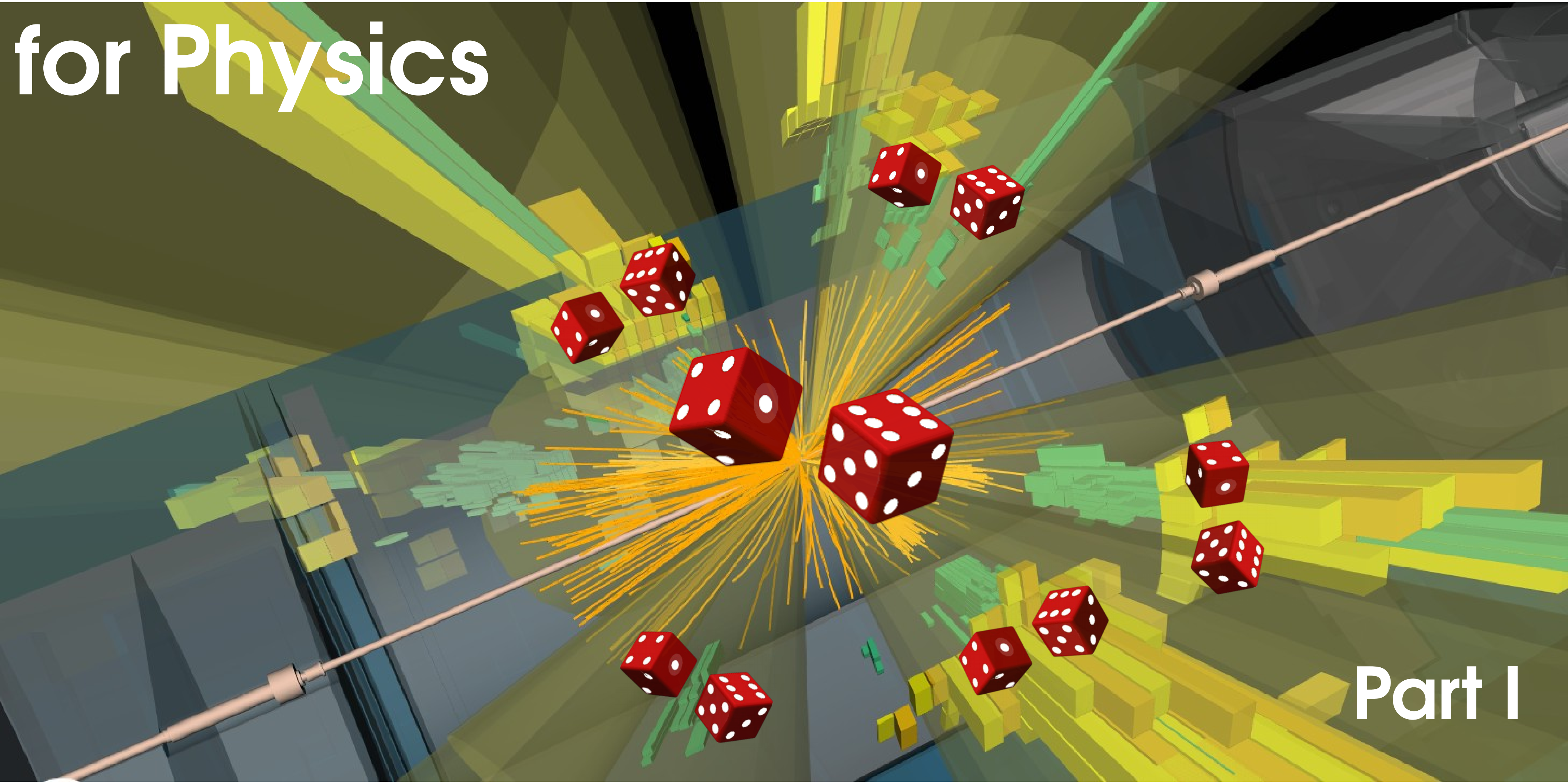
Statistical analysis methods for Physics



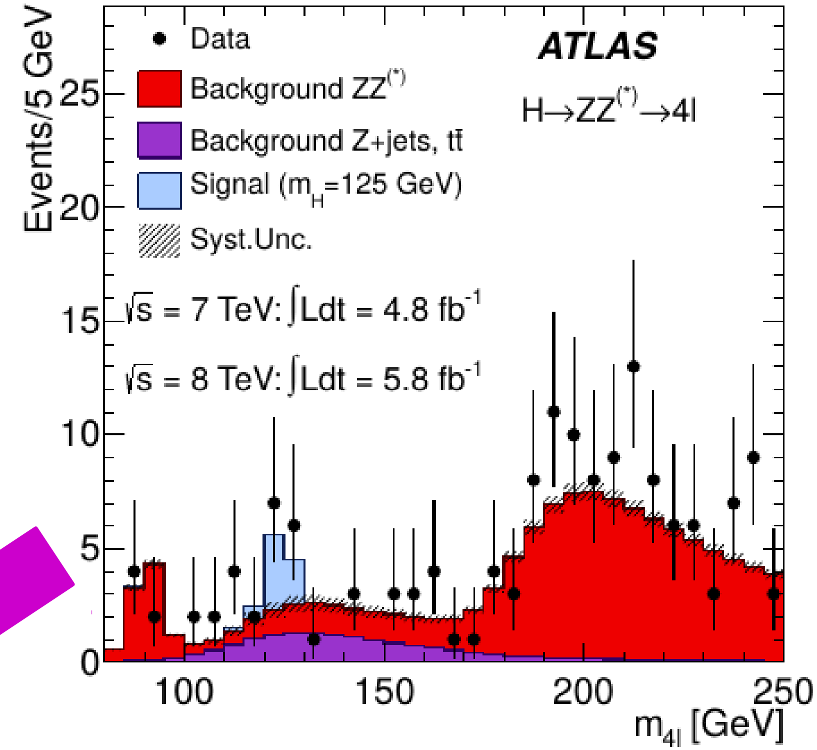
Part I

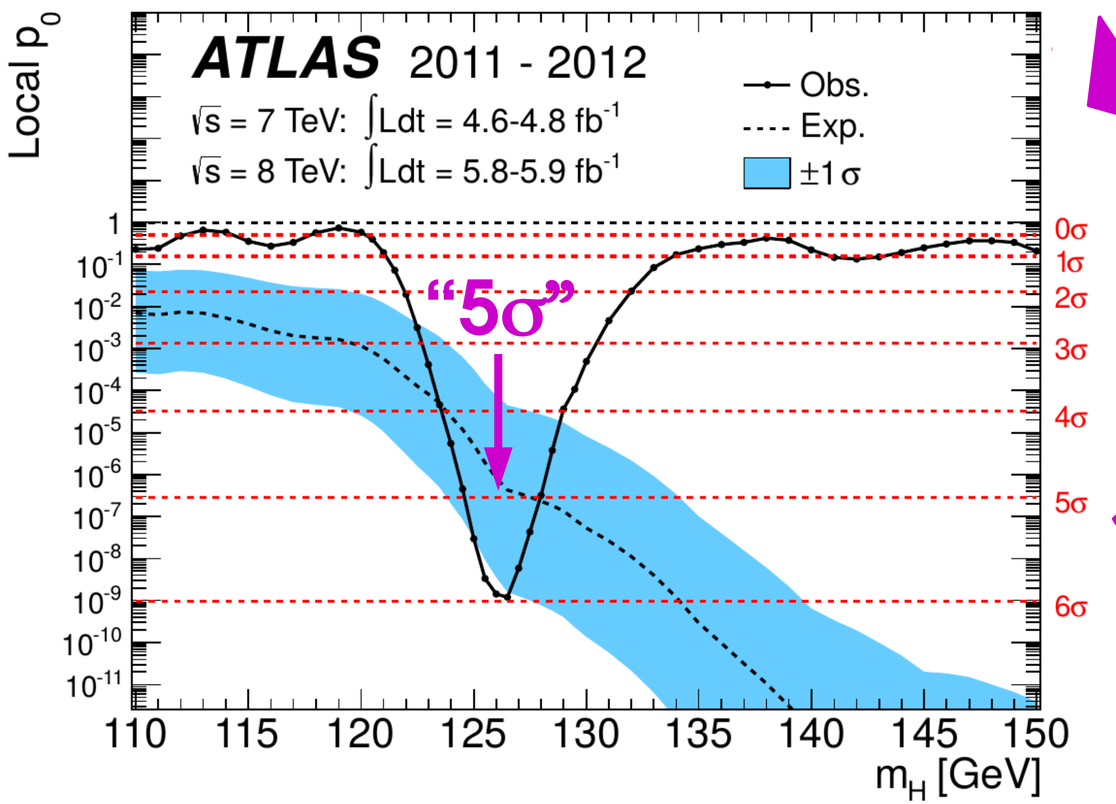
Nicolas Berger (LAPP)

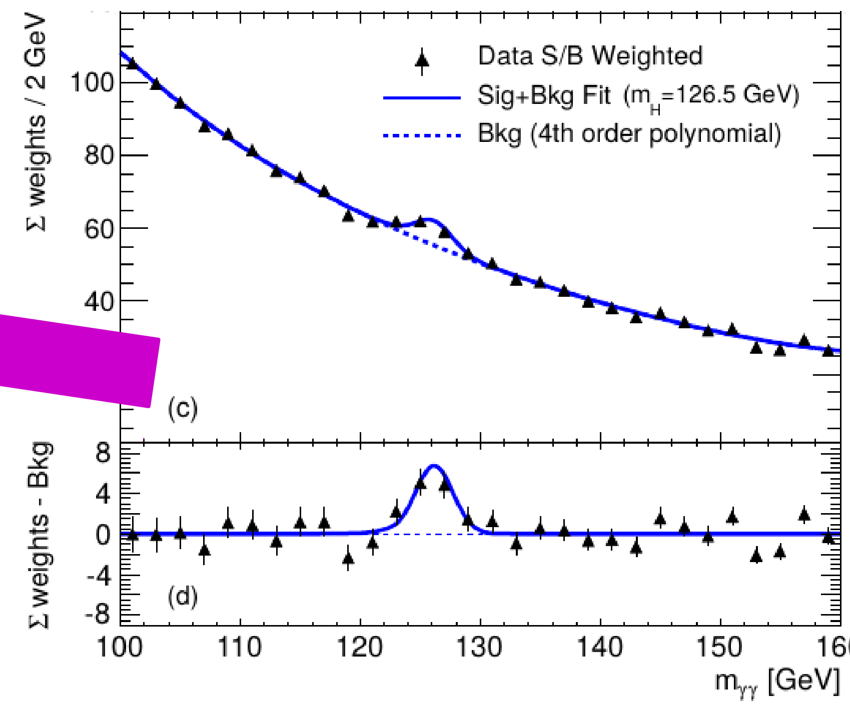
Introduction

Statistical methods play a critical role in many areas of physics

Higgs discovery : **"We have 5σ" !**



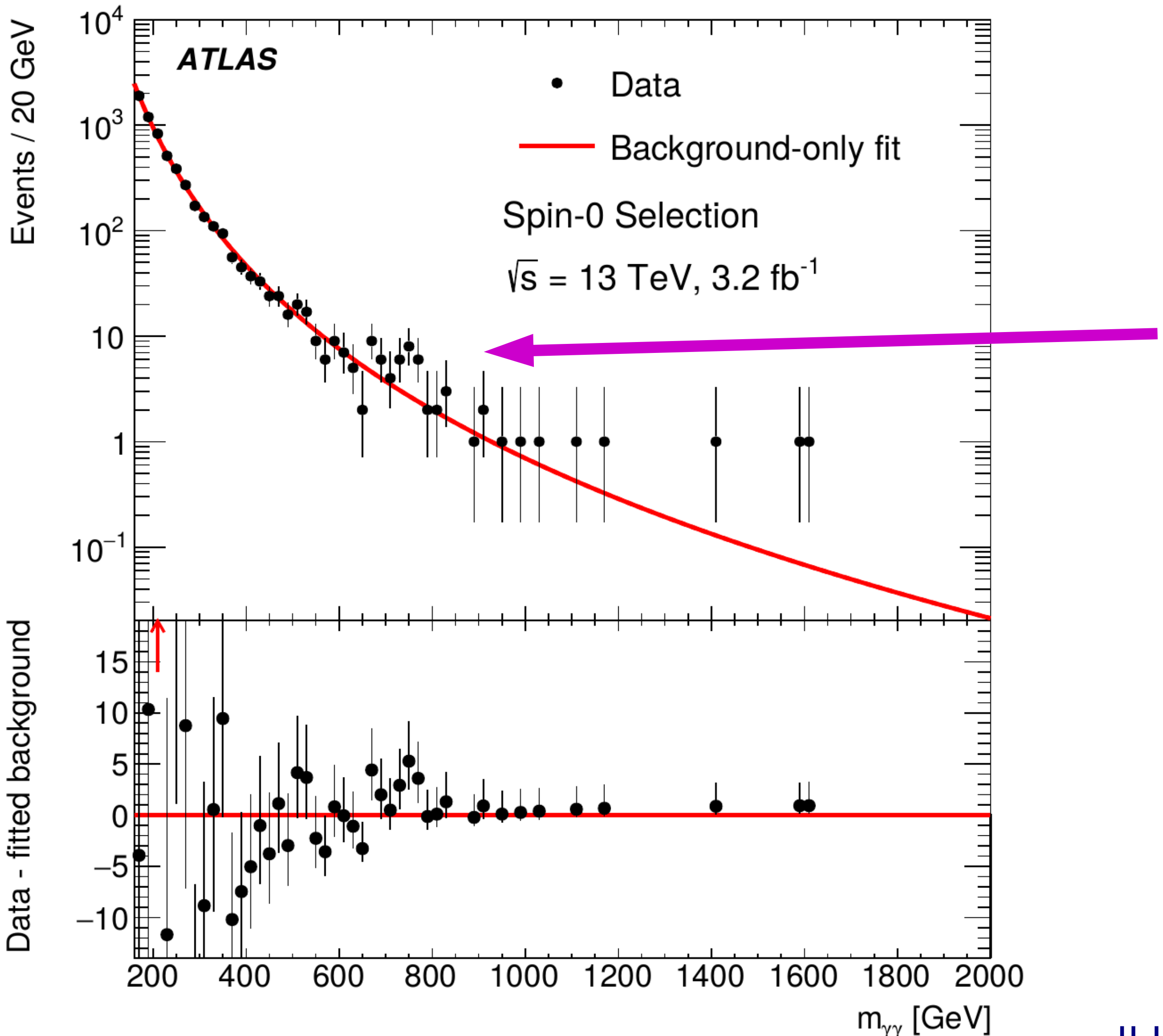




Phys. Lett. B 716 (2012) 1-29

Introduction

Sometimes difficult to distinguish a bona fide discovery from a **background fluctuation**…



New Physics ?

~~3.9σ !?~~ ... 2.1σ

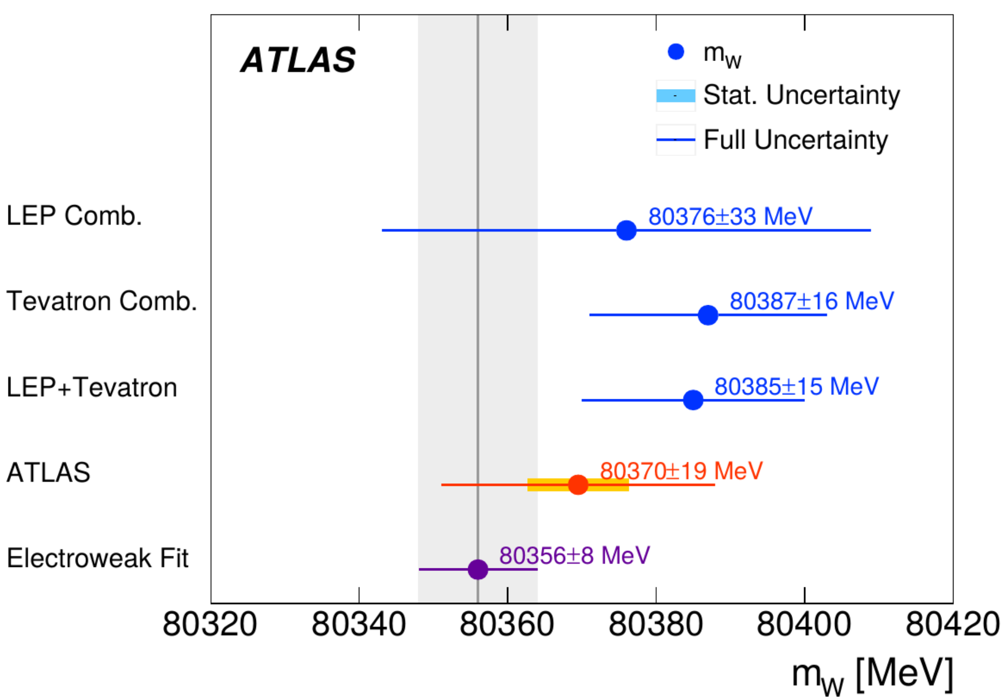
JHEP 09 (2016) 1

Uncertainties

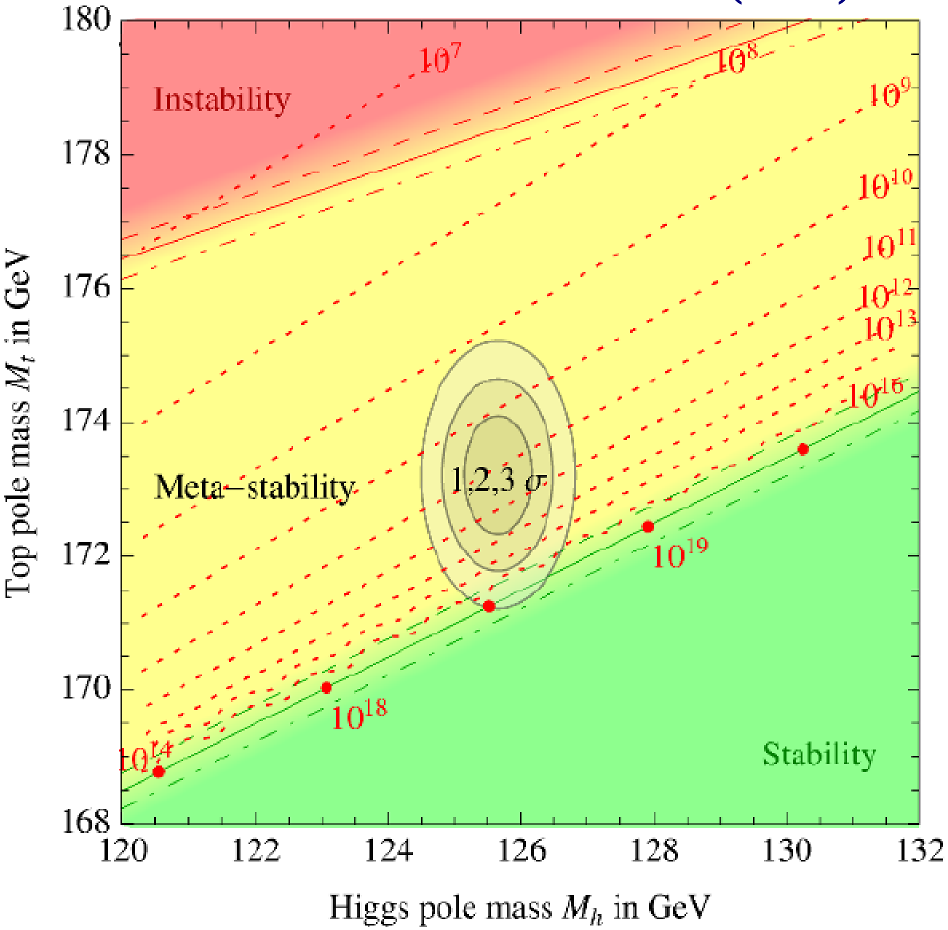
Many important questions answered by **precision measurements**, especially if no new peaks found at high mass…

Key point = determination of **uncertainties**

arXiv:1701.07240



JHEP12(2013)089



Consistency of the SM...

… or the fate of the universe

Overview

Topics covered:

- Computing statistics results
- Interpreting statistical results
- Understanding the measurement process (what is a systematic ?)

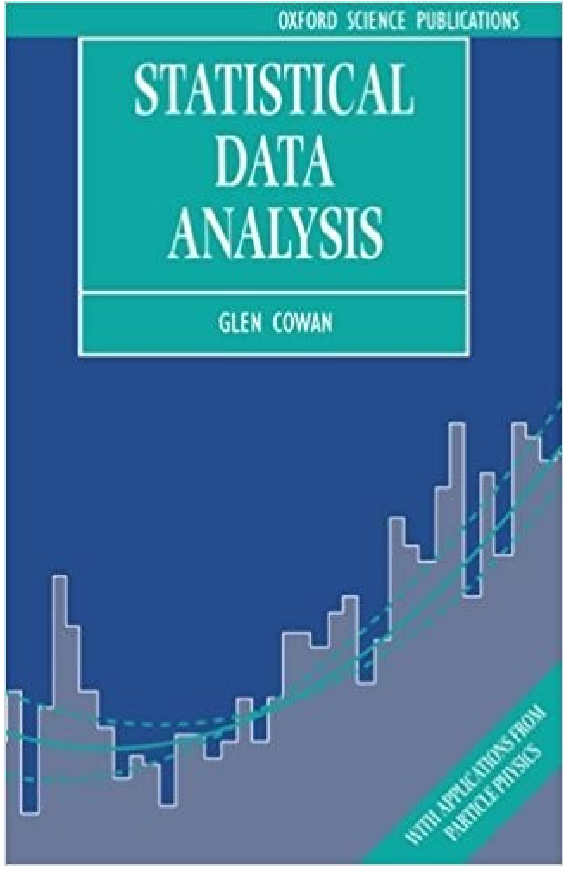
Prerequisites:

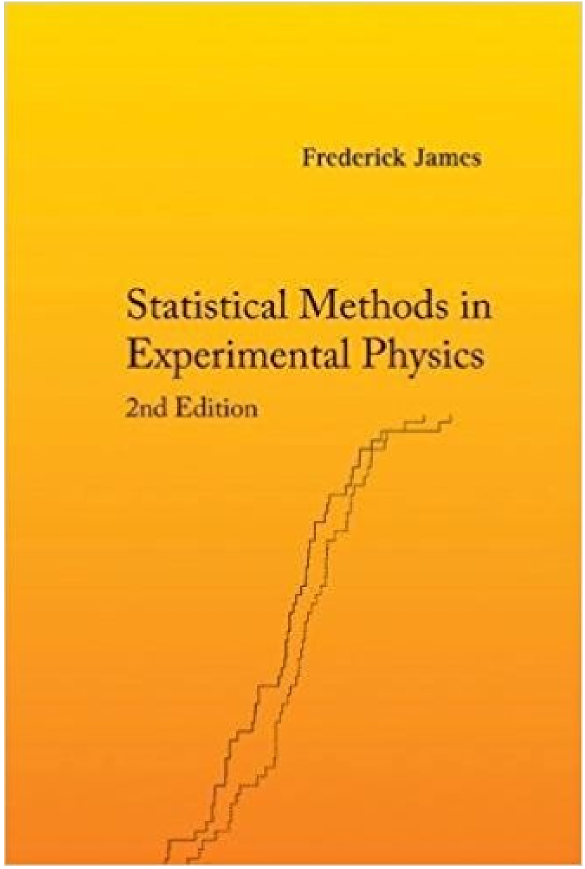
- Some background in High energy physics
- Some basic knowledge of statistics – but will review the basics.

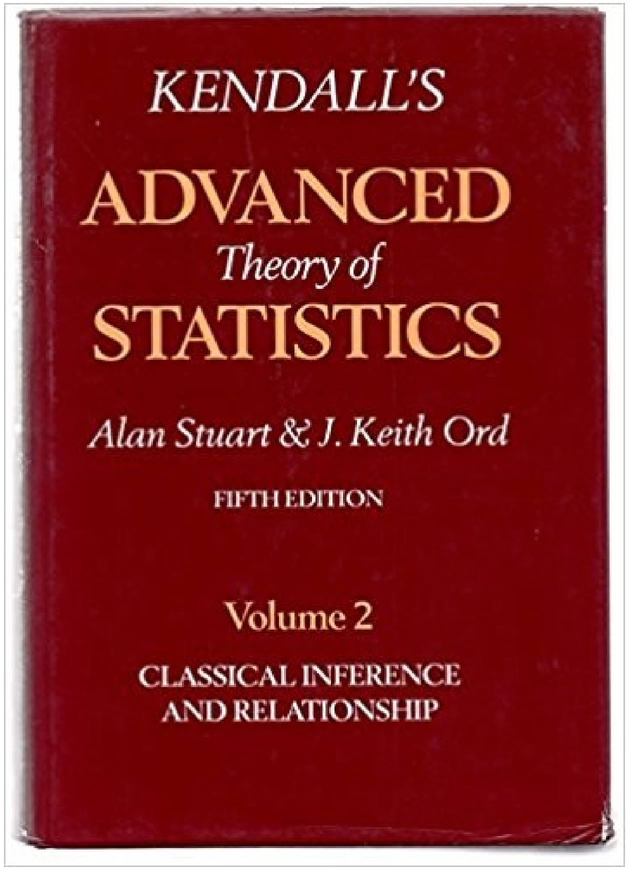
I will mostly use the "physics" names of statistical quantities, rather than those used in the statistics community ("significance" and not "size of a test", etc.)

Much of the discussion and examples have an ATLAS/CMS/LHC slant due to my limited experience… But hopefully the concepts should be generally applicable.

Books and Courses







Some 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

Outline

Statistics basics for HEP

Random processes

Probability distributions

Describing HEP measurements

Computing statistics results

Likelihoods

Estimating parameter values

Lecture 2: Testing hypotheses, Computing discovery significances, Limits

Lecture 3: Look-elsewhere effect, Profiling, Bayesian methods

Random Processes

Random Processes

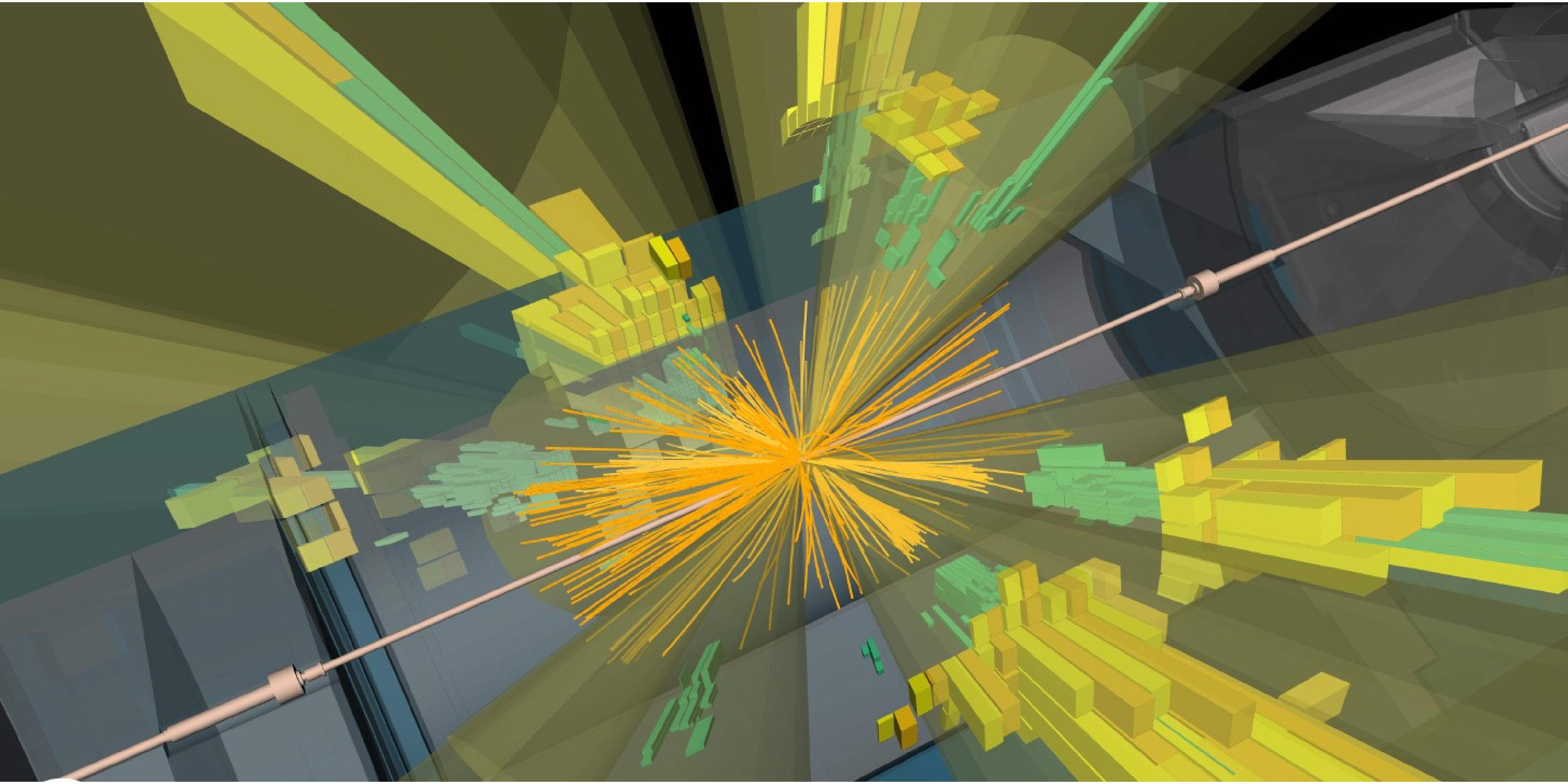
Statistics is the description of **random** processes. Where does this come into physics results ?

Measurement errors

Quantum Randomness

Randomness in High-Energy Physics

Experimental data is produced by incredibly complex processes



Randomness in High-Energy Physics

Experimental data is produced by incredibly complex processes

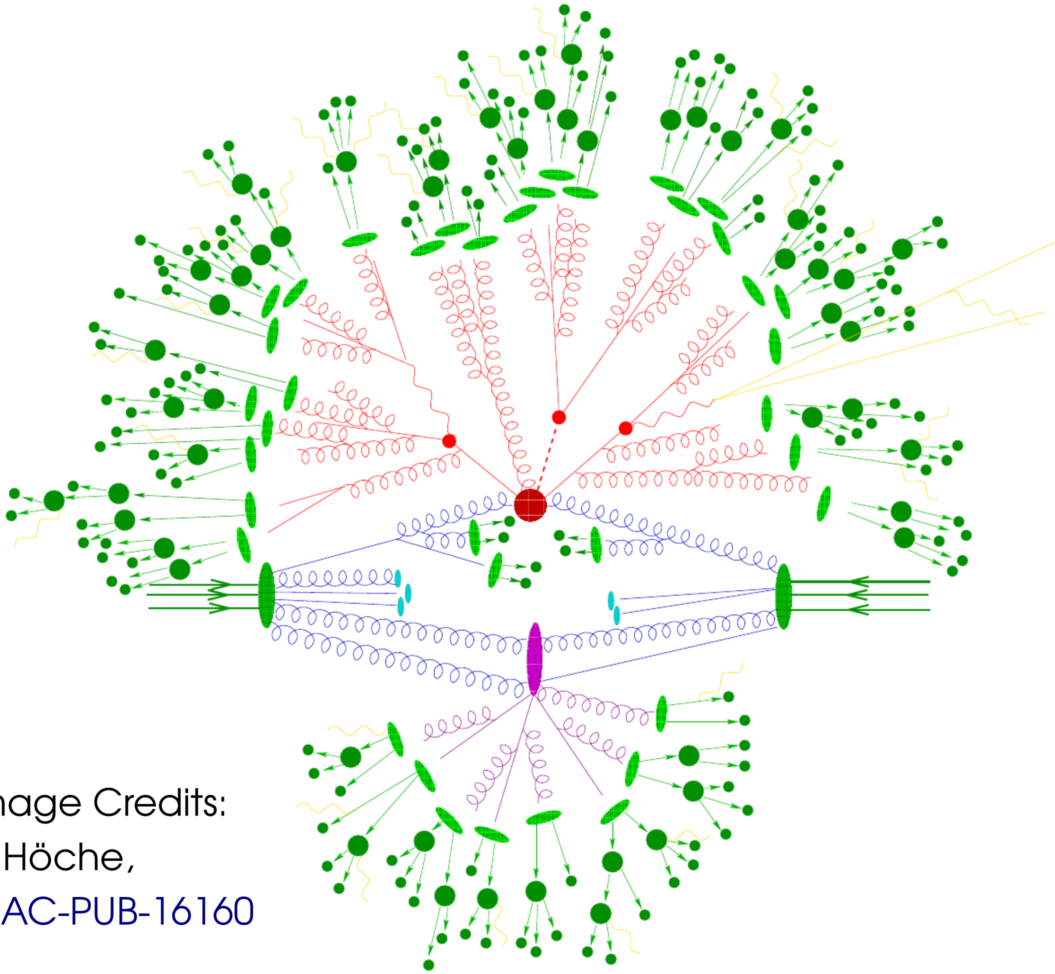


Image Credits:
S. Höche,
SLAC-PUB-16160

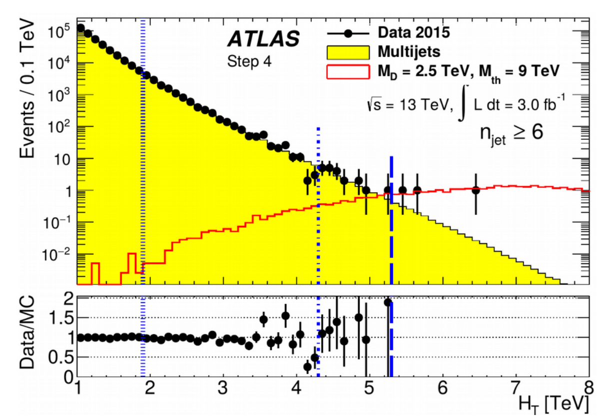
Hard scattering

PDFs, Parton shower, Pileup

Decays

Detector response

Reconstruction



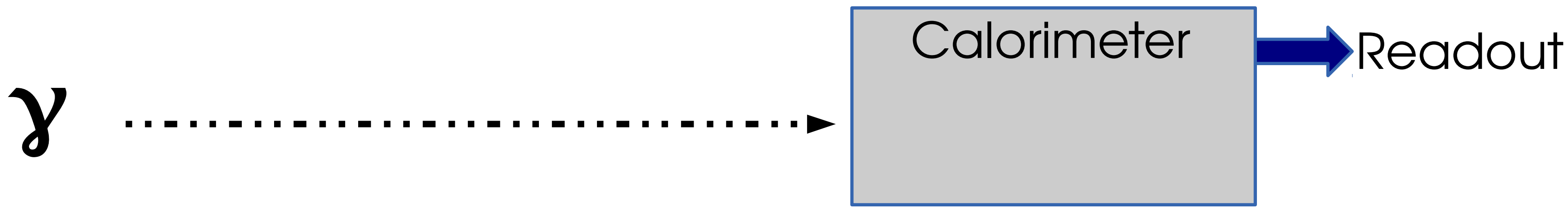
Randomness involved in all stages

→ **Classical** randomness: detector reponse

→ **Quantum** effects in production, decay

Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

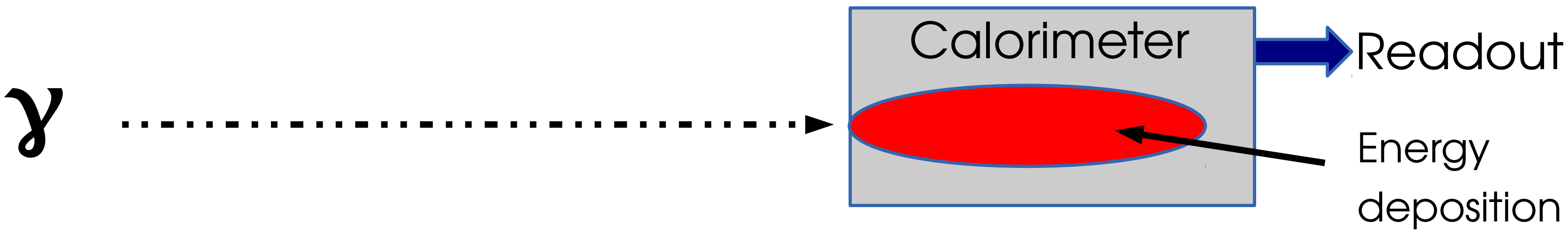


Cannot predict the measured value for a given event ⇒ **Random process**

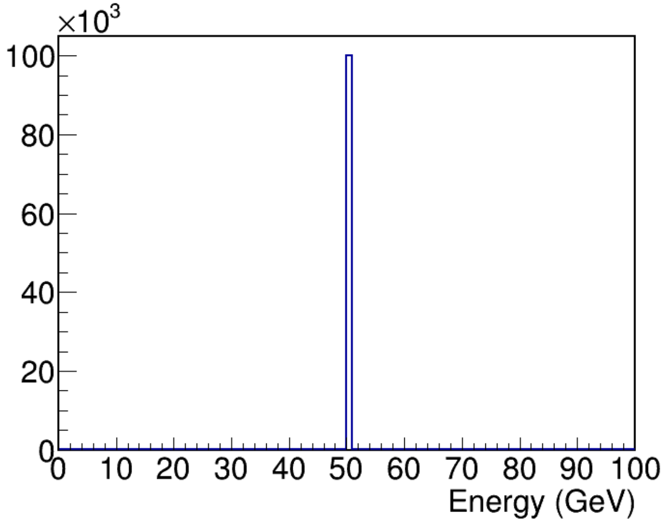
⇒ **Need a probabilistic description**

Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter



Perfect case

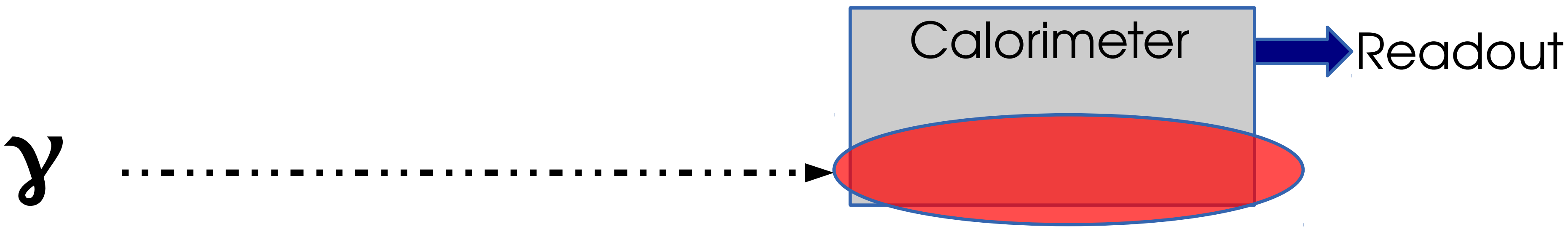


Cannot predict the measured value for a given event ⇒ **Random process**

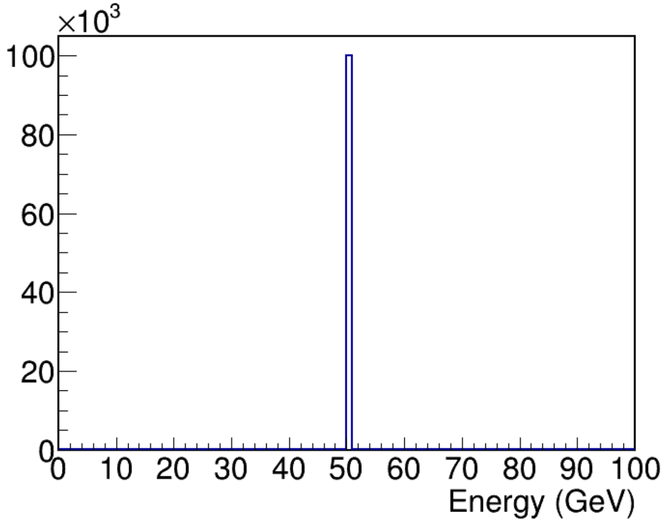
⇒ **Need a probabilistic description**

Measurement Errors: Energy measurement

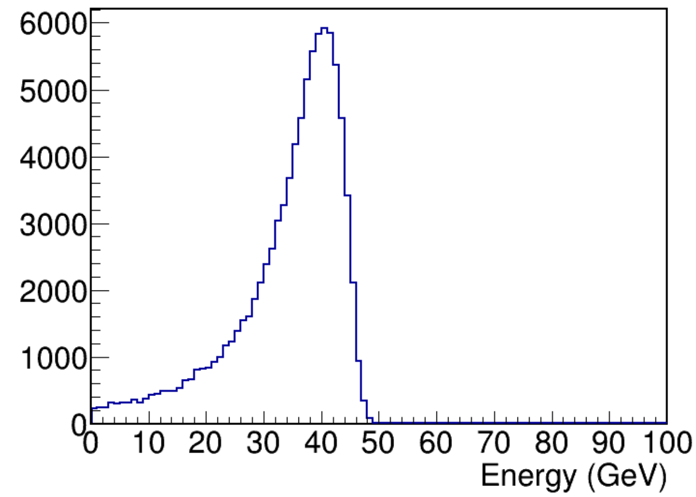
Example: measuring the energy of a photon in a calorimeter



Perfect case



Real life

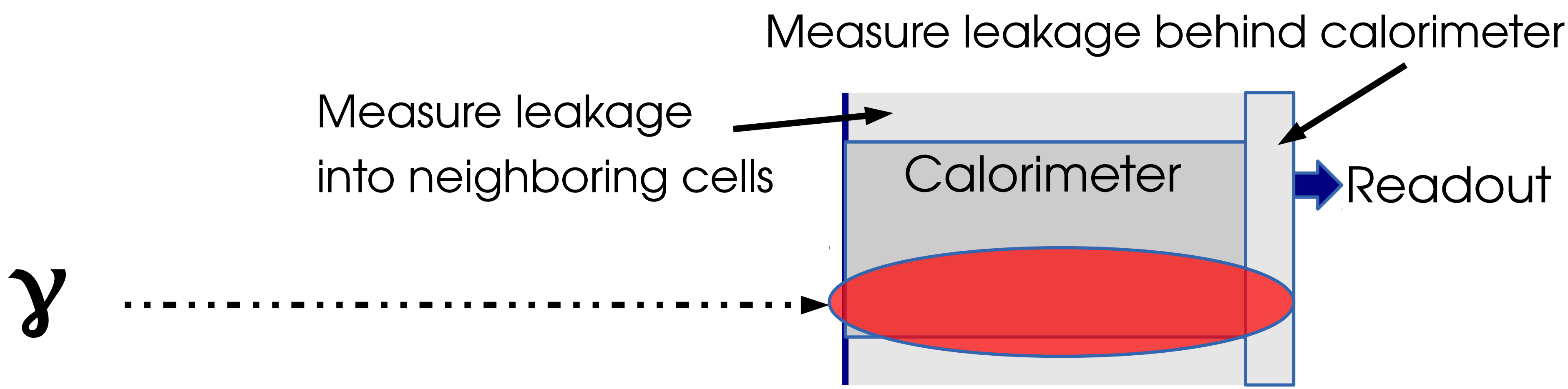


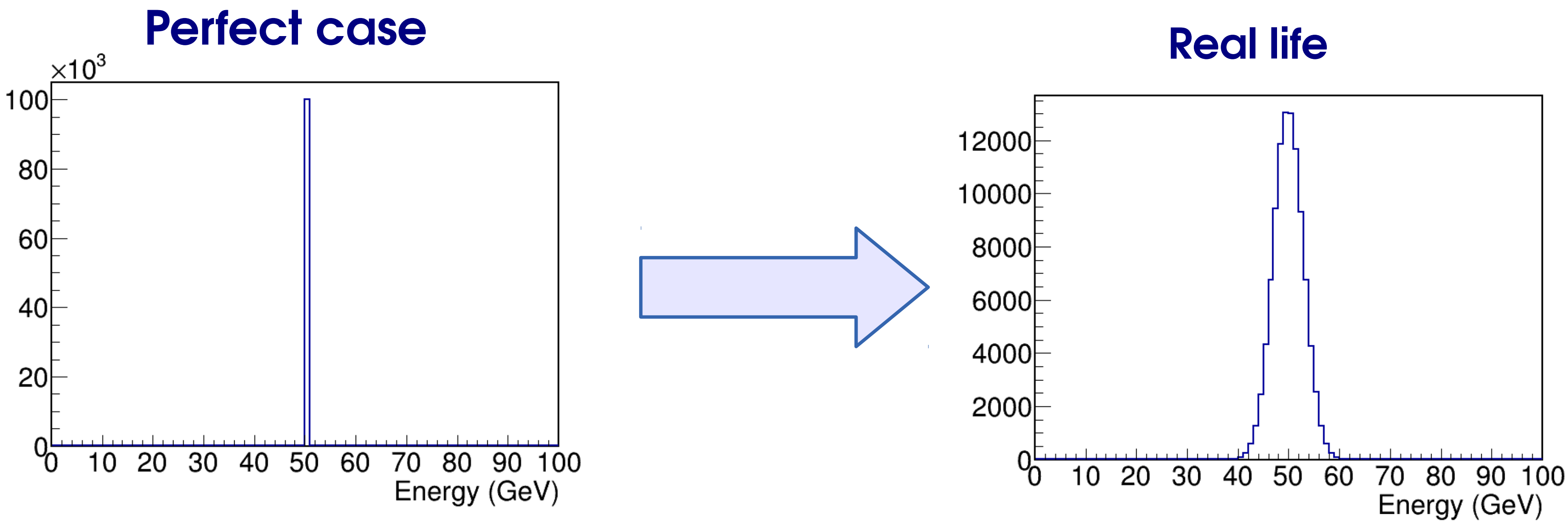
Cannot predict the measured value for a given event ⇒ **Random process**

⇒ Need a probabilistic description

Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter



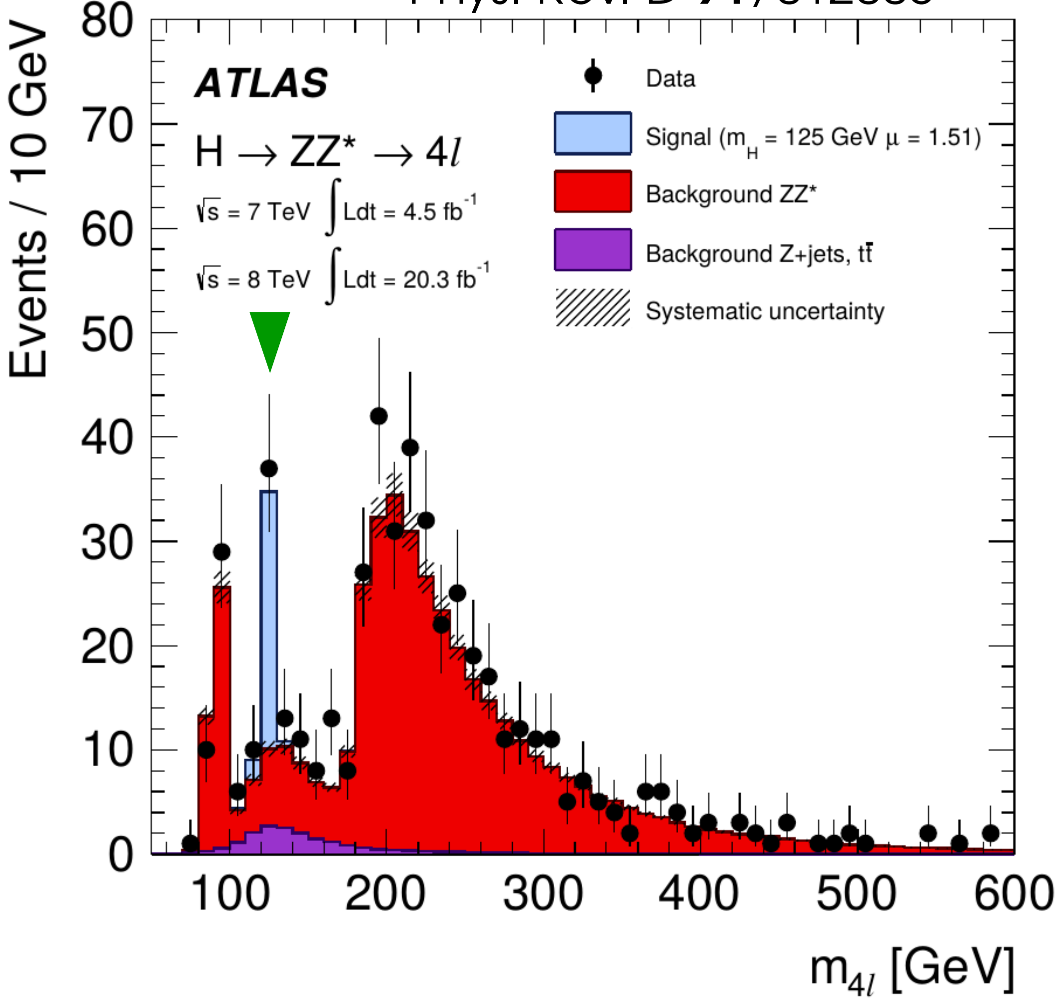


Cannot predict the measured value for a given event ⇒ **Random process**

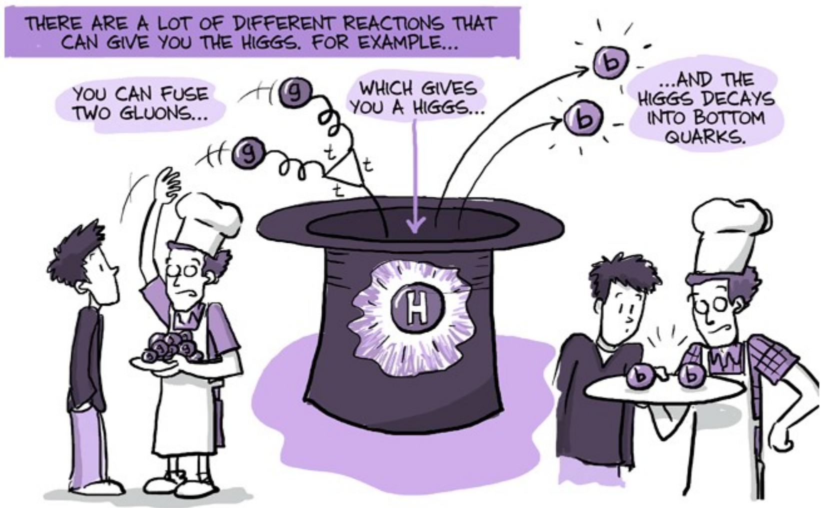
⇒ **Need a probabilistic description**

Quantum Randomness: H→ZZ*→4l

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



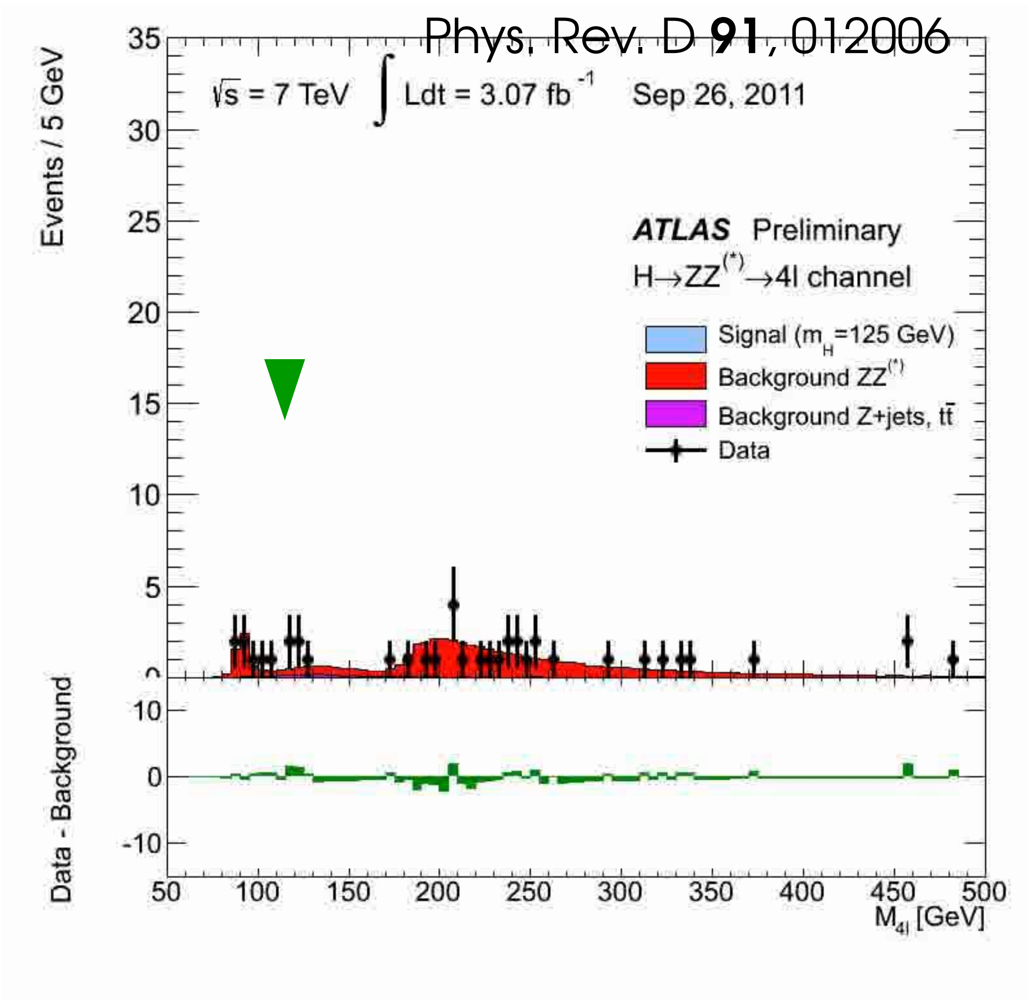
Rare process: Expect 1 signal event every **~6 days**



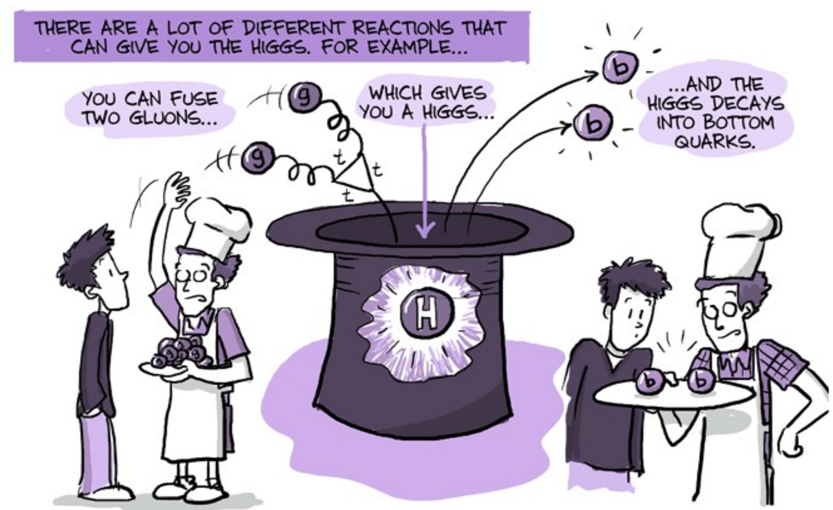
http://www.phdcomics.com/comics/archive.php?comicid=1489

Quantum Randomness: H→ZZ*→4l

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



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

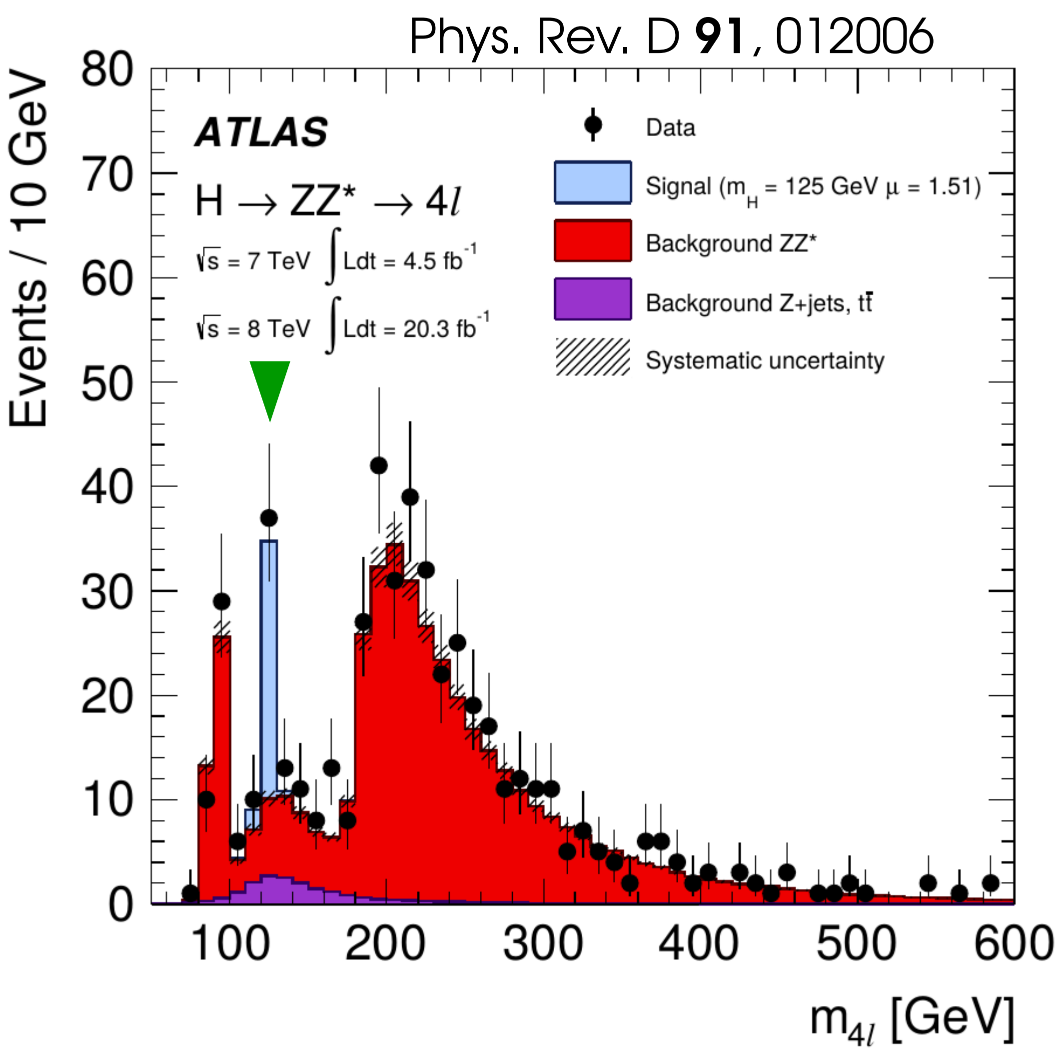


http://www.phdcomics.com/comics/archive.php?comicid=1489

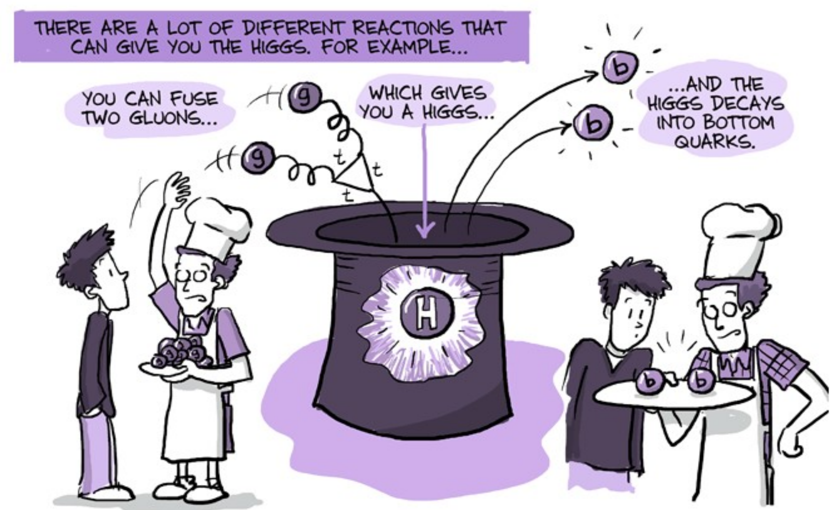
View online

Quantum Randomness: H→ZZ*→4l

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



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

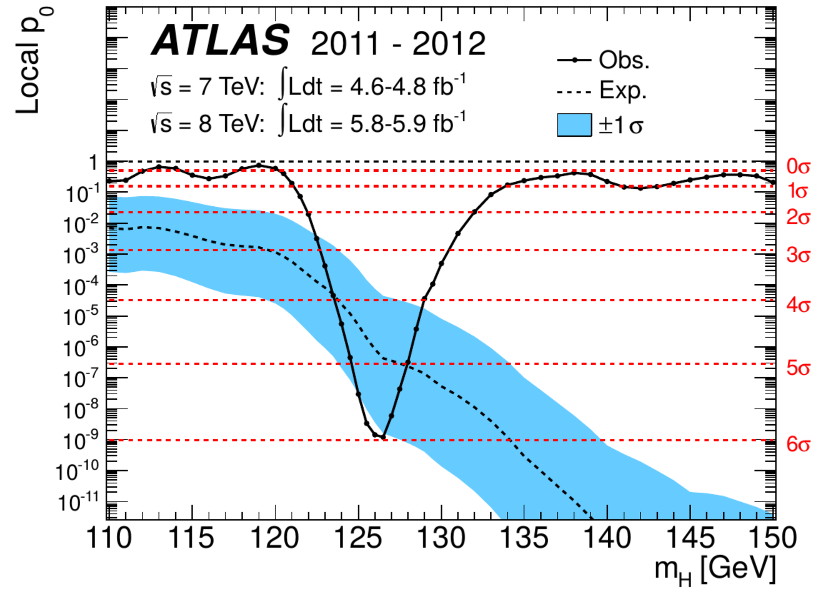


http://www.phdcomics.com/comics/archive.php?comicid=1489

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

Randomness in Physics

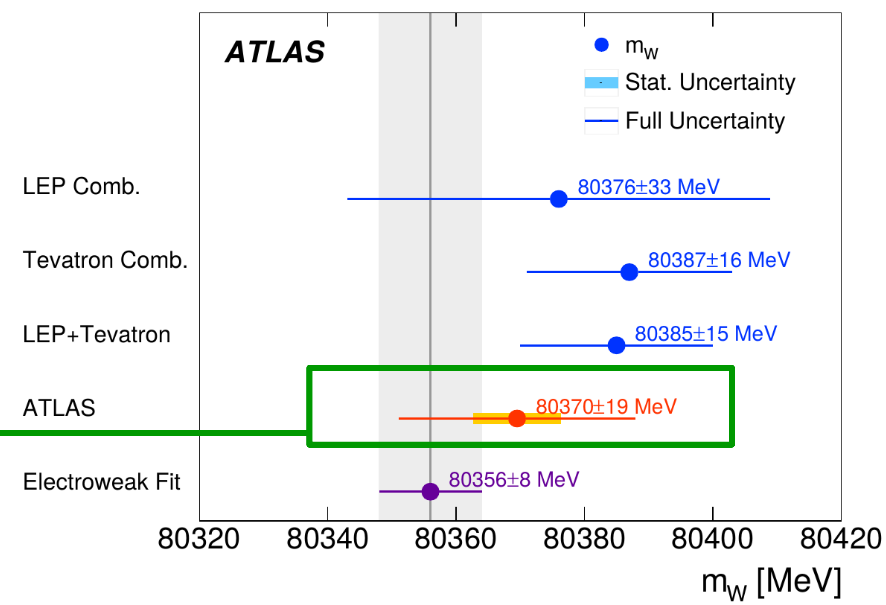
Questions with probabilistic answers:



- **Is my Higgs-like excess just a background fluctuation?**
  → associated with prob ~10^{-9} (by now ~10^{-24})
  ⇒ above the famous (and conventional) ***5σ***

- For measurements: probability that the **true value** of a parameter is within an interval:

68% chance that the true m_W is within the orange interval



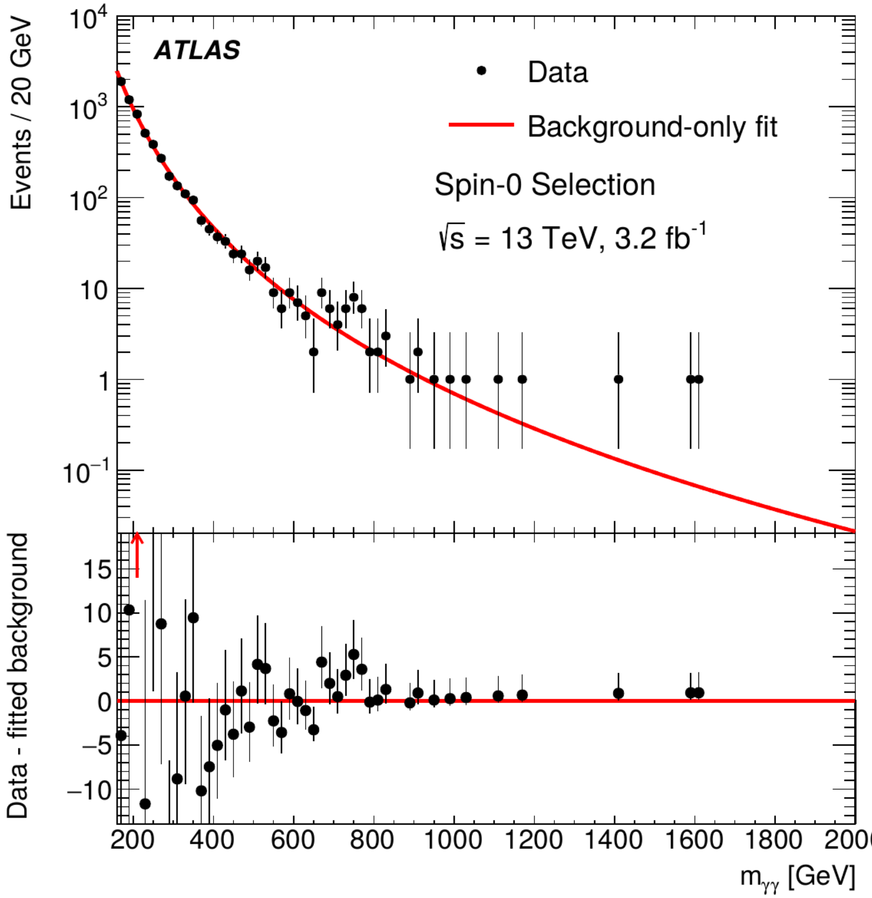
Randomness in Physics

Particularly important for searches for new phenomena:

→ Robust methods needed to control **spurious "discoveries"**…

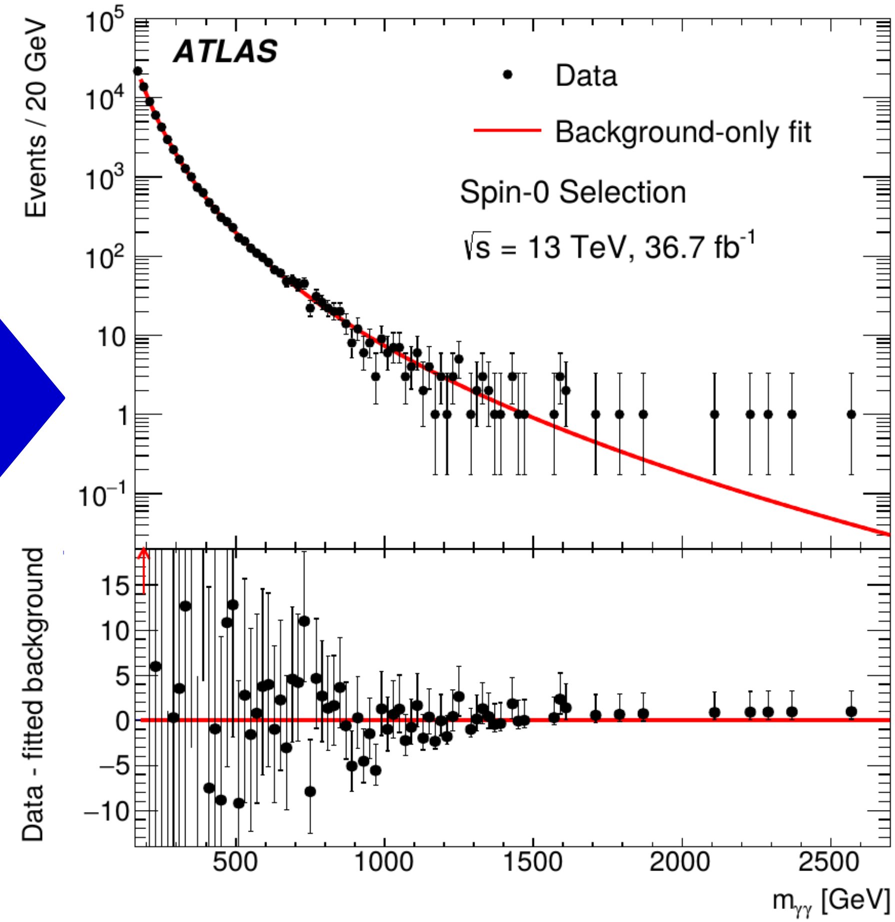
→ … and accurately **report the significance of excesses** in case of surprises

JHEP 09 (2016) 1



1 Year Later

Phys. Lett. B 775 (2017) 105



Example Analyses

Example 1: Z→ee Inclusive σ^{fid}

Measurement Principle:

Signal events	$34865 \pm 187 \pm 7 \pm 3$
Correction C	$0.552\ ^{+0.006}_{-0.005}$
σ^{fid}[nb]	$0.781 \pm 0.004 \pm 0.008 \pm 0.016$

Phys. Lett. B 759 (2016) 601

35000 ± (√35000 = 187)

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

175 ± 8

(81 ± 2) pb^{-1}

0.552 ± 0.006

Simple uncertainty propagation:

σ^{fid} **= 0.781 ± 0.004 (stat) ± 0.008 (syst) ± 0.016 (lumi) nb**

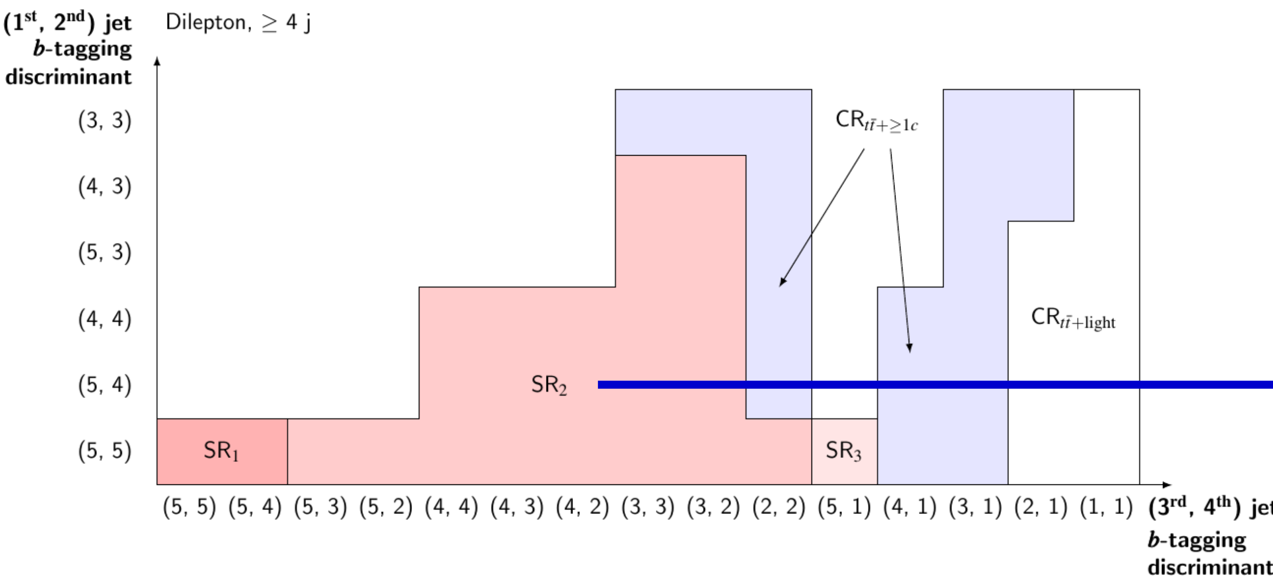
→ **Simplest possible example** in several ways

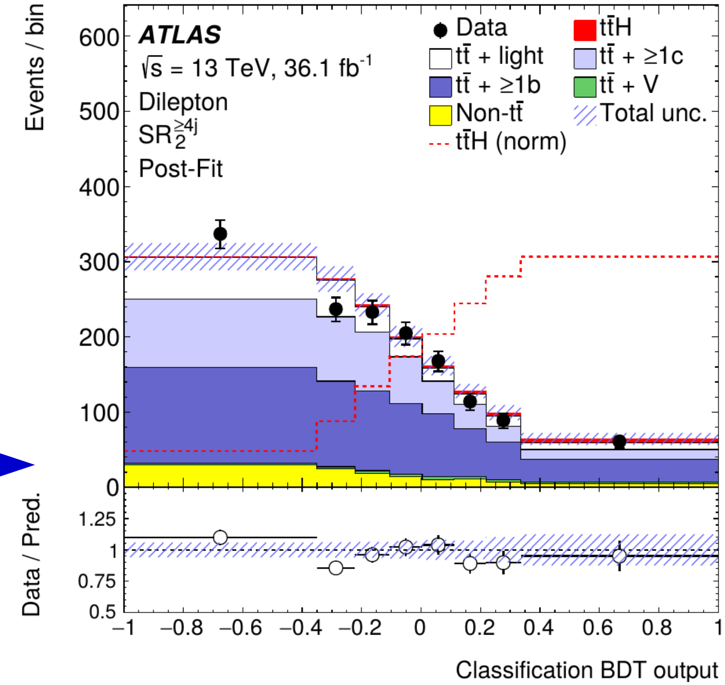
– **"Single bin counting"** : only data input is $\mathbf{N_{data}}$.

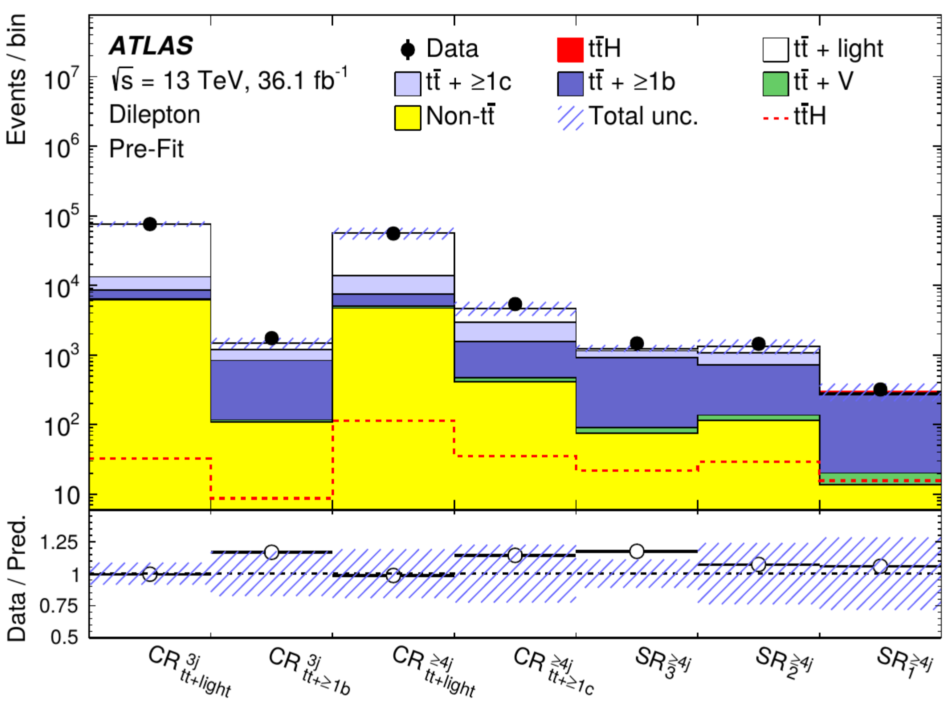
– Here **Gaussian** assumptions

Example 2: $t\bar{t}H \rightarrow b\bar{b}$

arXiv:1712.008895







Event counting in different regions:

Multiple-bin counting

Lots of information available

→ How to make optimal use of it ?

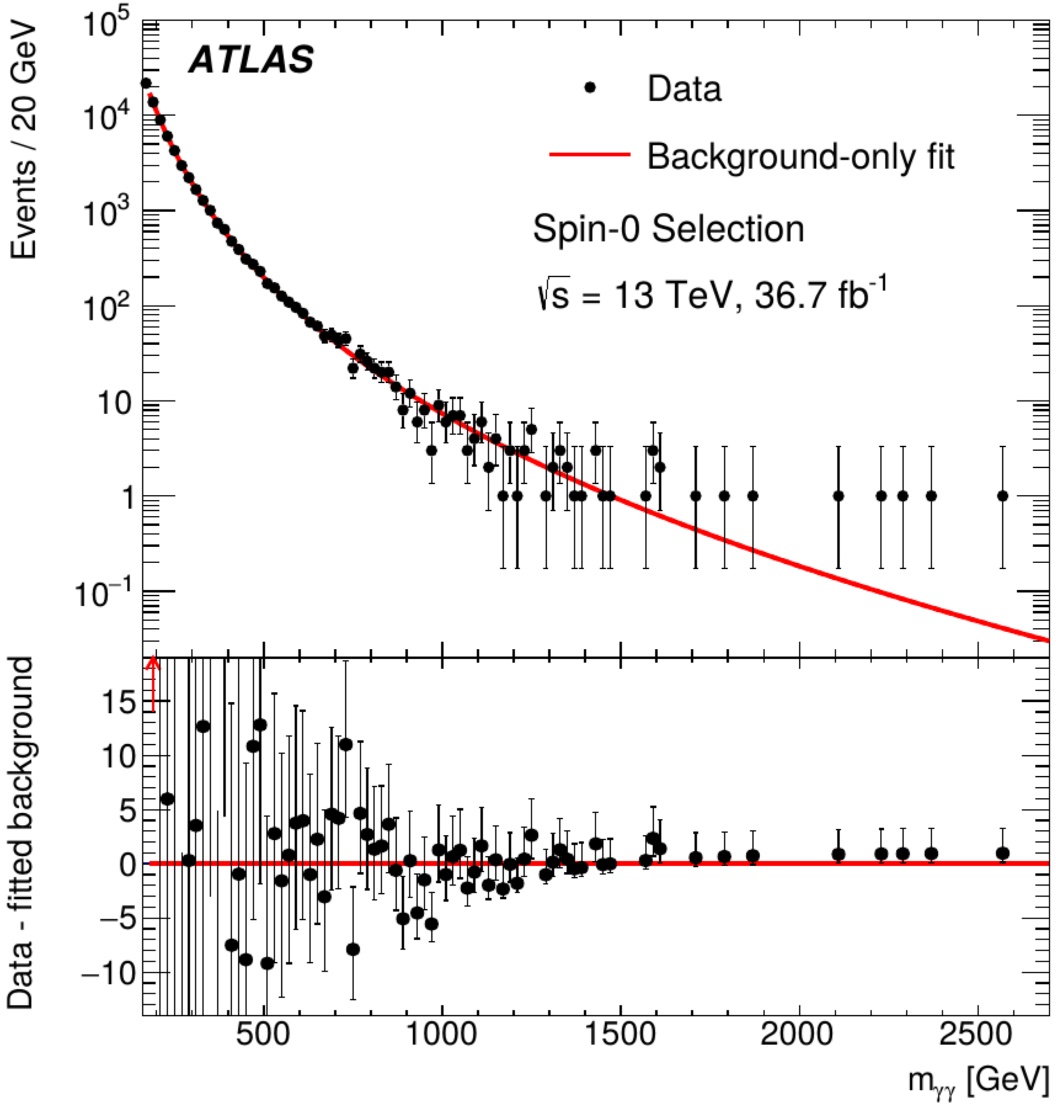
Goals:

→ discovery significance,

→ σ × BR measurement

Example 3: Unbinned shape analysis

Phys. Lett. B 775 (2017) 105



Describe spectrum without discrete binning
→ use smooth functions of a continuous variable.

Unbinned shape analysis

How to describe the shapes ?

Goals:
→ Discovery significance
→ $\sigma \times$ BR measurements
→ Upper limits.

Probability Distributions

Short reminder on Probability Distribution functions (PDFs)

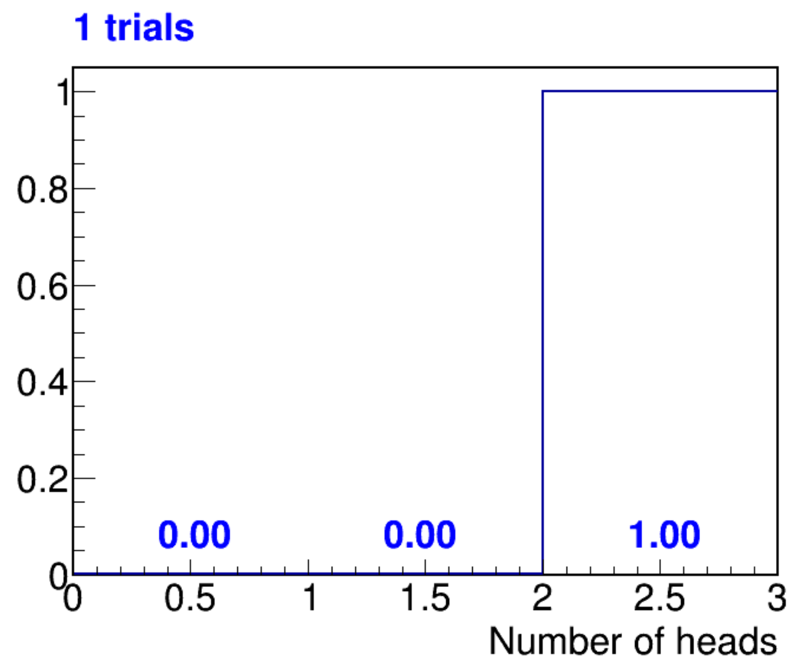
Probability Distributions

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

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

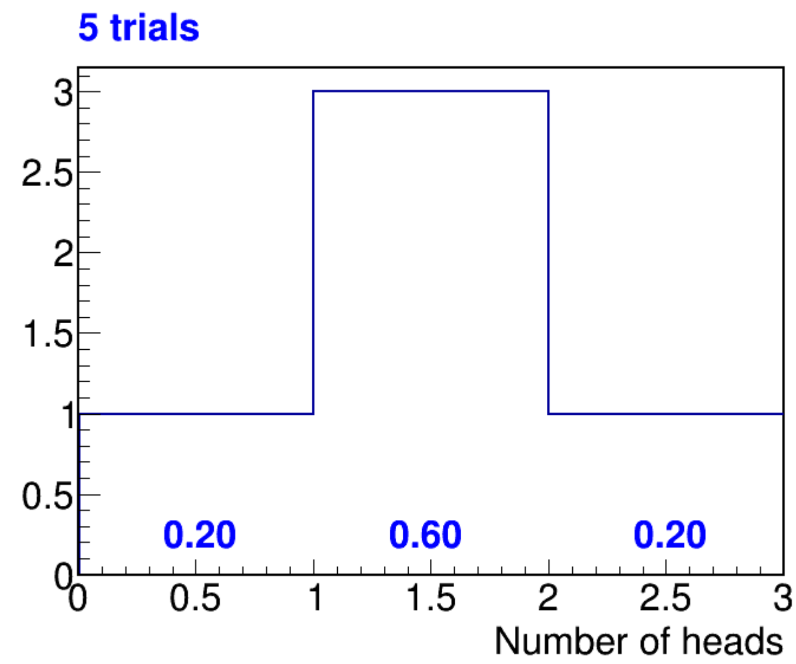
Probability Distributions

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

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

Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

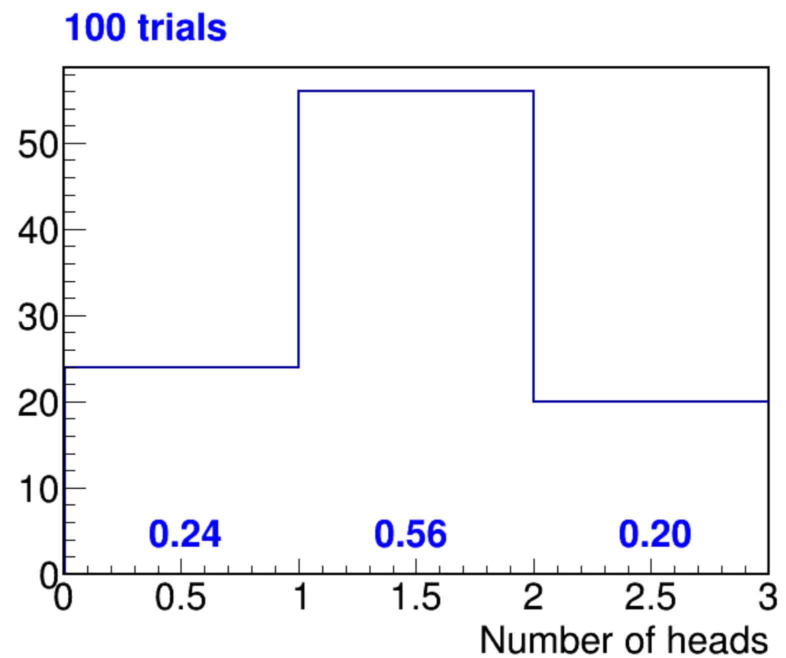
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

Properties

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



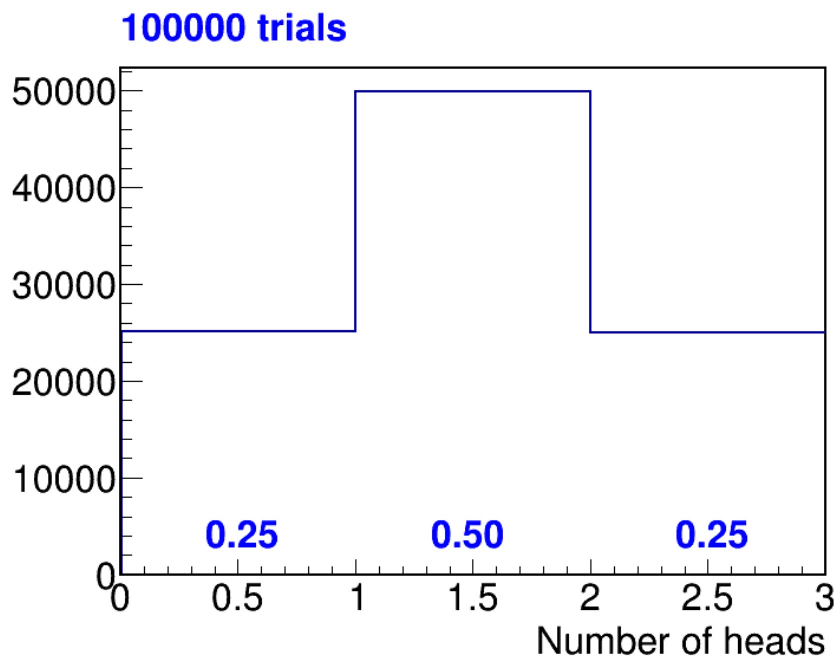
Probability Distributions

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 :

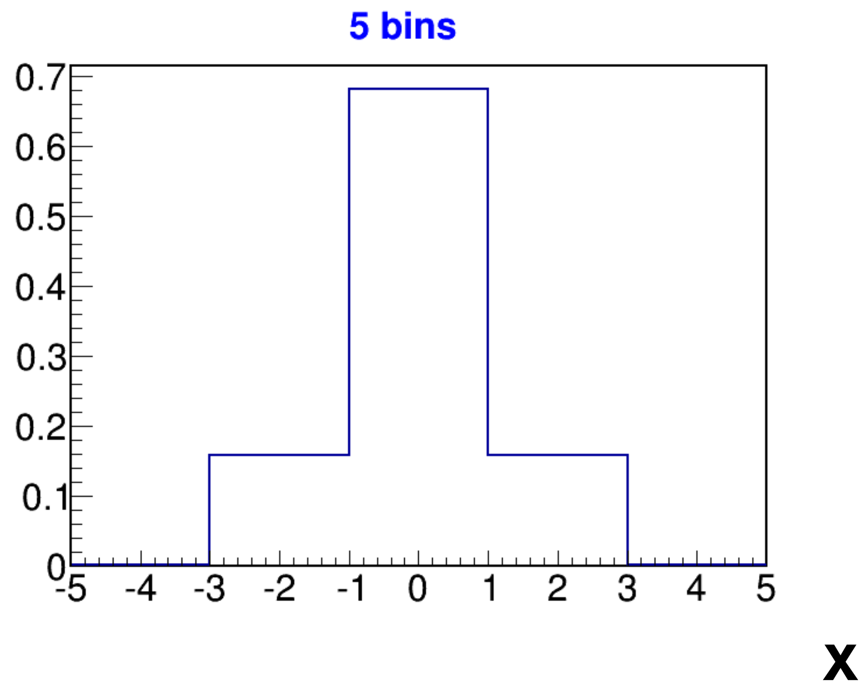
{ P_i } for i = 0, 1, 2

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 values ⇔ high chance to get a measurement here

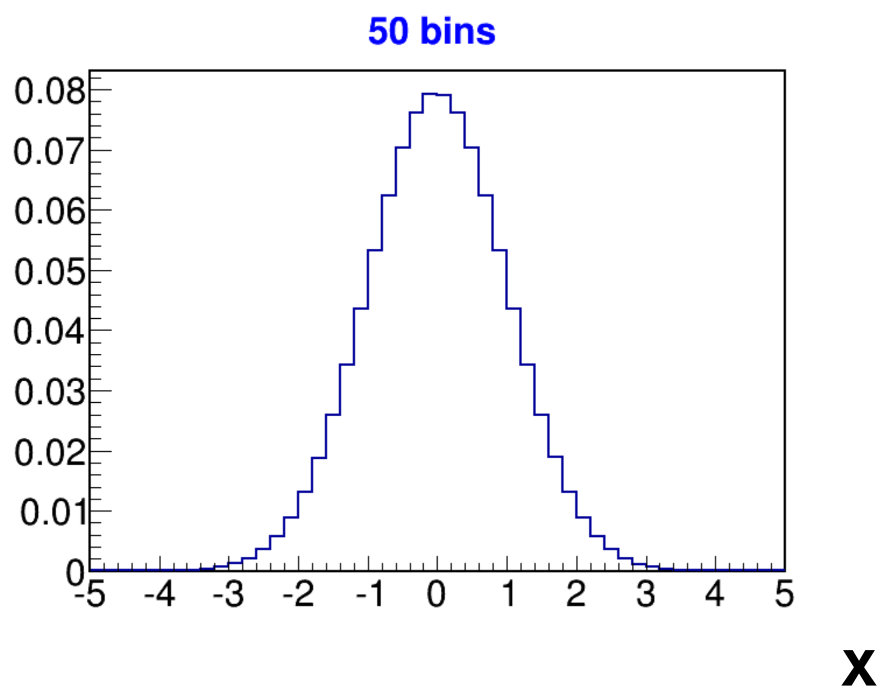
→ $P(x) > 0$

→ $\int P(x)\,dx = 1$

Generalizes to **multiple variables** : $\int P(x,y)\,dx\,dy = 1$

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 values ⇔ high chance to get a measurement here

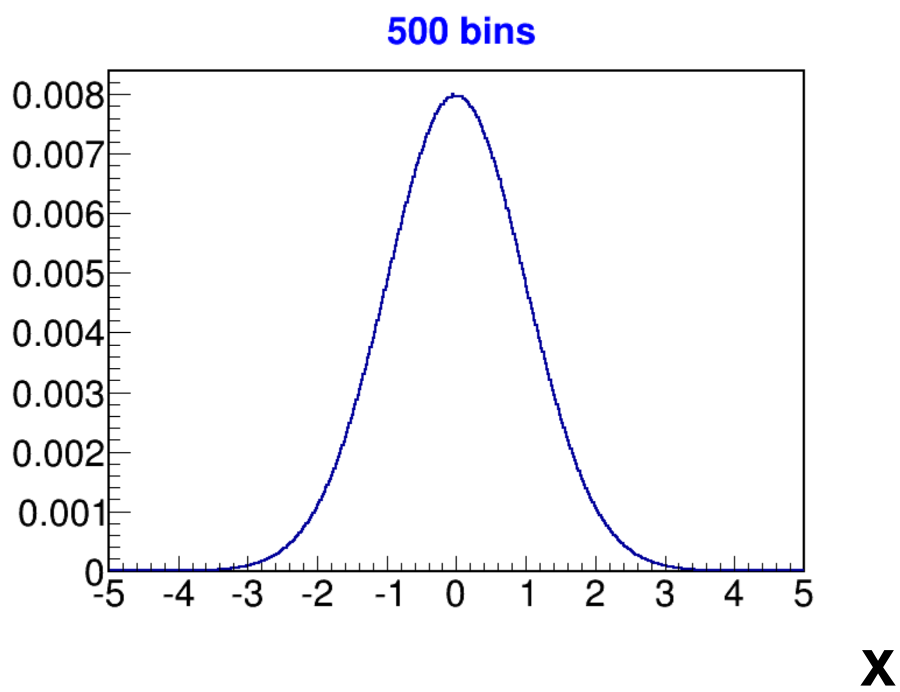
→ $P(x) > 0$

→ $\int P(x)\, dx = 1$

Generalizes to **multiple variables** : $\int P(x,y)\, dx\, dy = 1$

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 values ⇔ high chance to get a measurement here

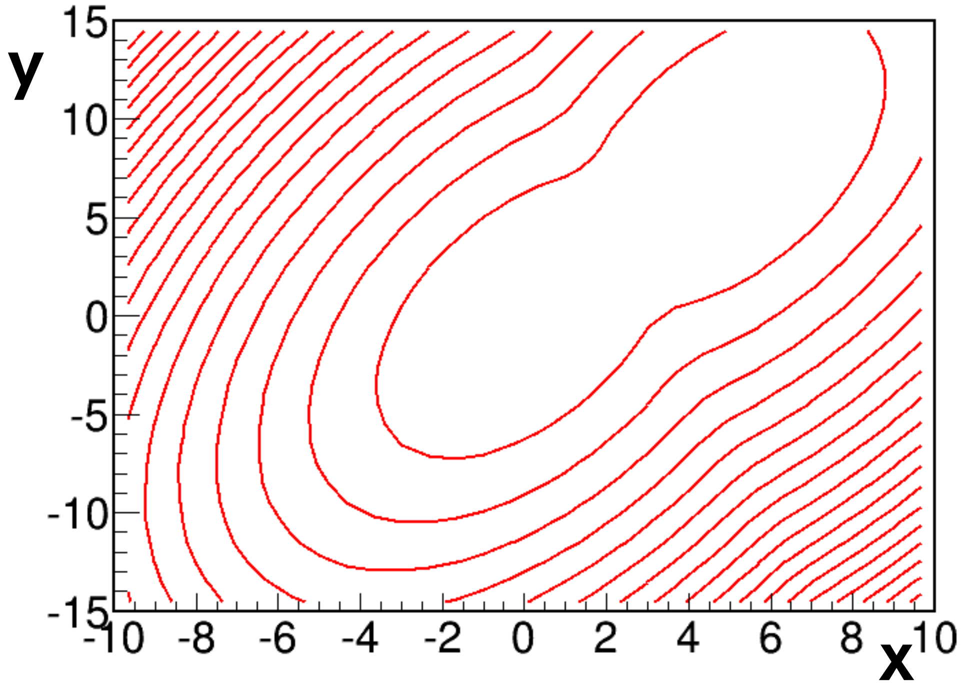
→ $P(x) > 0$

→ $\int P(x)\, dx = 1$

Generalizes to **multiple variables** : $\int P(x,y)\, dx\, dy = 1$

Continuous Variables: PDFs

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



Contours: P(x,y)

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

→ High values ⇔ high chance to get a measurement here

→ P(x) > 0

→ $\int$ P(x) dx = 1

Generalizes to **multiple variables** : $\int$ P(x,y) dx dy = 1

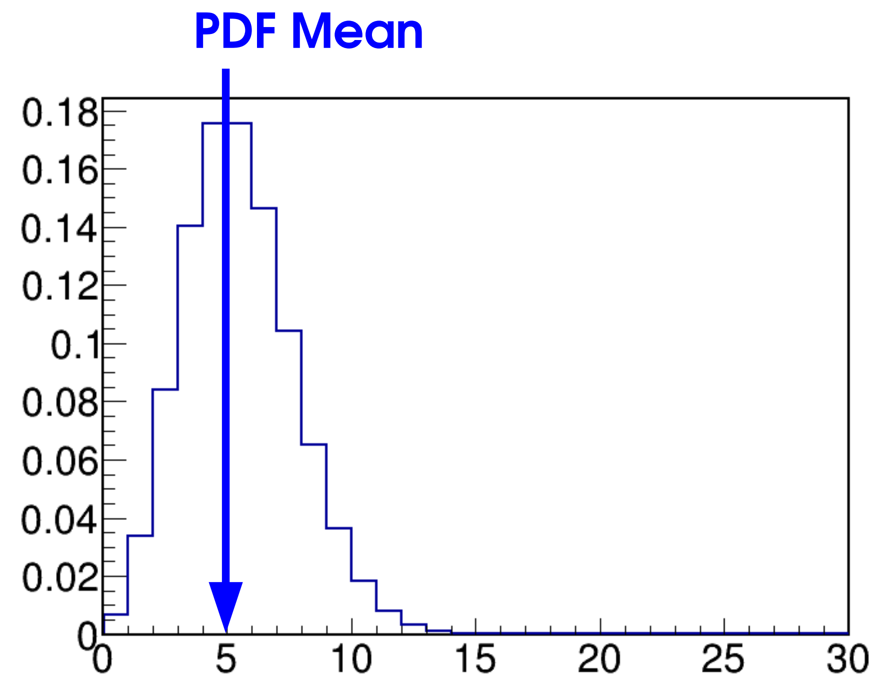
PDF Properties: Mean

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

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

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

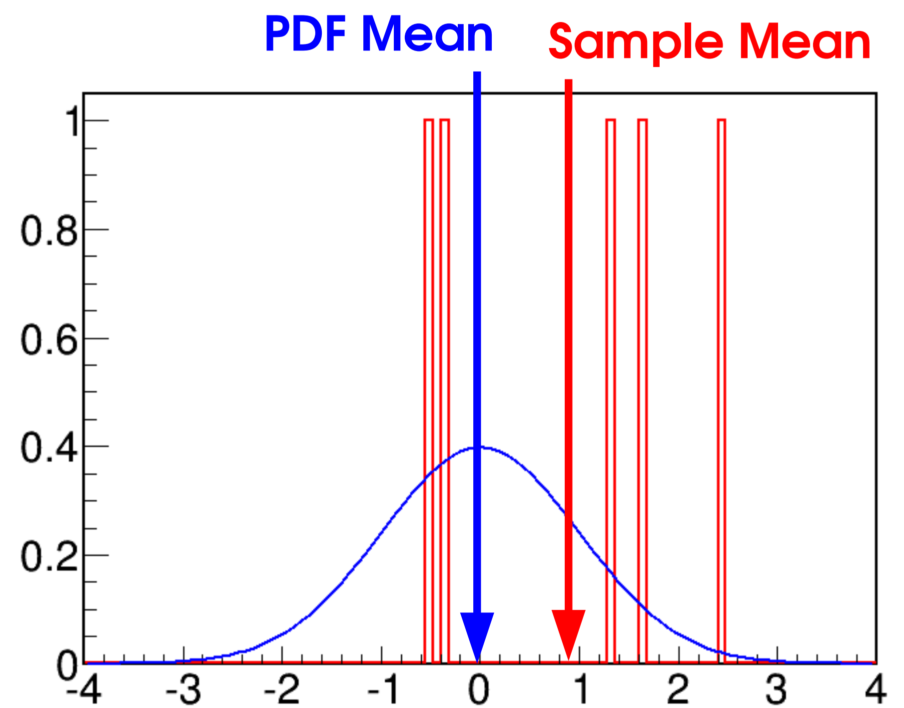
→ Property of the **PDF**



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.



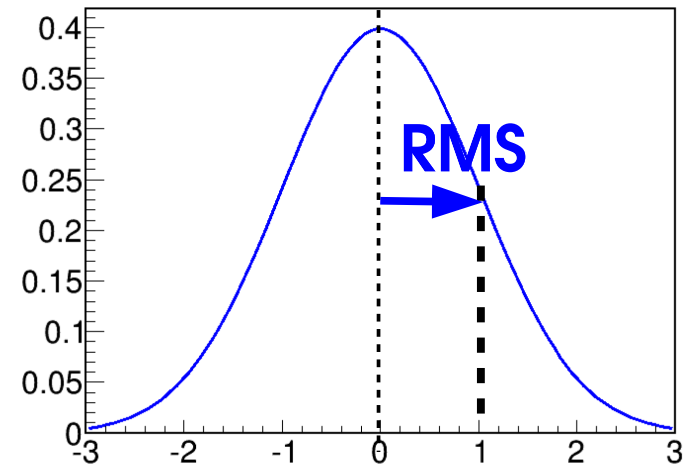
PDF Properties: Variance

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**:

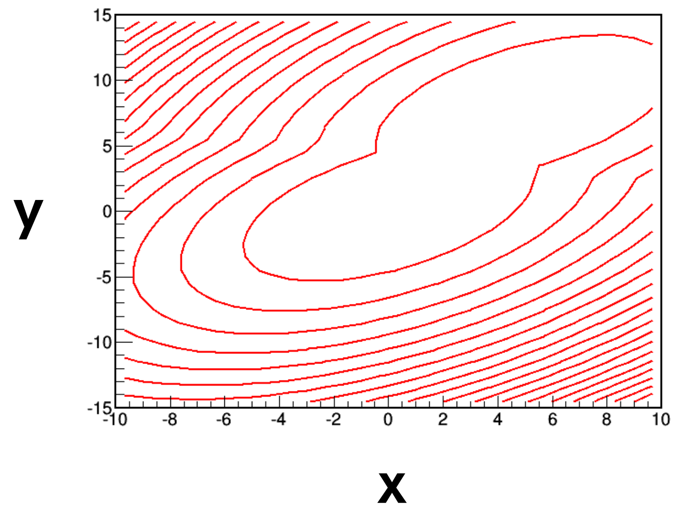
$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

Covariance of x and y:

$$\mathrm{Cov}(x) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$$

→ Large if variations of x, y are "synchronized"

- **Cov(x, y) > 0** if x and y vary in the **same** direction
- **Cov(x, y) < 0** if x and y vary in **opposite** direction
- **Cov(x, y) = 0** if x and y vary **independently**



Correlation coefficient

$$\gamma = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}}$$

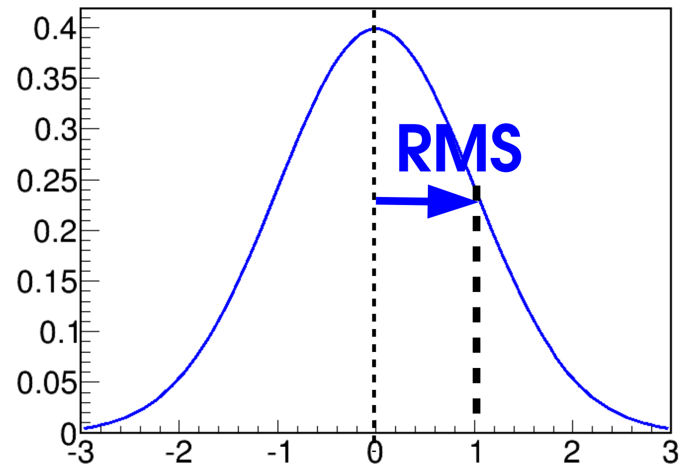
PDF Properties: Variance

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***

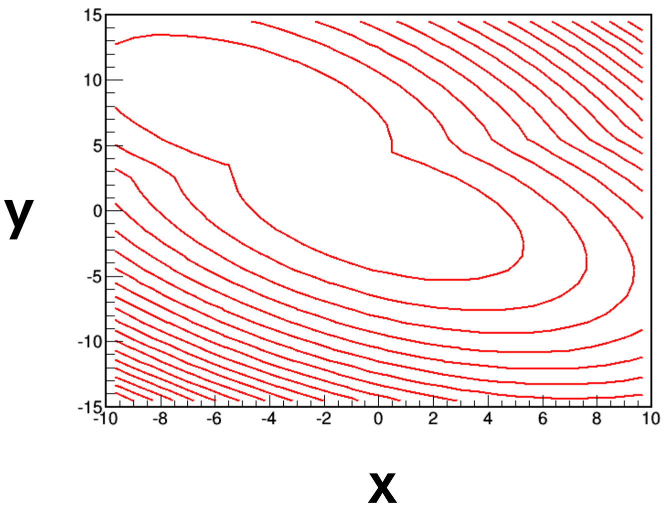


Can be approximated by **sample variance**:

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

Covariance of x and y:

$$\mathrm{Cov}(x) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$$



→ Large if variations of x, y are "synchronized"

- **Cov(x, y) > 0** if x and y vary in the **same** direction
- **Cov(x, y) < 0** if x and y vary in **opposite** direction
- **Cov(x, y) = 0** if x and y vary **independently**

Correlation coefficient

$$\gamma = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}}$$

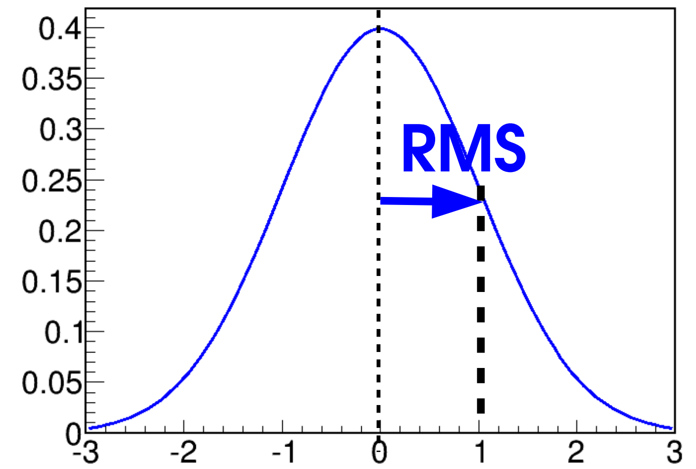
PDF Properties: Variance

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*

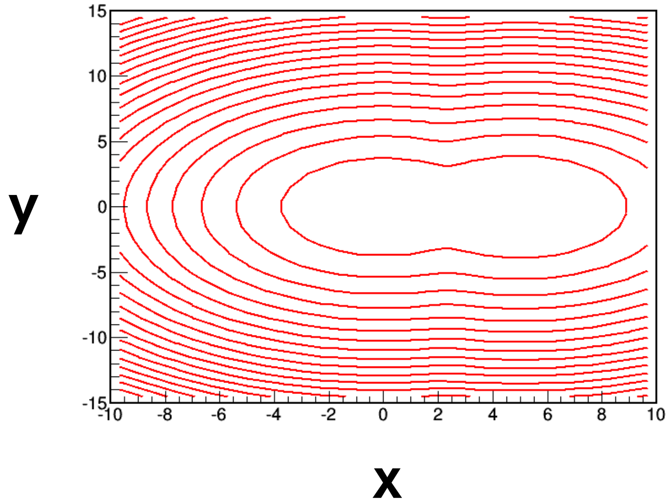


Can be approximated by **sample variance**:

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

Covariance of x and y:

$$\mathrm{Cov}(x) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$$



→ Large if variations of x, y are "synchronized"

- **Cov(x, y) > 0** if x and y vary in the **same** direction
- **Cov(x, y) < 0** if x and y vary in **opposite** direction
- **Cov(x, y) = 0** if x and y vary **independently**

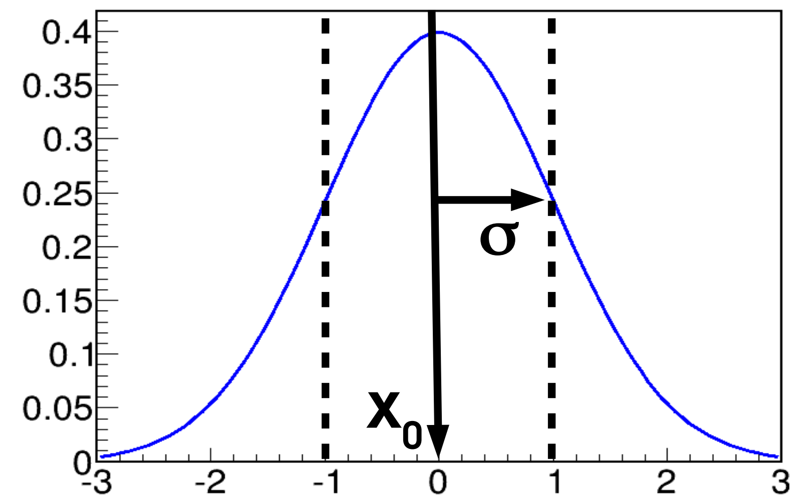
Correlation coefficient

$$\gamma = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}}$$

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**:

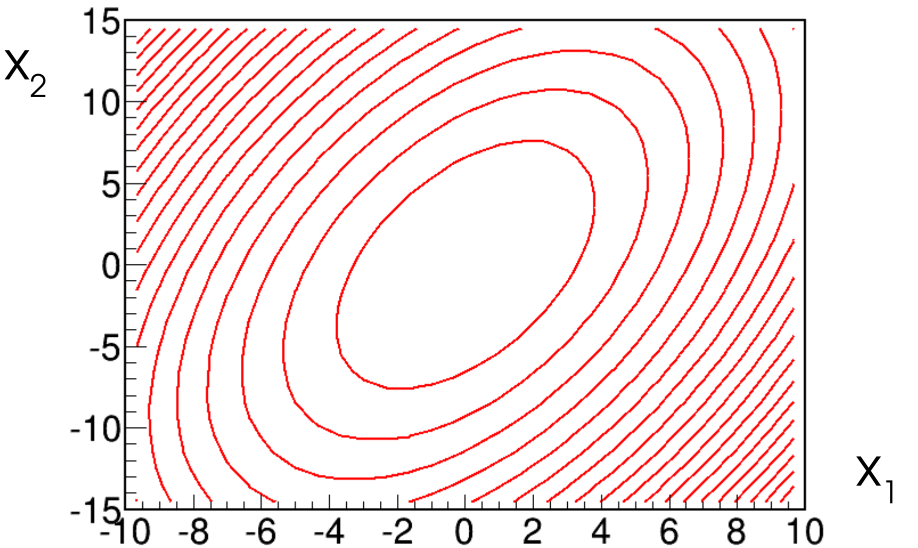
$$G(x; X_0, C) = \frac{1}{(2\pi|C|)^{N/2}} e^{-\frac{1}{2}(x-X_0)^T C^{-1}(x-X_0)}$$

→ **Mean** : X_0

→ **Covariance matrix** :

$$C = \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Var}(x_2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{x_1}^2 & \gamma\sigma_{x_1}\sigma_{x_2} \\ \gamma\sigma_{x_1}\sigma_{x_2} & \sigma_{x_2}^2 \end{bmatrix}$$



Gaussian Quantiles

Probability to be away from the Gaussian mean:

Consider $z = \left(\frac{x - x_0}{\sigma}\right)$ *"pull"*

P depends only on z ~ G(0,1)

Z	$P(\lvert x - x_0 \rvert > Z\sigma)$
1	**0.327**
2	**0.045**
3	**0.003**
5	**6 x 10^{-7}**

Gaussian **Cumulative Distribution Function (CDF)** :

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

In ROOT,

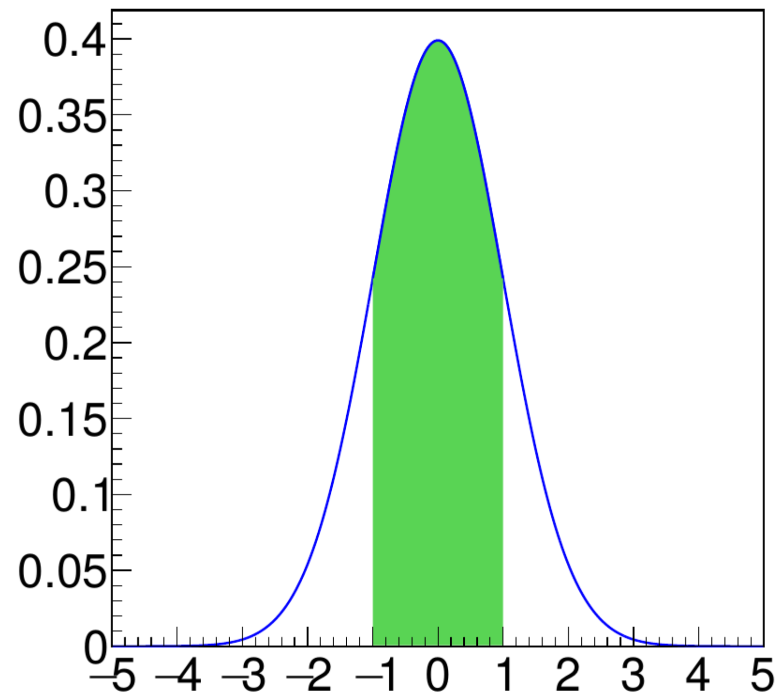
z→Φ : `ROOT::Math::gaussian_cdf(p)`

Φ→z : `ROOT::Math::gaussian_quantile(p,1)`

and add _c to use 1-Φ instead of Φ

```
root [0] ROOT::Math::gaussian_cdf(1) - ROOT::Math::gaussian_cdf(-1)
(double) 0.68268949
root [1] ROOT::Math::gaussian_quantile_c(0.05/2, 1)
(double) 1.9599640
```

$P(|x - x_0| < 1\sigma) = 68.3\ \%$



Gaussian Quantiles

Probability to be away from the Gaussian mean:

Consider $z = \left(\frac{x - x_0}{\sigma}\right)$ *"pull"*

P depends only on z ~ G(0,1)

Z	$P(\lvert x - x_0 \rvert > Z\sigma)$
1	**0.327**
2	**0.045**
3	**0.003**
5	**6 x 10^{-7}**

Gaussian **Cumulative Distribution Function (CDF)** :

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

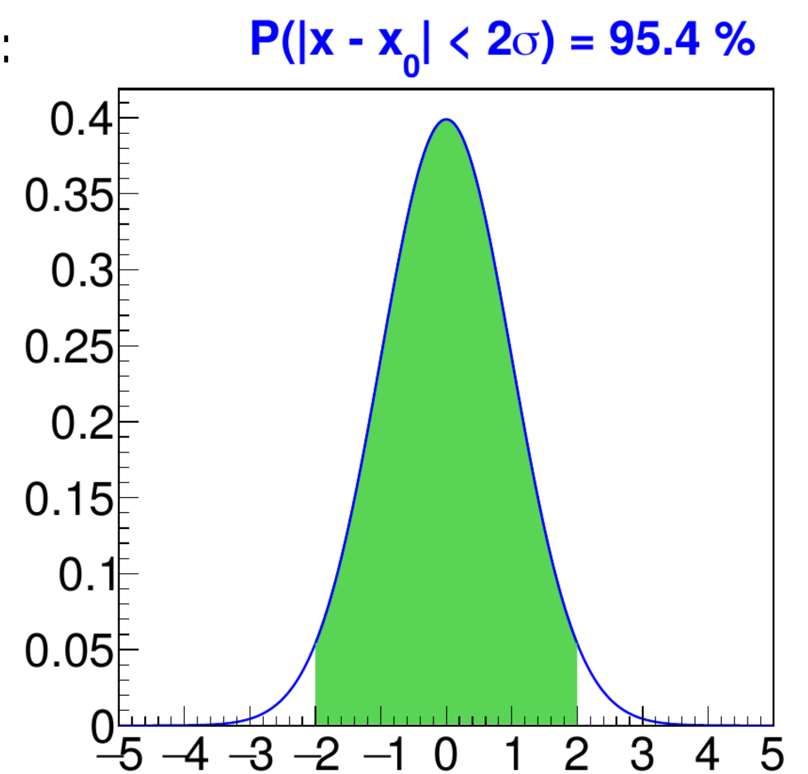
In ROOT,

z→Φ : `ROOT::Math::gaussian_cdf(p)`

Φ→z : `ROOT::Math::gaussian_quantile(p,1)`

and add _c to use 1-Φ instead of Φ

```
root [0] ROOT::Math::gaussian_cdf(1) - ROOT::Math::gaussian_cdf(-1)
(double) 0.68268949
root [1] ROOT::Math::gaussian_quantile_c(0.05/2, 1)
(double) 1.9599640
```



Gaussian Quantiles

Probability to be away from the Gaussian mean:

Consider $z = \left(\frac{x - x_0}{\sigma}\right)$ *"pull"*

P depends only on z ~ G(0,1)

Z	$P(\lvert x - x_0 \rvert > Z\sigma)$
1	**0.327**
2	**0.045**
3	**0.003**
5	**6×10^{-7}**

Gaussian **Cumulative Distribution Function (CDF)** :

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

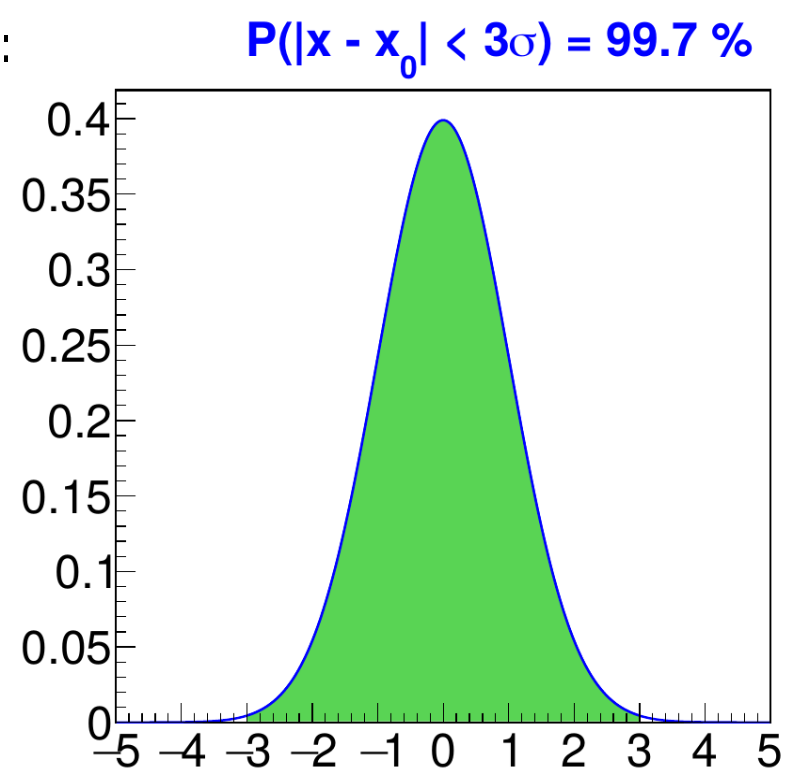
In ROOT,

z→Φ : `ROOT::Math::gaussian_cdf(p)`

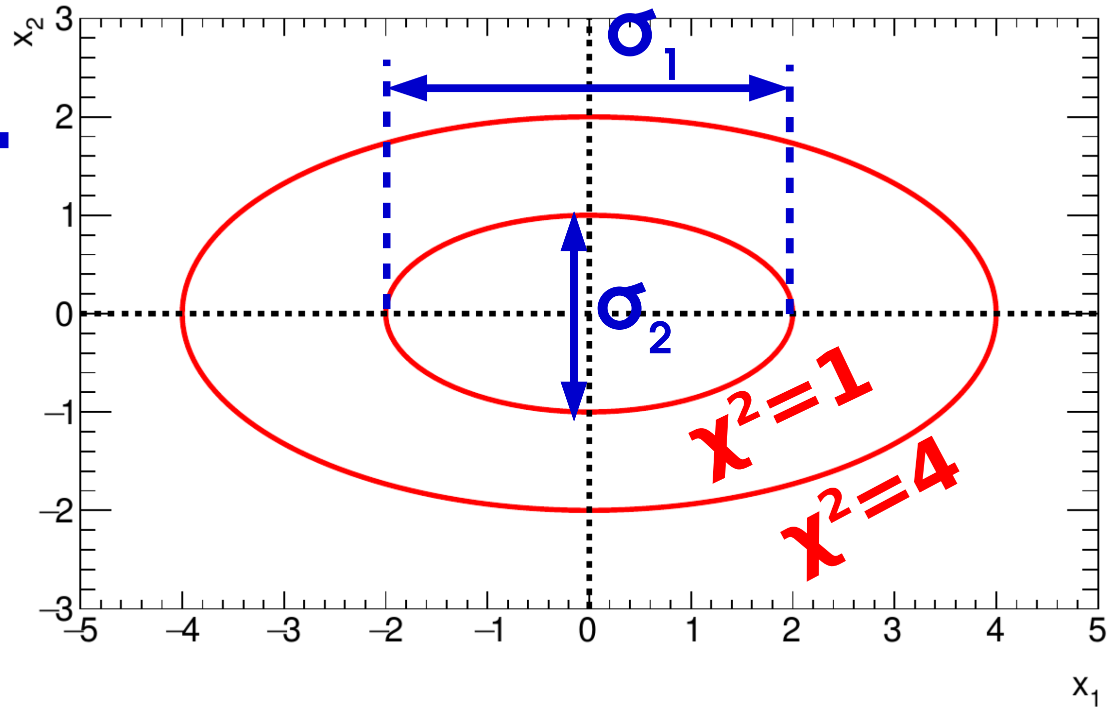
Φ→z : `ROOT::Math::gaussian_quantile(p,1)`

and add _c to use 1-Φ instead of Φ

```
root [0] ROOT::Math::gaussian_cdf(1) - ROOT::Math::gaussian_cdf(-1)
(double) 0.68268949
root [1] ROOT::Math::gaussian_quantile_c(0.05/2, 1)
(double) 1.9599640
```



Chi-squared



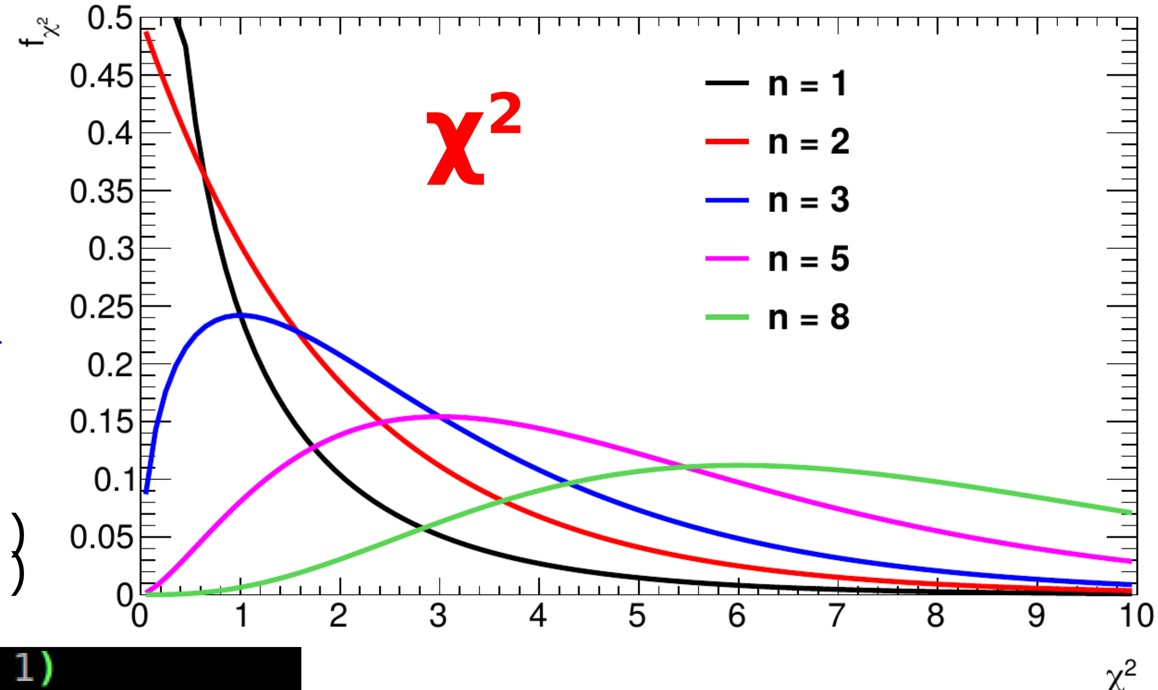
Multiple Independent Gaussians:

Define

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

Measures global distance from reference point $(x_1^0 \ldots x_n^0)$

Distribution depends on n :



Rule of thumb: χ^2/n **should be** $\lesssim 1$

Exact distributions in ROOT:

```
ROOT::Math::chisquared_pdf(x, n)
ROOT::Math::chisquared_cdf(x, n)
```

```
root [0] ROOT::Math::chisquared_cdf(1, 1)
(double) 0.68268949
root [1] ROOT::Math::chisquared_cdf(4, 1)
(double) 0.95449974
```

Chi-squared

Multiple Independent Gaussians:

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)

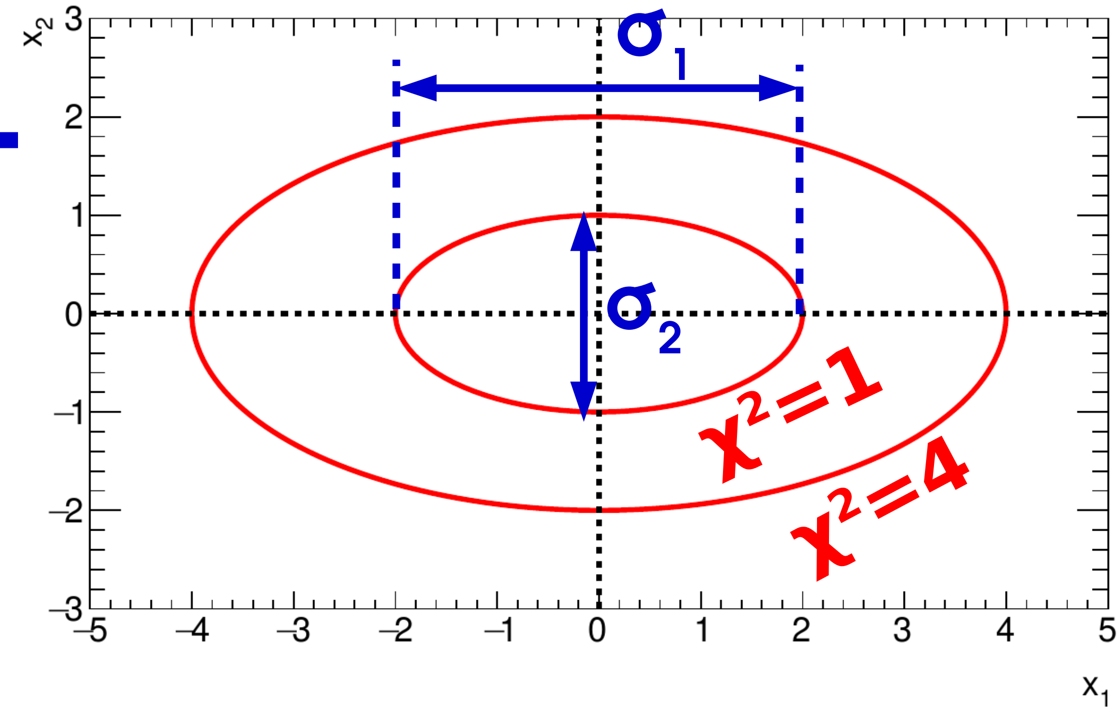
Distribution depends on n :

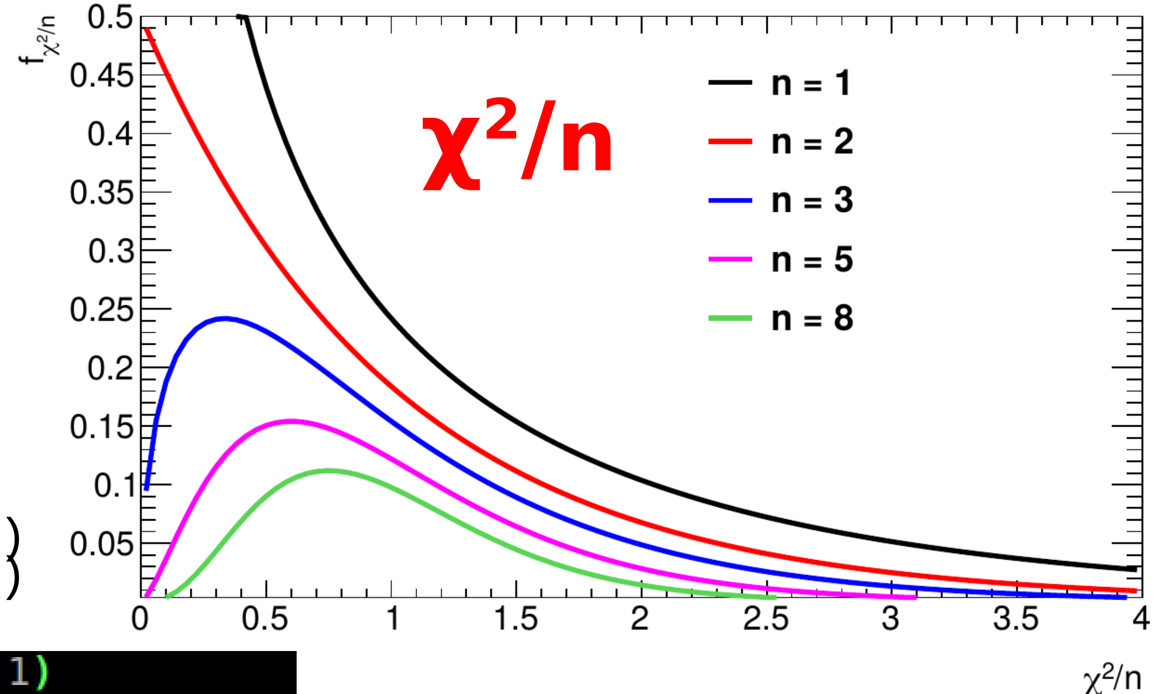
Rule of thumb: χ^2/n **should be** $\lesssim 1$

Exact distributions in ROOT:

```
ROOT::Math::chisquared_pdf(x, n)
ROOT::Math::chisquared_cdf(x, n)
```

```
root [0] ROOT::Math::chisquared_cdf(1, 1)
(double) 0.68268949
root [1] ROOT::Math::chisquared_cdf(4, 1)
(double) 0.95449974
```

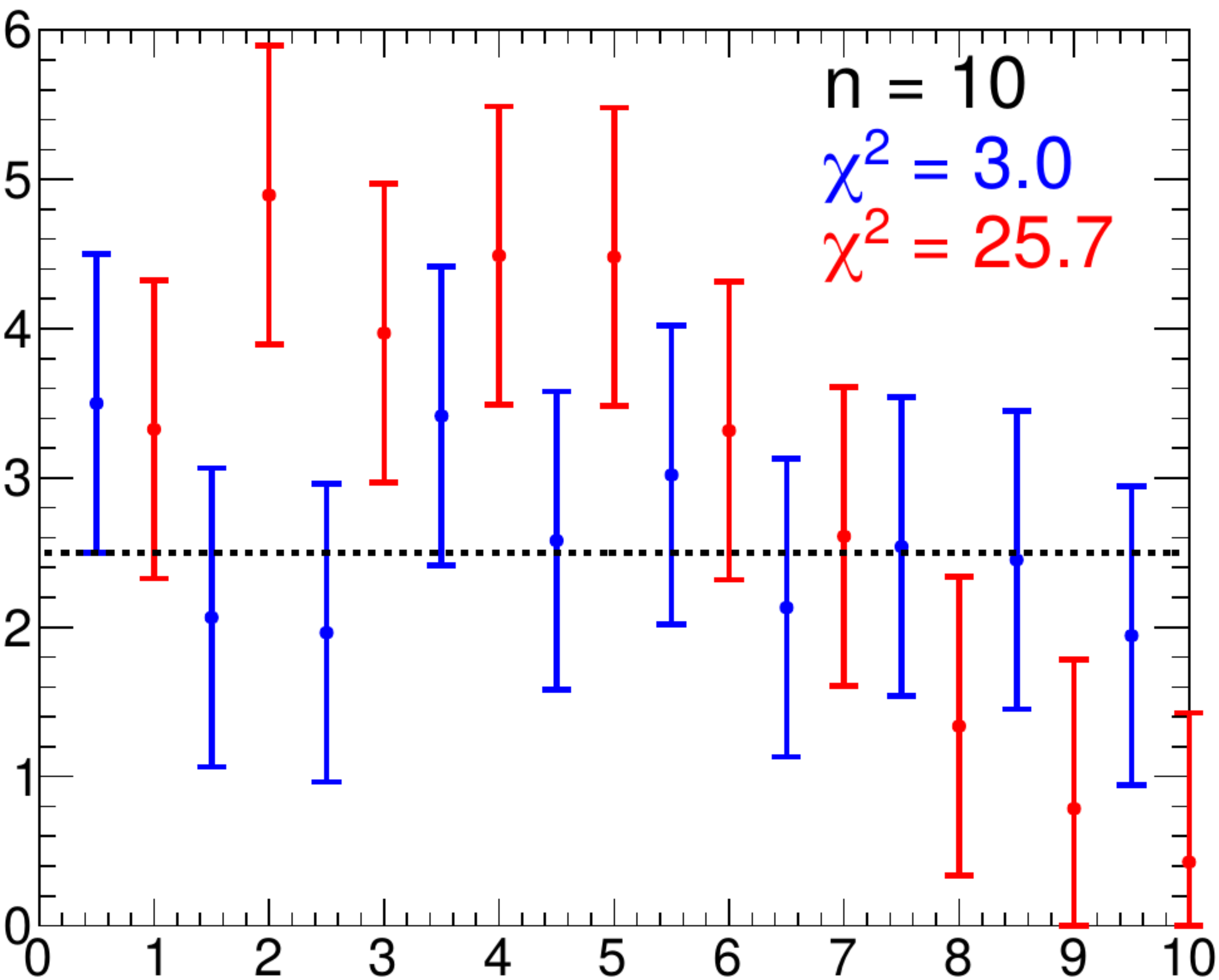




Histogram Chi-squared

Histogram χ2 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

$\chi^2 = 3.0$

$p(\chi^2=3.0, n=10) =$ **98%** ✔

BLUE histogram

$\chi^2 = 25.7$

$p(\chi^2=25.7, n=10) =$ **0.4%** ✘

```
root [0] ROOT::Math::chisquared_cdf_c(3, 10)
(double) 0.98142406
root [1] ROOT::Math::chisquared_cdf_c(25.7, 10)
(double) 0.0041653244
```

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** : less fluctuations when averaging over many measurements

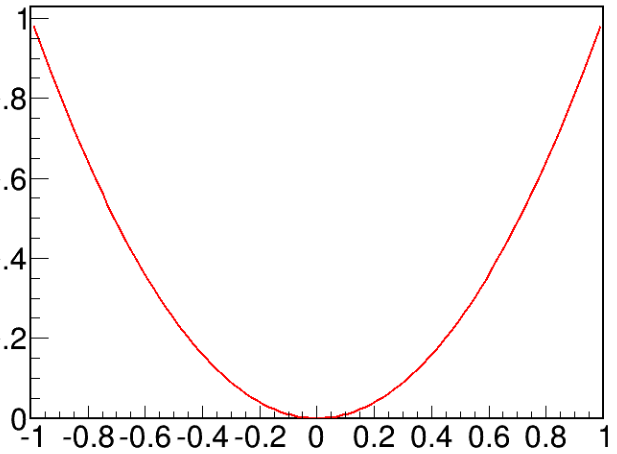
Another version, for the sum:

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

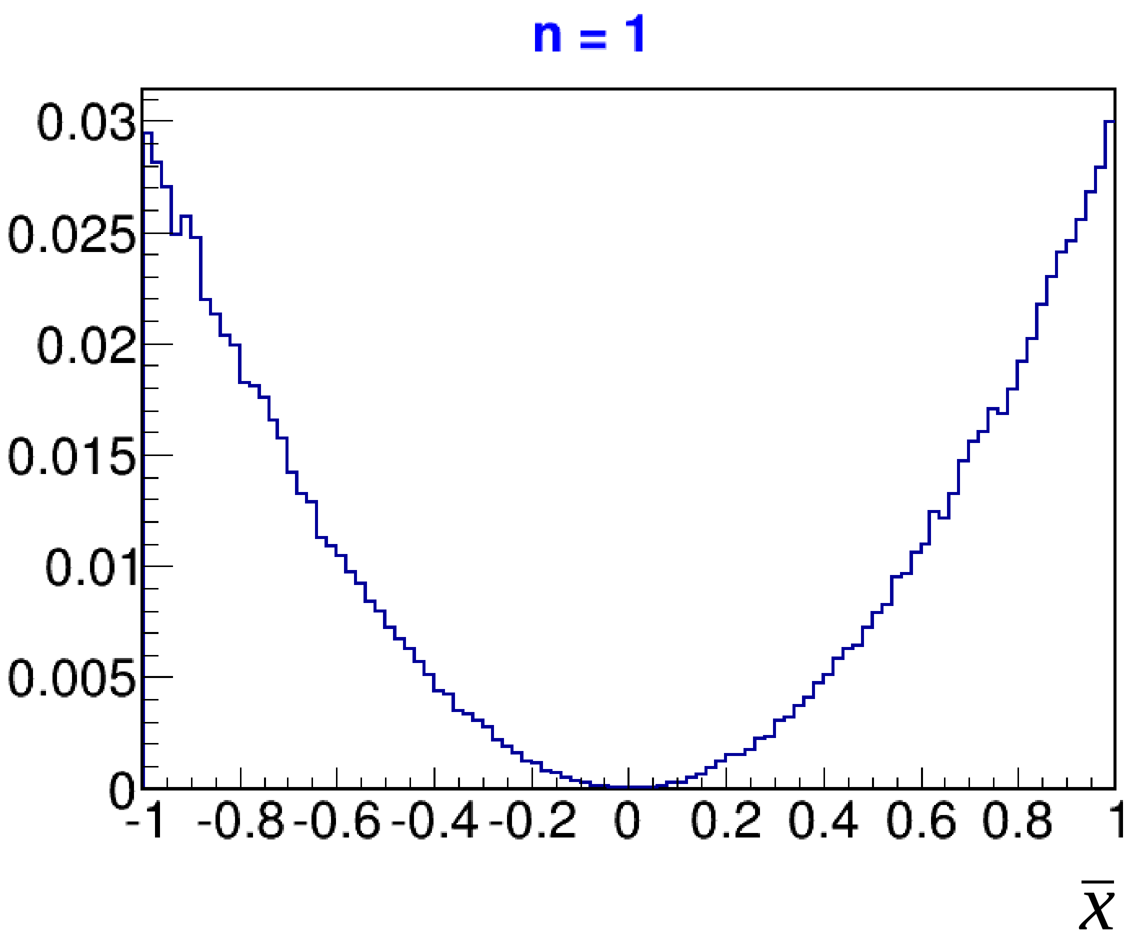
Mean scales like n, but RMS only like √n

Central Limit Theorem in action

Draw events from a x^2 distribution (for illustration only)



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

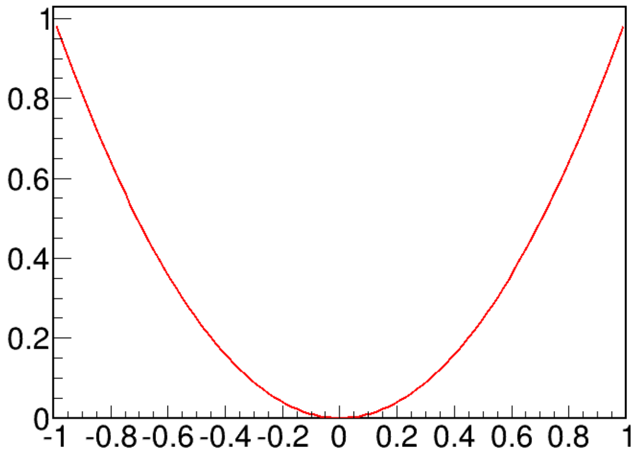


Distribution becomes Gaussian, although very non-Gaussian originally

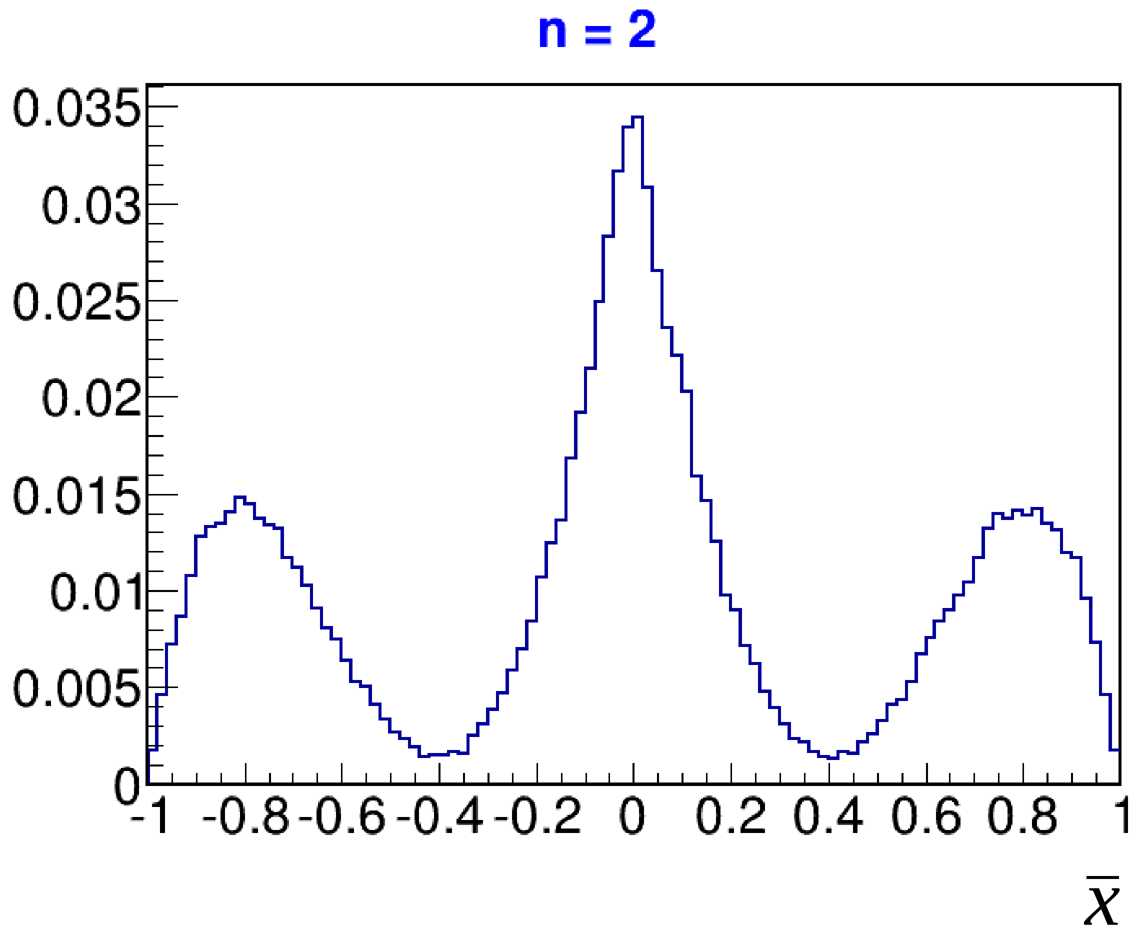
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a x^2 distribution (for illustration only)



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

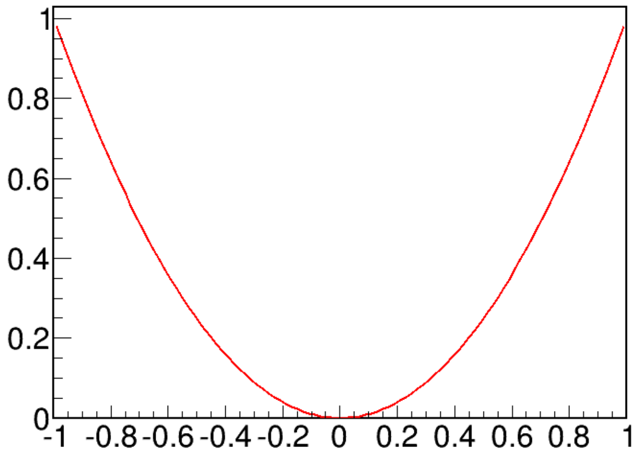


Distribution becomes Gaussian, although very non-Gaussian originally

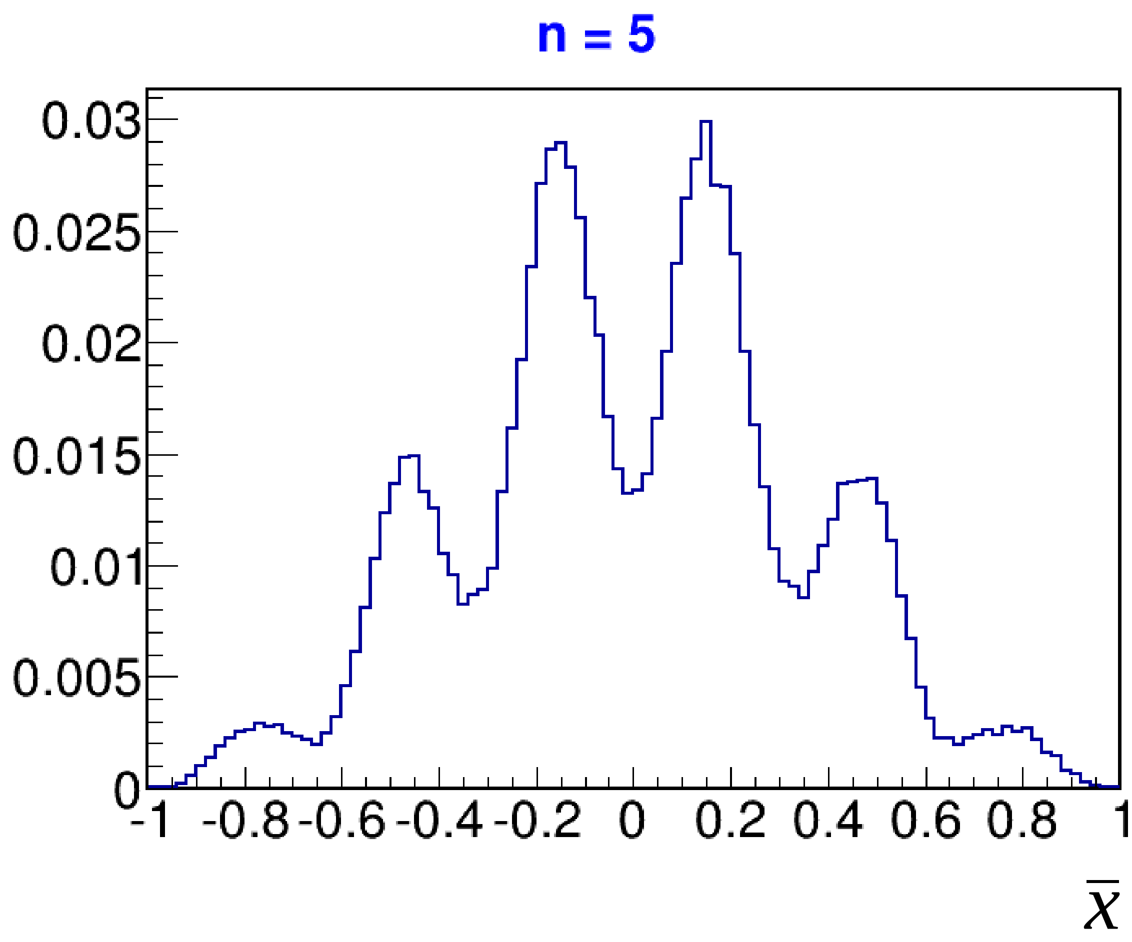
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a x^2 distribution (for illustration only)



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

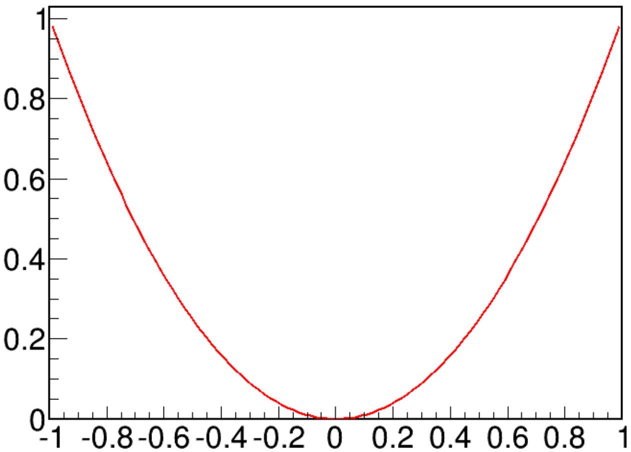


Distribution becomes Gaussian, although very non-Gaussian originally

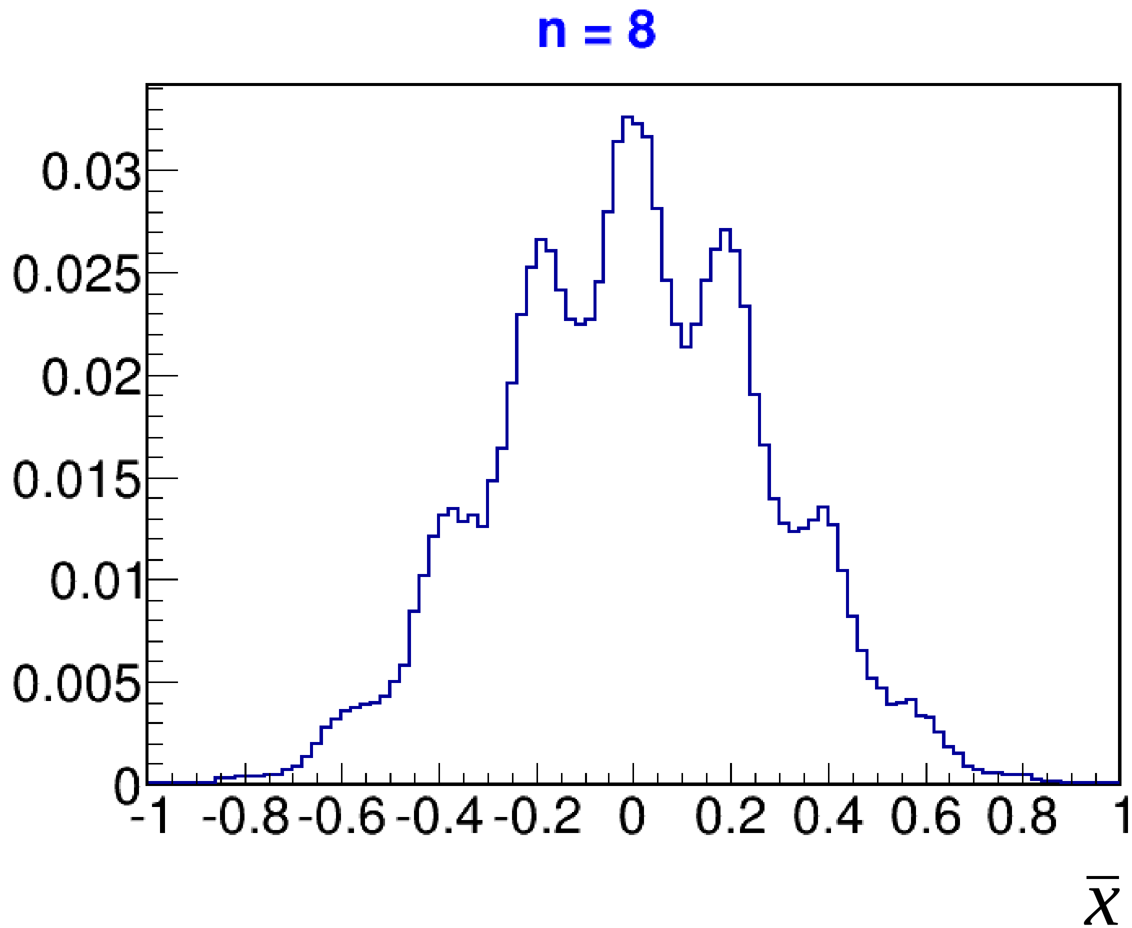
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a x^2 distribution (for illustration only)



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

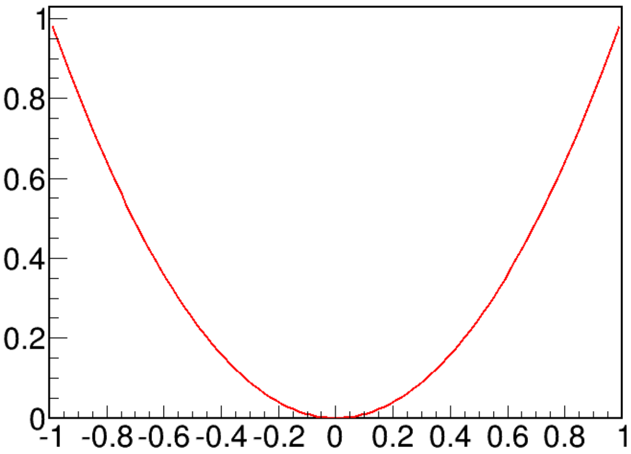


Distribution becomes Gaussian, although very non-Gaussian originally

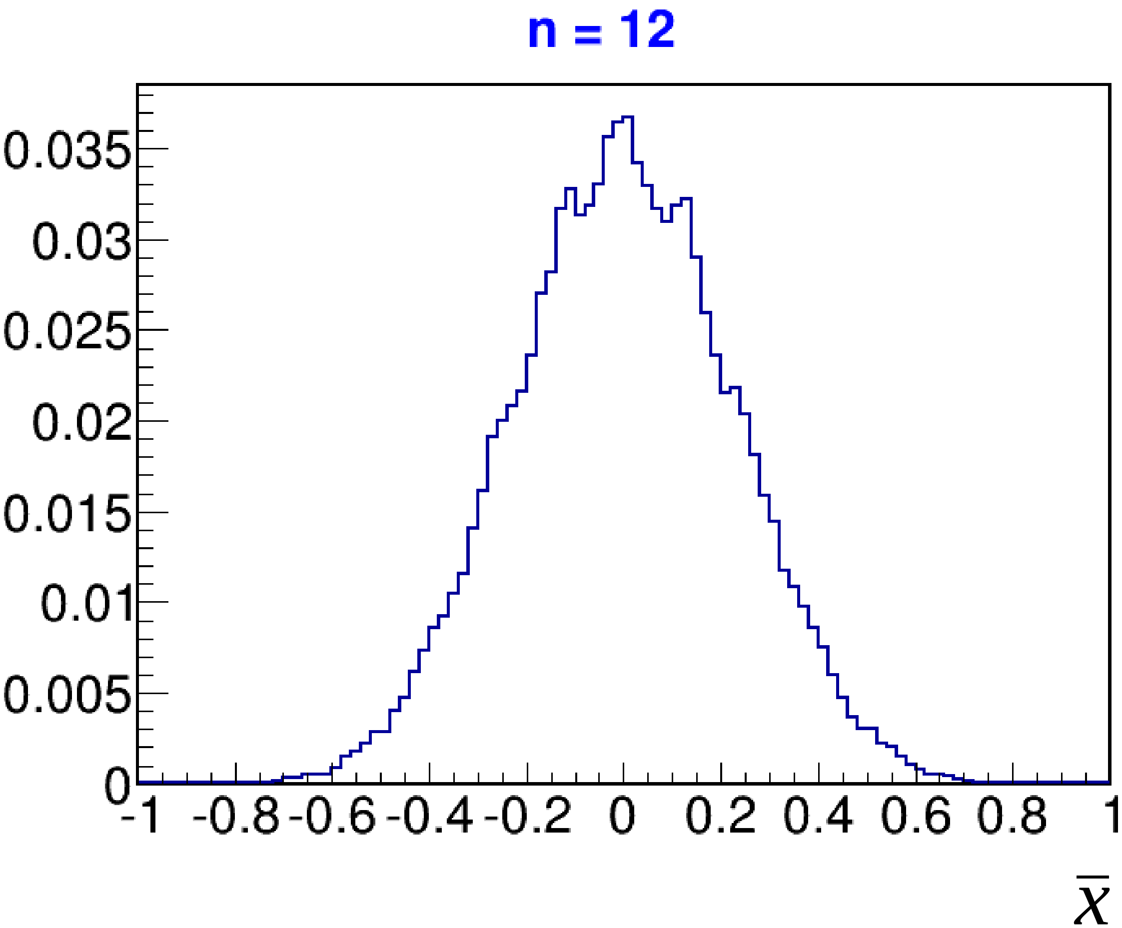
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a x^2 distribution (for illustration only)



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

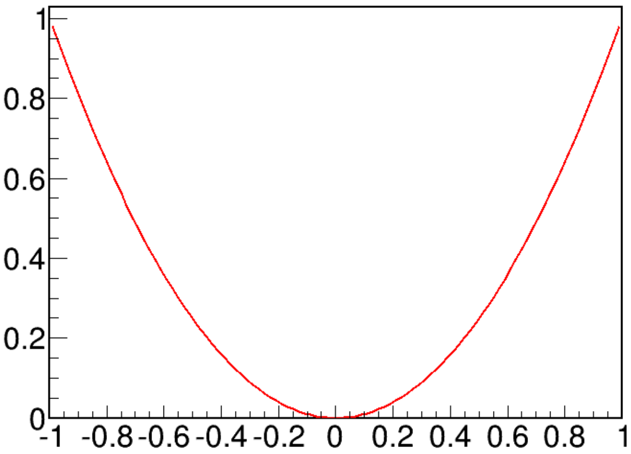


Distribution becomes Gaussian, although very non-Gaussian originally

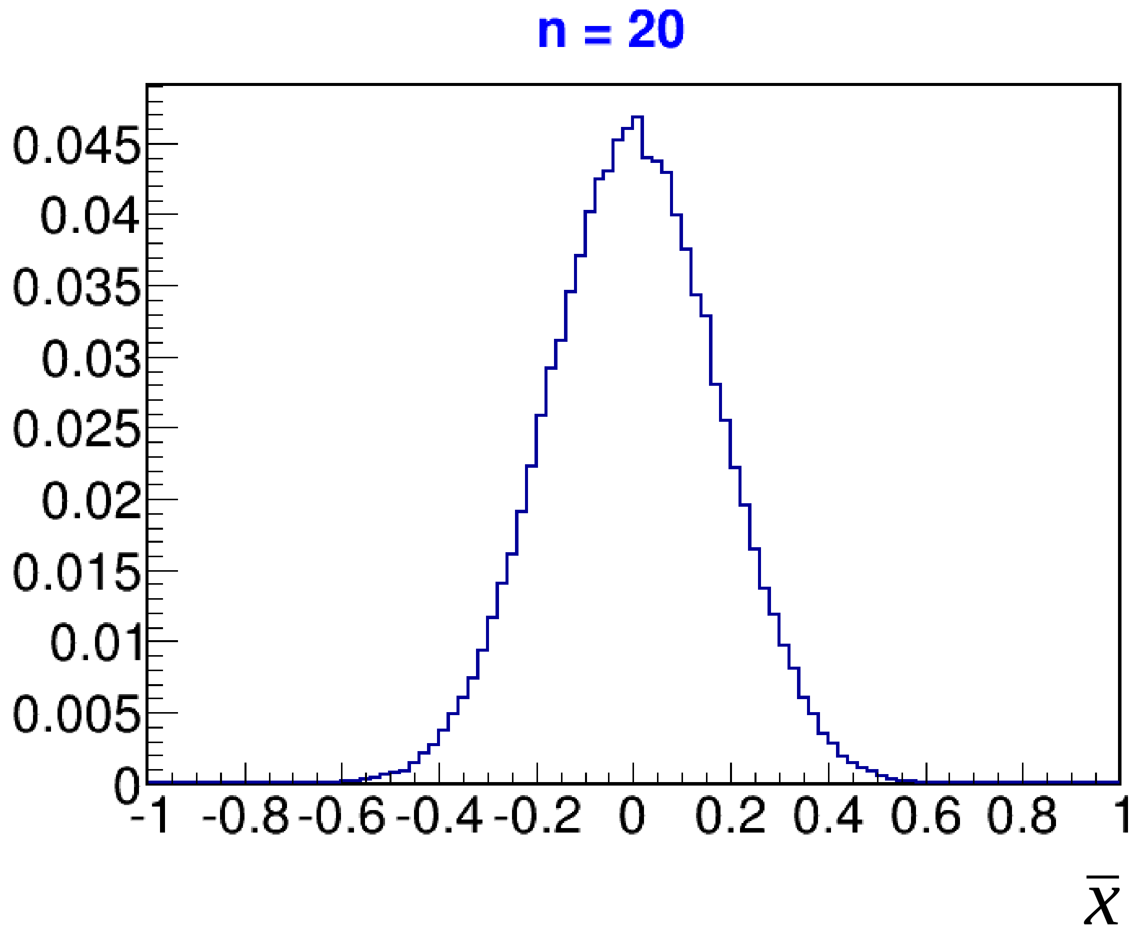
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a x^2 distribution (for illustration only)



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

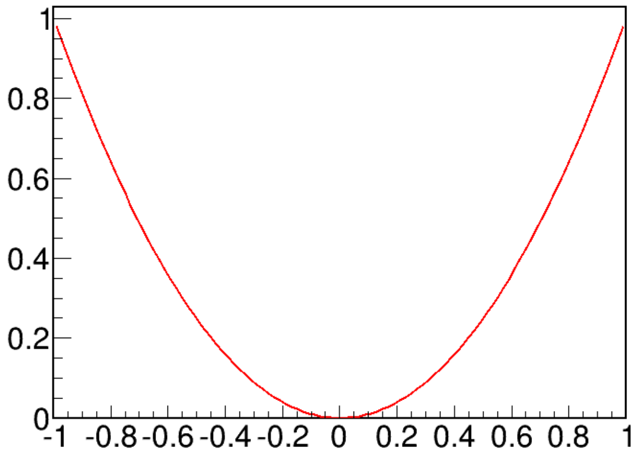


Distribution becomes Gaussian, although very non-Gaussian originally

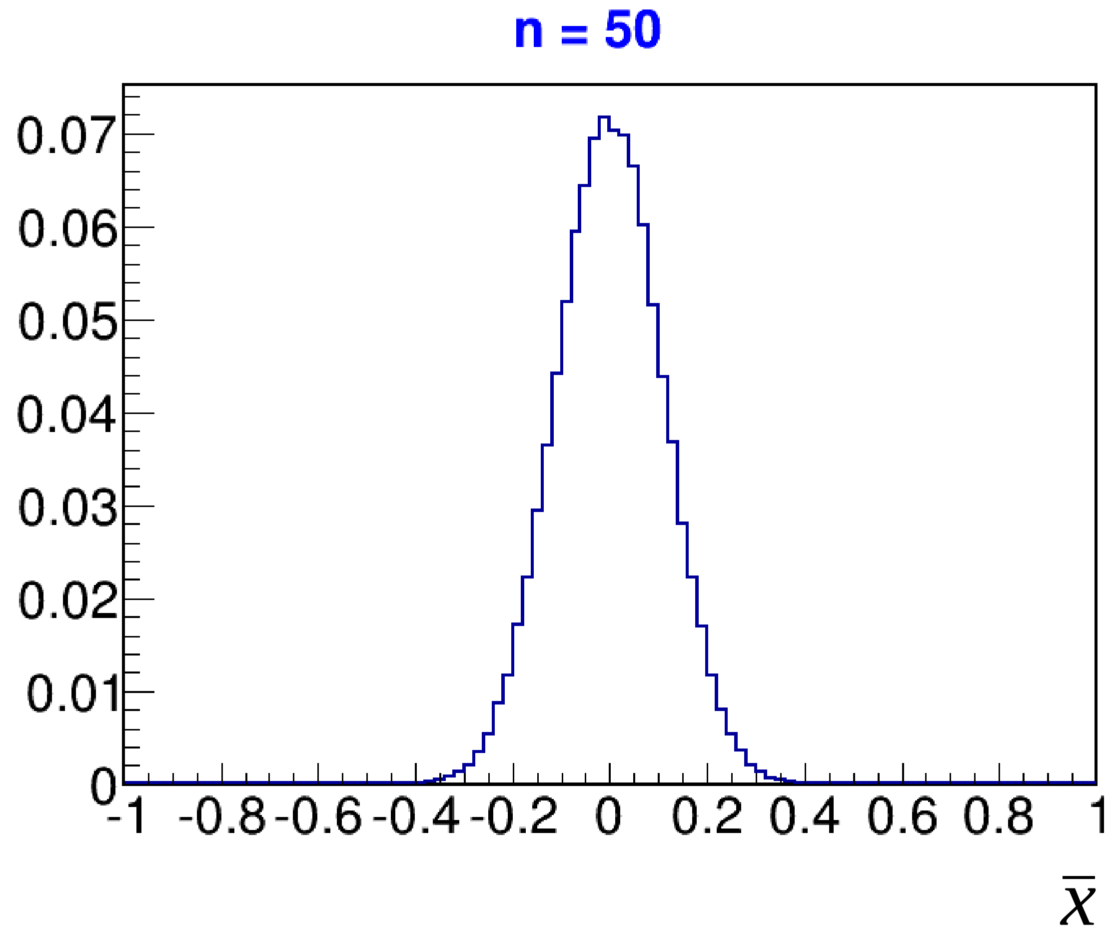
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a x^2 distribution (for illustration only)



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

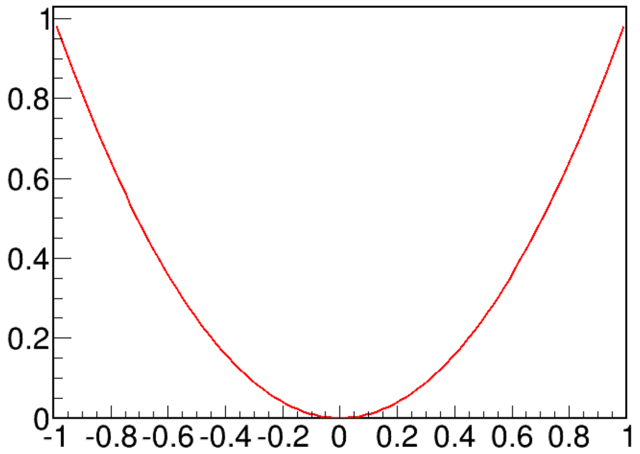


Distribution becomes Gaussian, although very non-Gaussian originally

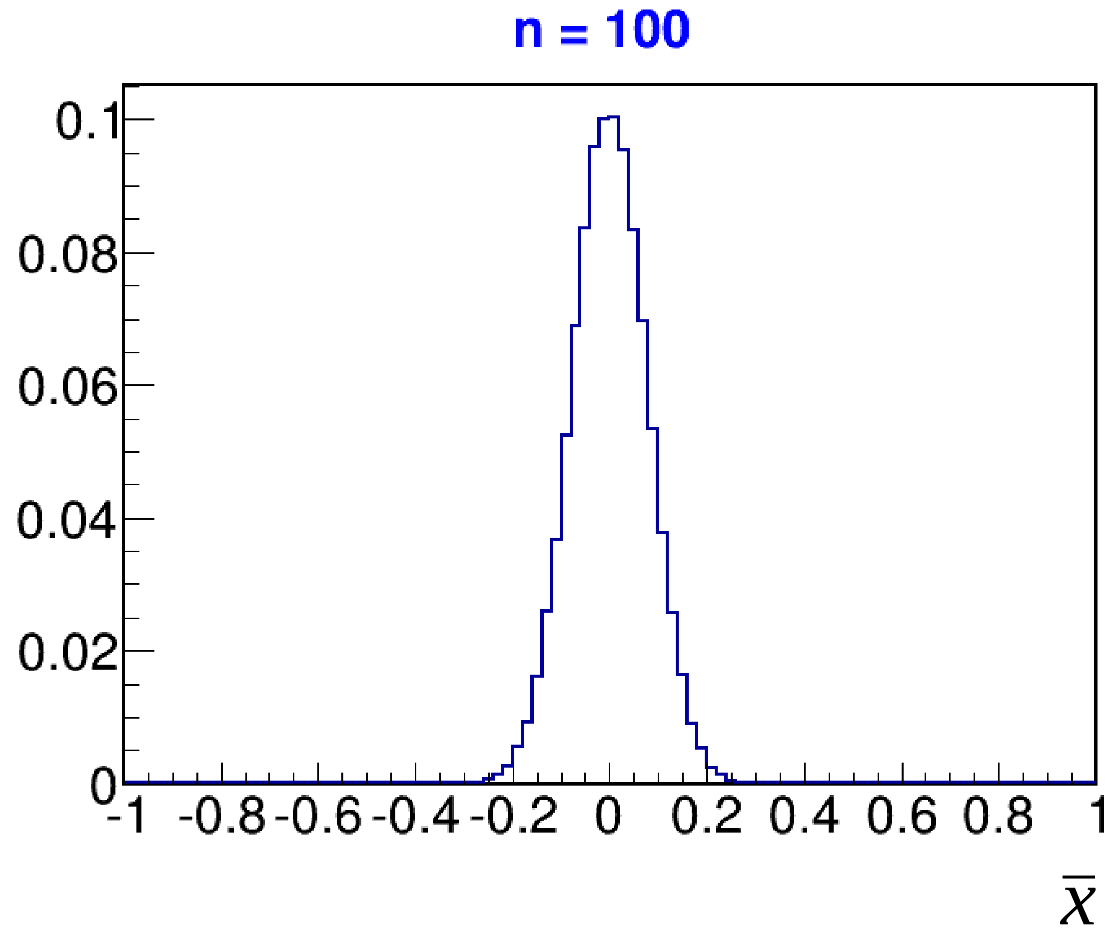
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a x^2 distribution (for illustration only)



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



Distribution becomes Gaussian, although very non-Gaussian originally

Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Outline

Statistics basics for HEP

- Random processes
- Probability distributions

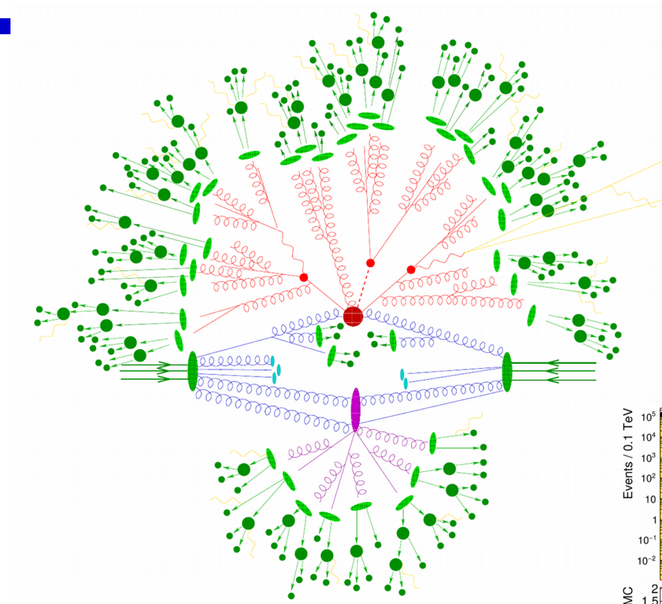
Describing HEP measurements

Computing statistics results

- Likelihoods
- Estimating parameter values
- Testing hypotheses
- Computing discovery significance

Describing HEP measurements

Statistical Model

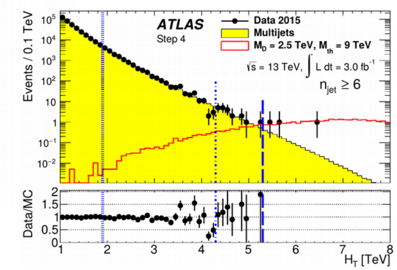


Hard scattering

Decays

Detector response

Reconstruction



Goal:
Describe the random process by which the data was obtained.

→ Build a **Statistical Model**

Ingredients:

1. **Statistical description** of the random aspects
   ⇒ **Probability distributions**
2. **Assumptions** on the underlying statistical processes (physics, etc.)
   → Uncertainties on the assumptions themselves: **systematic uncertainties**

"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 ?***

Statistical results can only be as accurate as the model itself !

Counting events

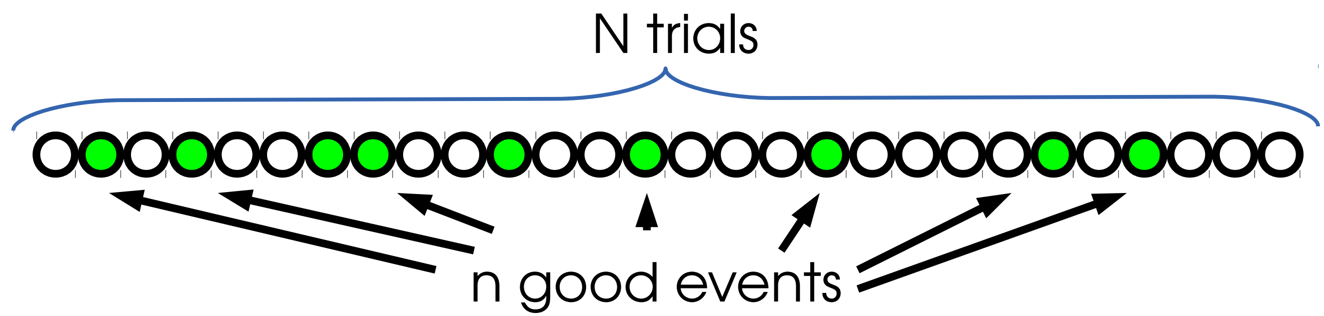
Consider N total events, select ***good*** events with probability P.
Probability to get **n good events** ?

Binomial distribution :

$$P(n; N, P) = C_N^n P^n (1-P)^{N-n}$$

Mean = **N·P**

Variance = **N·P(1 - P)**



However suppose **P ≪ 1, N ≫ 1**, and let **λ = N·P :**

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

Poisson distribution:

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

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

Mean = **λ**

Variance = **λ** ⇒ **RMS = √λ**

↑

Uncertainty of √N on N expected events

Rare Processes ?

HEP : almost always use Poisson distributions. Why ?

ATLAS :

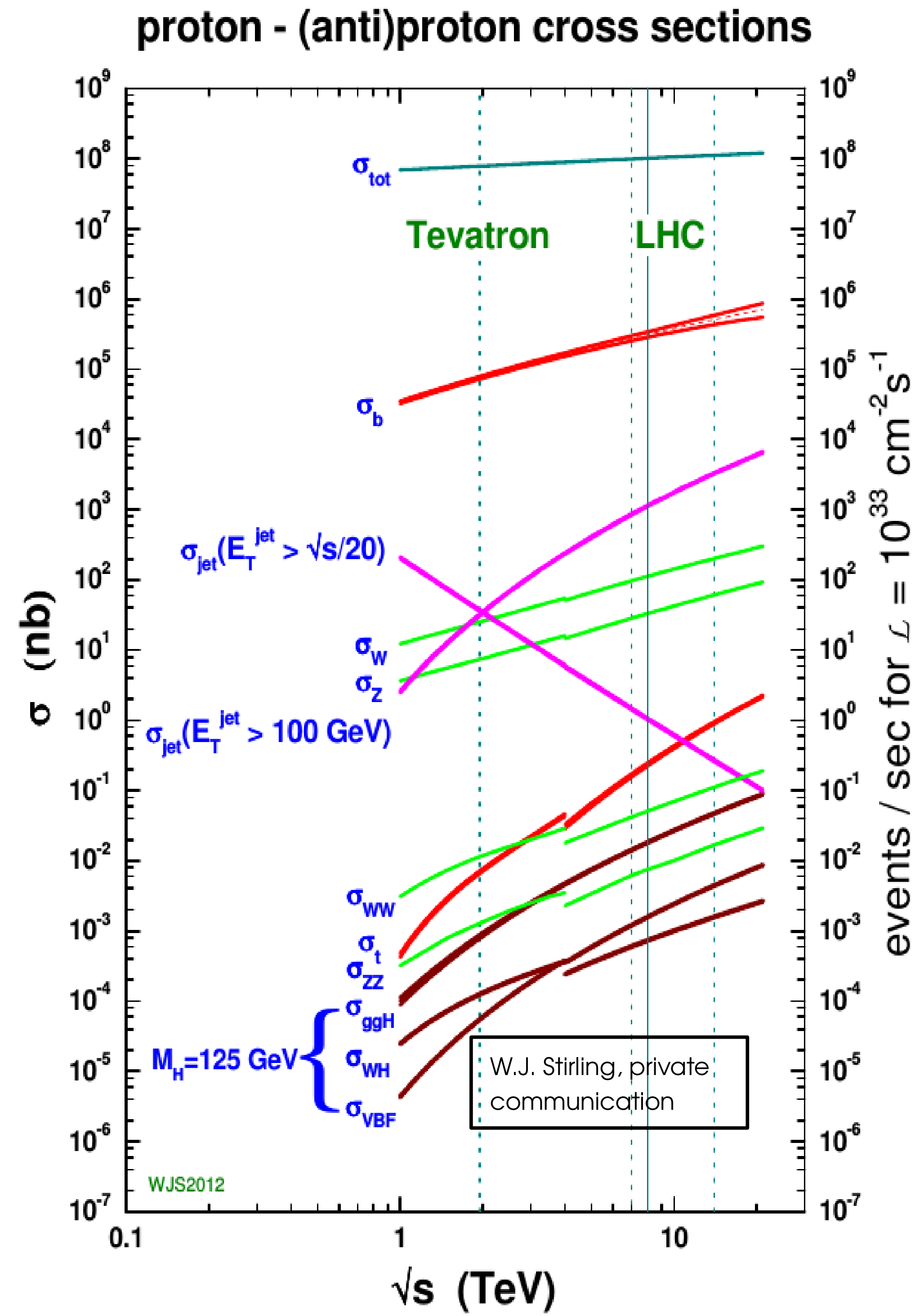
- **Event rate ~ 1 GHz**
  ($L \sim 10^{34}\ cm^{-2}s^{-1} \sim 10\ nb^{-1}/s$, $\sigma_{tot} \sim 10^{8}$ nb,)
- **Trigger rate ~ 1 kHz**
  (Higgs rate **~ 0.1 Hz**)

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

A day of data: **N ~ 10^{14} ≫ 1**

⇒ **Poisson regime!**

(Large N = design requirement, to get not-too-small λ=NP...)

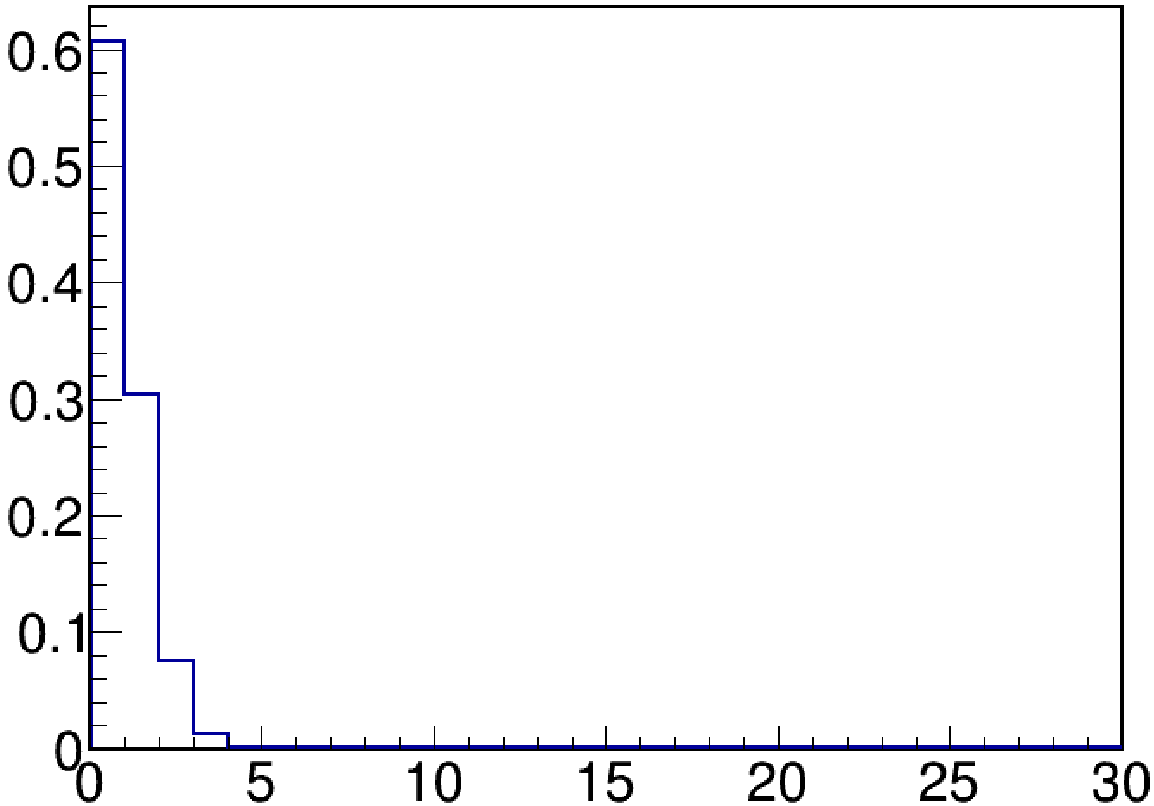


Poisson Distributions

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

λ : expected number of events

λ = 0.5



Mean = λ

Variance = λ

$\boldsymbol{\sigma} = \sqrt{\lambda}$

- **Discrete distribution** (integers only), **asymmetric for small λ**
- Typical variation (RMS) of n events is **$\sqrt{n}$**
- Central limit theorem : becomes **Gaussian for large λ** :

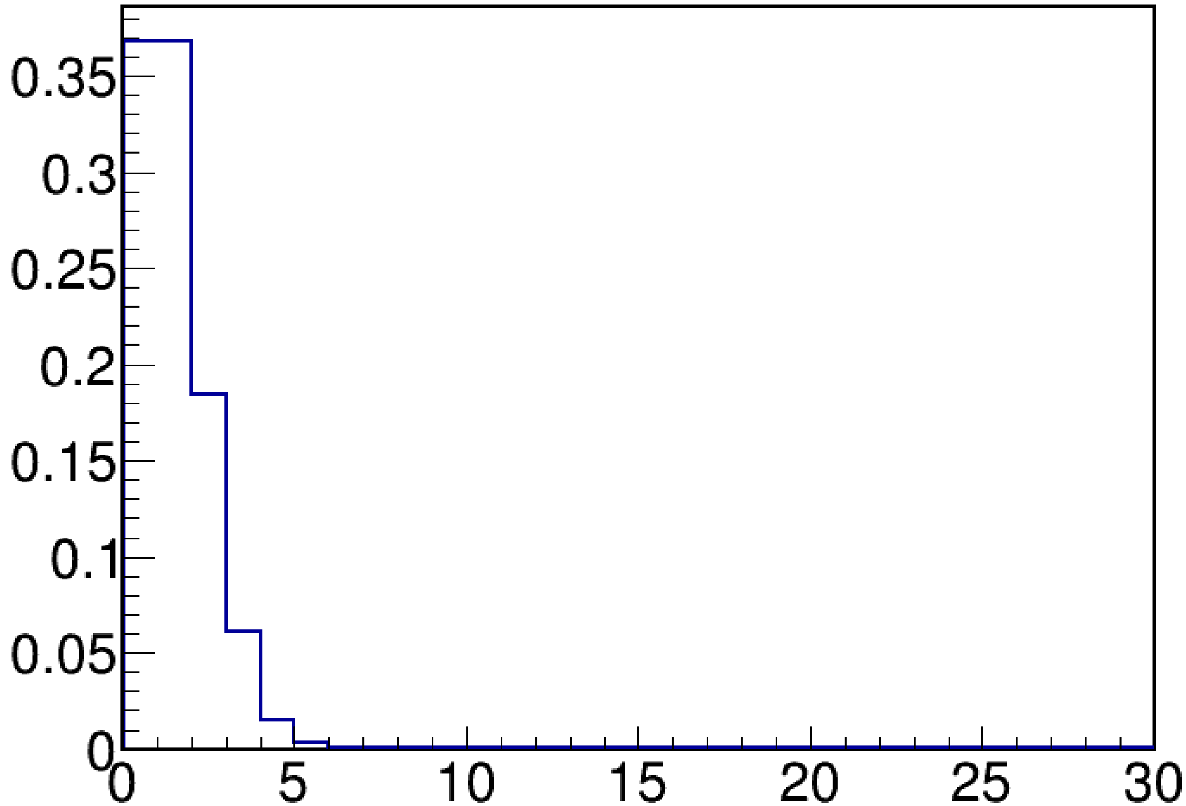
$$P(\lambda) \xrightarrow{\lambda \to \infty} G(\lambda, \sqrt{\lambda})$$

Poisson Distributions

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

λ : expected number of events

λ = 1



Mean = λ

Variance = λ

σ = √λ

- **Discrete distribution** (integers only), **asymmetric for small λ**
- Typical variation (RMS) of n events is **√n**
- Central limit theorem : becomes **Gaussian for large λ** :

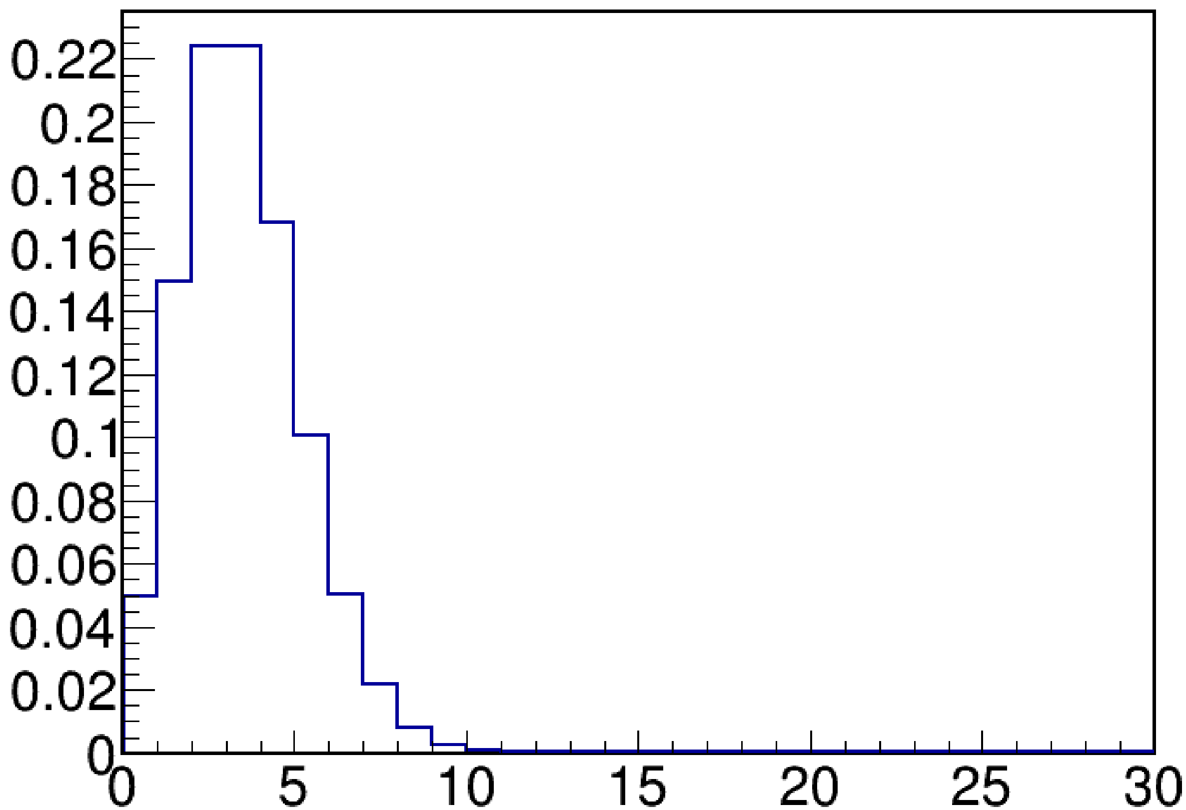
$$P(\lambda) \xrightarrow{\lambda \to \infty} G(\lambda, \sqrt{\lambda})$$

Poisson Distributions

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

λ : expected number of events

λ = 3



Mean = λ

Variance = λ

σ = √λ

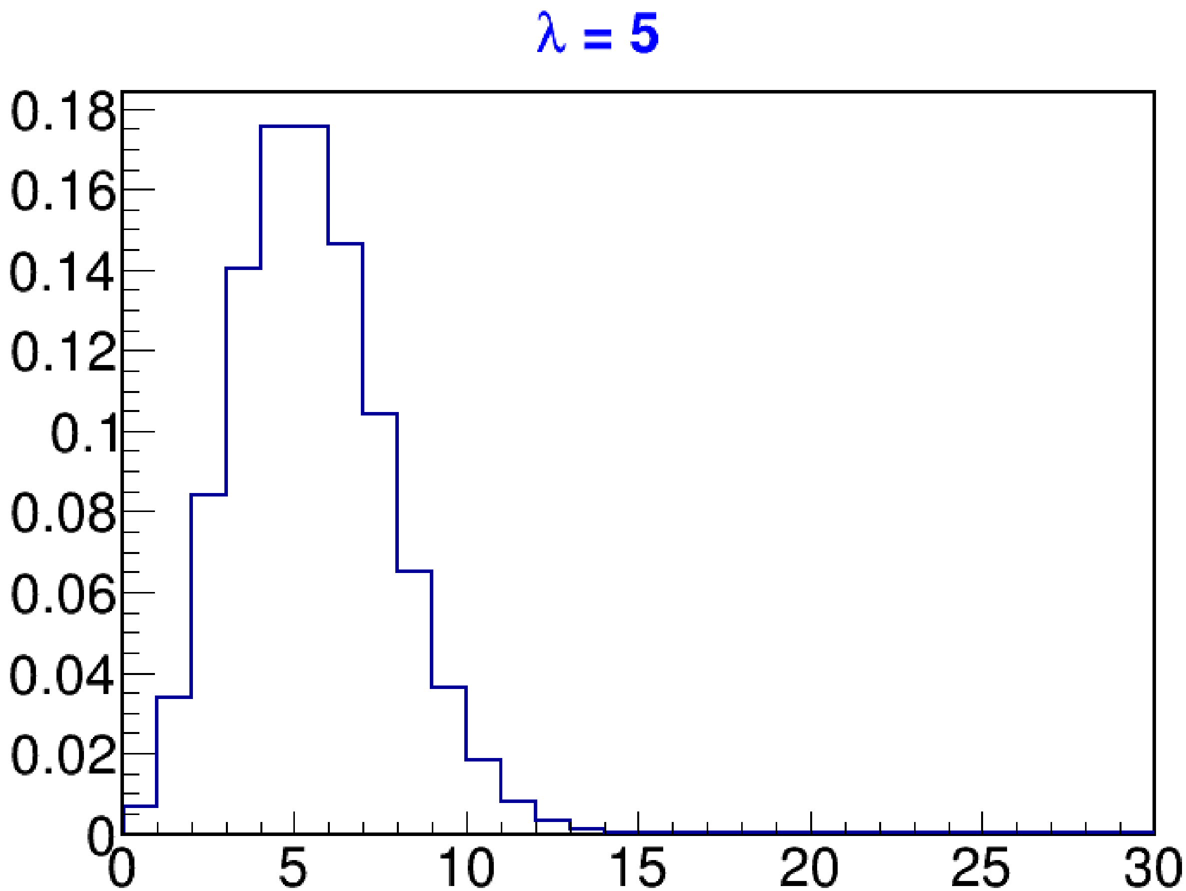
- **Discrete distribution** (integers only), **asymmetric for small λ**
- Typical variation (RMS) of n events is **√n**
- Central limit theorem : becomes **Gaussian for large λ** :

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

Poisson Distributions

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

λ : expected number of events



Mean = λ

Variance = λ

σ = √λ

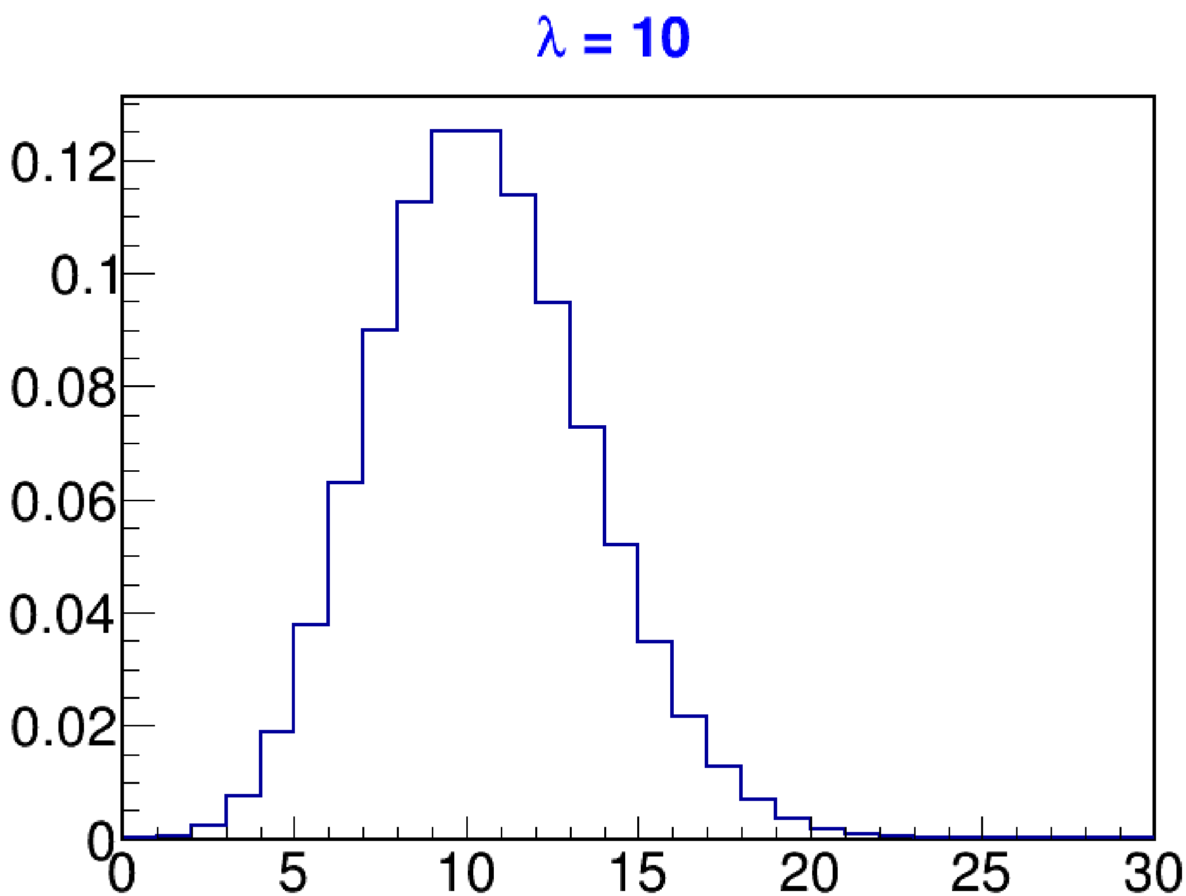
- **Discrete distribution** (integers only), **asymmetric for small λ**
- Typical variation (RMS) of n events is **√n**
- Central limit theorem : becomes **Gaussian for large λ** :

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

Poisson Distributions

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

λ : expected number of events



Mean = λ

Variance = λ

σ = √λ

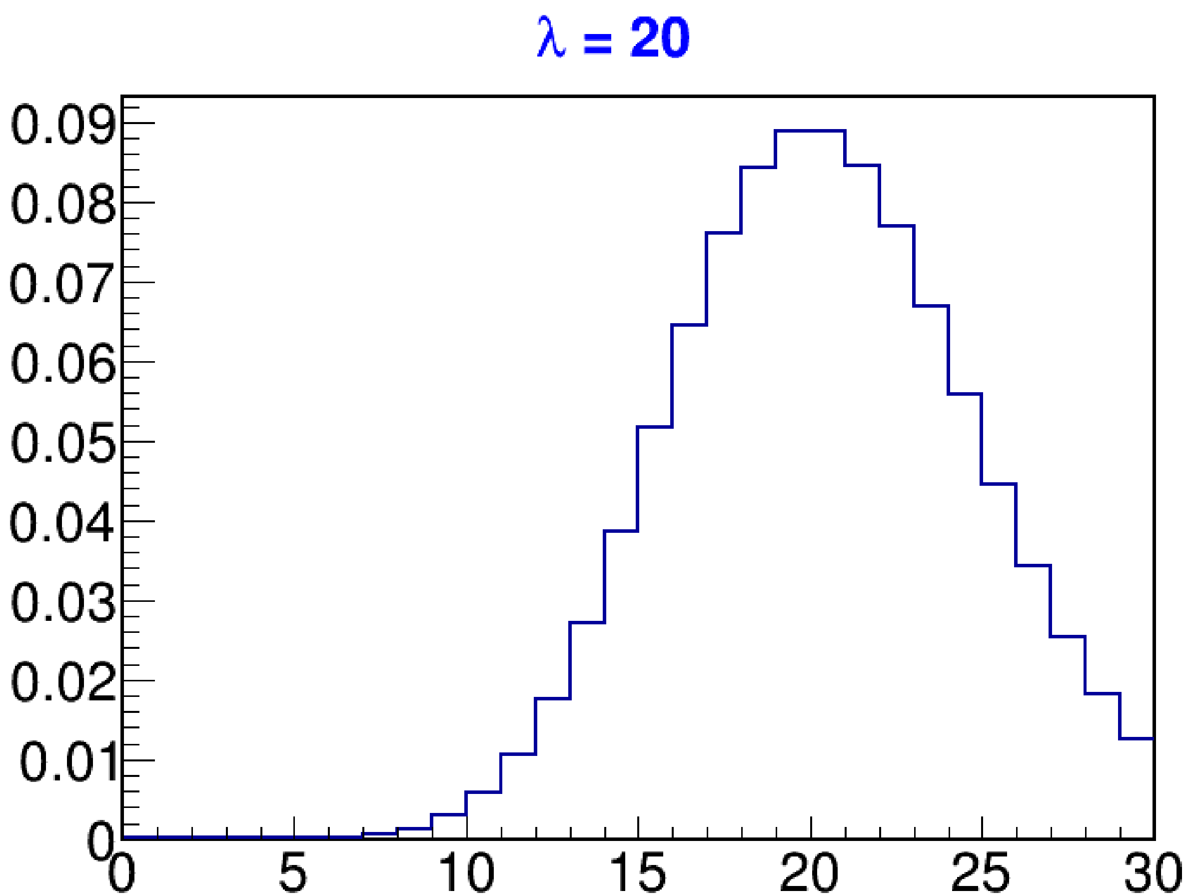
- **Discrete distribution** (integers only), **asymmetric for small λ**
- Typical variation (RMS) of n events is **√n**
- Central limit theorem : becomes **Gaussian for large λ** :

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

Poisson Distributions

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

λ : expected number of events



Mean = λ

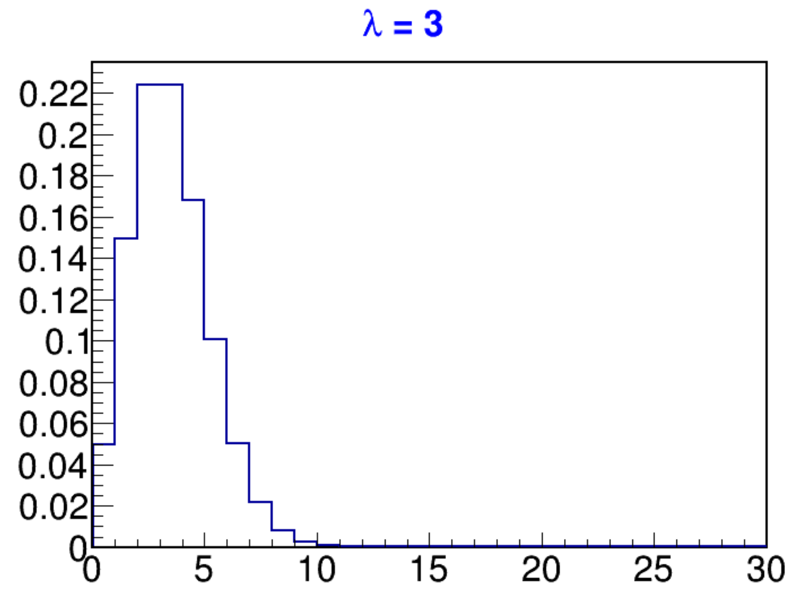
Variance = λ

σ = √λ

- **Discrete distribution** (integers only), **asymmetric for small λ**
- Typical variation (RMS) of n events is **√n**
- Central limit theorem : becomes **Gaussian for large λ** :

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

Statistical Model for Counting



Counting experiment:

observable: a **number of events n**

→ describe by a **Poisson distribution**

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

Typically both signal and background expected:

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

S : # of events from signal process

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

We have **assumed** a Poisson distribution for n : This is our model, based on physics knowledge (but usually a very safe one).

Model has **parameters S** and **B.** B can be known a priori or not (S usually not...)

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

→ usually up to uncertainties → **systematics**

$Z \rightarrow ee$ Inclusive σ^{fid}

Measurement Principle:

35000 ± (√35000 = 187)

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

175 ± 8

(81 ± 2) pb^{-1}

0.552 ± 0.006

Signal events	34865 ± 187 ± 7 ± 3
Correction C	$0.552\ ^{+0.006}_{-0.005}$
σ^{fid}[nb]	0.781 ± 0.004 ± 0.008 ± 0.016

Phys. Lett. B 759 (2016) 601

Simple uncertainty propagation:

σ^{fid} = 0.781 ± 0.004 (stat) ± 0.008 (syst) ± 0.016 (lumi) nb

→ **Simplest possible example** in several ways

– **"Single bin counting"** : only data input is $\mathbf{N}_{\mathbf{data}}$.

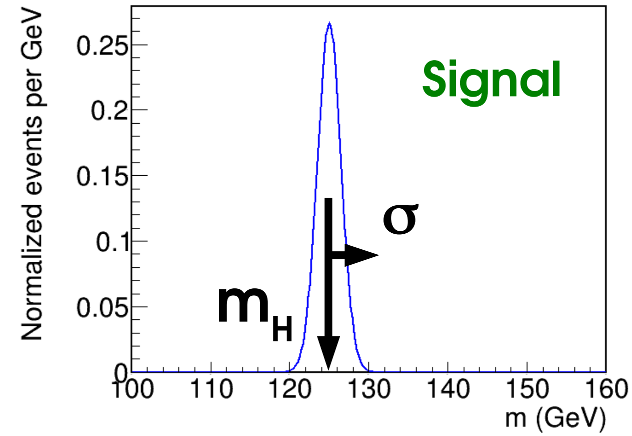
– Describe using **Poisson** distribution, or **Gaussian** for large n_{data}

Unbinned Shape Analysis

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

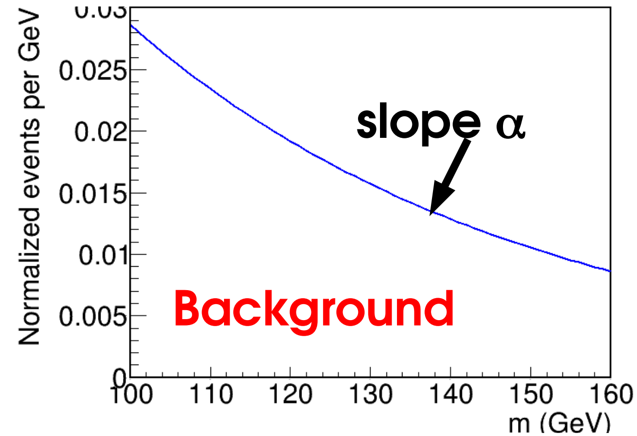
→ Describe shape of the **distribution of m**

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



H→γγ-inspired example:

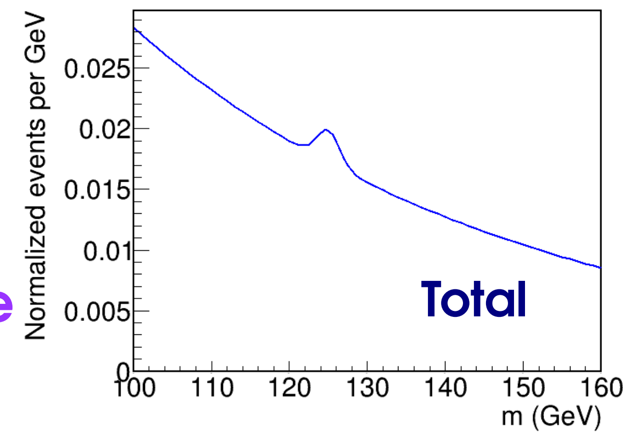
- **Gaussian signal** $P_{\text{signal}}(m) = G(m; m_H, \sigma)$
- **Exponential bkg** $P_{\text{bkg}}(m) = \alpha\, e^{-\alpha m}$



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**

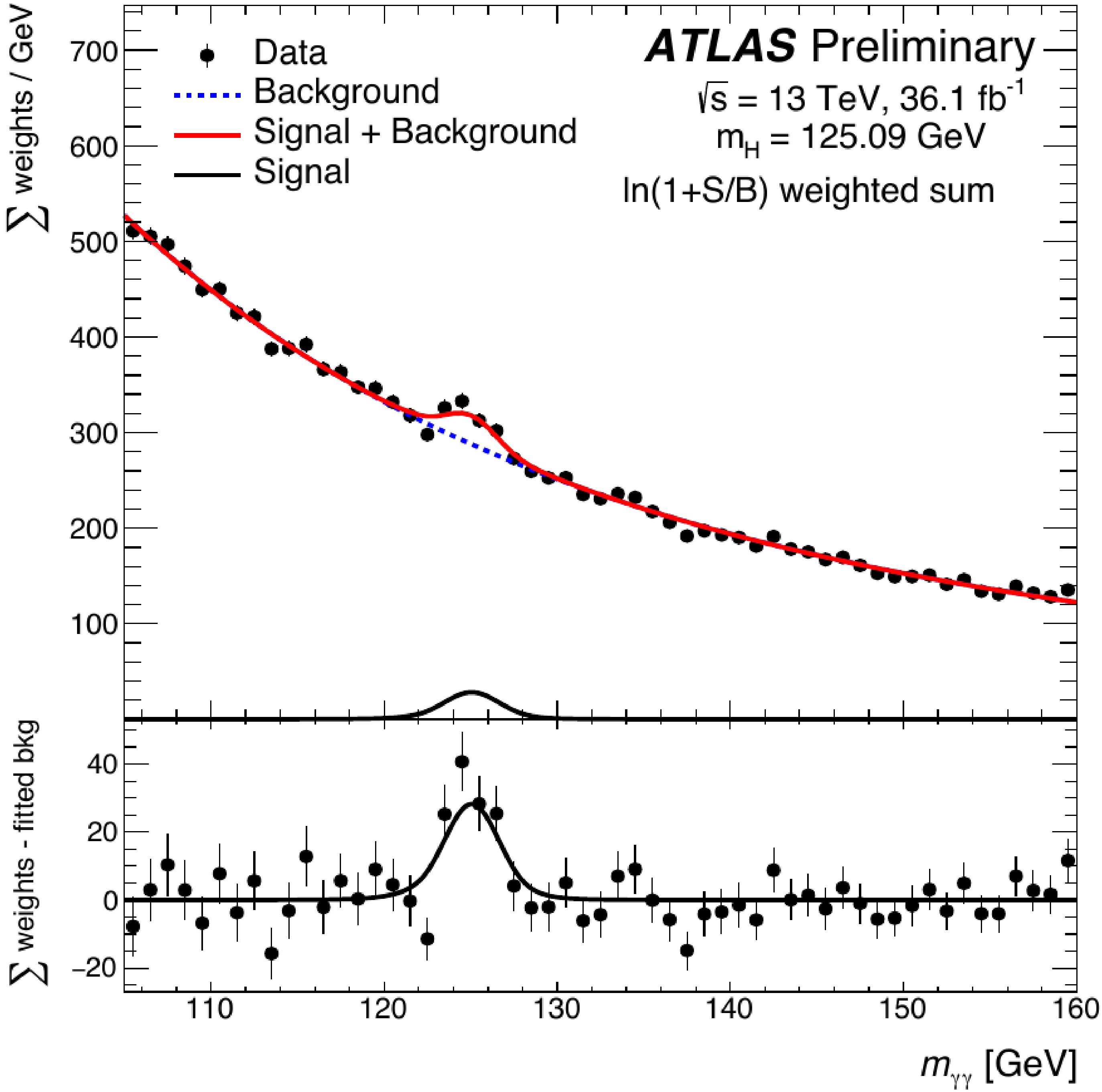
Probability to observe n events

Probability to observe the value m_i

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

H→γγ

ATLAS-CONF-2017-045



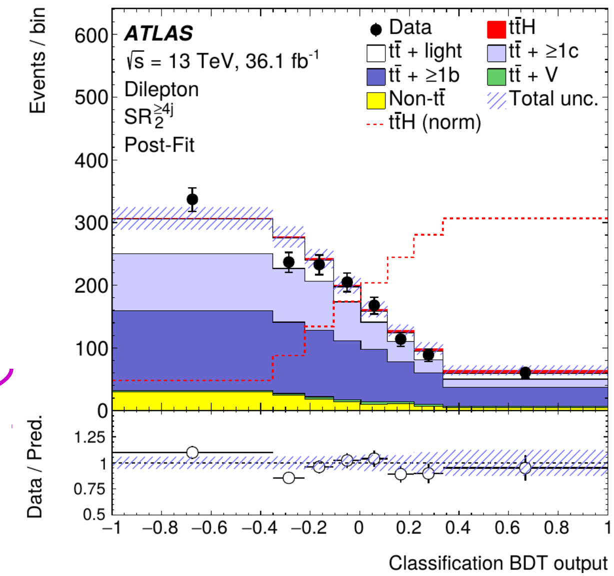
The Halfway Option: Binned Shape Analysis

Instead of using $m_1 ... m_n$ directly, can build a histogram $n_1 ... n_N$.
→ N : number of bins

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

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

Poisson distribution in each bin



N=1: Counting analysis
N→∞: Unbinned shape analysis (the fractions become PDF values)

Shapes specified through $f_{S,i}$, $f_{B,i}$ rather than $P_{signal}(m)$, $P_{bkg}(m)$

⊕ **Obtained directly from MC**, no need to define continuous PDFs.

⊖ **MC stat fluctuations** can create artefacts, especially for S≪B.

→ discussed in more detail on Wednesday

Summary: How to describe data

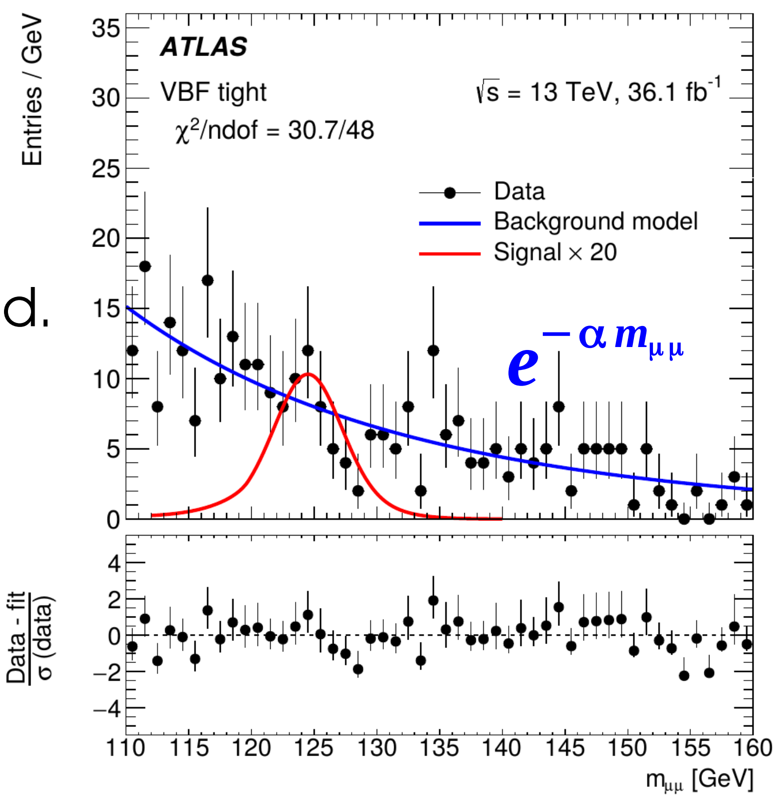
Description	Observable	Likelihood
Counting	**n** : measured number of events	**Poisson** $$P(n;S,B) = e^{-(S+B)}\frac{(S+B)^n}{n!}$$ **S**, **B** : expected signal & background
Binned shape analysis	n_i, i=1..N_{bins} : measured events in each bin.	**Poisson product** $$P(n_i;S,B) = \prod_{i=1}^{n_{bins}} e^{-(S f_i^{sig} + B f_i^{bkg})}\frac{(S f_i^{sig} + B f_i^{bkg})^{n_i}}{n_i!}$$ **S**, **B** : expected signal & background f^{sig}_i, f^{bkg}_i : fraction of sig & bkg in each bin
Unbinned shape analysis	m_i, i=1..n_{evts} : observable value for each event	**Extended Unbinned Likelihood** $$P(m_i;S,B) = \frac{e^{-(S+B)}}{n_{evts}!}\prod_{i=1}^{n_{evts}} S\, P_{sig}(m_i) + B\, P_{bkg}(m_i)$$ **S**, **B** : expected signal & background P_{sig}, P_{bkg} : PDFs for **m** in signal and bkg.

Model Parameters

Model typically includes:

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

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



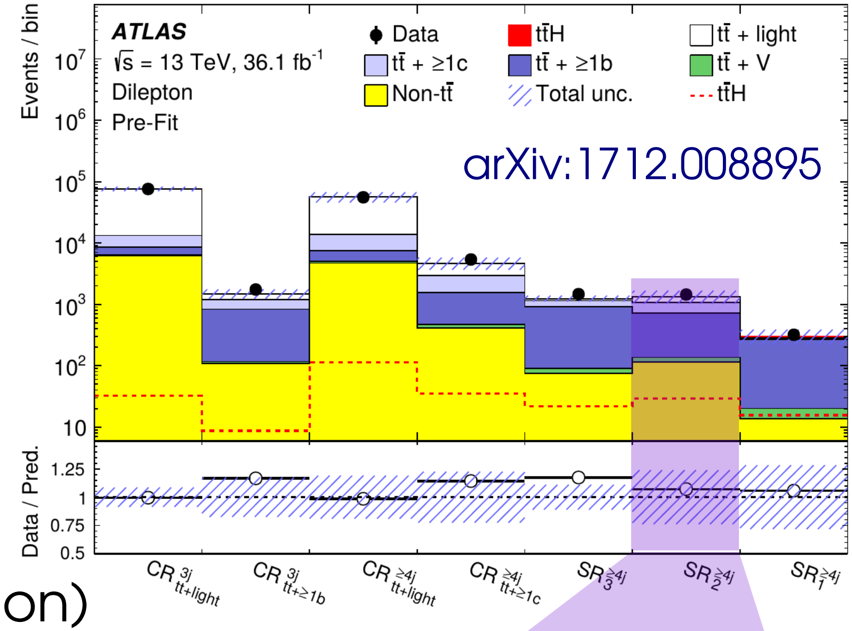
NPs must be either
→ **known a priori** (possibly within systematics) or
→ **constrained by the data** (e.g. in sidebands)

Phys. Rev. Lett. 119 (2017) 051802

Categories

Multiple analysis regions often used:

- Multiple decay modes
- Multiple kinematic selections, etc.



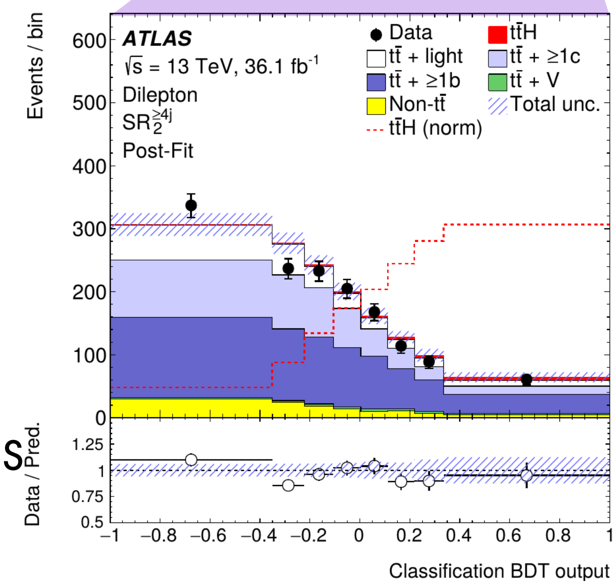
→ Useful to model these separately if

- Better sensitivity in some regions (avoids dilution)
- Some regions can constrain NPs
  - e.g. ***Control regions*** for backgrounds

⇒ ***Analysis categories*** :

PDF for category k

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



No overlaps between categories ⇒ No stat. correlations
⇒ can simply take product of PDFs.

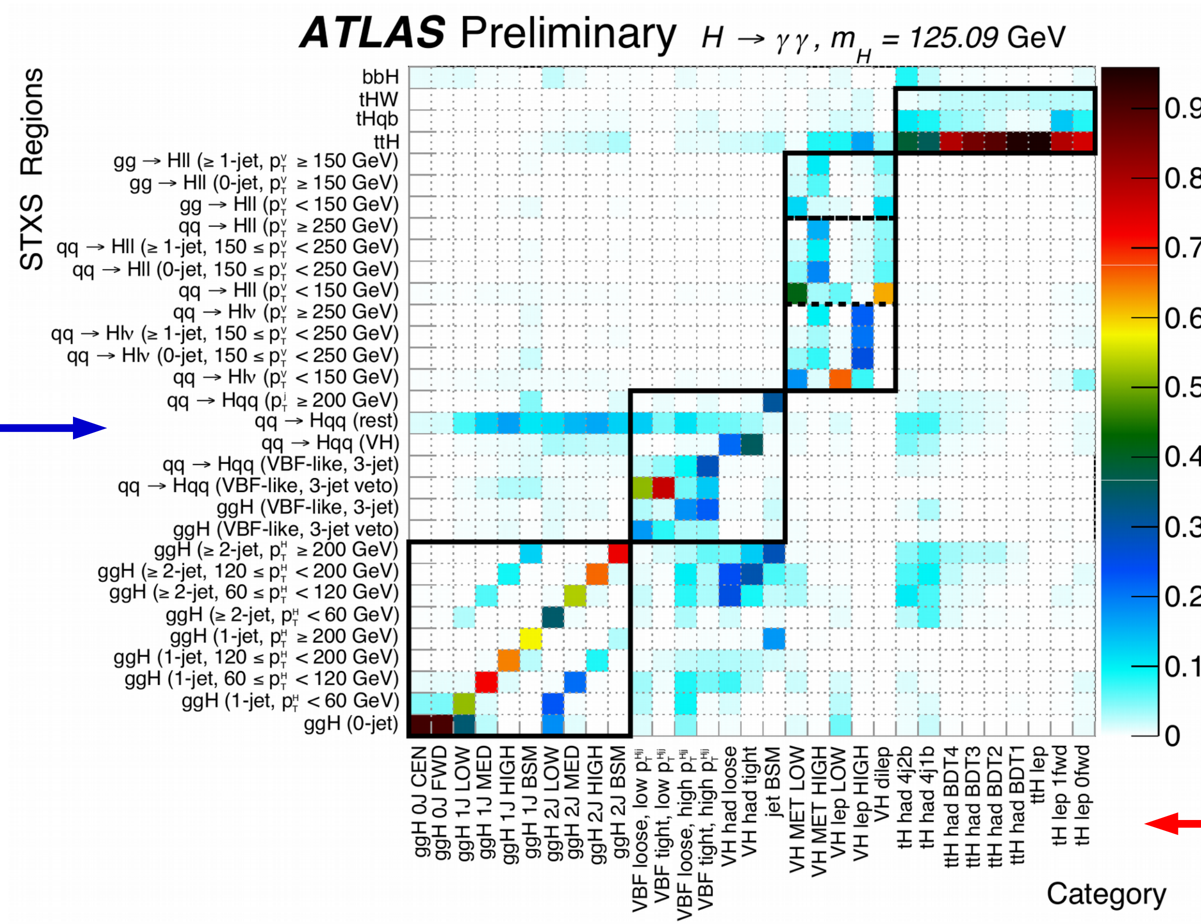
→ Similar to a-posteriori combination of the various regions, but allows proper handling of correlated parameters (e.g. systematics).

Categories for H→γγ Property Measurements

Categories also useful to provide measurements of separate kinematic regions

→ e.g. differential cross-section measurements

Targeted truth regions →



H→ γγ Properties Measurement (ATLAS-CONF-2017-045)

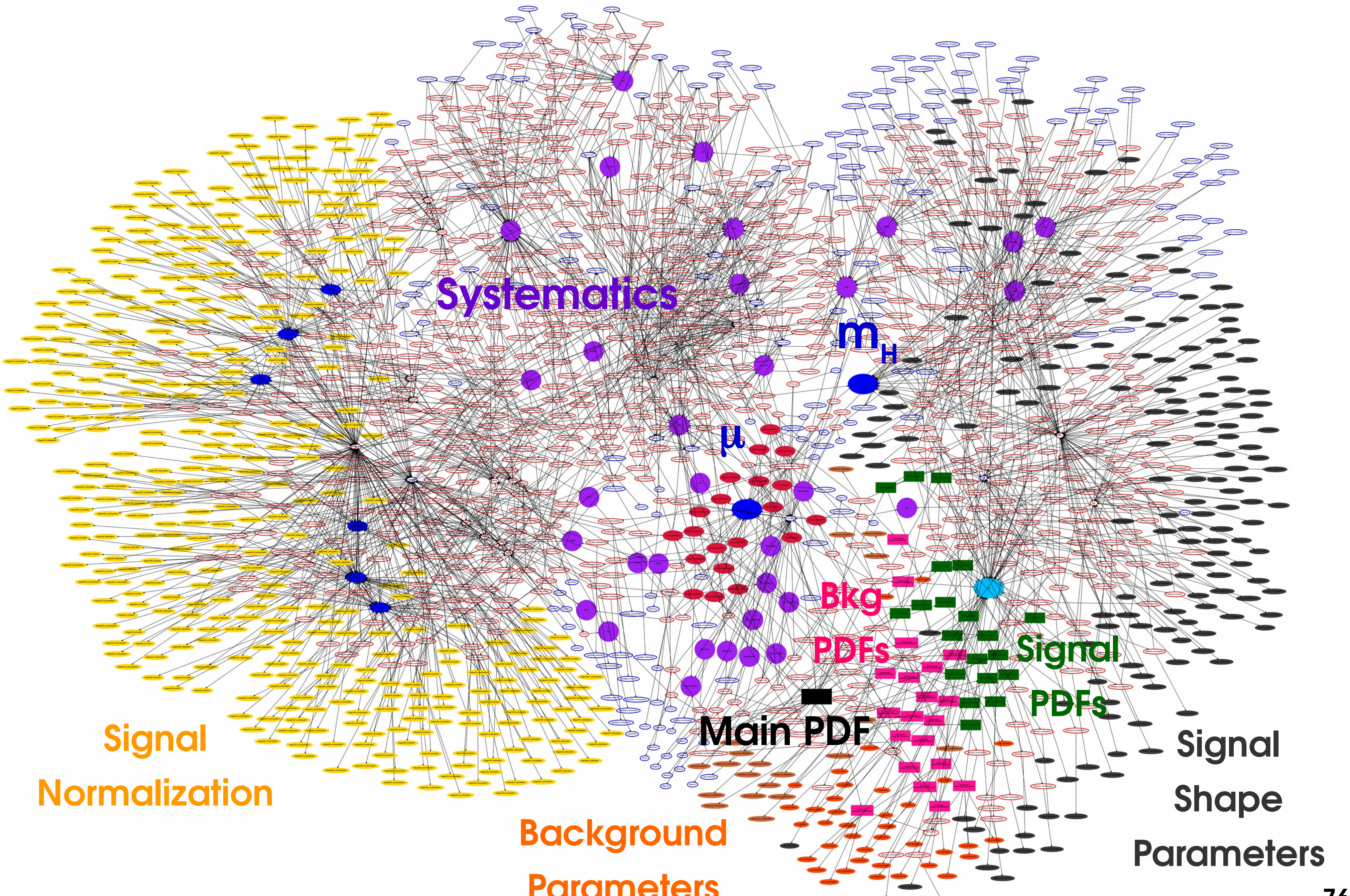
← **Analysis Selections**

Most **categories** aimed at one particular **truth region**

→ also cross-feed from other regions (detector acceptance, pileup, etc.)

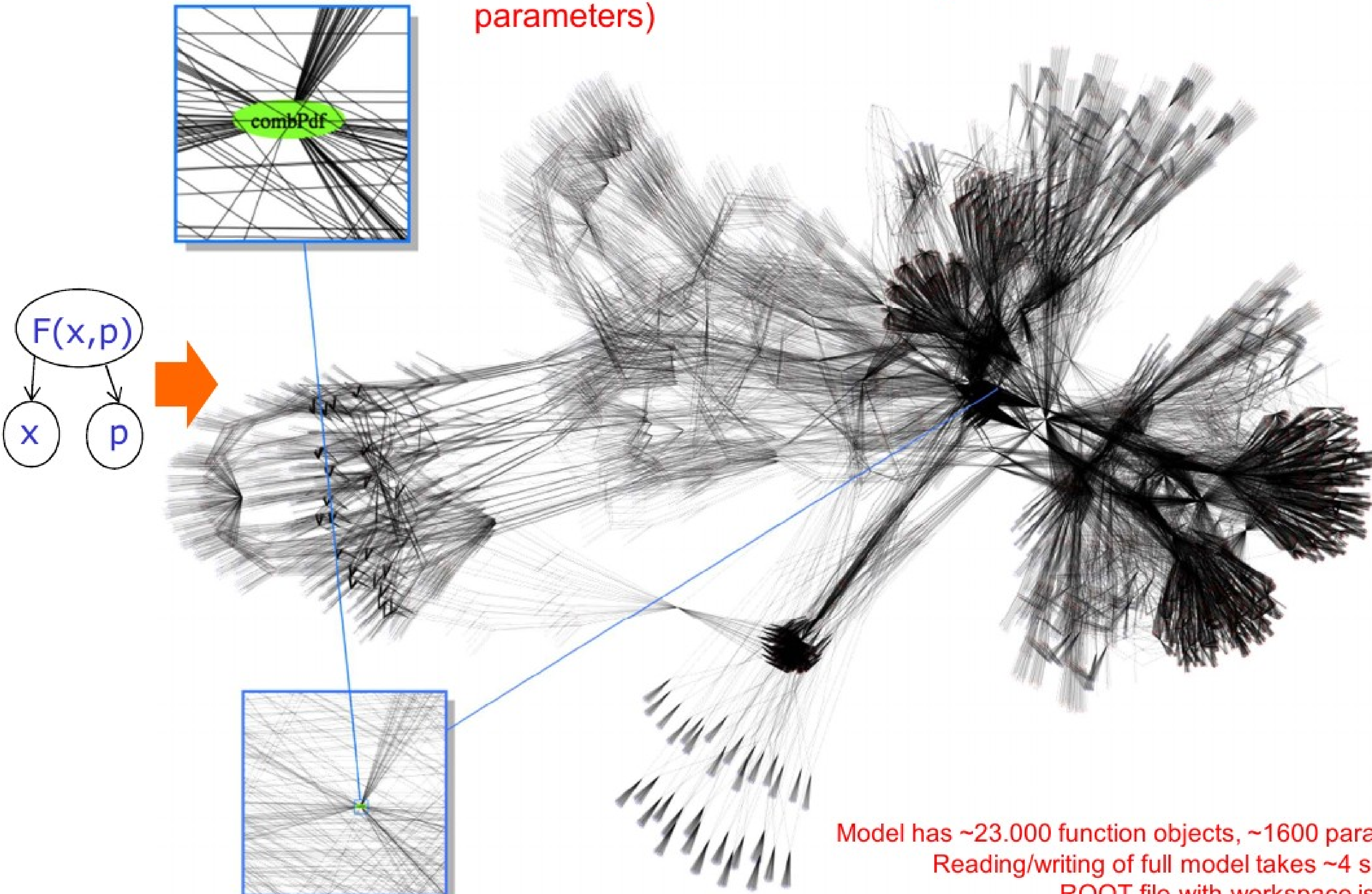
⇒ **Combined analysis for optimal use of all information**

Model Example: H→γγ Discovery Analysis



ATLAS Higgs Combination Model

Atlas Higgs combination model (23.000 functions, 1600 parameters)

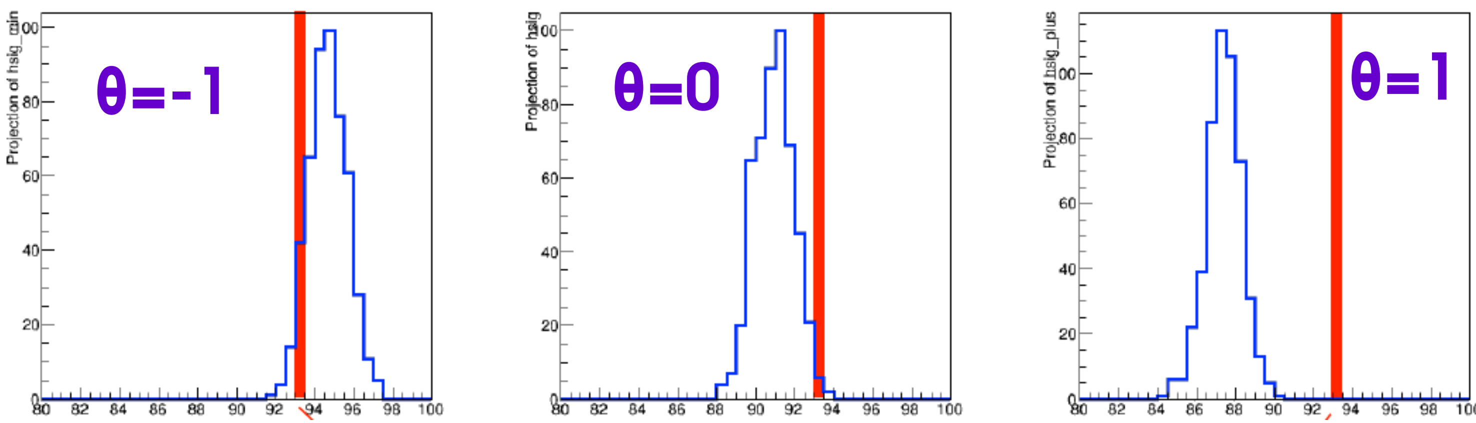


Model has ~23.000 function objects, ~1600 parameters
Reading/writing of full model takes ~4 seconds
ROOT file with workspace is ~6 Mb

Technical Implementation

Implemented in **ROOT** using the **RooFit/RooStats/HistFactory** toolkits

- **C++ classes** for PDFs, formulas, variables, etc.
- **Numerical methods**: convolutions, automatic computation of normalization factors. **Analytical evaluation** used when possible
- **Template morphing**



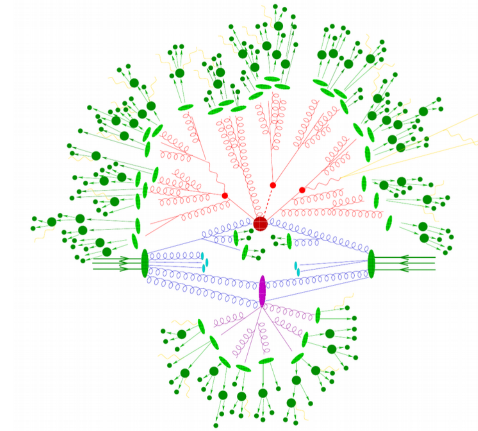
- Storage in **RooWorkspace** structures within ROOT files

  → Standard tools in LHC experiments, used in similar ways in ATLAS and CMS

Realistic models can be quite complex: ATLAS+CMS Higgs couplings comb. :

- **20** POIs, **4200** parameters, **600** categories
- > 7 GB memory footprint
- Time for 1 MINUIT fit ~ **O(few hours)**

Takeaways



HEP data is produced through **random processes**,
Need to be described using a statistical model:

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

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

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

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

Outline

Statistics basics for HEP

Random processes

Probability distributions

Describing HEP measurements

Computing statistics results

Likelihoods

Estimating parameter values

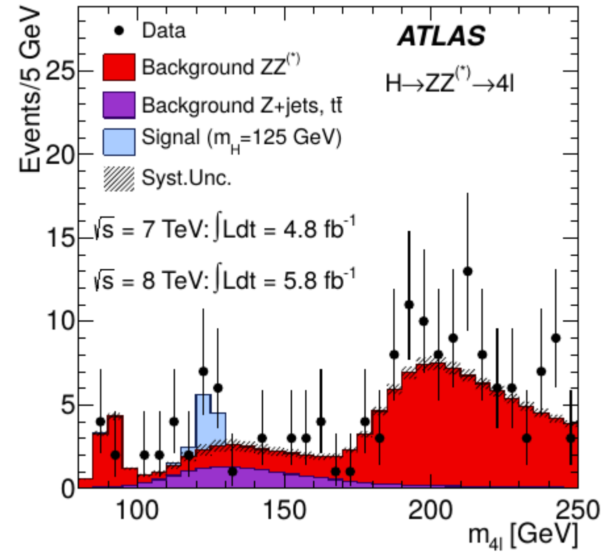
Testing hypotheses

Computing discovery significance

Computing Statistical Results

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

Overview



What we have so far:

- **Observed data**

- **Statistical model** : **P(data; parameters)**
  description of the random process producing the data
  → includes parameters that we want to measure (**S**, **σ × B**, **m_W**, ...)

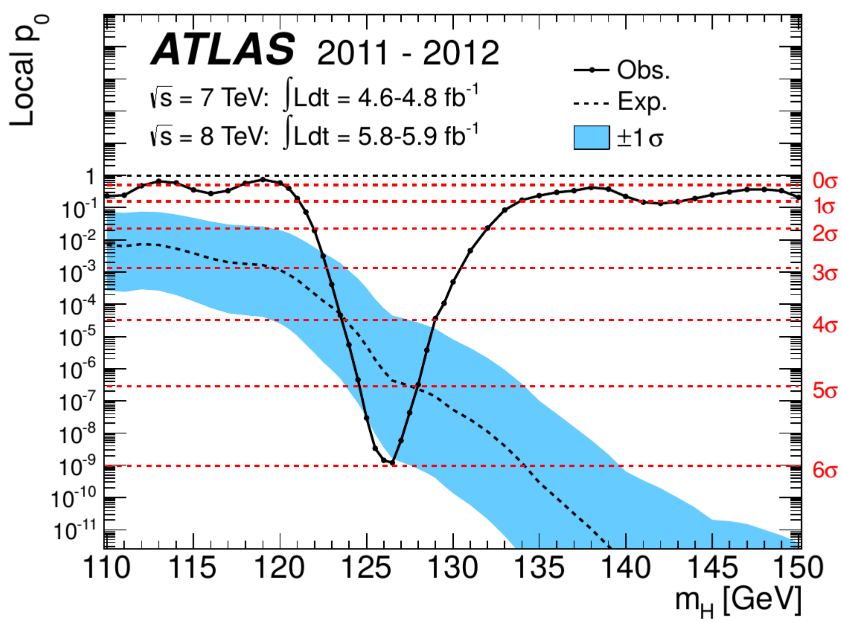
What we want : *Statistical Results*

- **Parameter measurement**: x_0 ± uncertainty

- **Upper limits** on signal yields, etc.

- **Discovery significance**

- ...



Computing Statistical Results
I. Parameter Estimation

Using the PDF

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

⇒ Possible outcomes of the experiment, for given parameter values

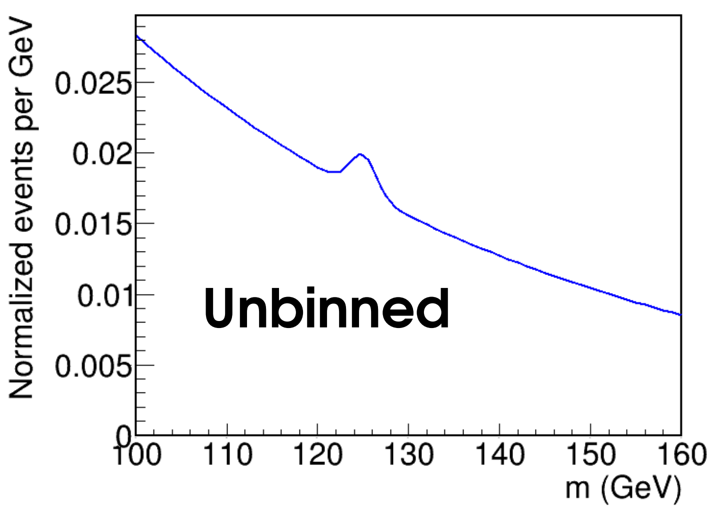
Can draw random events according to PDF : **generate** ***pseudo-data***

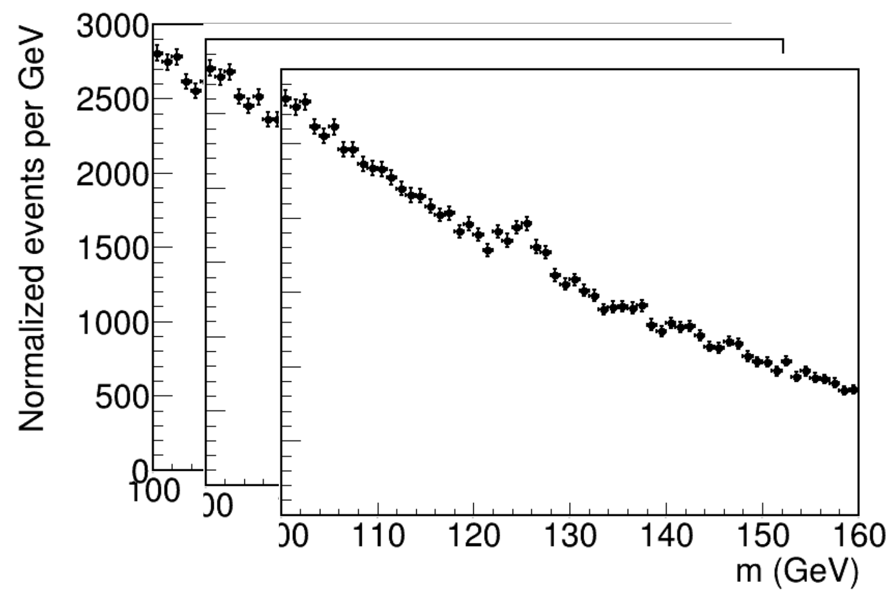
$$P(\lambda=5)$$

2, 5, 3, 7, 4, 9, ….

Each entry = separate "experiment"

Generate





Likelihood

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

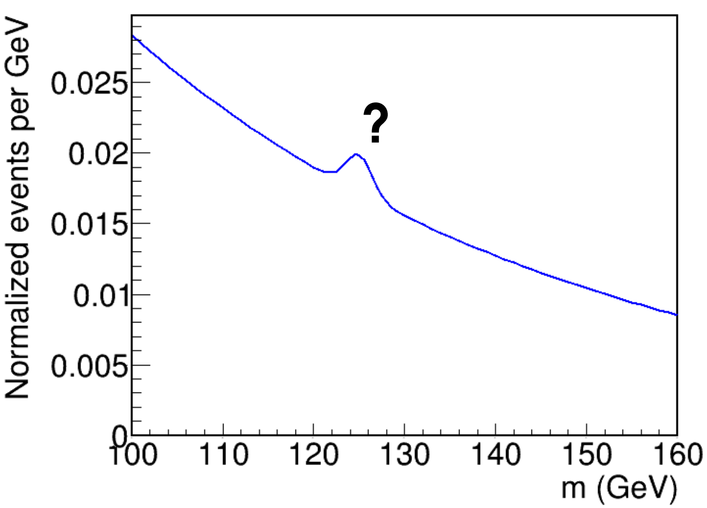
⇒ Possible outcomes of the experiment, for given parameter values

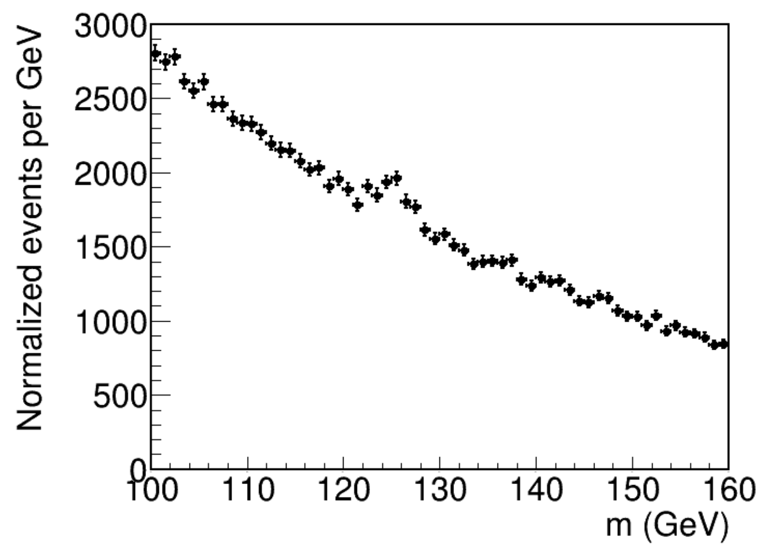
We want the other direction: **use data to get information on parameters**

$$P(\lambda = ?)$$

2

Estimate





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

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

Poisson Example

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

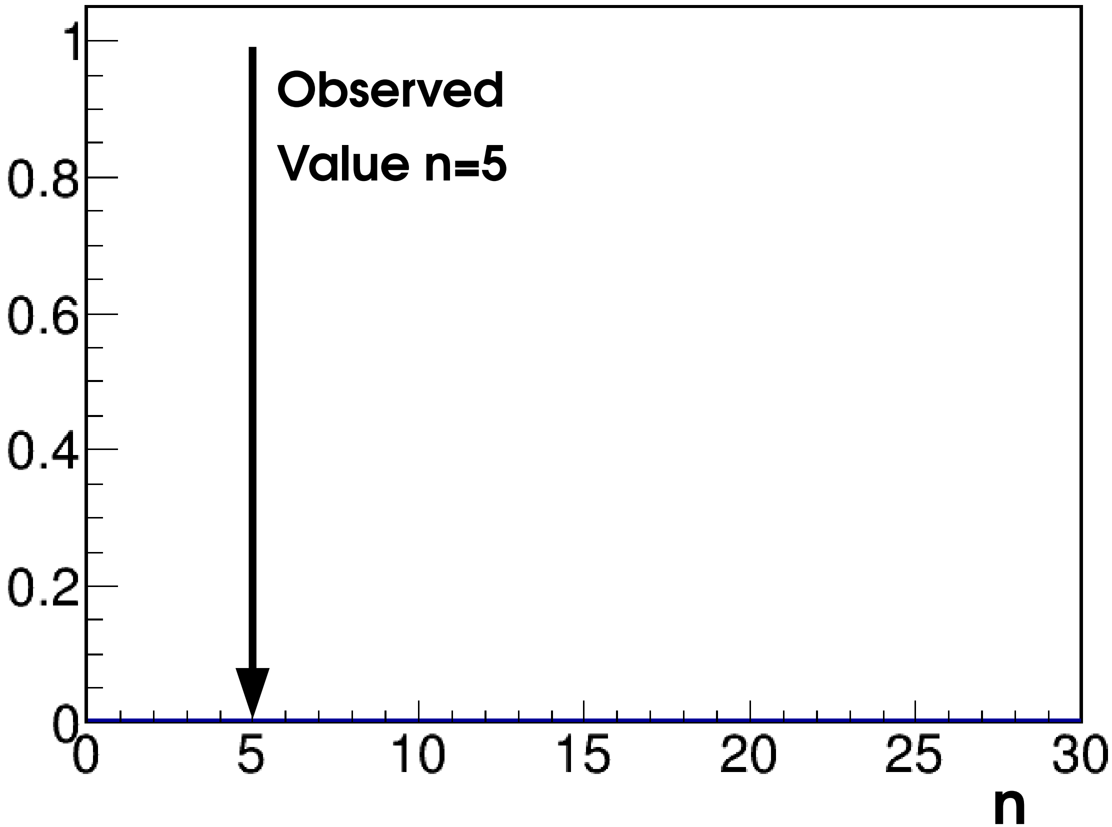
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

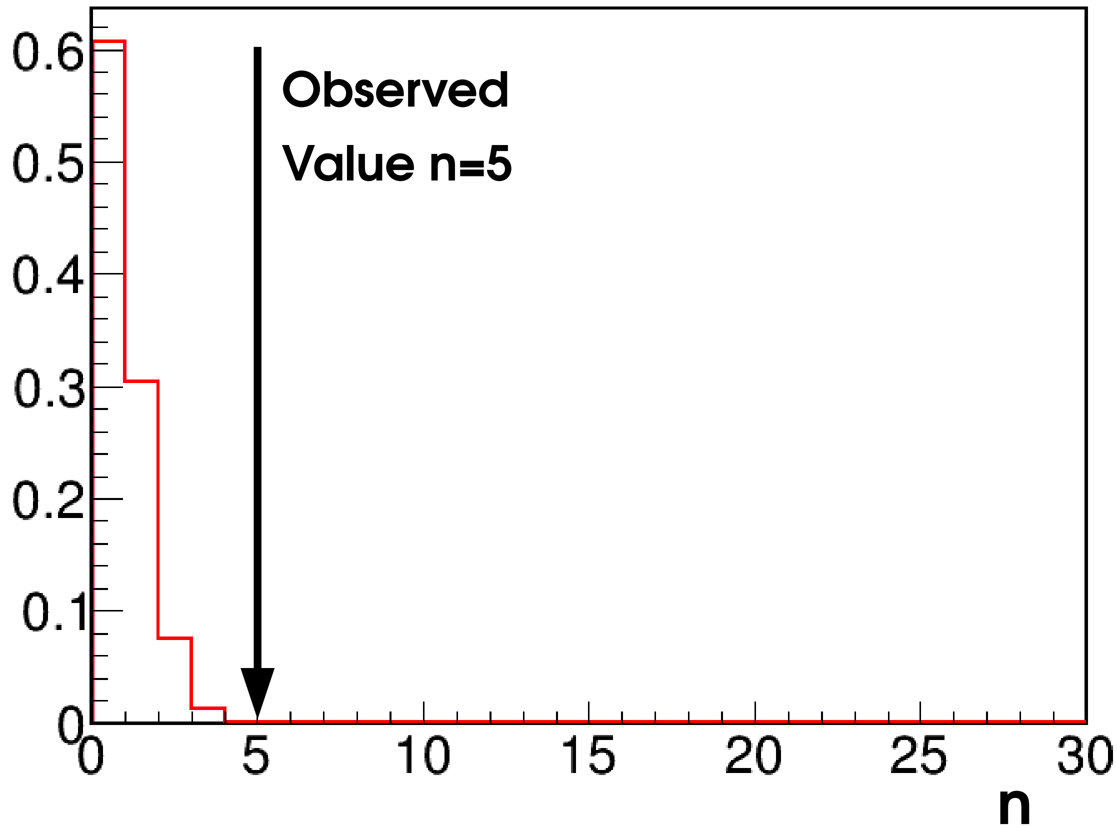
→ Varying parameter, fixed data: **likelihood**

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

P(S = 0.5)

Low

likelihood



Poisson Example

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

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

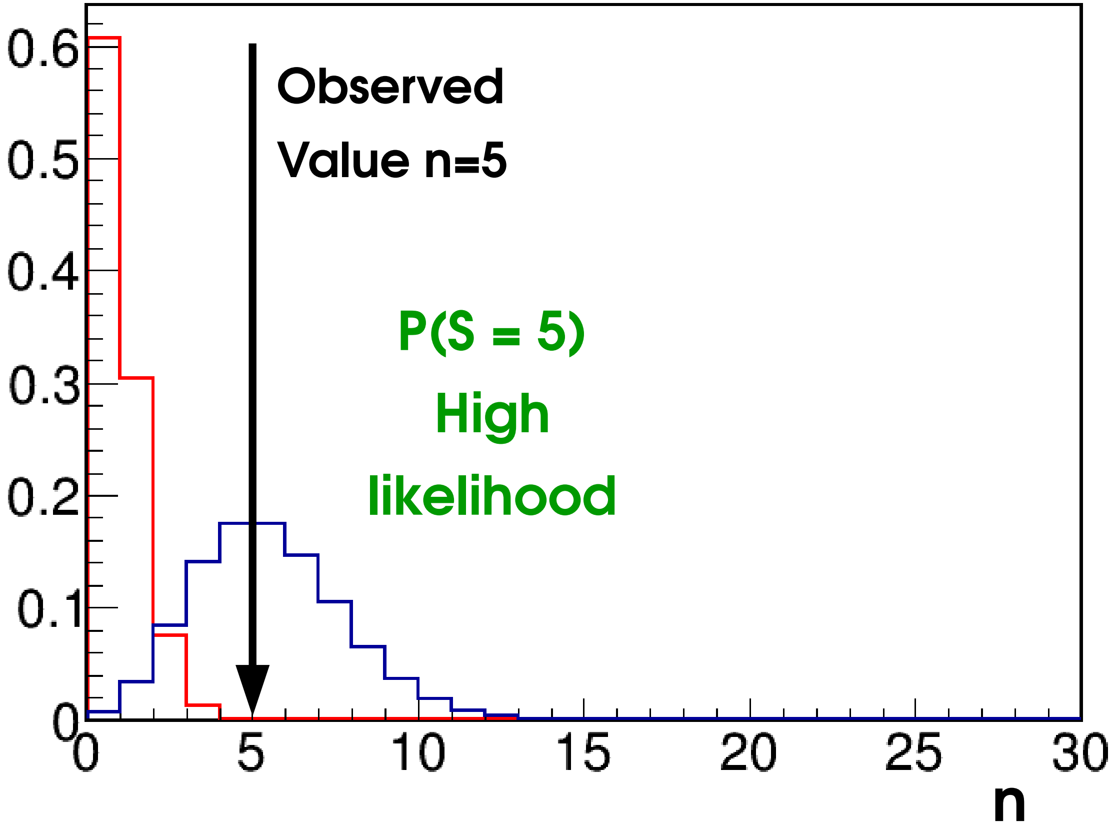
Say we **observe n=5**, want to infer information on the parameter **S**

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

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

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

P(S = 0.5)
Low
likelihood



Poisson Example

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

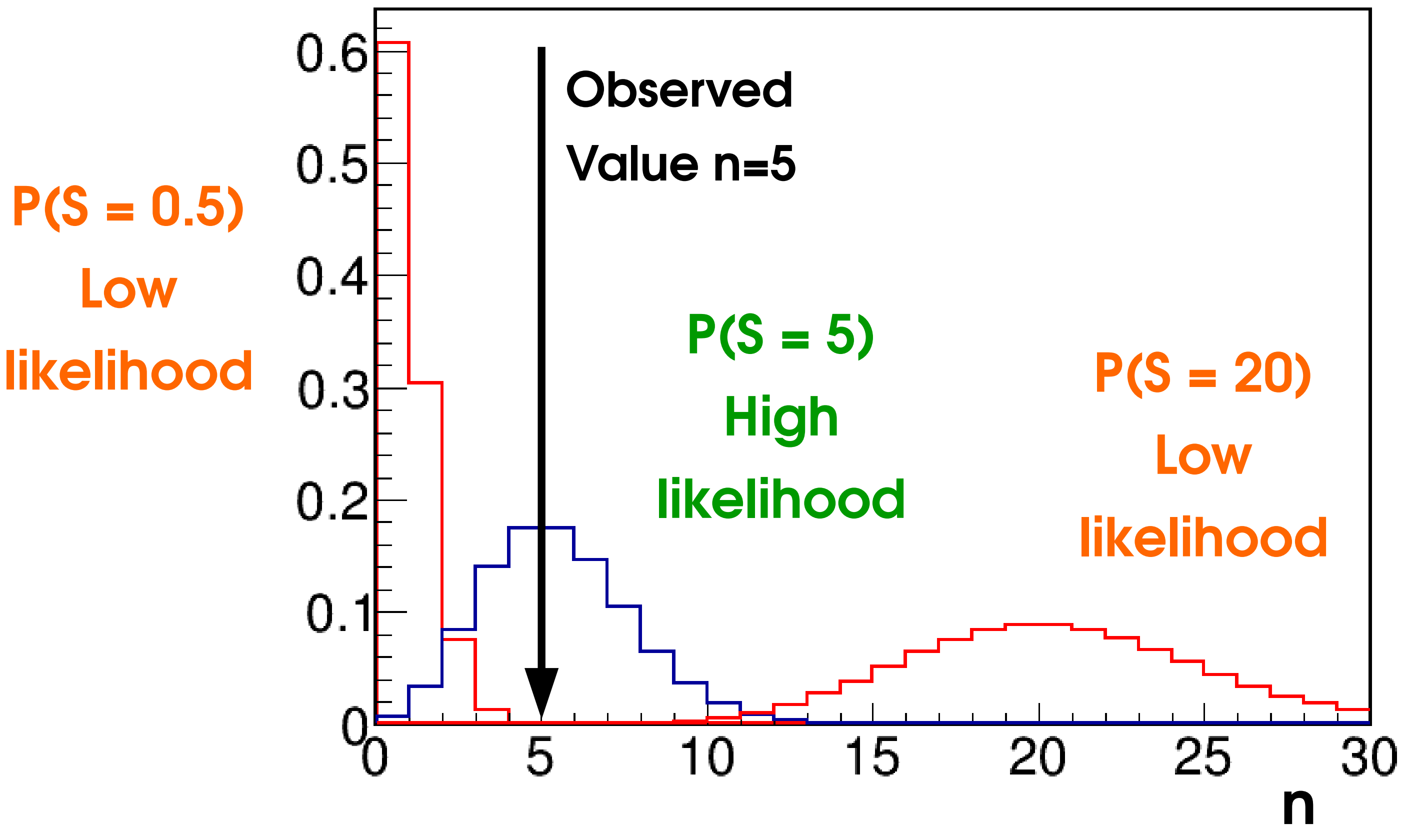
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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 :

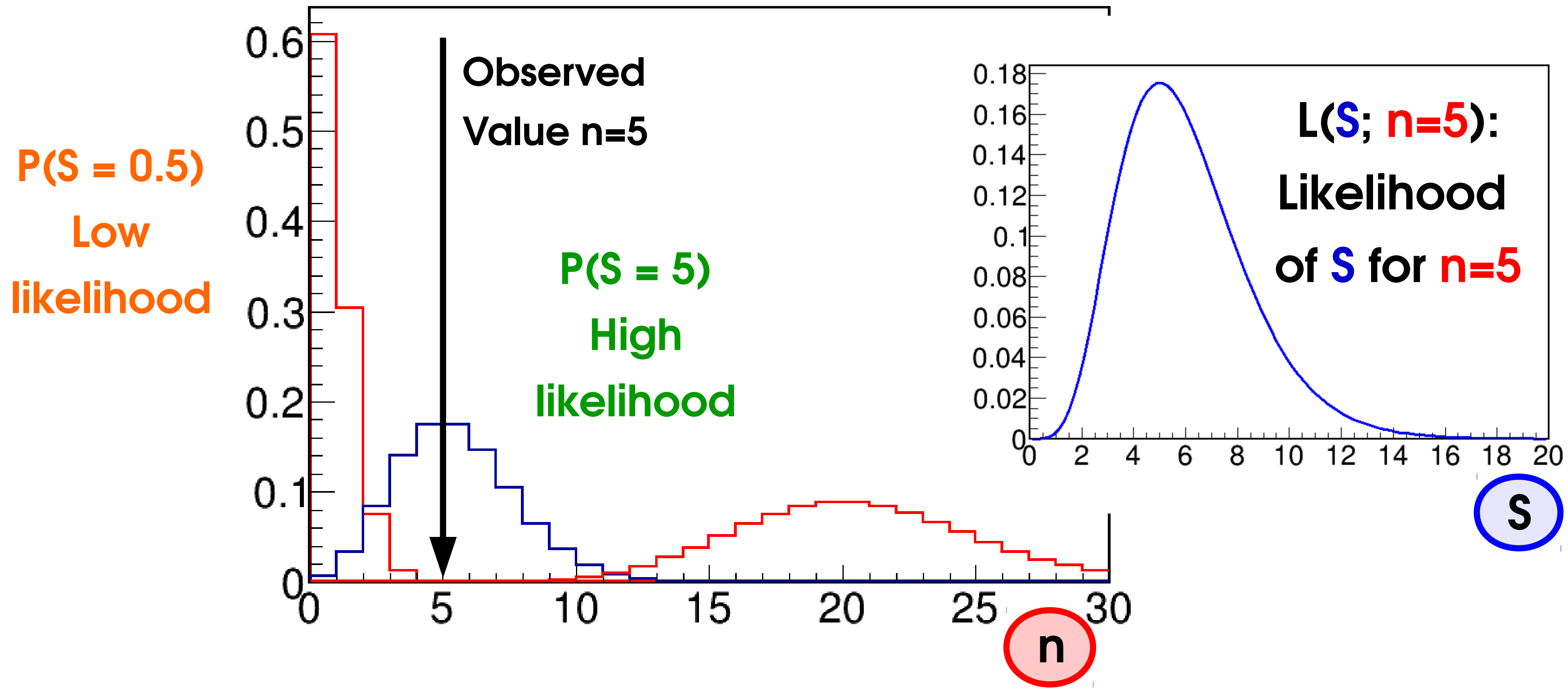
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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**

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



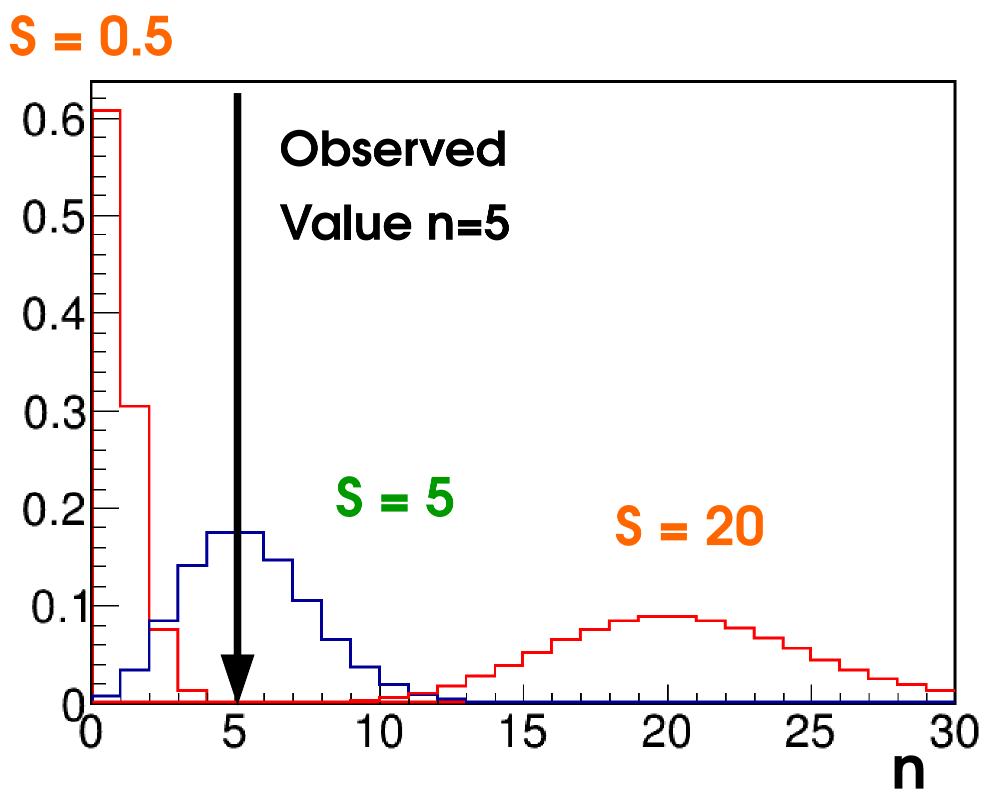
Maximum Likelihood Estimation

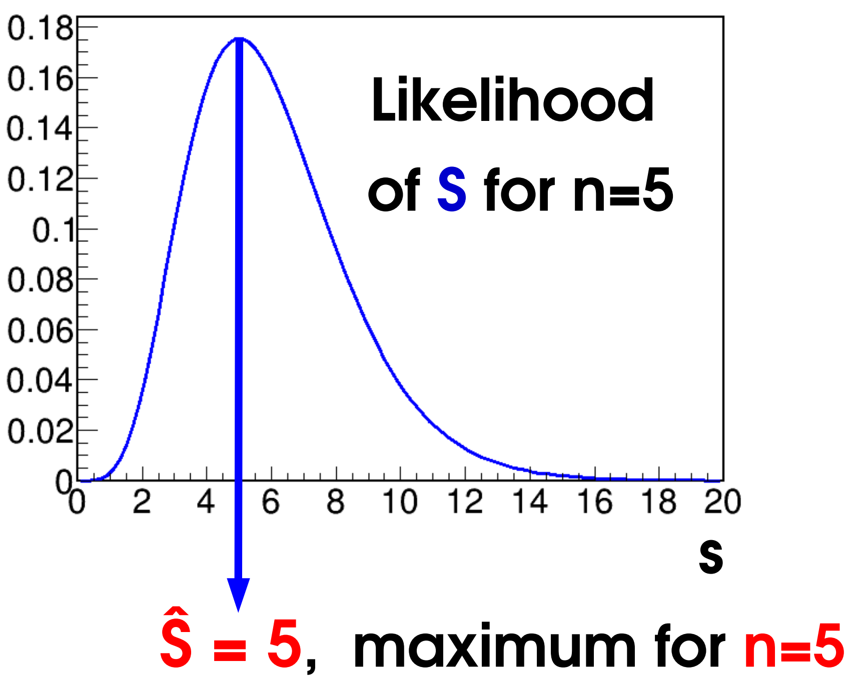
Estimate a parameter μ : Find the **value that maximizes** L(μ)

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

→ **Maximum Likelihood Estimator, $\hat{\mu}$**

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

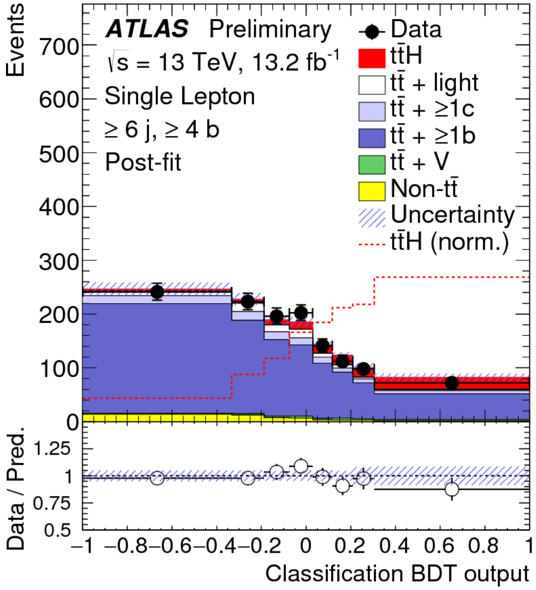




The MLE is a function of the data – itself an **observable**

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

MLEs in Shape Analyses



Binned shape analysis:

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

Need to maximize L(S) :
in practice easier to minimize

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

Or in the Gaussian limit

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

χ^2 formula!

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

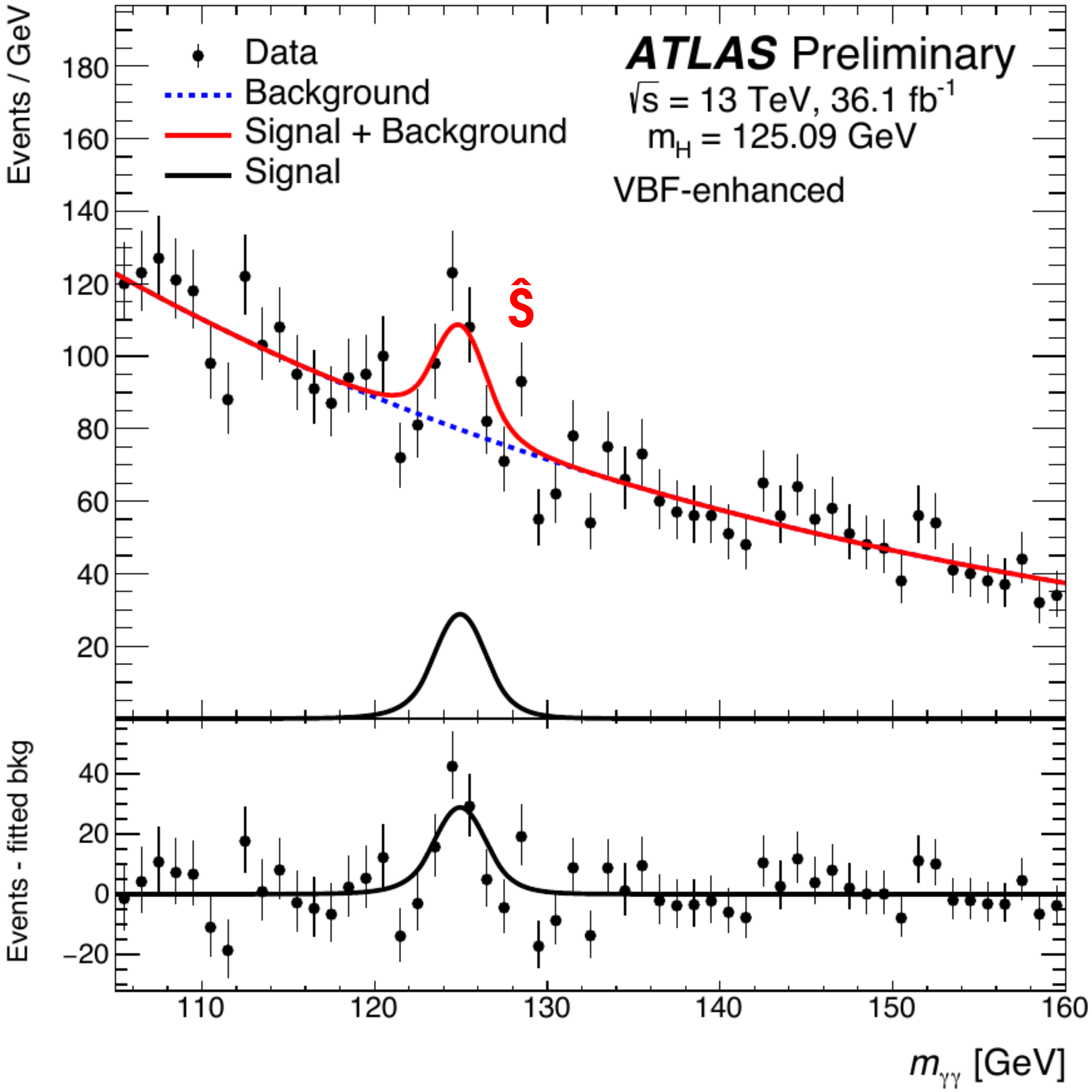
→ **Poisson MLE** (min λ_{Pois}) : ***Best fit value*** in a *likelihood* fit (in `ROOT`, fit option "`L`")

In RooFit, $\boldsymbol{\lambda}_{\mathbf{Pois}}$ ⇒ `RooAbsPdf::fitTo()`, $\boldsymbol{\lambda}_{\mathbf{Gaus}}$ ⇒ `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 ***S*** *using MLE* ***Ŝ*** *?*

→ Just perform (likelihood) best-fit of model to data

⇒ fit result for S is the desired **Ŝ**.

ATLAS-CONF-2017-045

MLE Properties

- **Consistent**: $\hat{\mu}$ converges to the true value for large n, $\hat{\mu} \xrightarrow{n\to\infty} \mu^*$
- **Asymptotically Gaussian** : $P(\hat{\mu}) \propto \exp\left(-\frac{(\hat{\mu}-\mu^*)^2}{2\sigma_{\hat{\mu}}^2}\right)$ for $n \to \infty$

  ↑ for large datasets

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

- **Asymptotically Efficient** : $\sigma_{\hat{\mu}}$ is the **lowest possible value** (in the limit n→∞) among consistent estimators.
  → MLE captures all the available information in the data

- **Log-likelihood :** Can also **minimize** λ **= -2 log L**
  - → Usually more efficient numerically
  - → For Gaussian L, λ is parabolic: $\lambda(\mu) = \left(\frac{\hat{\mu}-\mu}{\sigma_\mu}\right)^2$

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

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

→ smaller $\sigma_{\text{Likelihood}}$ ⇒ more information.

Cramer-Rao bound:

For any estimator $\hat{\mu}$,

$$\text{Var}(\hat{\mu}) \geq \frac{1}{I(\mu)}$$

→ cannot be more precise than information allows.

Gaussian: for any estimator $\hat{\mu}$ with

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

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

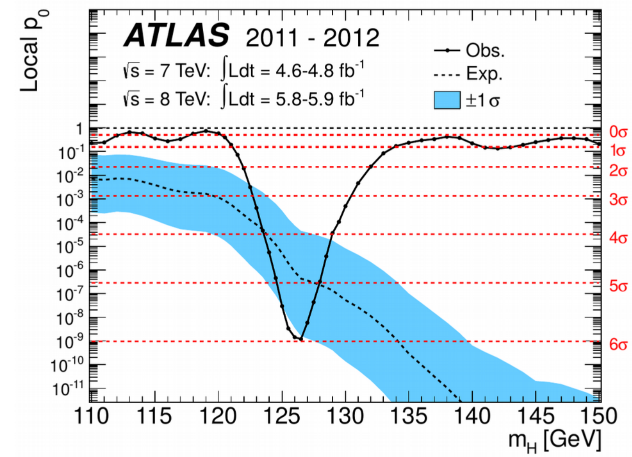
$\sigma_{\hat{\mu}}^2 \geq \sigma_{\text{Likelihood}}^2 = \sigma_{\text{MLE}}^2$

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

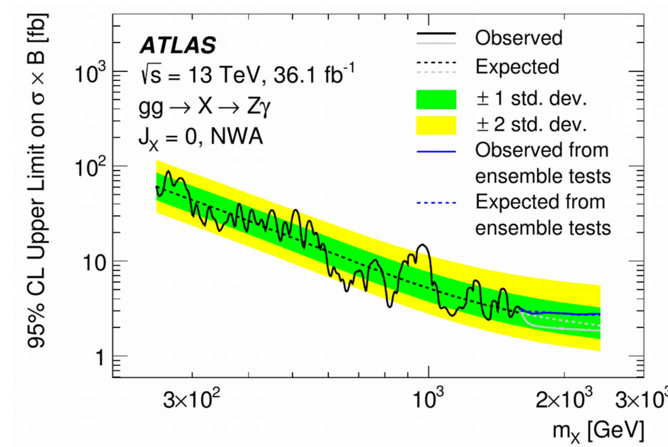
What's next ? Usual Statistical Results

We need more than just best-fit values:

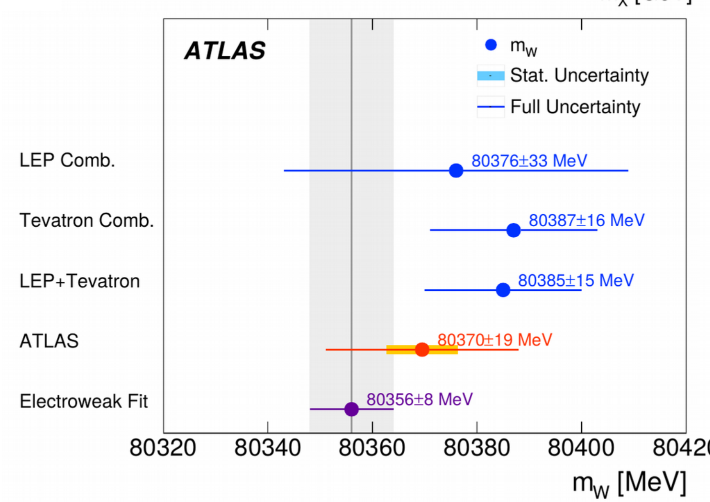
- **Discovery**: we see an excess –
  is it a (new) signal, or a background fluctuation ?



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



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



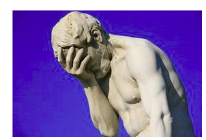
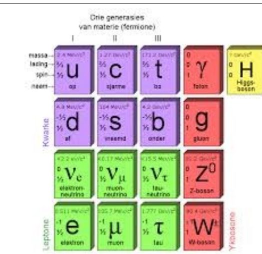
The Statistical Model already contains all the necessary information – how to use it ?

Computing Statistical Results
II. Testing Hypotheses

Hypothesis Testing

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

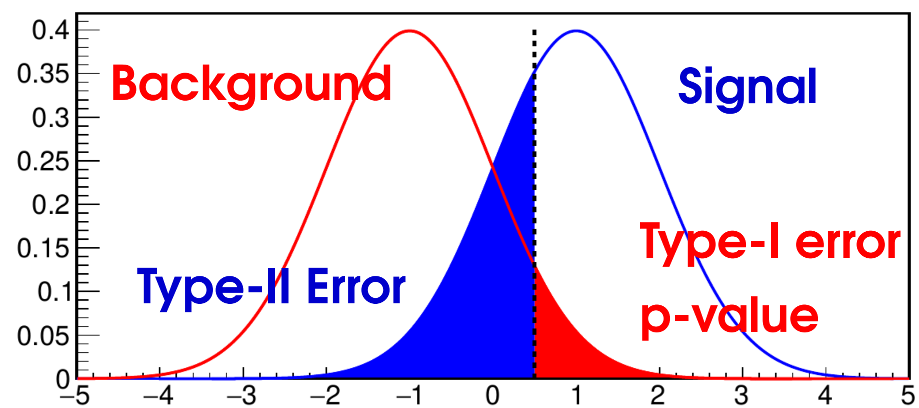
→ Goal : determine if H_0 is true or false using a test based on the data

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

Stringent discovery criteria

⇒ **lower Type-I errors**, **higher Type-II errors**

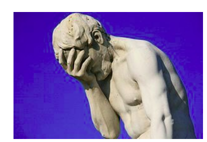
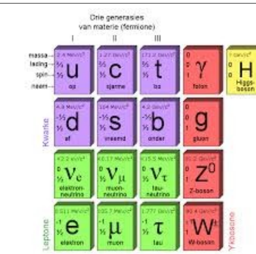
→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.



Hypothesis Testing

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

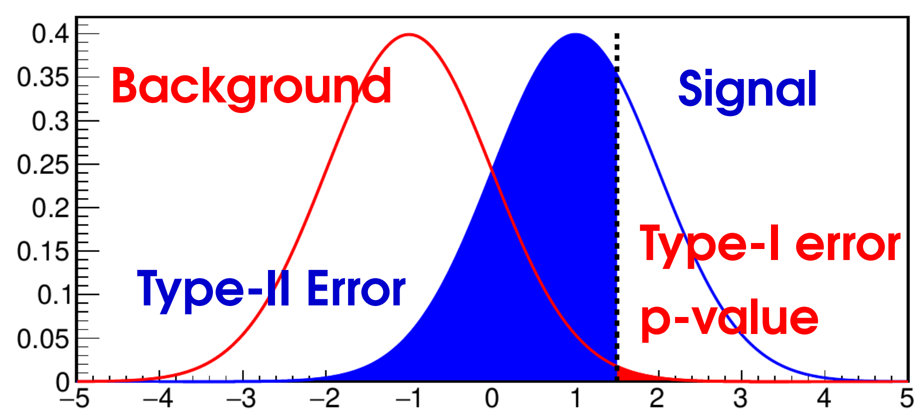
→ Goal : determine if H_0 is true or false using a test based on the data

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

Stringent discovery criteria

⇒ **lower Type-I errors**, **higher Type-II errors**

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



Hypothesis Testing with Likelihoods

Neyman-Pearson Lemma

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

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

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

→ **Minimizes Type-II uncertainties** for given level of Type-I uncertainties
→ Always need an **alternate hypothesis** to test against.

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

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

Statistical Results as Hypothesis Tests

Usual HEP results can be recast in terms of **hypothesis testing**:

- **Discovery**: is the data compatible with background-only ?
  → H_0 : only background is present
  → How well can we **reject** H_0 ? → **p-value (significance)**

- **Upper limits**: no excess observed – how small must the signal be ?
  → $H_0(S)$: B + some signal S
  → How small can we make S, and still reject $H_0(S)$ at 95% C.L. (p=5%) ?

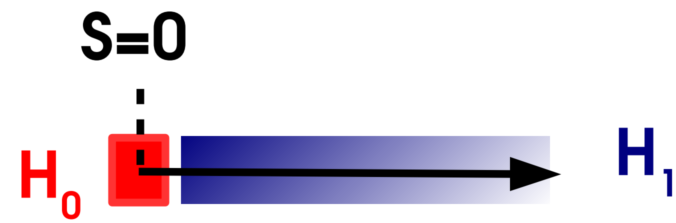
- **Parameter measurement**
  → $H_0(\mu)$: some parameter value μ
  → What values μ are **not** rejected at 68% C.L. (p=32%) ?
  ⇒ **1σ confidence interval on μ**

In all cases, H_0 : ***null hypothesis*** – what we are trying to disprove

Computing Statistical Results
III. Discovery

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

Discovery: Test Statistic



Discovery :

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

→ For H_1, any S≠0 is possible, which to use ? **The one preferred by the data, Ŝ.**

⇒ Use LR $\frac{L(S=0)}{L(\hat{S})}$

→ In fact use the **test statistic** $$t_0 = -2\log\frac{L(S=0)}{L(\hat{S})}$$

→ t_0 is computed from the observed data – fit to data to get Ŝ.

→ **t_0 always ≥ 0**, t_0 = 0 reached for Ŝ = 0.

→ t_0 measures the relative *likelihood* of H_1 vs. H_0 in data:

Large values of t_0 ⇔ large observed S

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

Discovery p-value

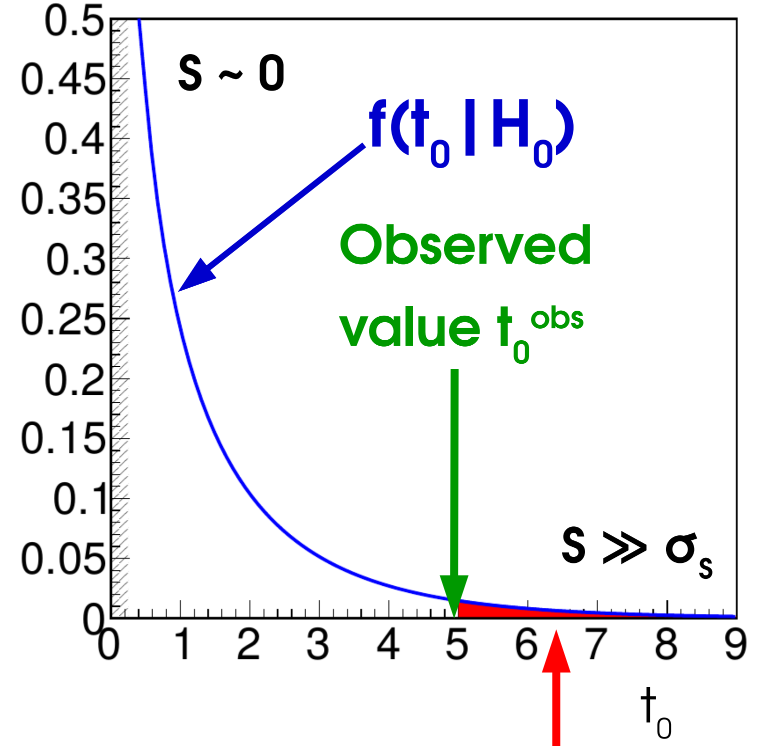
Large values of $t_0 = -2\log\frac{L(S=0)}{L(\hat{S})}$

⇒ large observed $\hat{S}$

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

data prefer H_0 ⟷ data prefer H_1

How large t_0 before we can exclude H_0 ?
(and claim a discovery!)



p-value **:** *Fraction of outcomes that are* ***at least as H_1-like*** *(signal-like)* ***as data****, when* ***H_0 is true*** *(no signal present).*

→ Smaller p-value ⇒ Stronger case for discovery

→ Compute from distribution $\mathbf{f(t_0 | H_0)}$ of t_0 if H_0 is true:

$$p_0 = \int_{t_0^{obs}}^{\infty} f(t_0|H_0)\, dt_0$$

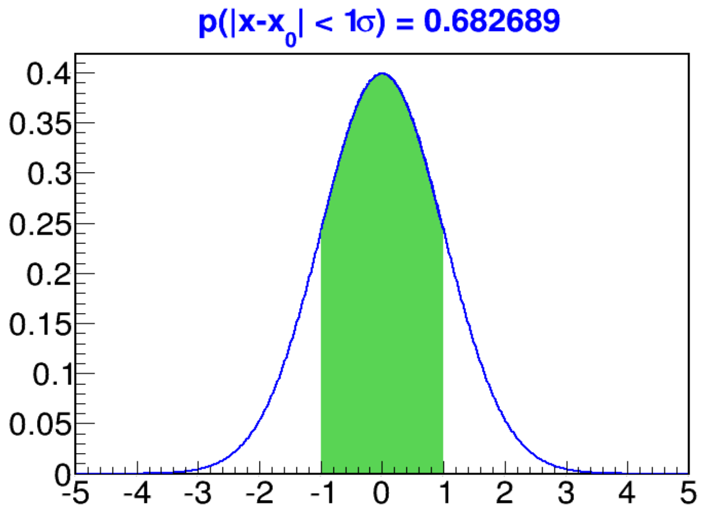
Discovery significance

Interesting p-values are quite small

⇒ express in terms of Gaussian quantiles

→ **Significance Z**

$$p_0 = 1 - \int_{-Z}^{+Z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$
$$= 1 - 2\,\Phi(Z)$$



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

Z	p-value
1	0.32
2	0.045
3	0.003
5	6 x 10^{-7}

In ROOT:

p_0 → Z (Φ) : `ROOT::Math::gaussian_quantile_c`

Z → p_0 (Φ^{-1}) : `ROOT::Math::gaussian_cdf_c`

⇒ How small is small enough ?

→ Conventionally, discovery for **p_0= 6 10^{-7} ⇔ Z = 5σ**

Asymptotic Approximation

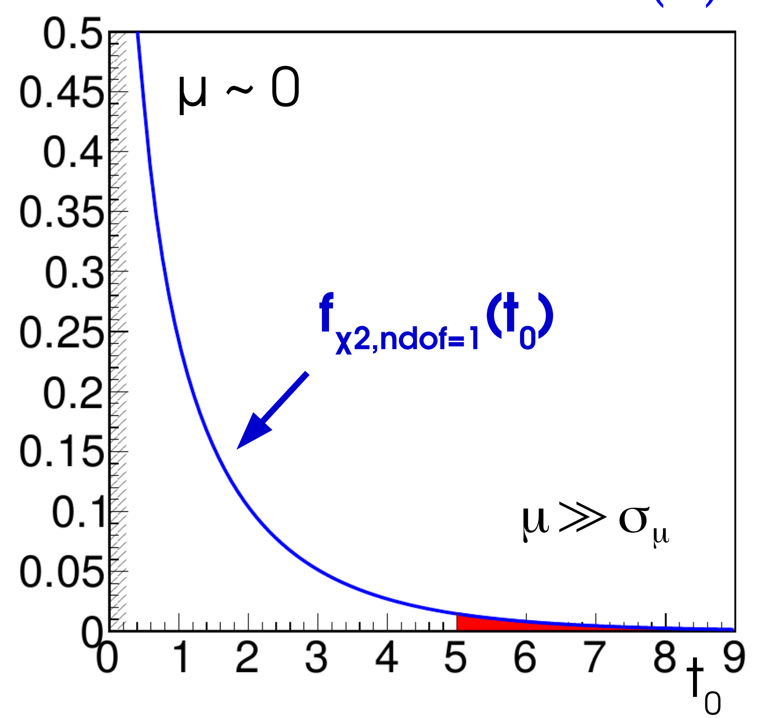
Cowan, Cranmer, Gross & Vitells
Eur.Phys.J.C71:1554,2011

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

⇒ **t_0 is distributed as a χ^2** under the hypothesis H_0

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

$$f(t_0 \mid H_0) = f_{\chi^2(n_{dof}=1)}(t_0)$$



In particular, significance:

$$Z = \sqrt{t_0}$$

By definition,
$t_0 \sim \chi^2 \Rightarrow \sqrt{t_0} \sim G(0,1)$

Typically works well for for event counts O(5) and above (5 already "large"...)

The 1-line "proof" : asymptotically L and S are Gaussian, so

$$L(S) = \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}}{\sigma}\right)^2\right] \Rightarrow t_0 = \left(\frac{\hat{S}}{\sigma}\right)^2 \Rightarrow t_0 \sim \chi^2(n_{dof}=1) \text{ since } \hat{S} \sim G(0,\sigma)$$

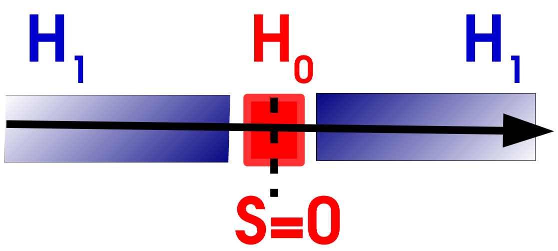
One-sided vs. Two-Sided

If $\hat{S}$ < 0, is it a *discovery* ? (does reject the S=0 hypothesis…)

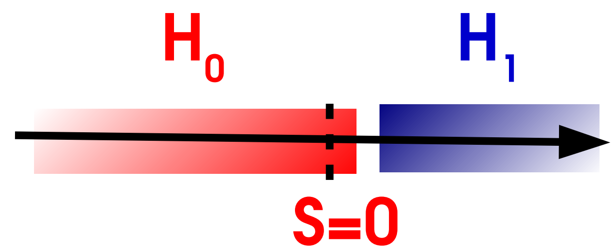
Usual assumption : only $\hat{S}$ > 0 is a *bona fide* signal

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

Two-sided



One-sided



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

Test Statistic

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

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

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

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

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

⇒ Same Z in both cases for a given signal S

One-Sided Asymptotics

→ **One-sided test**:

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

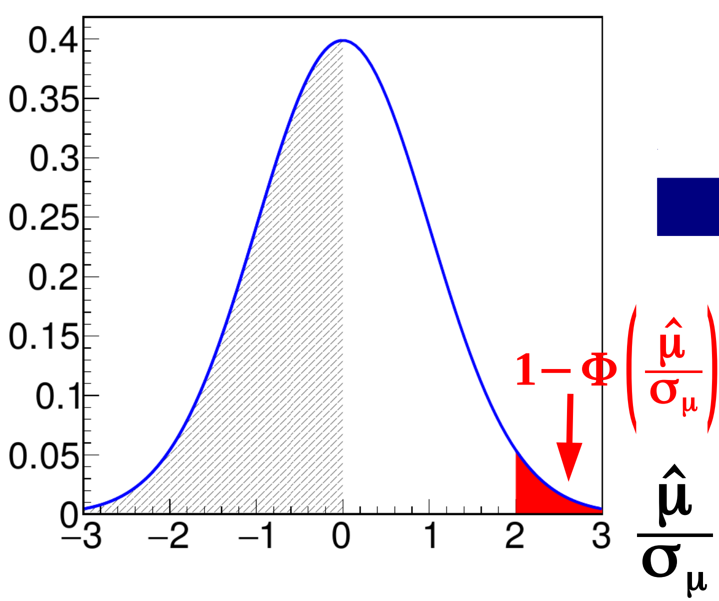
Asymptotics: "half-χ²" distribution:

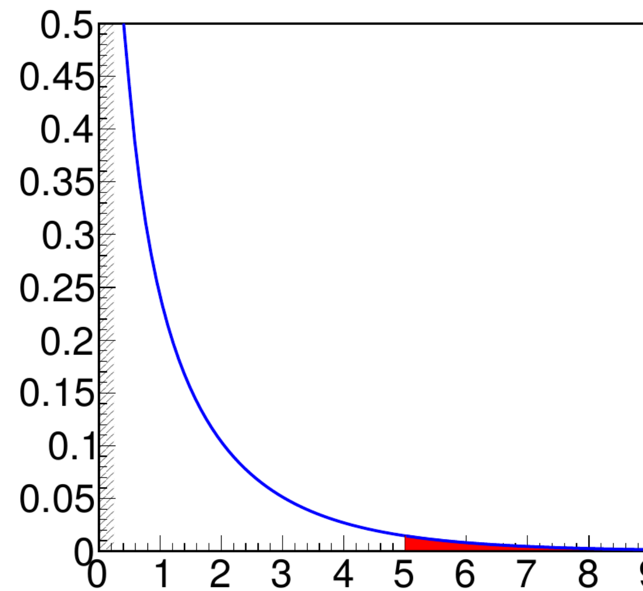
$$f(q_0 \mid S=0) = \frac{1}{2}\delta(q_0) + \frac{1}{2} f_{\chi^2(n_{dof}=1)}(q_0)$$

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

Φ : normal CDF

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





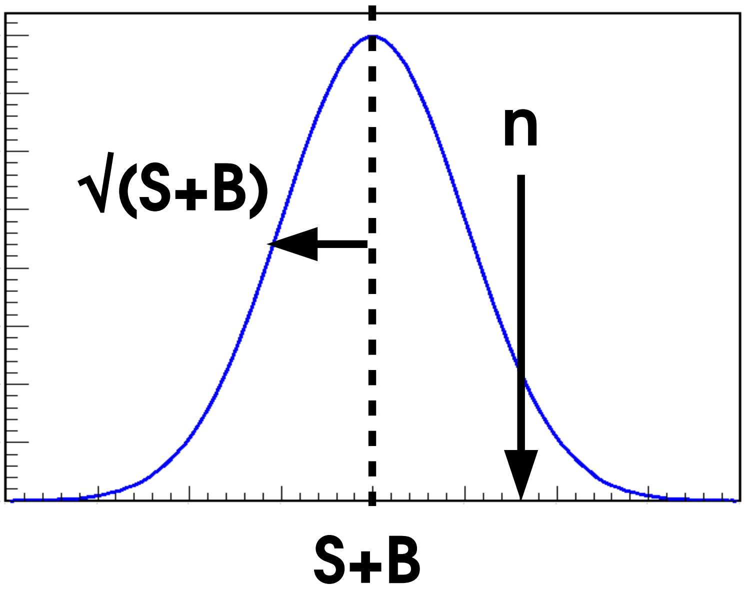
$$q_0 = \left(\frac{\hat{\mu}}{\sigma_\mu}\right)^2$$

Example: Gaussian Counting

Count number of events n in data

→ assume n large enough so process is Gaussian

→ assume B is known, measure S



Likelihood :

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

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

MLE for S : **Ŝ = n – B**

Test statistic: assume Ŝ > 0,

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

Finally:

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

Known formula!

→ Strictly speaking only valid in Gaussian regimge

Example: Poisson Counting

Same problem but now not assuming Gaussianity

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

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

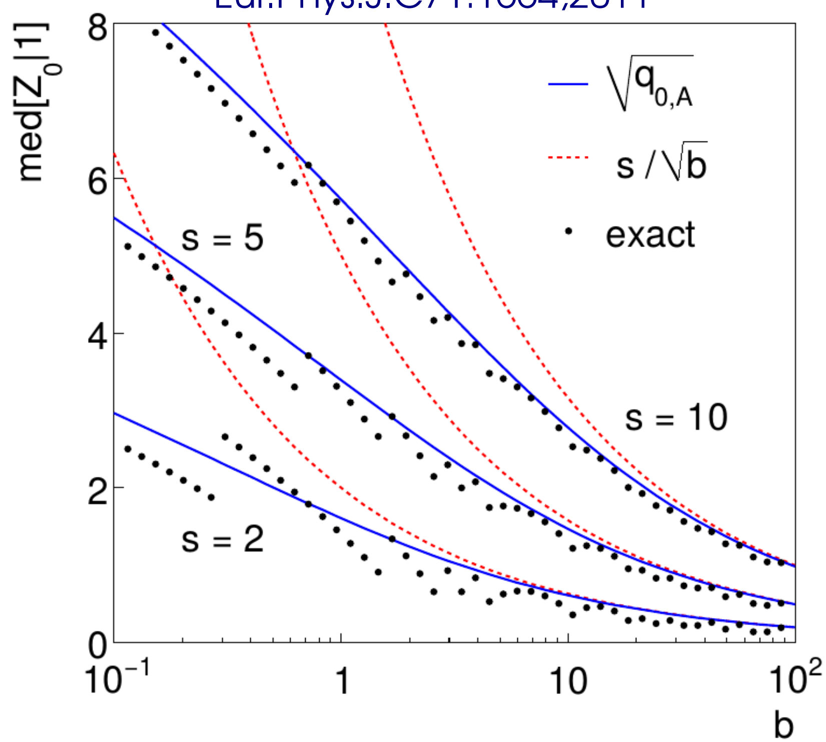
Test statistic (for $\hat{S} > 0$): $$q_0 = \lambda(S=0) - \lambda(\hat{S}) = -2\hat{S} - 2(\hat{S}+B)\log\frac{B}{\hat{S}+B}$$

Assuming asymptotic distribution for q_0,

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

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

Eur.Phys.J.C71:1554,2011



Asymptotic formulas justified by Gaussian regime, but remain valid even for small values of S+B (5!)

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

Example: Multi-bin counting

Likelihood :

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

Assume Gaussianity:

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

$$\hat{S} = \frac{\sum_{i=1}^{N} f_i \frac{n_i - B_i}{B_i}}{\sum_{i=1}^{N} \frac{f_i^2}{B_i}}$$

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

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

Asymptotics:

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

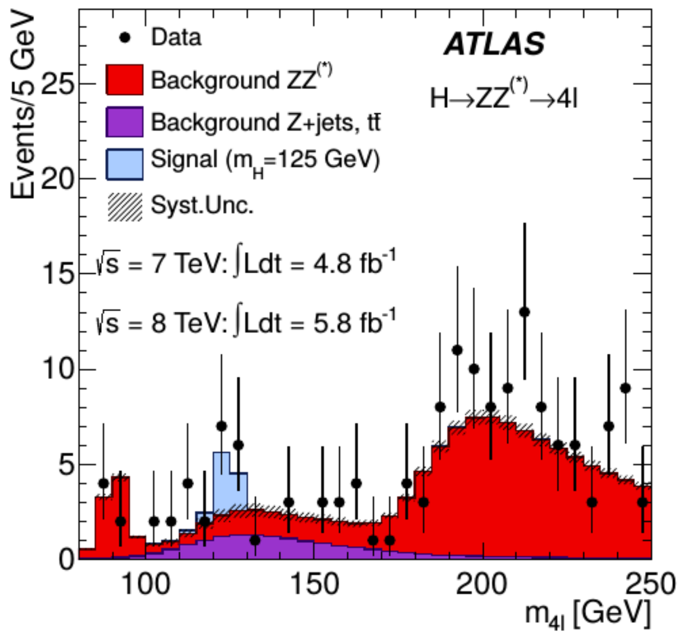
Combined uncertainty on $\hat{S}$ from all the bins

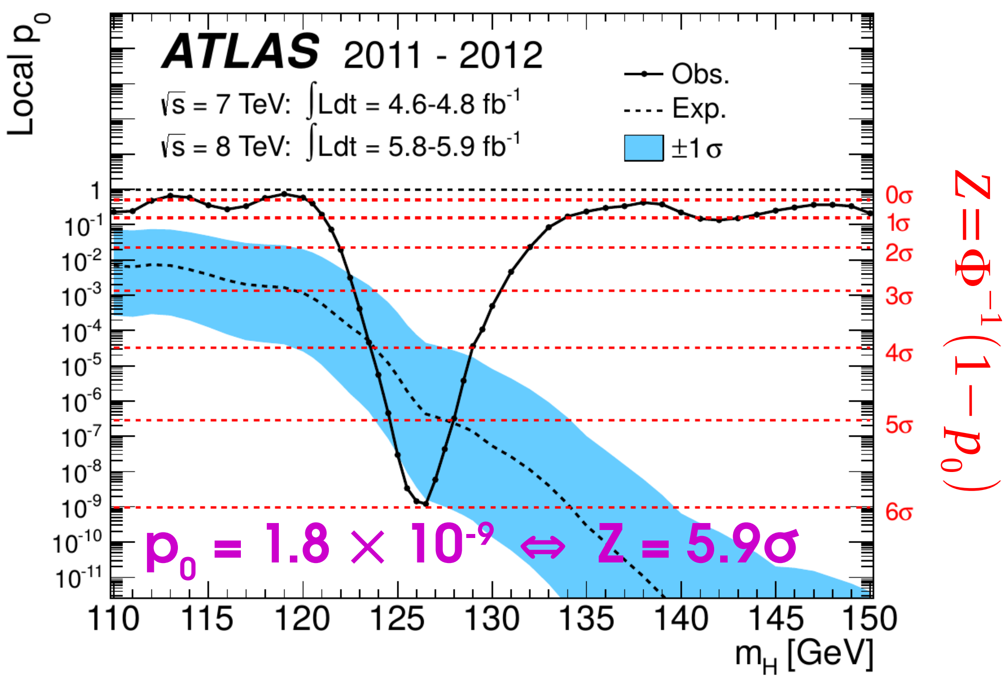
Always better than

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

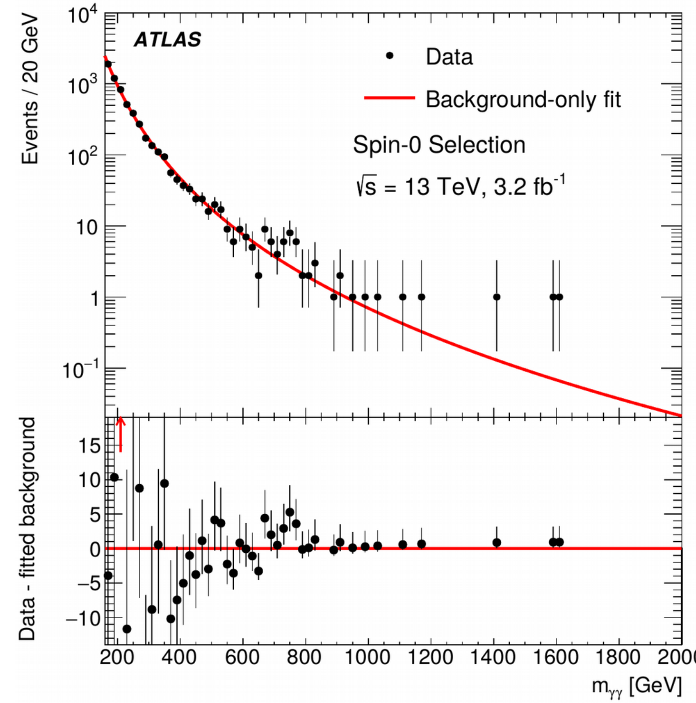
Some Examples

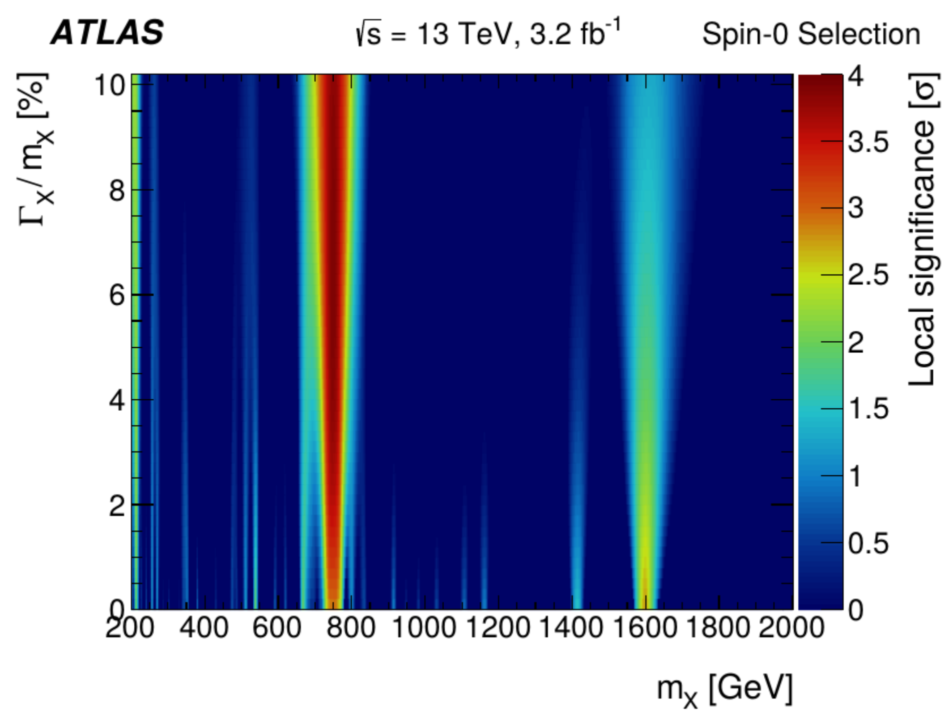
Higgs Discovery: Phys. Lett. B 716 (2012) 1-29





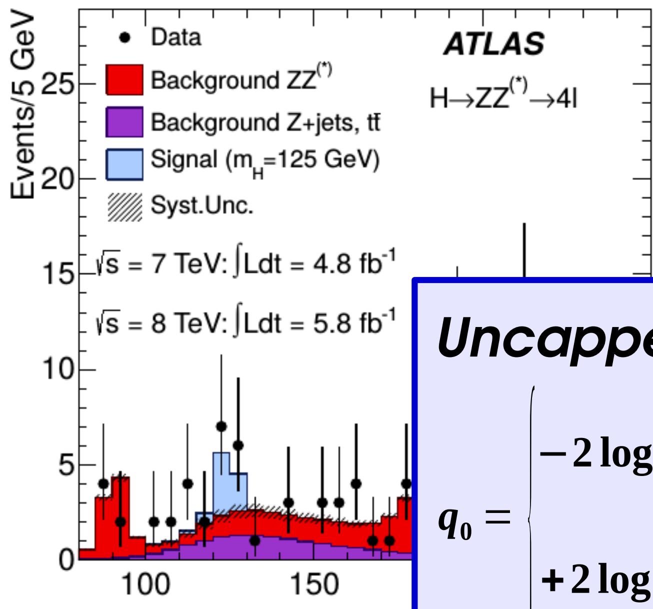
High-mass X→γγ Search: JHEP 09 (2016) 1





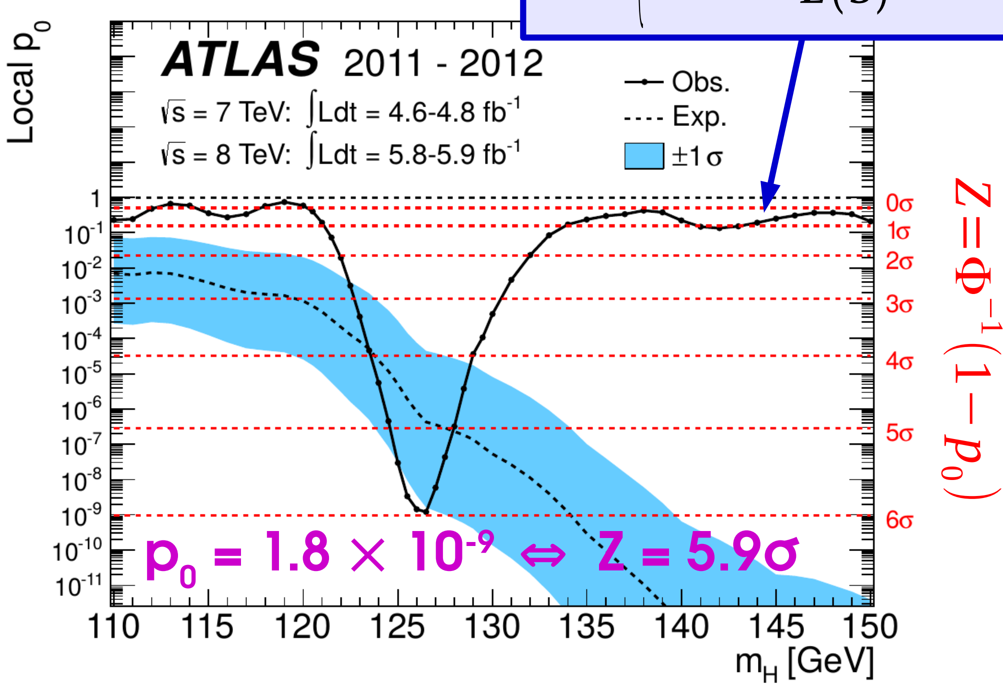
Some Examples

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

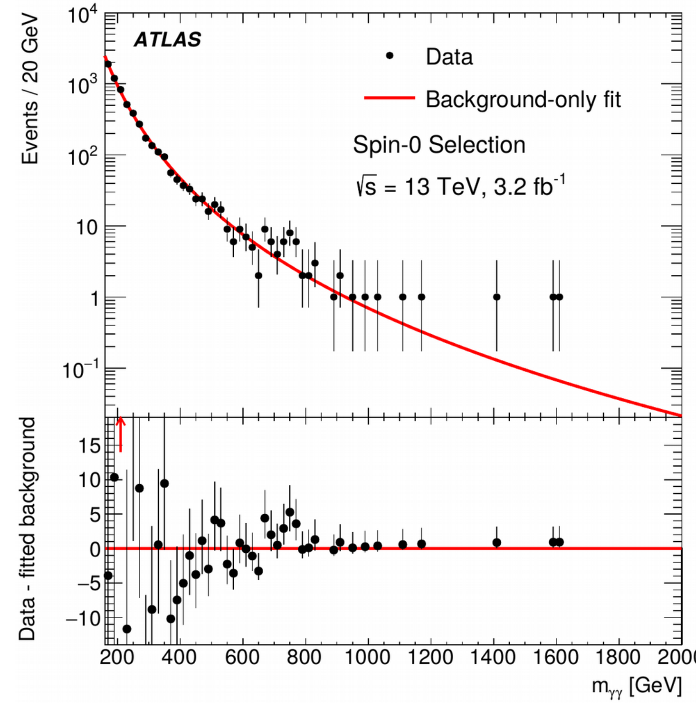


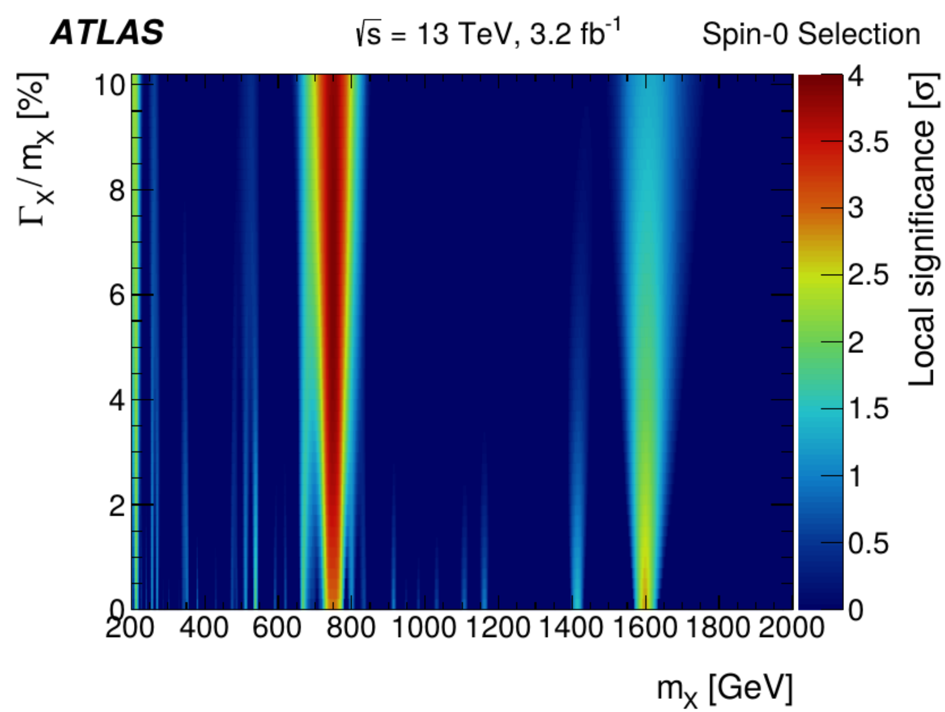
Uncapped q_0:

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



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





Takeaways

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

To estimate a parameter, use value **μ̂** that maximizes L(μ).

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

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

For large enough datasets, $$Z = \sqrt{q_0}$$

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

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

What was the question ?

Definition of the p-value:

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

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

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

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

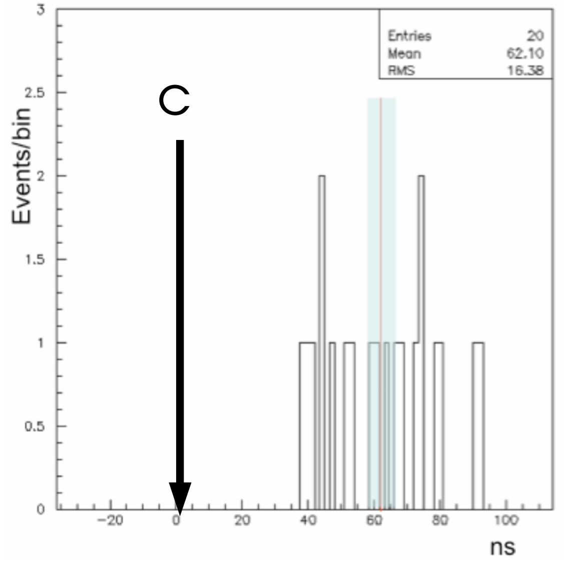
The second is on **P(H_0)** itself – not addressed in the framework described so far
→ makes sense in a ***Bayesian*** context, more on this tomorrow.

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

What was the question ?

e.g. Faster-than-light neutrino anomaly

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



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

"Extraordinary claims require extraordinary evidence"

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

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

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

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

→ In frequentist context, only have $\mathbf{p_0 = P(fluct|B)}$ (and P(fluct | S) = power ~ 1)

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