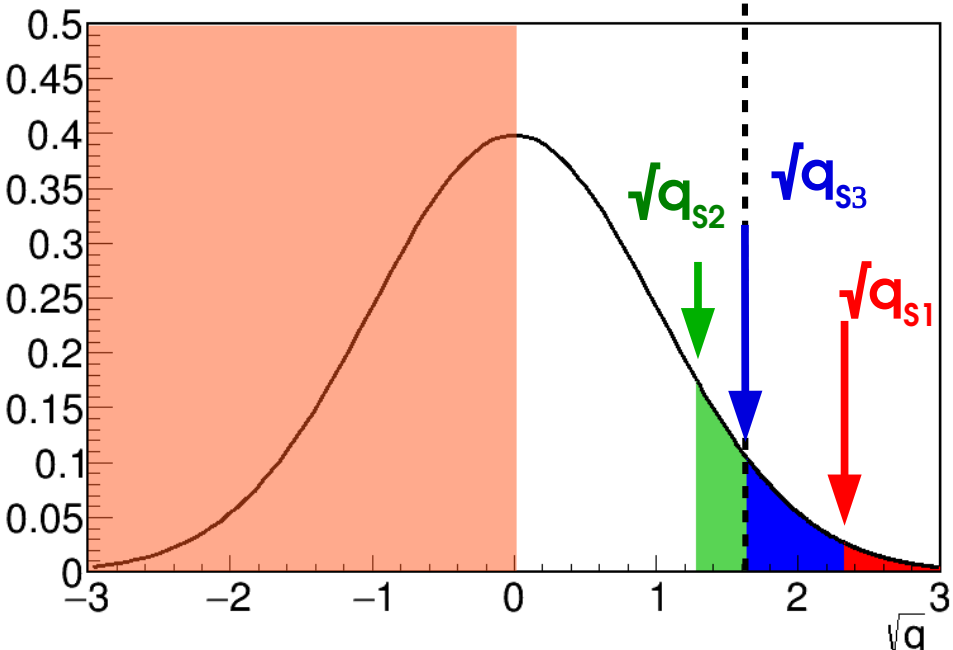
Asymptotics: $$p(S_0) = 1 - \Phi\left(\sqrt{q(S_0)}\right)$$

CL	p	Region
90%	10%	√q(S) > 1.28
95%	5%	√q(S) > 1.64
99%	1%	√q(S) > 2.33

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

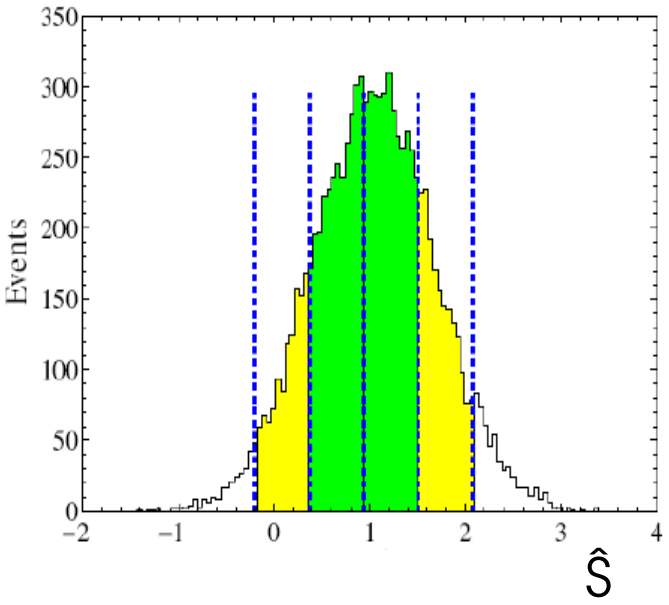


√q(S) = 1.64
(p = 5%)



CL_s : Gaussian Bands

Eur.Phys.J.C71:1554,2011



Usual Gaussian counting example with known B:

95% CL_s upper limit on S:

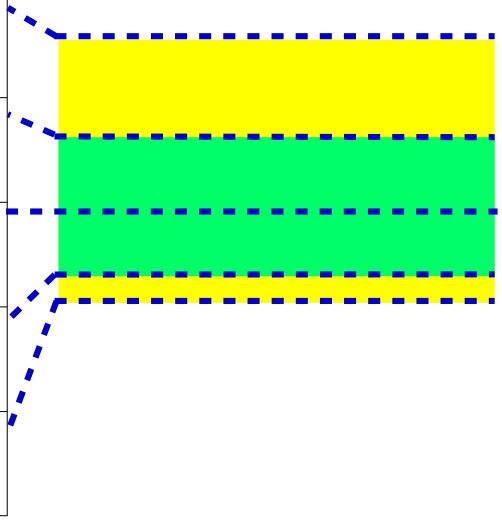
$$S_{up} = \hat{S} + \left[\Phi^{-1}\left(1 - 0.05\ \Phi\left(\hat{S}/\sigma_S\right)\right)\right]\sigma_S \qquad \text{with} \quad \sigma_S = \sqrt{B}$$

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

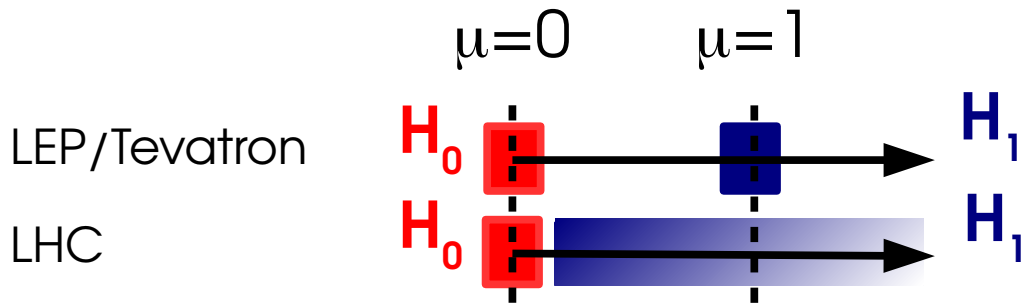
Likelihood ratios are not a new idea:

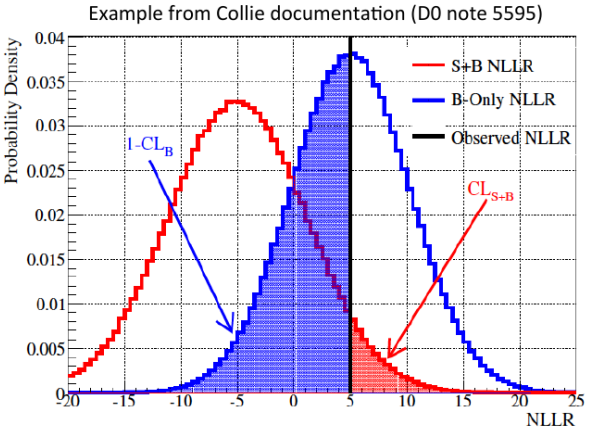
- **LEP**: Simple LR with NPs from MC
  - Compare μ=0 and μ=1
- **Tevatron**: PLR with profiled NPs

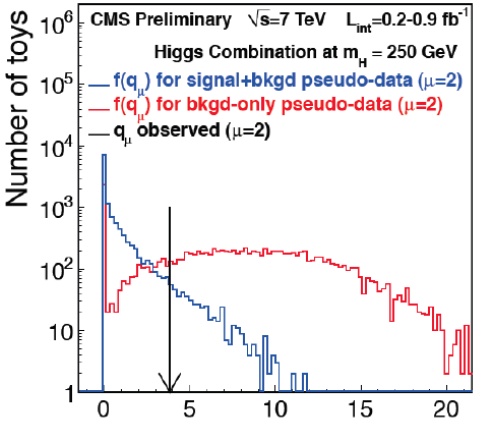
$$q_{LEP} = -2\log\frac{L(\mu=0,\tilde{\theta})}{L(\mu=1,\tilde{\theta})}$$

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

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







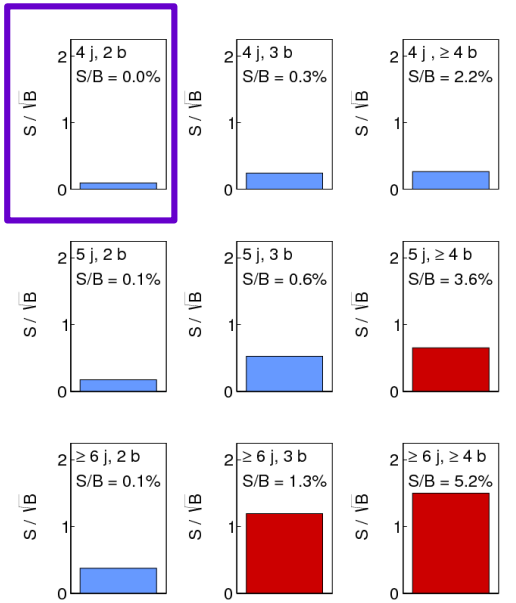
→ Asymptotically:

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

→ Still use TeVatron-style for discrete cases

Systematics and Profiling

Profiling Example: ttH→bb

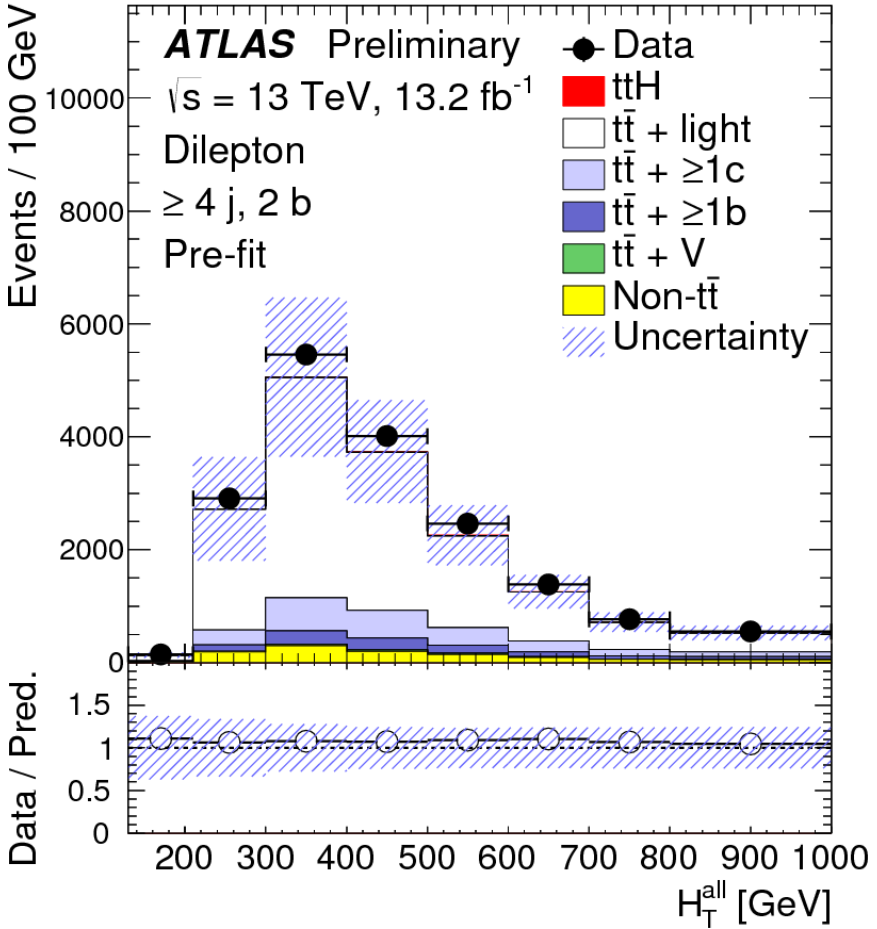


Analysis uses low-S/B categories to constrain backgrounds.

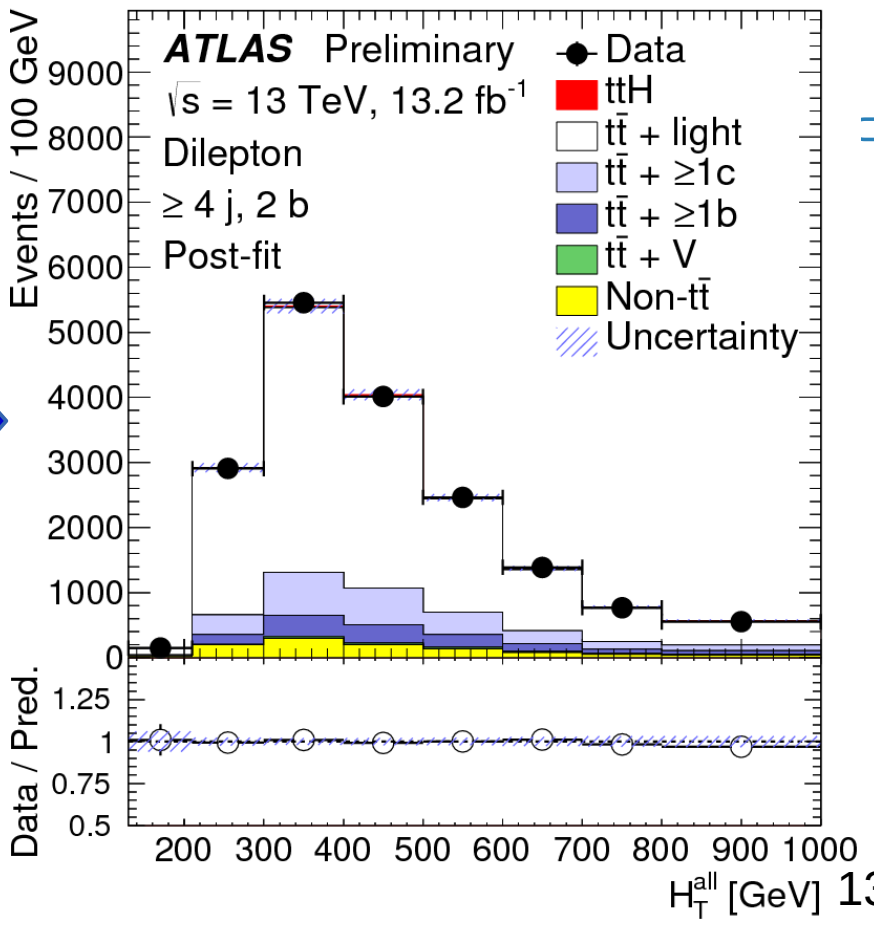
→ **Reduction in large uncertainties on tt bkg**

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

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



Fit

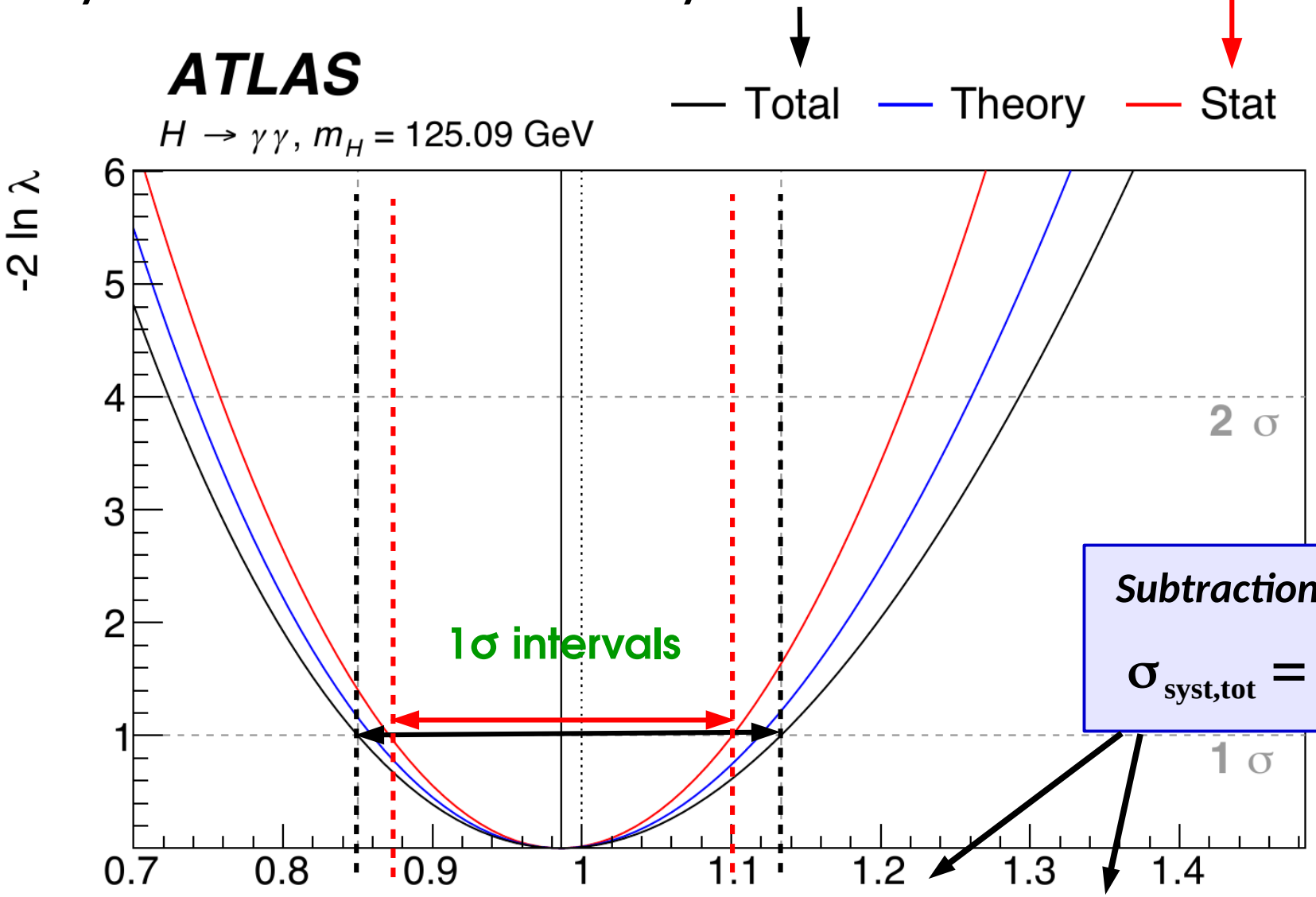


ATLAS-CONF-2016-08

Uncertainty decomposition

All systematics NPs excluded : statistical uncertainty only

All systematics NPs included: stat+syst uncertainties



Subtraction in quadrature

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

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

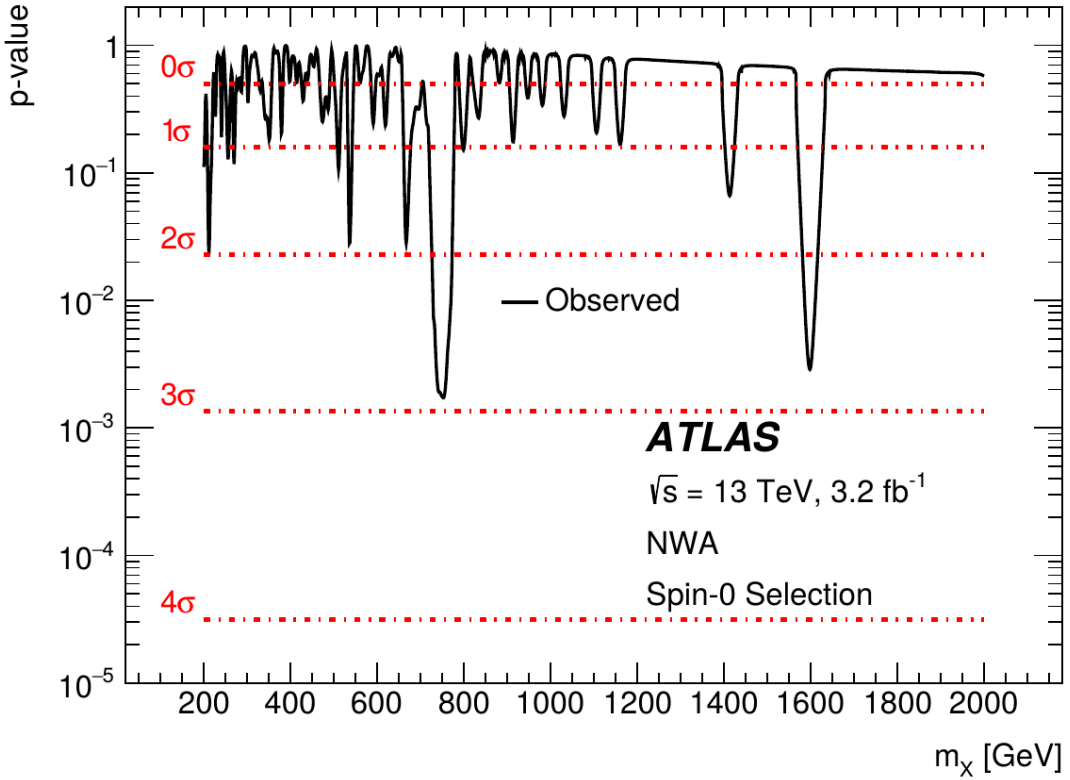
Look-Elsewhere Effect

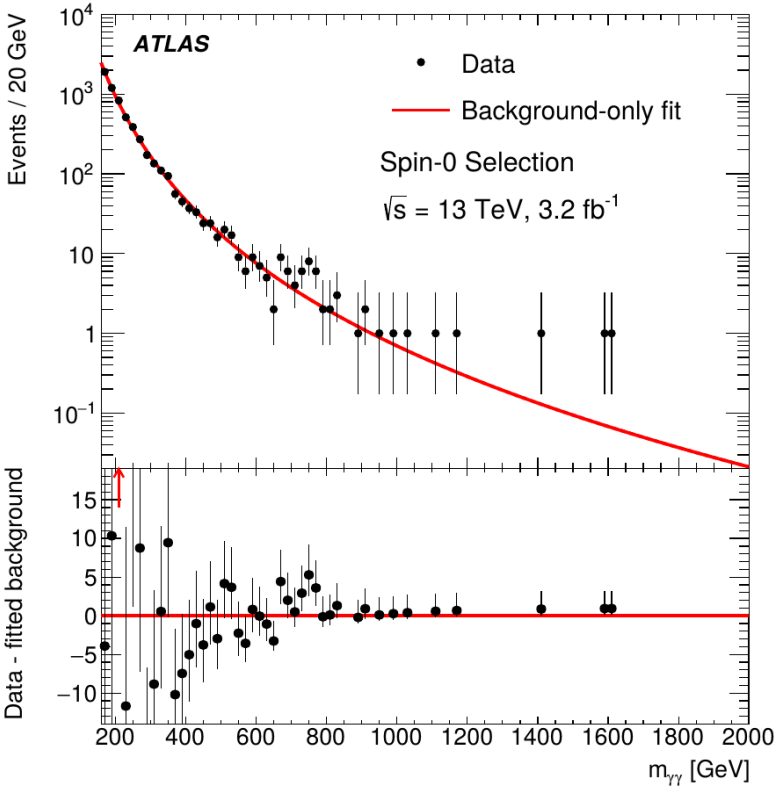
Look-Elsewhere effect

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

⇒ Effectively: **multiple, simultaneous searches**

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





→ More likely to find an excess **anywhere in the range**, rather than in a **predefined** location

⇒ ***Look-elsewhere effect*** (LEE)

Global Significance

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

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

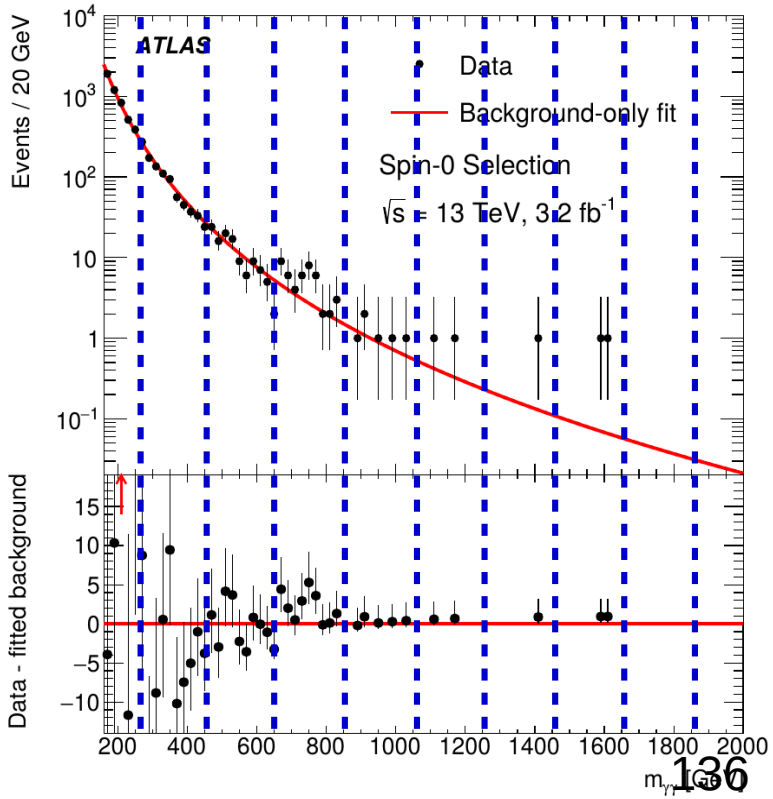
Trials factor

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

Trials factor : **naively** = # of independent intervals:

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

However this is usually **wrong** – more on this later



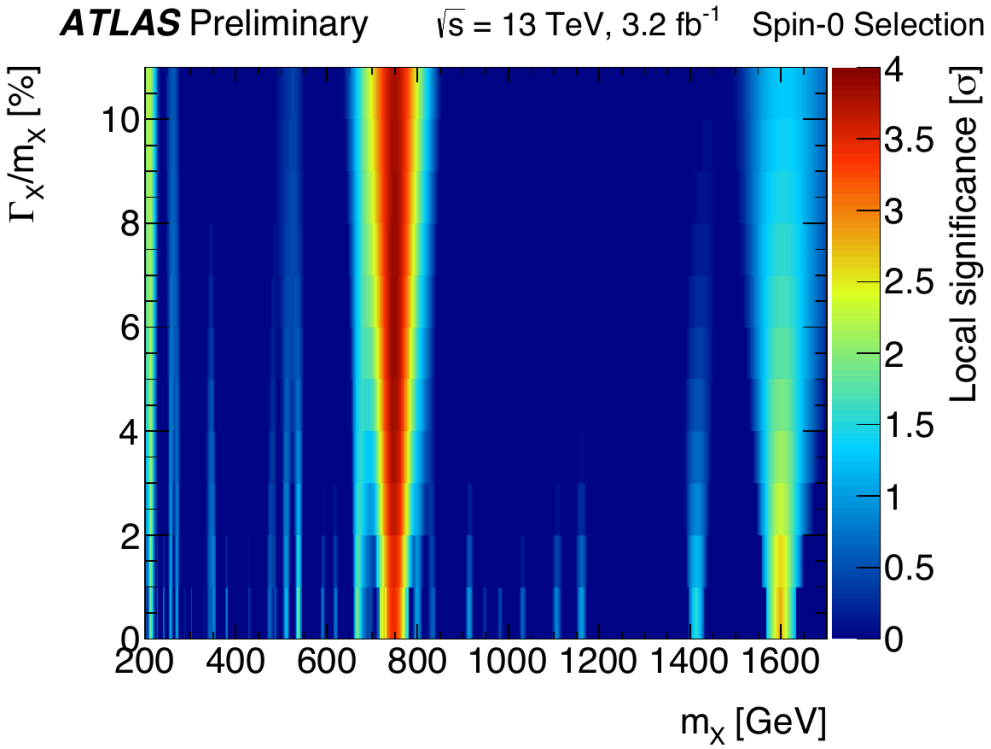
Global Significance

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

For searches over a parameter range, **the global p-value is the relevant one**

→ Accounts for the actual search procedure: look for an excess anywhere in the scanned range

→ Depends on the scanned parameter ranges



e.g. X→γγ :

- $200 < m_X < 2000$ GeV
- $0 < \Gamma_X < 10\% \, m_X$.

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

How to compute p_{global} (or N_{trials}) ?

Trials Factor

Trials factor N = # of independent searches:

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

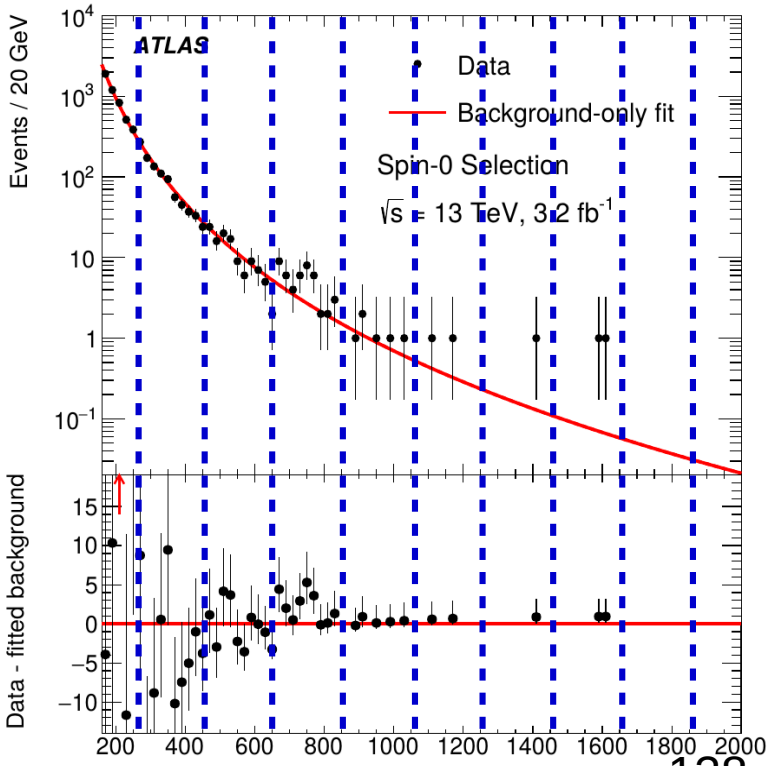
Trials factor

Naively, one could expect

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

However this is only correct for a discrete

Number of experiments (i.e. 10 different regions)



Trials Factor for continuous variables

Gross & Vitells, EPJ.C70:525-530,2010

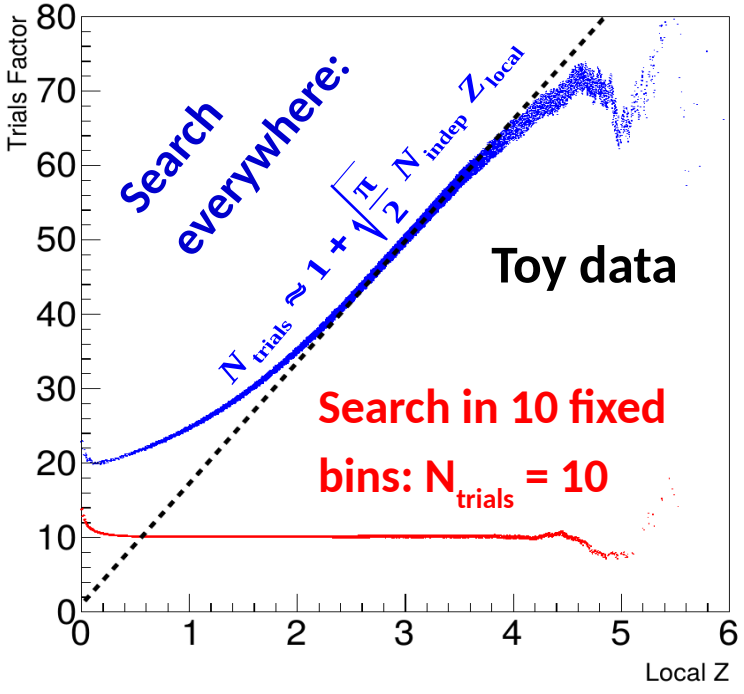
Asymptotic limit : trials factor (1 POI) is

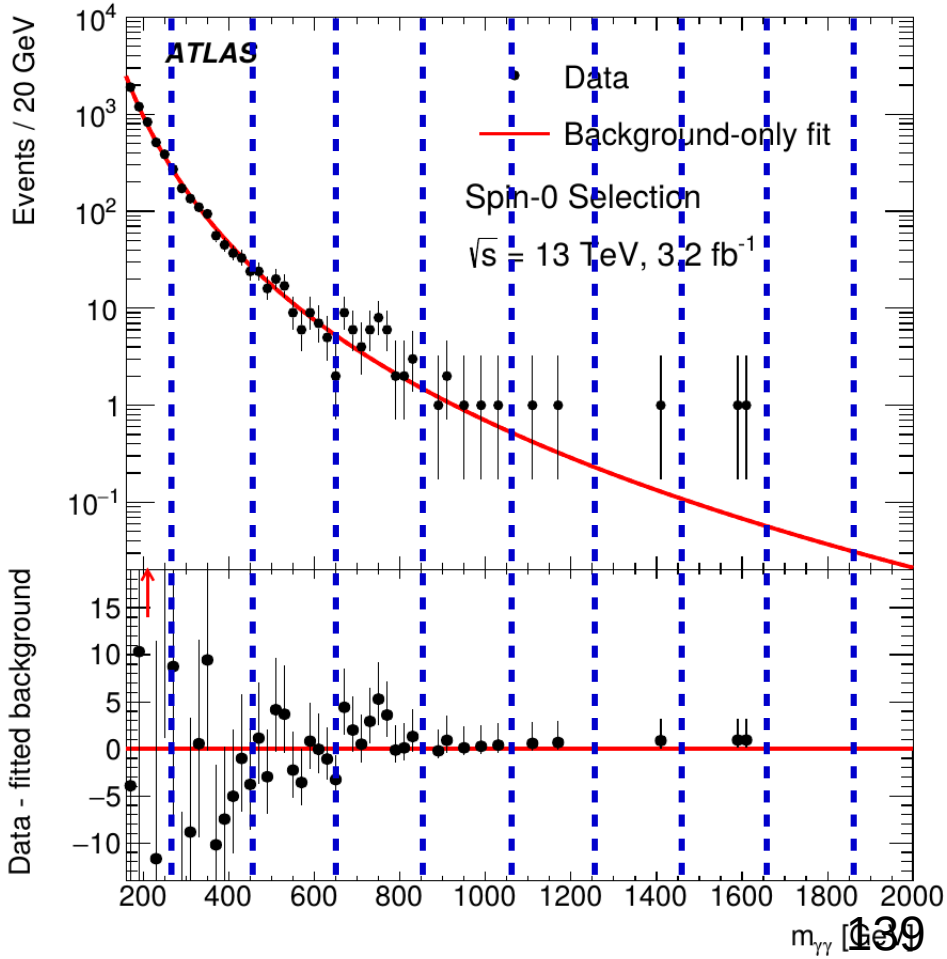
$$N_{\text{trials}} = 1 + \sqrt{\frac{\pi}{2}}\, N_{\text{indep}}\, Z_{\text{local}}$$

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

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

Why ? Slicing range into N_{indep} regions misses peaks sitting on **edges between regions**
⇒ true N_{trials} is **> N_{indep}**!





Trials Factor for continuous variables

Gross & Vitells, EPJ.C70:525-530,2010

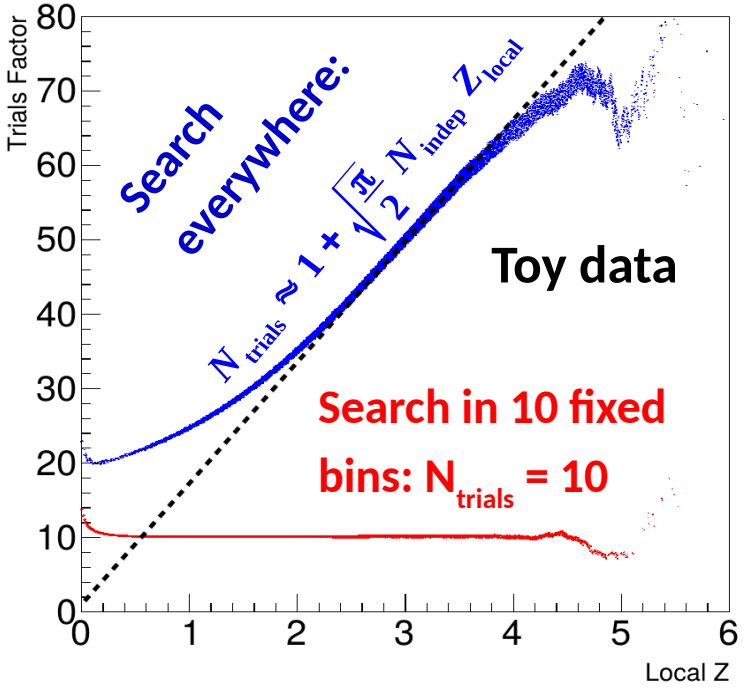
Asymptotic limit : trials factor (1 POI) is

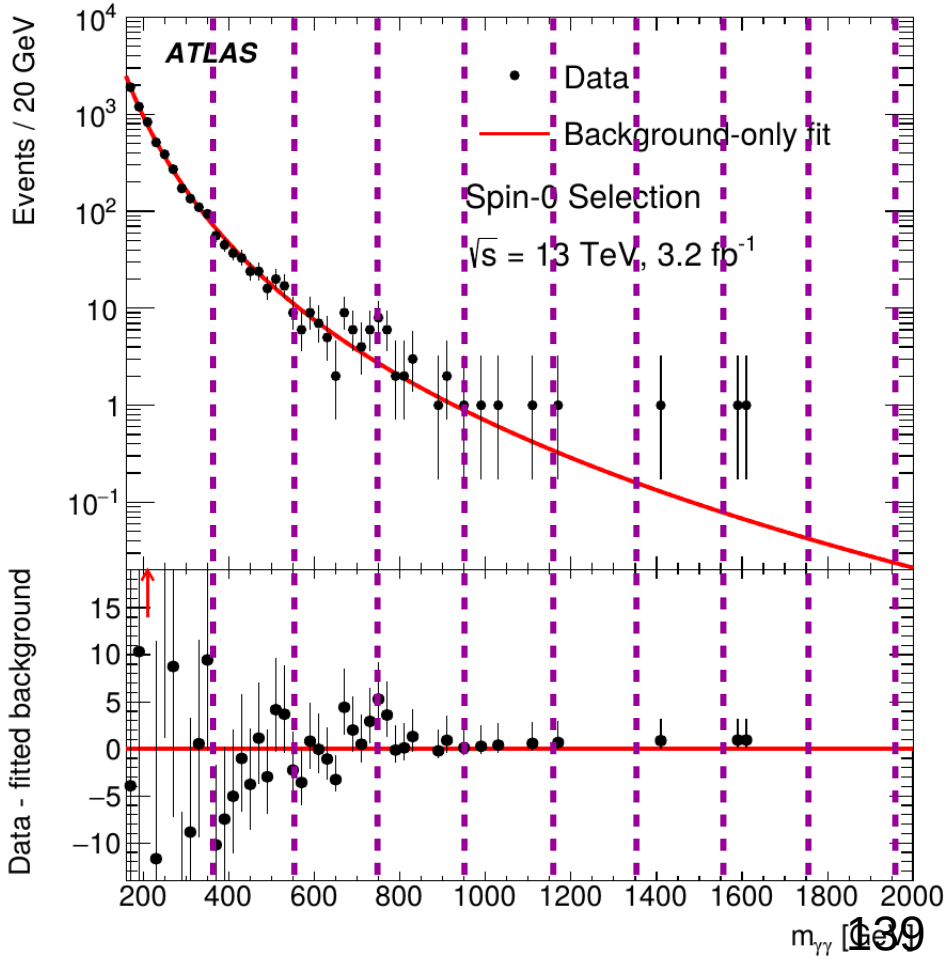
$$N_{\text{trials}} = 1 + \sqrt{\frac{\pi}{2}}\, N_{\text{indep}}\, Z_{\text{local}}$$

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

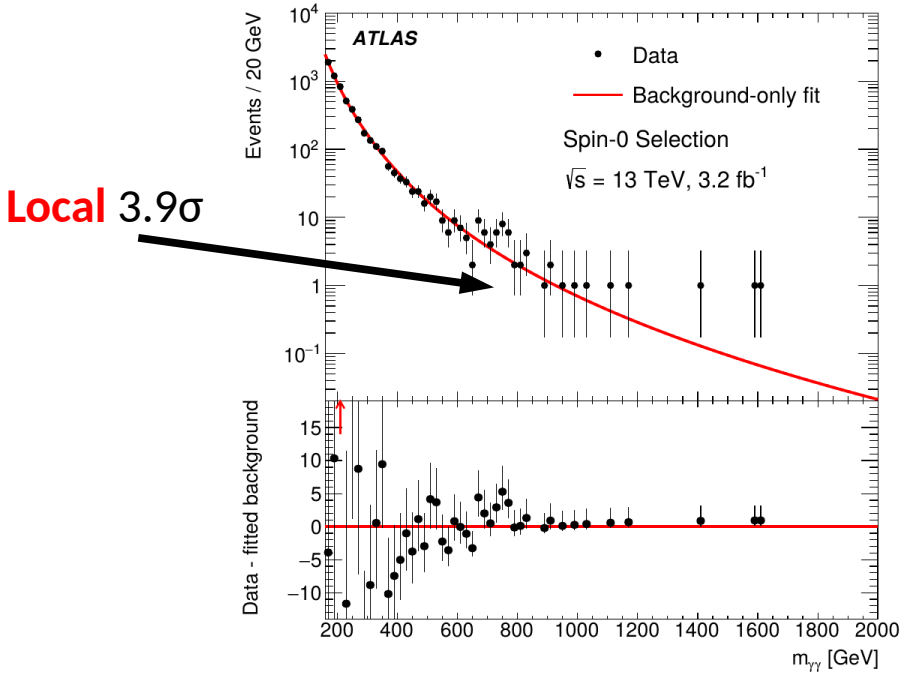
→ Trials factor is **not just N_{indep}**, also depends on Z_{local} !

Why ? Slicing range into N_{indep} regions misses peaks sitting on **edges between regions**
⇒ true N_{trials} is **> N_{indep}**!





Global Significance from Toys



Principle: repeat the analysis in toy data:

→ generate pseudo-dataset

→ perform the search, scanning over parameters as in the data

→ report the largest significance found

→ repeat many times

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

e.g. **X→γγ Search**: Z_{local} **= 3.9σ** (⇒ $p_{local} \sim 5\ 10^{-5}$),

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

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

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

⊕ **Exact treatment**

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

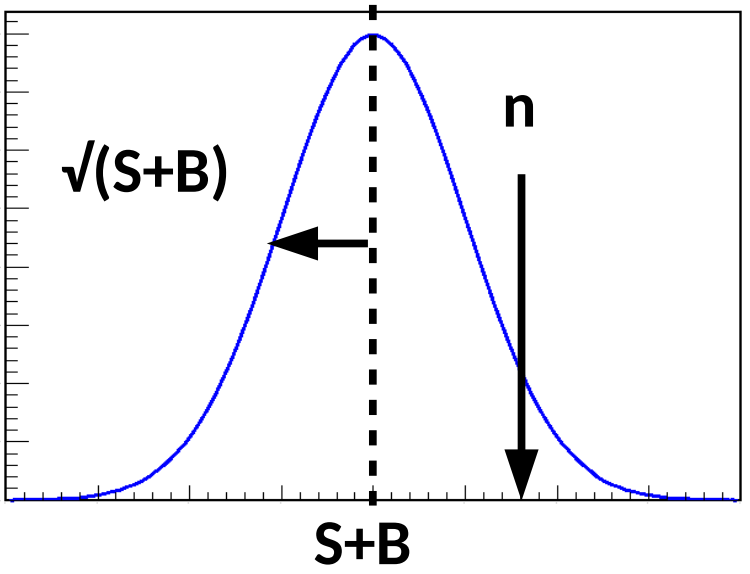
Homework Solutions

Homework 1: Gaussian Counting

Count number of events n in data

→ assume n large enough so process is Gaussian

→ assume B is known, measure S



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

Likelihood :

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

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

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

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

Finally:

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

Famous formula!

→ Strictly speaking only valid in Gaussian regime

Homework 2: Poisson Counting

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

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

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

Test statistic (for $\hat{S} > 0$):

$$q_0 = \lambda(S=0) - \lambda(\hat{S}) = -2\hat{S} - 2(\hat{S}+B)\log\frac{B}{\hat{S}+B}$$

Assuming asymptotic distribution for q_0,

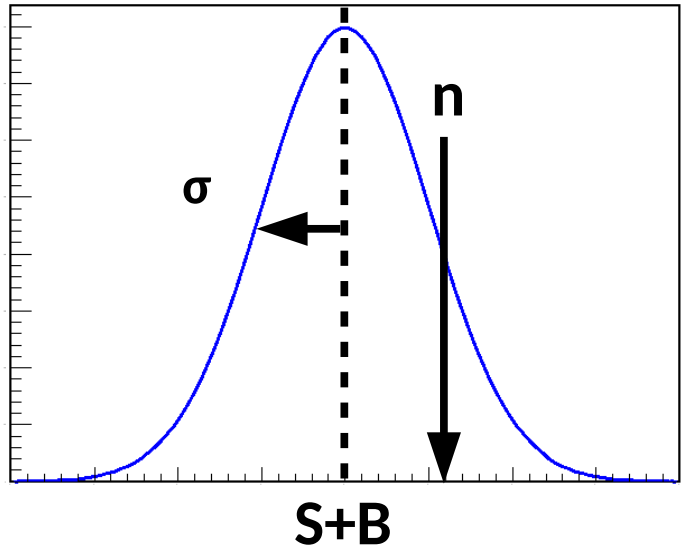
$$Z = \sqrt{2\left[(\hat{S}+B)\log\left(1+\frac{\hat{S}}{B}\right) - \hat{S}\right]}$$

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

Homework 4 : Gaussian CL_{s+b}

Usual Gaussian counting example with known B:

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



Reminder:

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

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

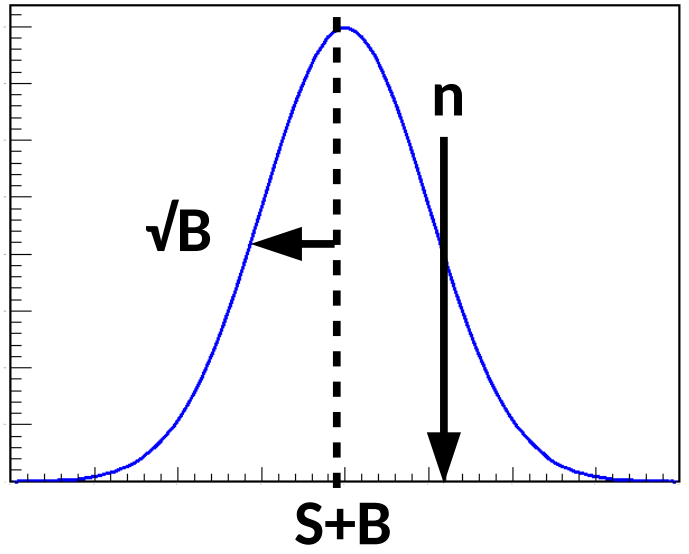
Compute the 95% CL upper limit on S:

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

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

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

Homework 5 : Gaussian CL_s



Usual Gaussian counting example with known B:

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

Reminder

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

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

CL_s upper limit : still have $q_{S_0} = \left| \frac{S_0 - \hat{S}}{\sigma_S} \right|^2$ (for $S_0 > \hat{S}$)

so need to solve

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

for $\hat{S} = 0$,

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

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

Under $H_0(S = S_0)$:

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

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

Under $H_0(S = 0)$:

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

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

Homework 6 : Poisson CL_s

Same exercise, for the Poisson case

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

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

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

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

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

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

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

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

⇒ Asymptotics not valid in this case (n=0) – need to use exact results, or toys

Homework 7: Gaussian Profiling

Counting experiment with background uncertainty: $n = S + \theta$:

→ Signal region: $n \sim G(S + \theta, \sigma_{stat})$

→ Control region: $\theta^{obs} \sim G(\theta, \sigma_{syst})$

$$L(S,\theta) = G(n; S+\theta, \sigma_{stat})\, G(\theta^{obs}; \theta, \sigma_{syst})$$

Then: $$\lambda(S,\theta) = \left(\frac{n-(S+\theta)}{\sigma_{stat}}\right)^2 + \left(\frac{\theta^{obs}-\theta}{\sigma_{syst}}\right)^2$$

For $S = \hat{S}$, matches MLE as it should

MLEs: $\hat{S} = n - \theta^{obs}$ $\quad \hat{\theta} = \theta^{obs}$

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

PLR: $$t_S = -2\log\frac{L(S,\hat{\hat{\theta}}(S))}{L(\hat{S},\hat{\theta})} = \lambda(S,\hat{\hat{\theta}}(S)) - \lambda(\hat{S},\hat{\theta}) = \frac{(S-\hat{S})^2}{\sigma_{stat}^2 + \sigma_{syst}^2}$$

1σ interval $$S = \hat{S} \pm \sqrt{\sigma_{stat}^2 + \sigma_{syst}^2} \qquad \sigma_S = \sqrt{\sigma_{stat}^2 + \sigma_{syst}^2}$$

Stat uncertainty (on n) and systematic (on θ) add in quadrature

Homework 8: CL_s computation

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

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

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

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

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

This boils down to the Gaussian case of HW 6, so the CL_s limit is

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

Homework 8: Bayesian computation

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

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

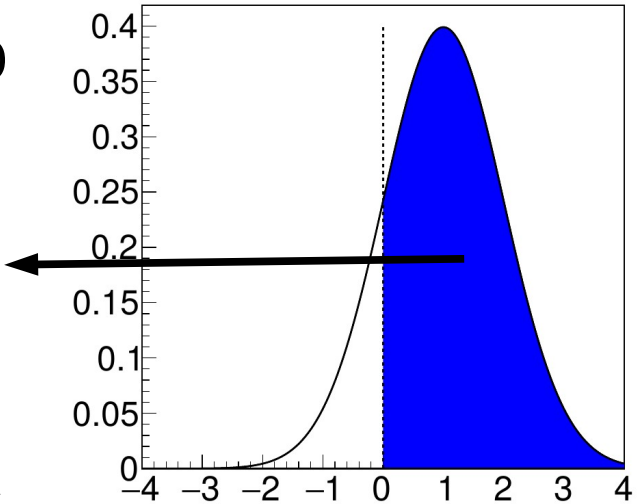
Bayesian: G(θ) is actually a ***prior*** on θ ⇒ perform integral (***marginalization***)

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

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

⇒ Truncate Gaussian at S=0: $P(S \mid n) = P(n \mid S)\; P(S)$

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



Bayesian Limit:

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

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

same result as CL_s!