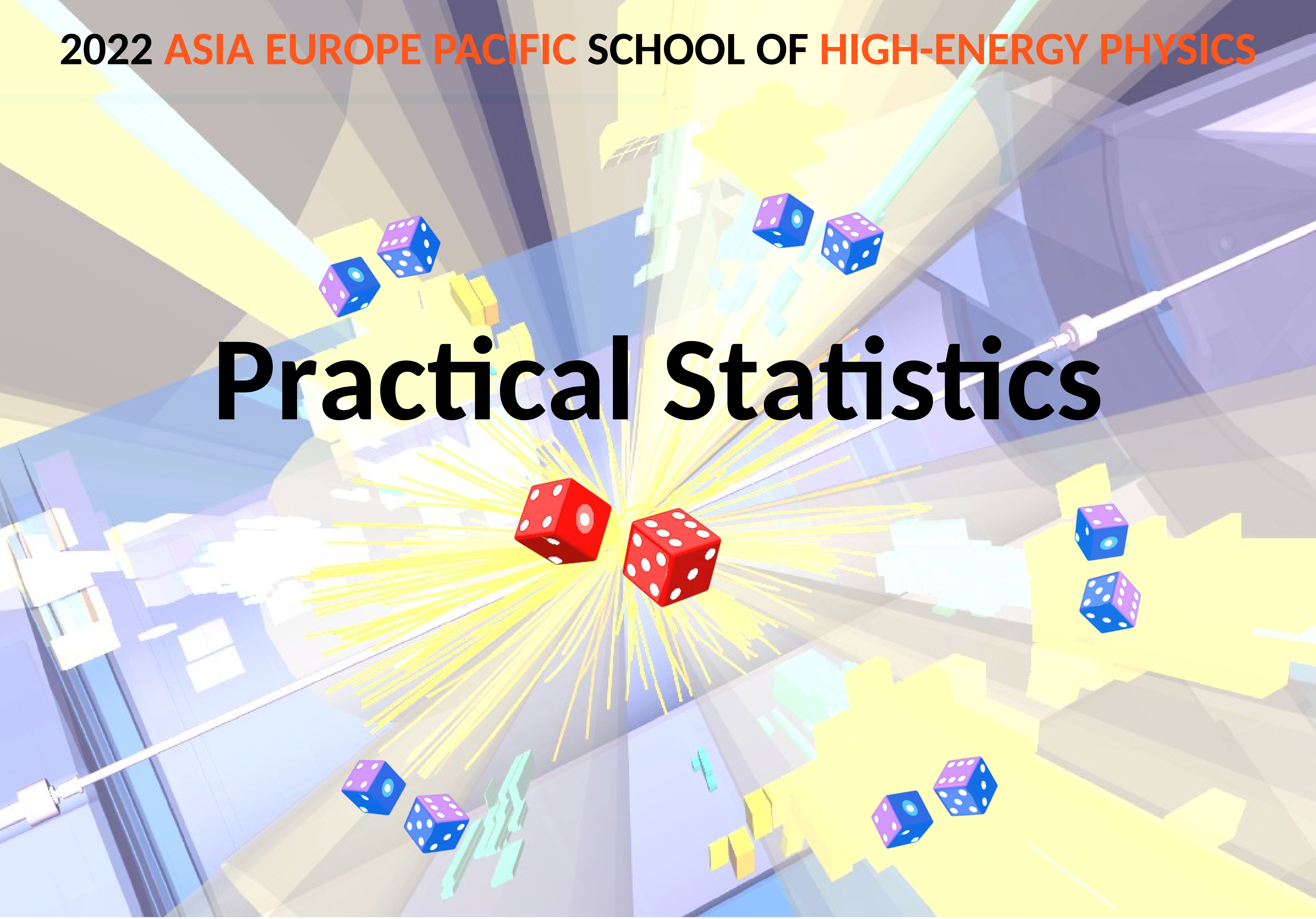
2022 ASIA EUROPE PACIFIC SCHOOL OF HIGH-ENERGY PHYSICS

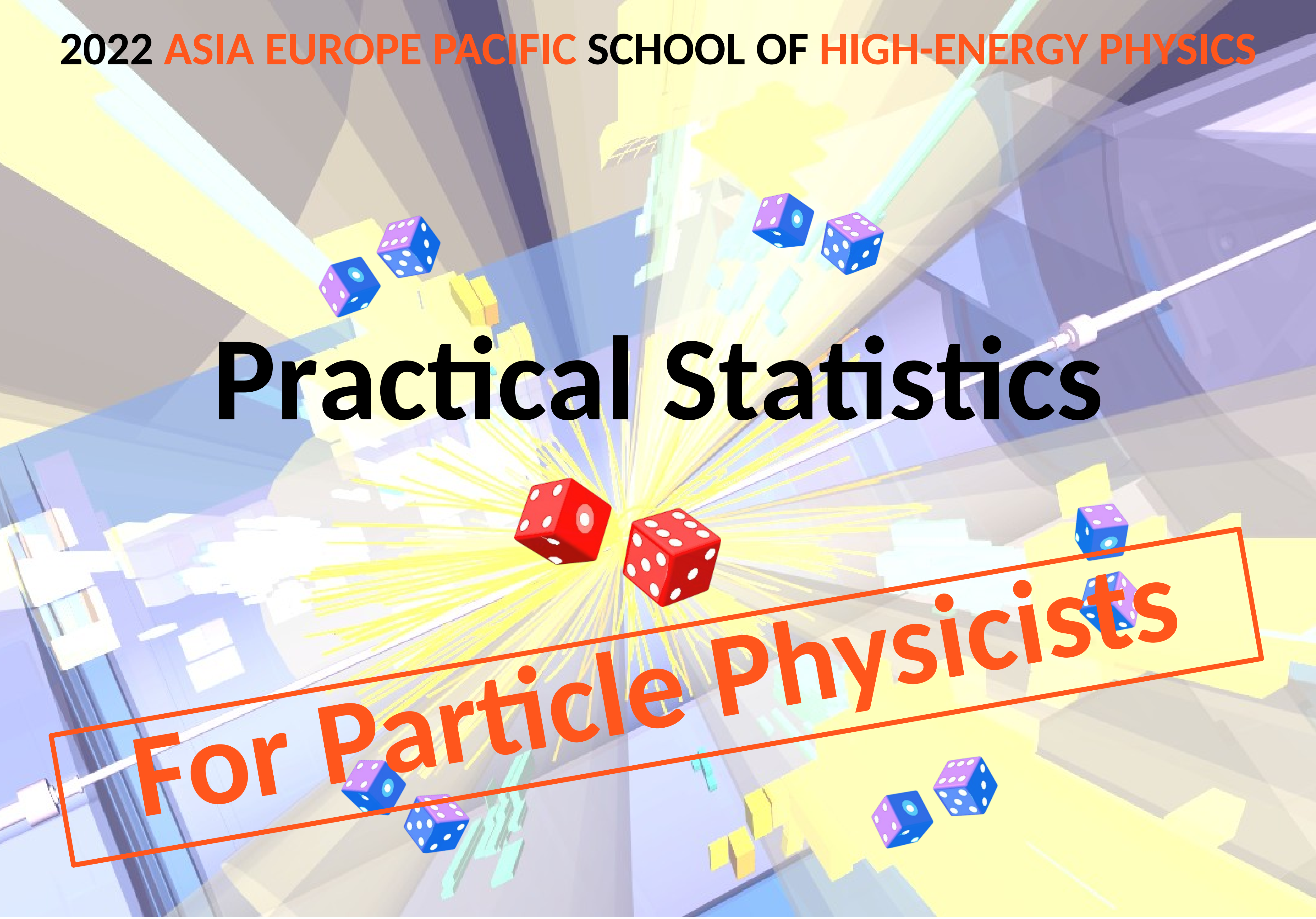
Practical Statistics



Nicolas Berger (LAPP Annecy)

Lecture 1

2022 ASIA EUROPE PACIFIC SCHOOL OF HIGH-ENERGY PHYSICS



Practical Statistics

For Particle Physicists

Nicolas Berger (LAPP Annecy)

Lecture 1

Statistics are everywhere

“There are three types of lies - lies, damn lies, and statistics.” – Benjamin Disraeli

Credits: mattbuck / wikimedia

Credits: StatLab

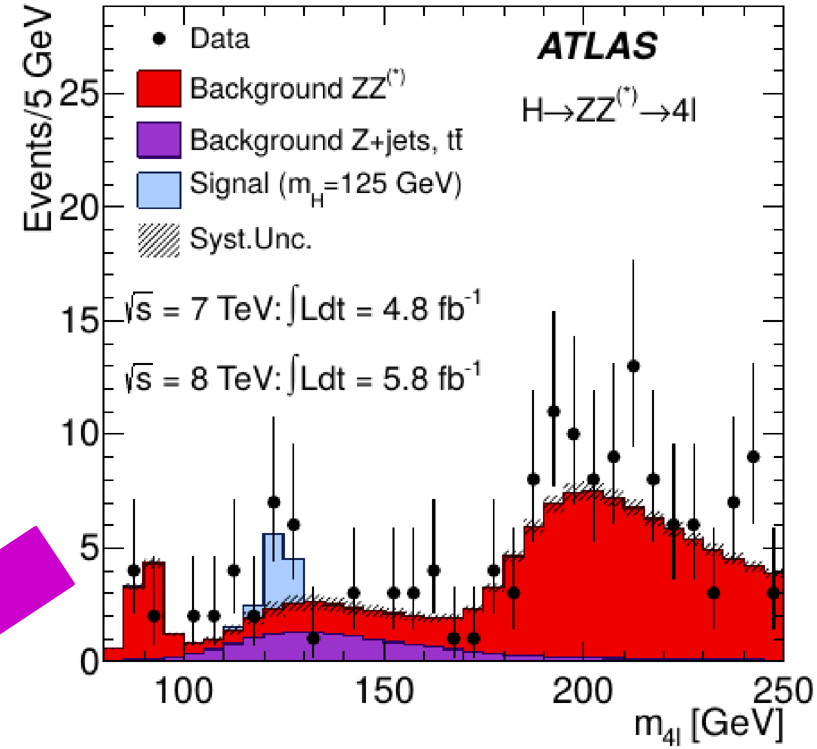
And Physics ?

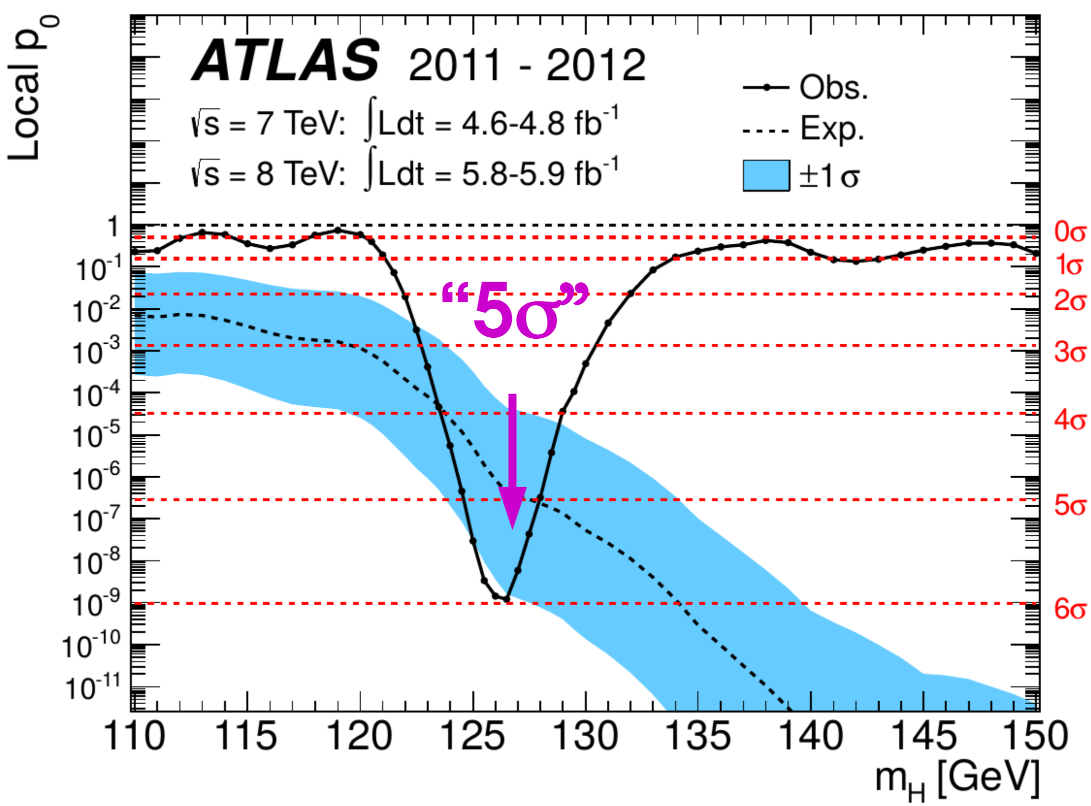
“If your experiment needs statistics, you ought to have done a better experiment” – E. Rutherford

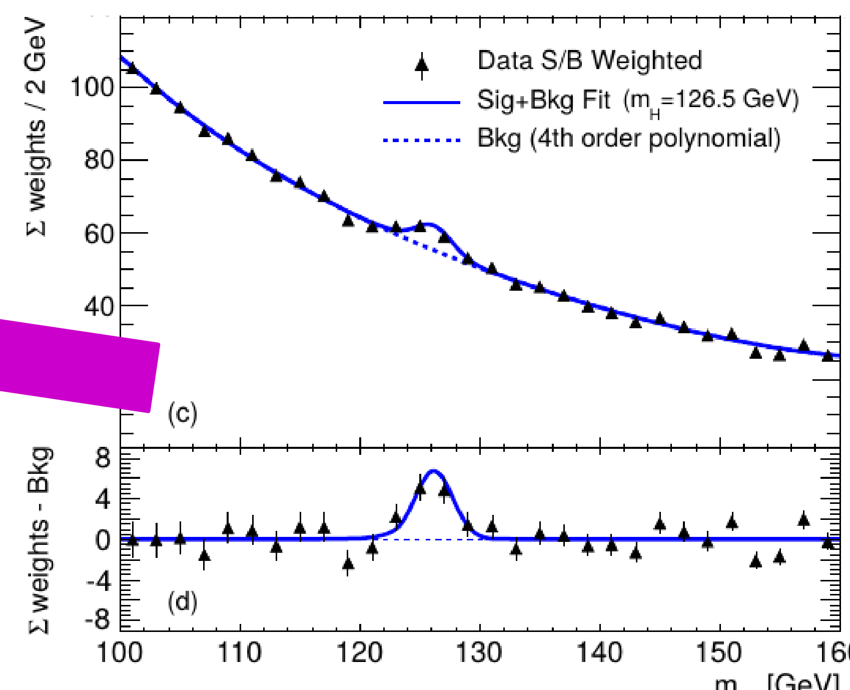
Introduction

Statistical methods play a critical role in many areas of physics

Higgs discovery : **"We have 5σ" !**



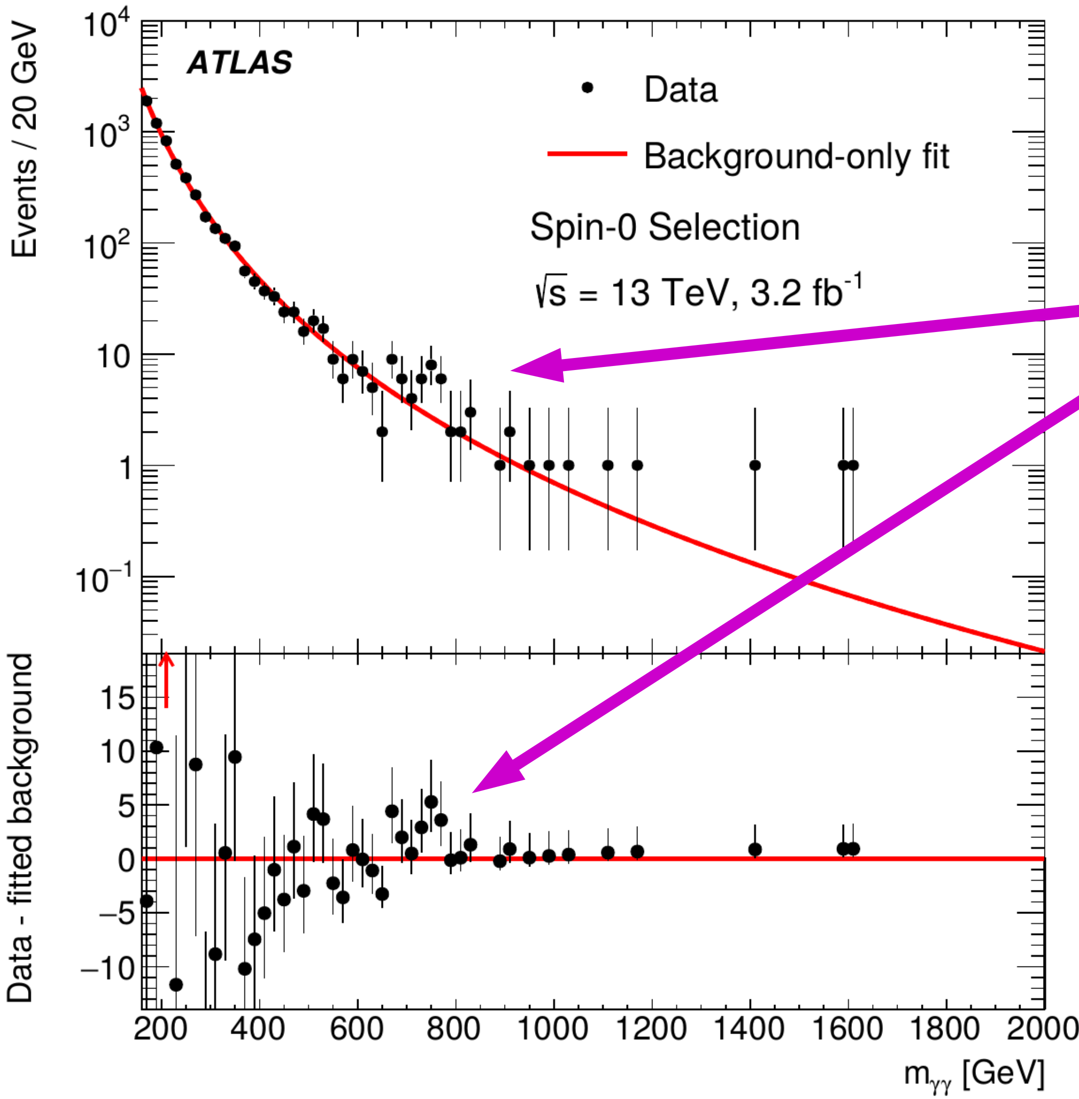




Phys. Lett. B 716 (2012) 1-29

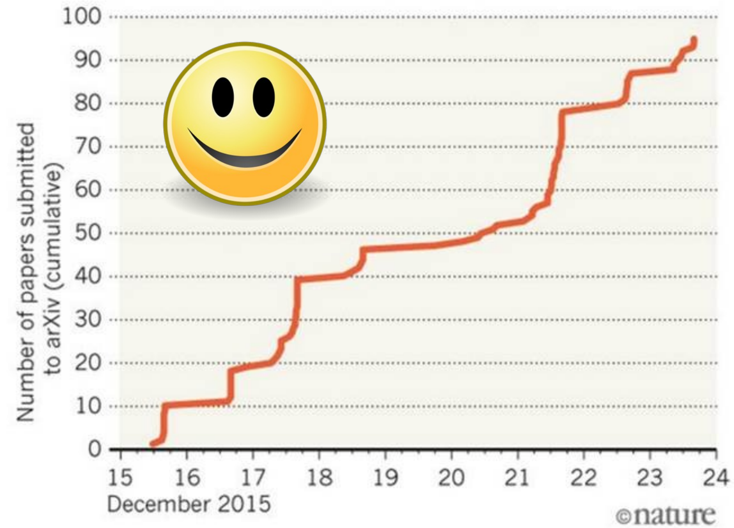
Introduction

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



New Physics ?

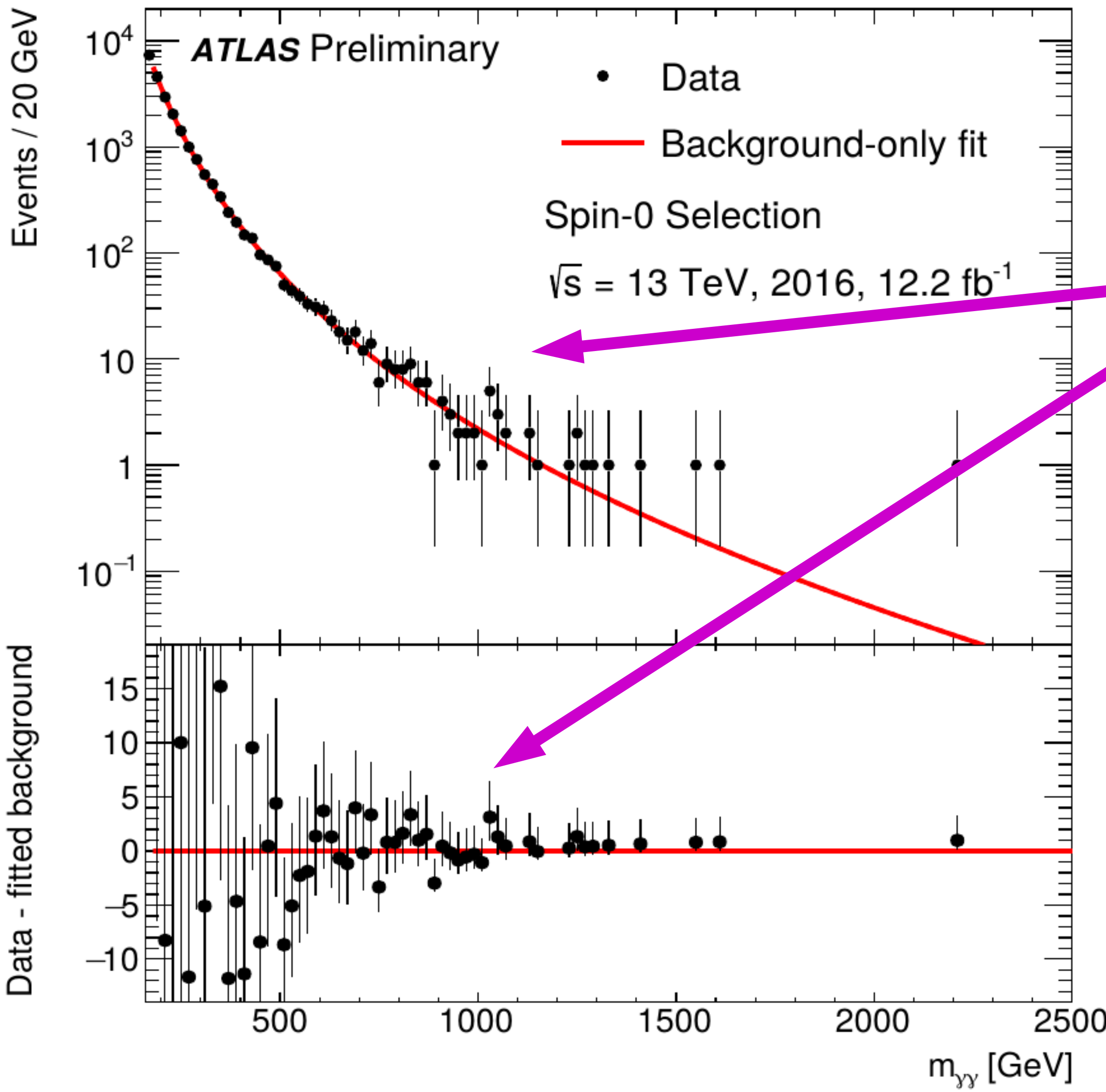
3.9σ ? 2.1σ ?



JHEP 09 (2016) 1

Introduction

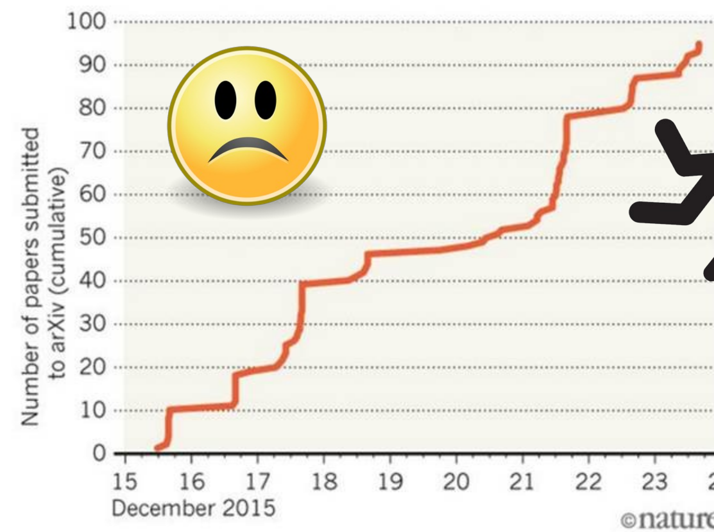
Sometimes difficult to distinguish a bona fide discovery from a **background fluctuation**…



A few months later…

~~**New Physics ?**~~

~~**3.9σ ? 2.1σ ?**~~



JHEP 09 (2016) 1

Introduction

Precision measurements are another window into BSM effects

→ How to compute (and interpret) measurement intervals

→ How to model systematic uncertainties ?

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

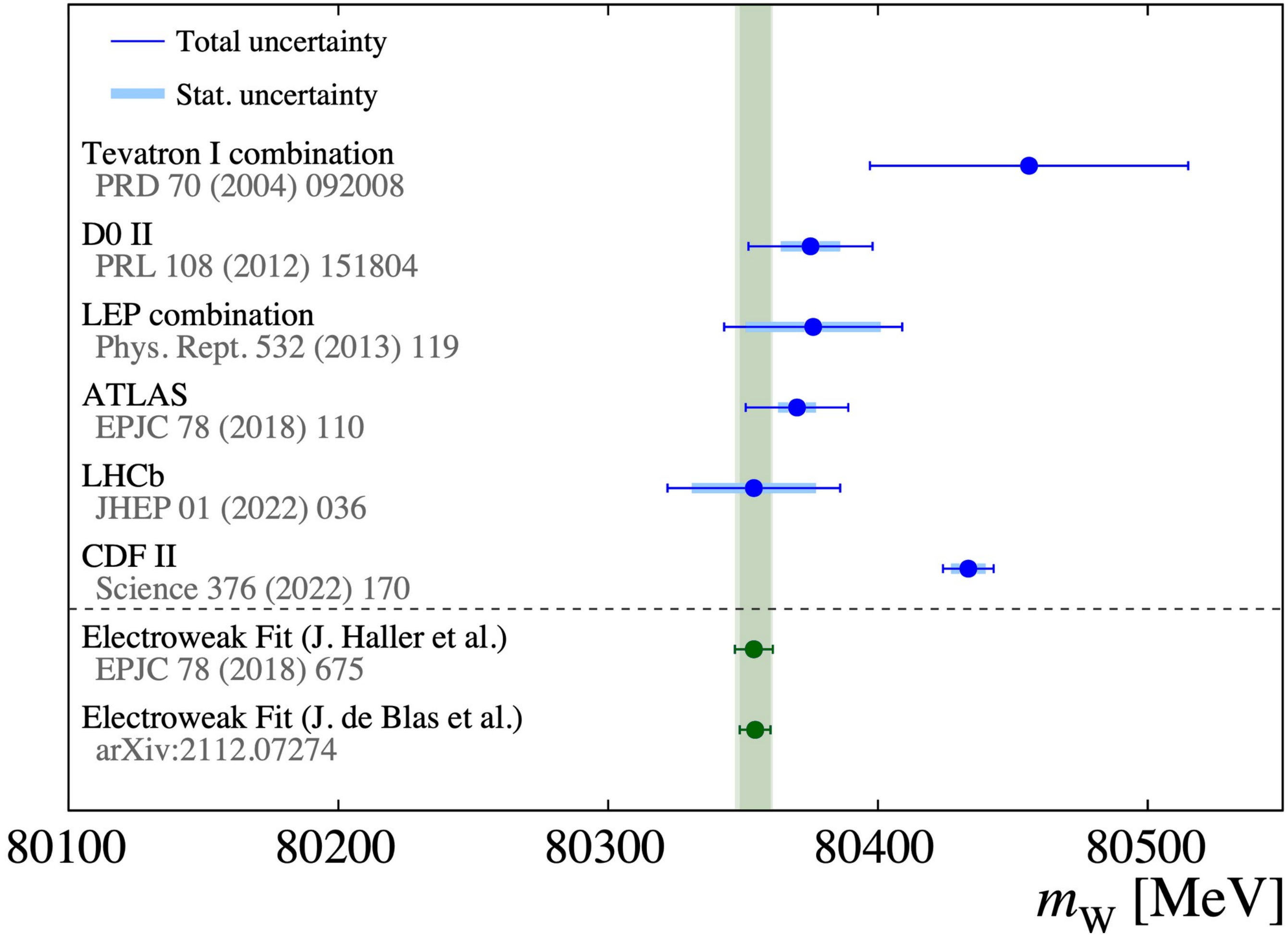


Image credits: CERN courier, LHCb

Lecture Plan

Statistics basic concepts (Today)

- **[Basic ingredients** (PDFs, etc.)**]**
- **Statistical Modeling** (PDFs for particle physics measurements)
- **Parameter estimation** (maximum likelihood, least-squares, ...)

Computing statistical results (Tomorrow)

- **Model testing** (χ^2 tests, hypothesis testing, p-values, ...)
- **Discovery testing**
- **Confidence intervals**
- **Upper limits**

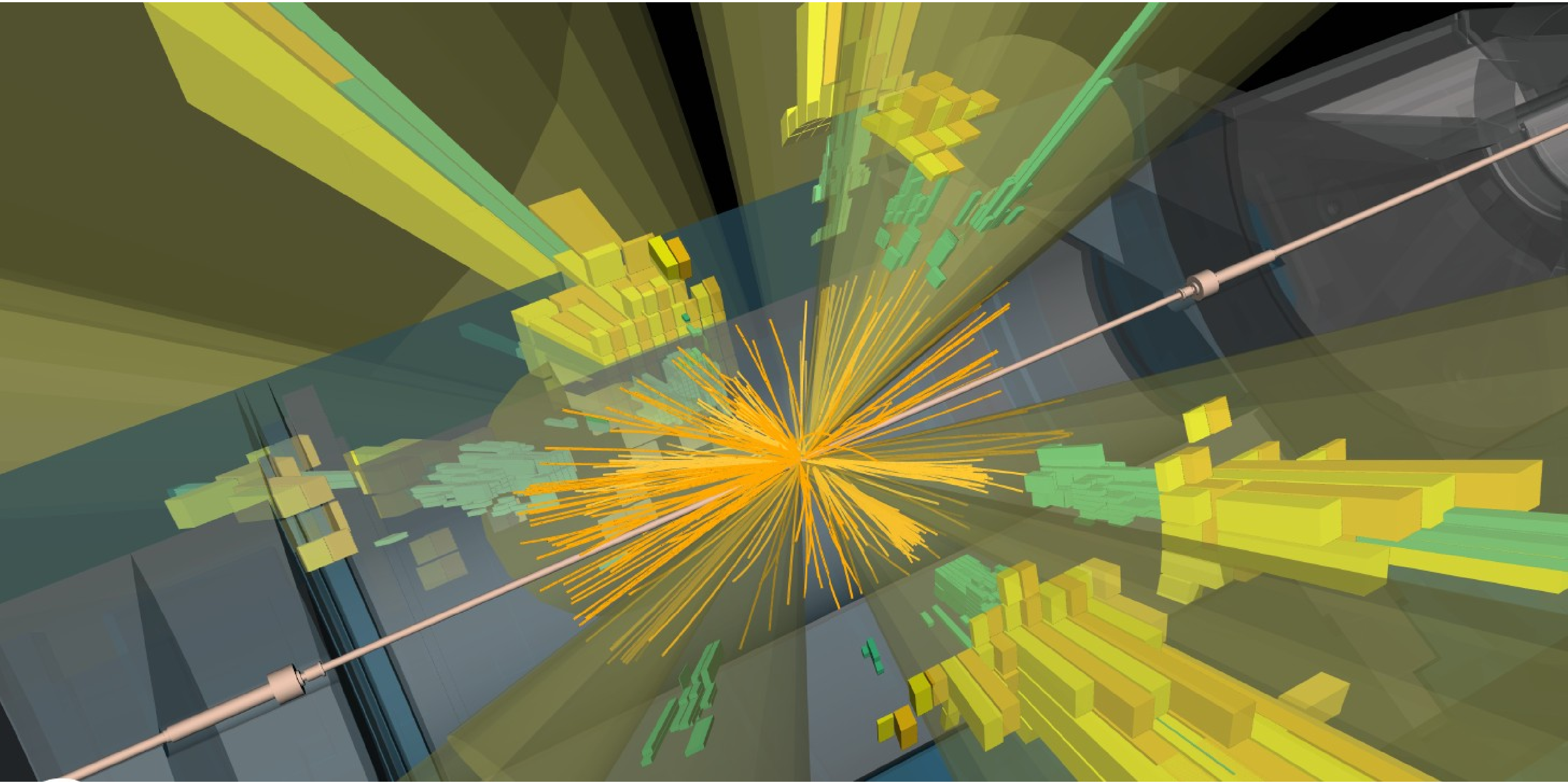
Systematics and further topics (Saturday)

- **Systematics and profiling**
- **[Bayesian techniques]**

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

The class will be based on both lectures and hands-on tutorials

Randomness in High-Energy Physics



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

Randomness in High-Energy Physics

More details in other lectures!

Experimental data is produced by incredibly complex processes

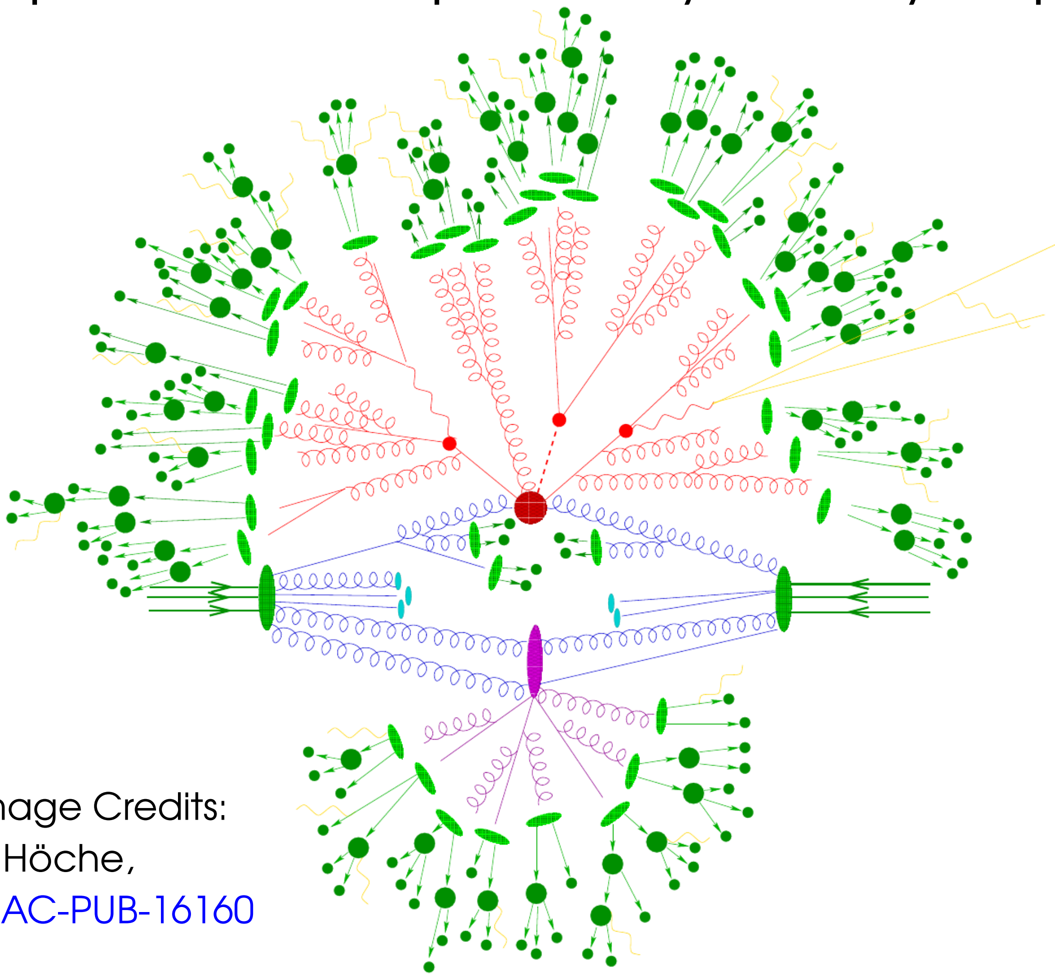


Image Credits:
S. Höche,
SLAC-PUB-16160

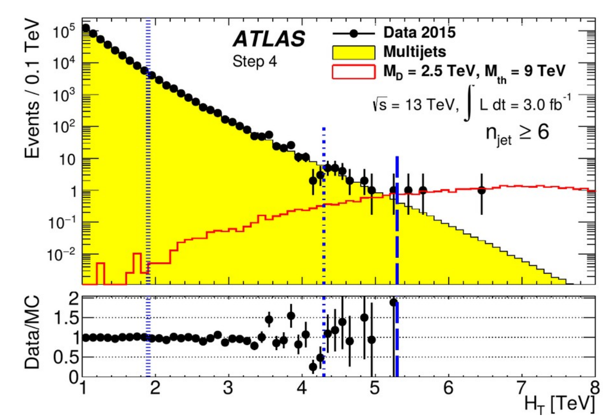
Hard scattering

PDFs, Parton shower, Pileup

Decays

Detector response

Reconstruction



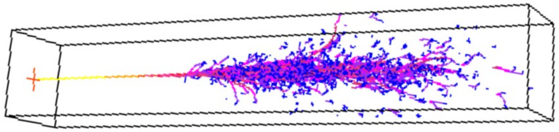
Randomness involved in all stages

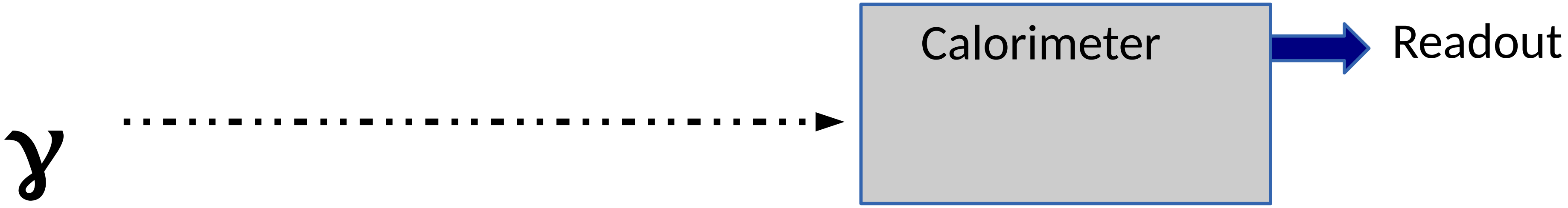
→ **Classical** randomness: detector response

→ **Quantum** effects in particle production, decay

Measurement Errors: Energy measurement

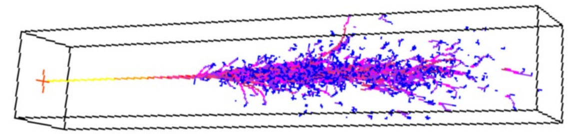
Example: measuring the energy of a photon in a calorimeter

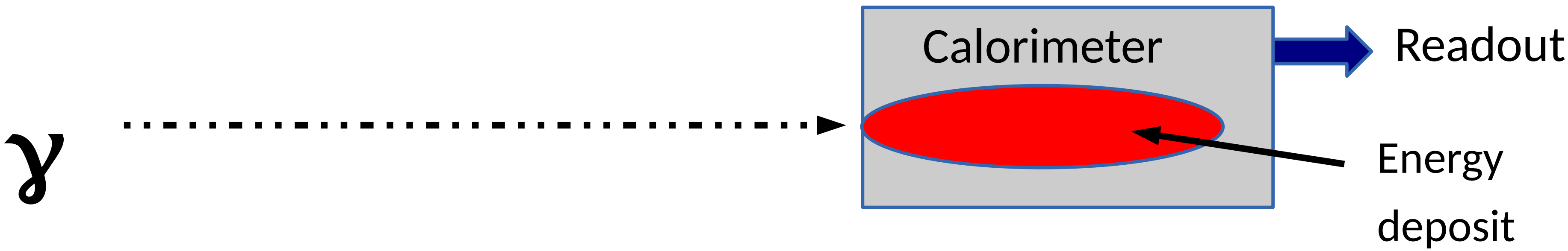


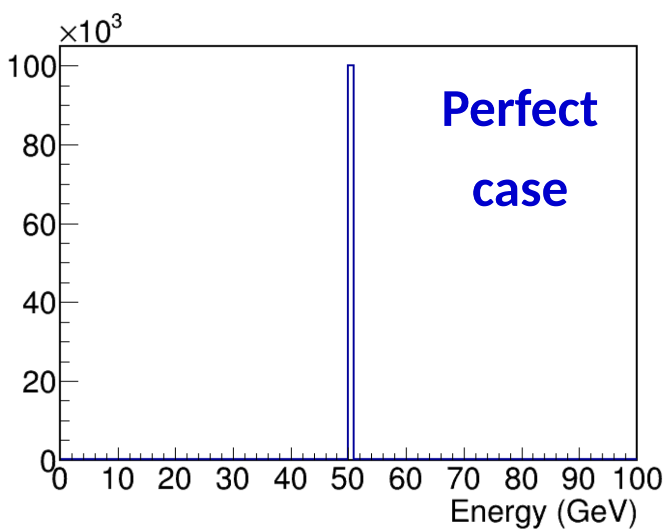


Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

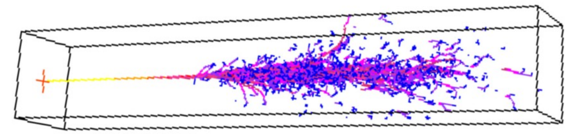


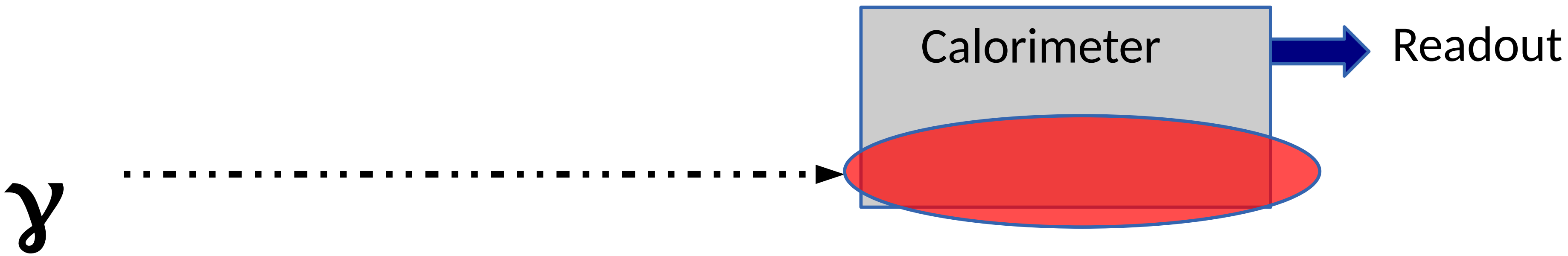


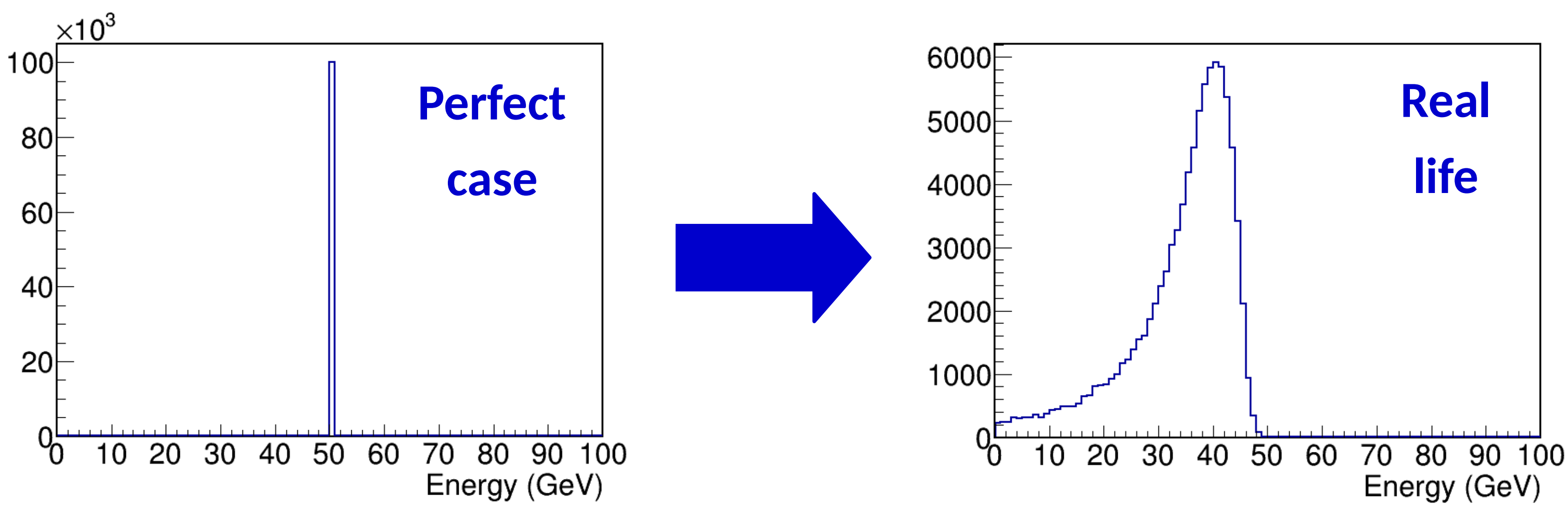


Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

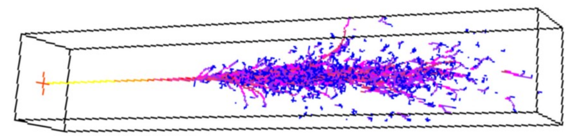


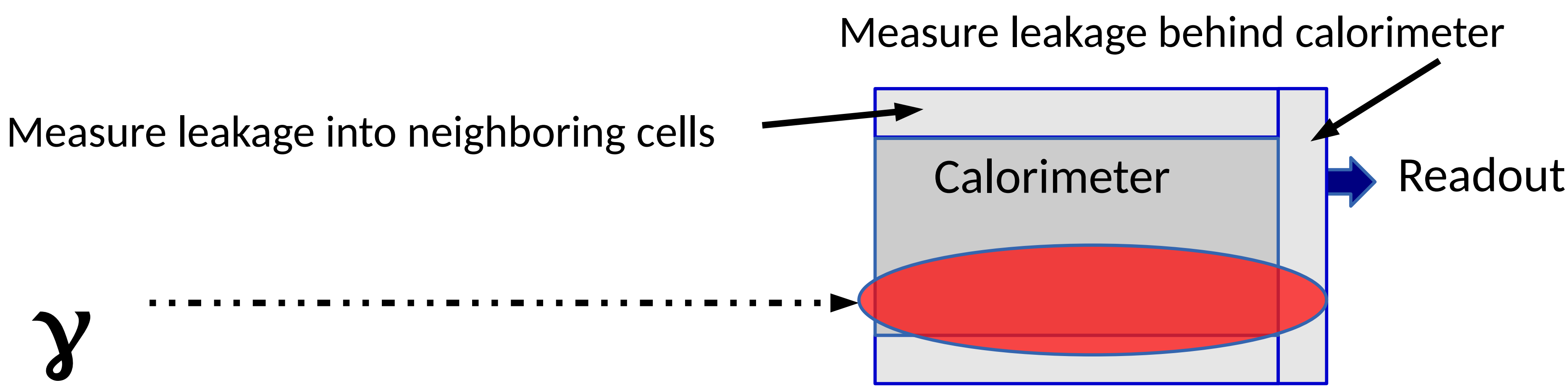


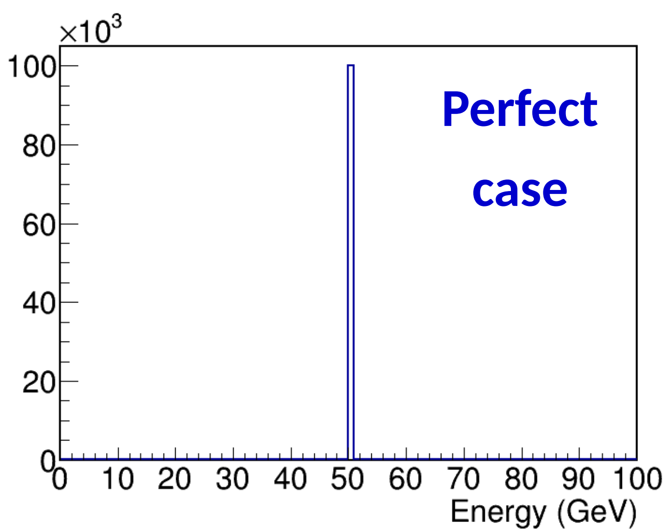


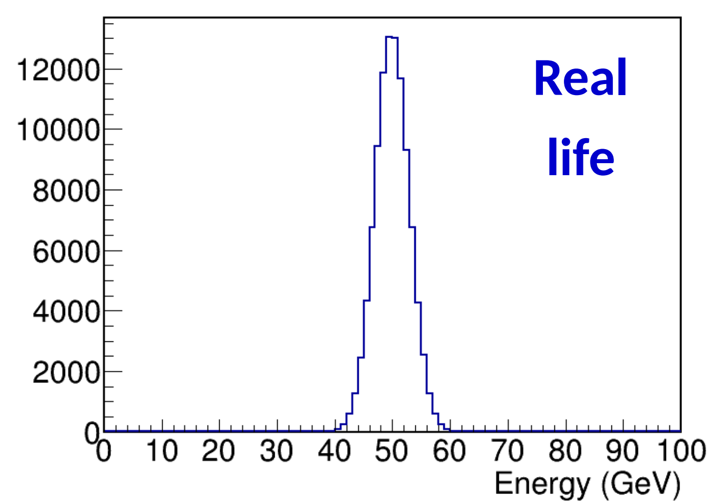
Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter



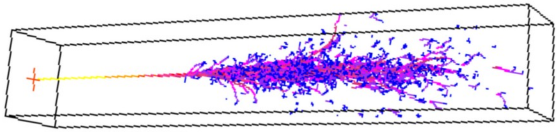


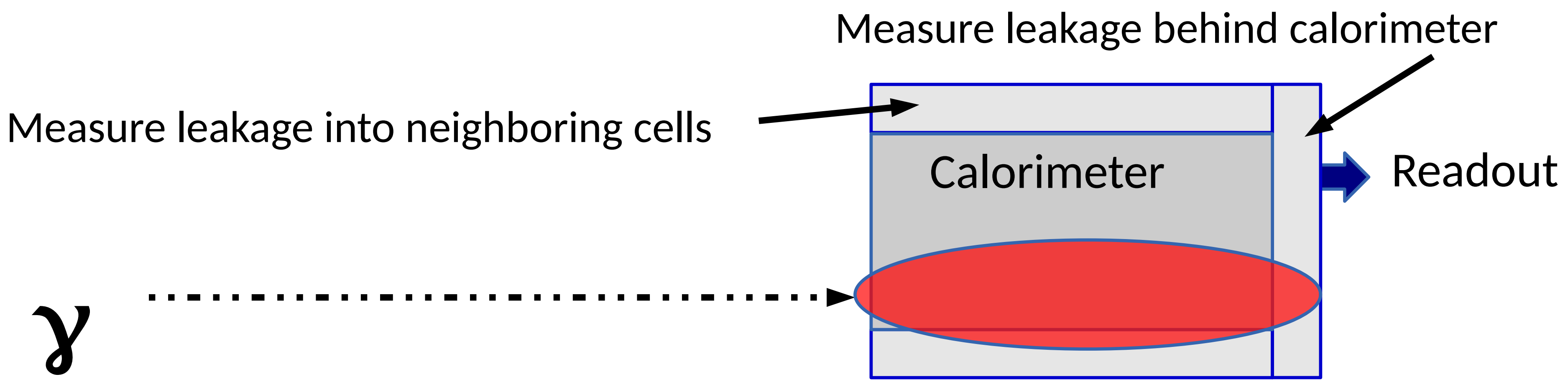


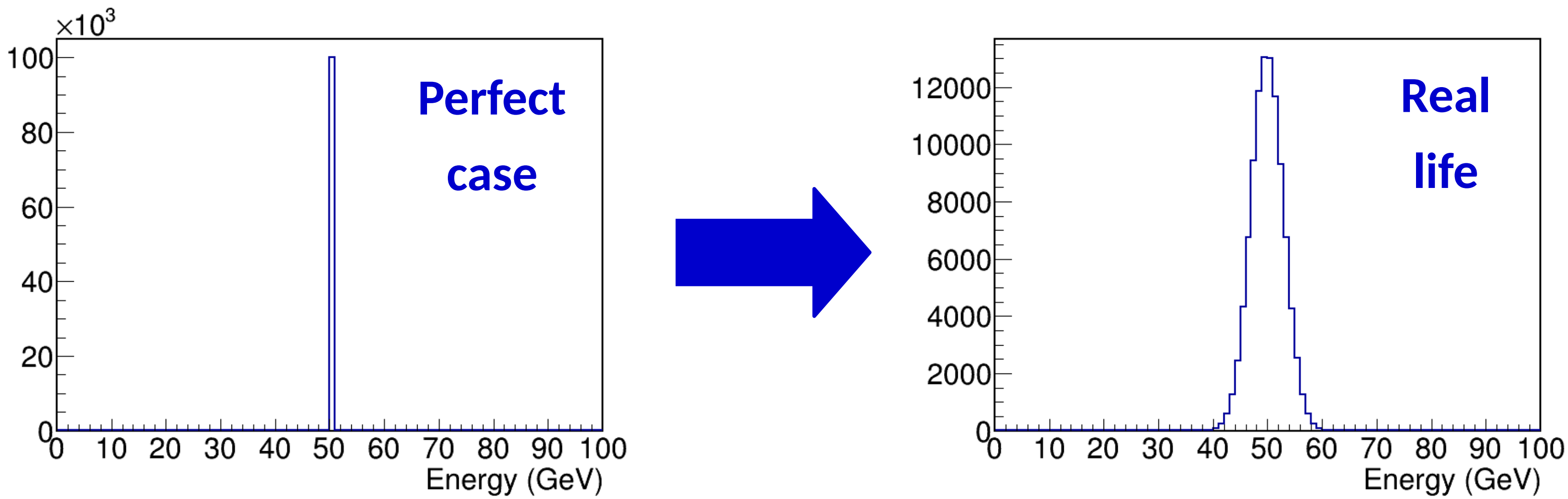


Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter





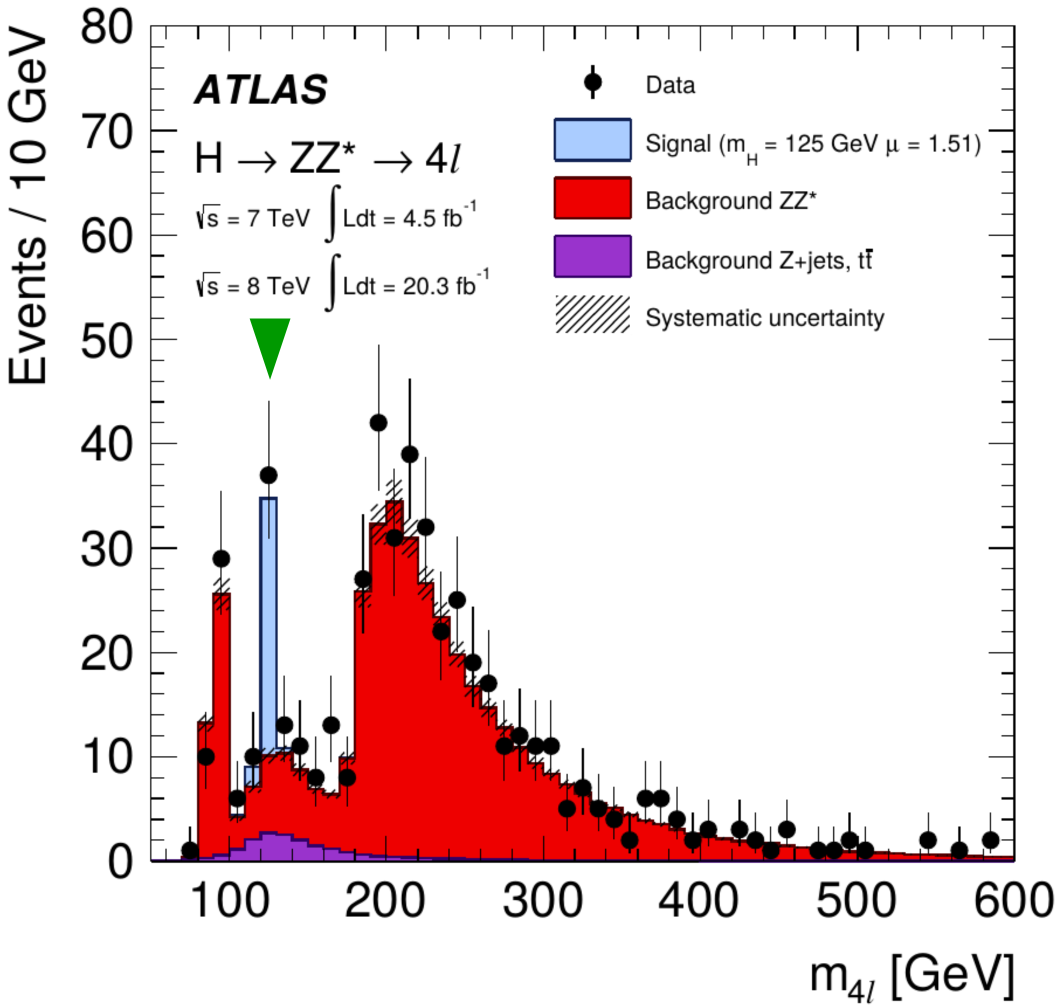


Cannot predict the measured value for a given event

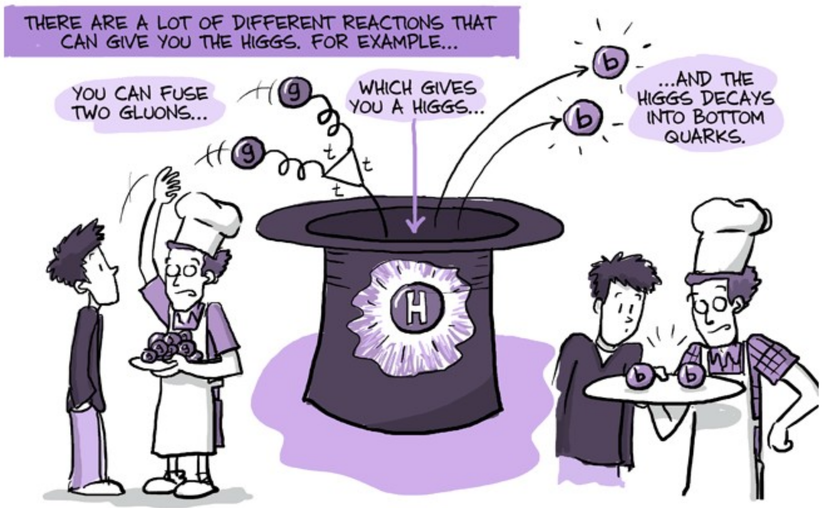
⇒ **Random process** ⇒ **Need a probabilistic description**

Quantum Randomness: H→ZZ*→4l

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

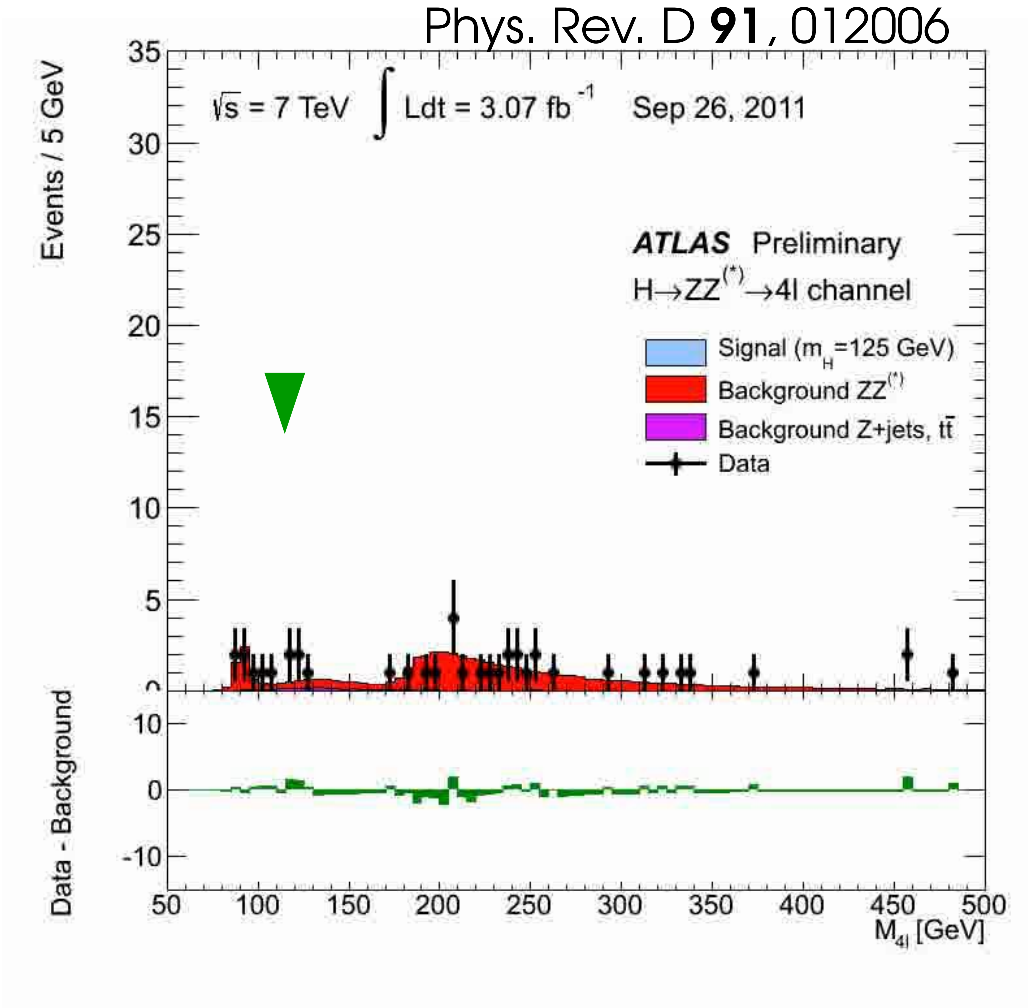


Rare process: Expect 1 signal event every **~6 days**

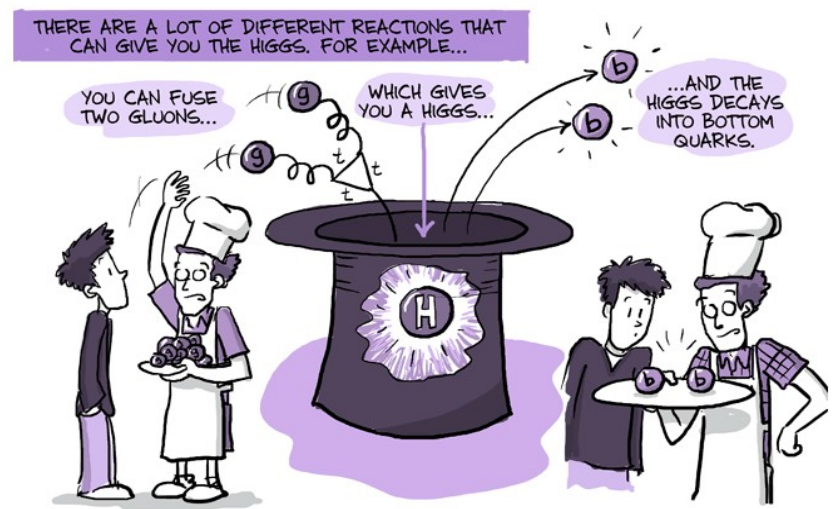


http://www.phdcomics.com/comics/archive.php?comicid=1489

Quantum Randomness: H→ZZ*→4l



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

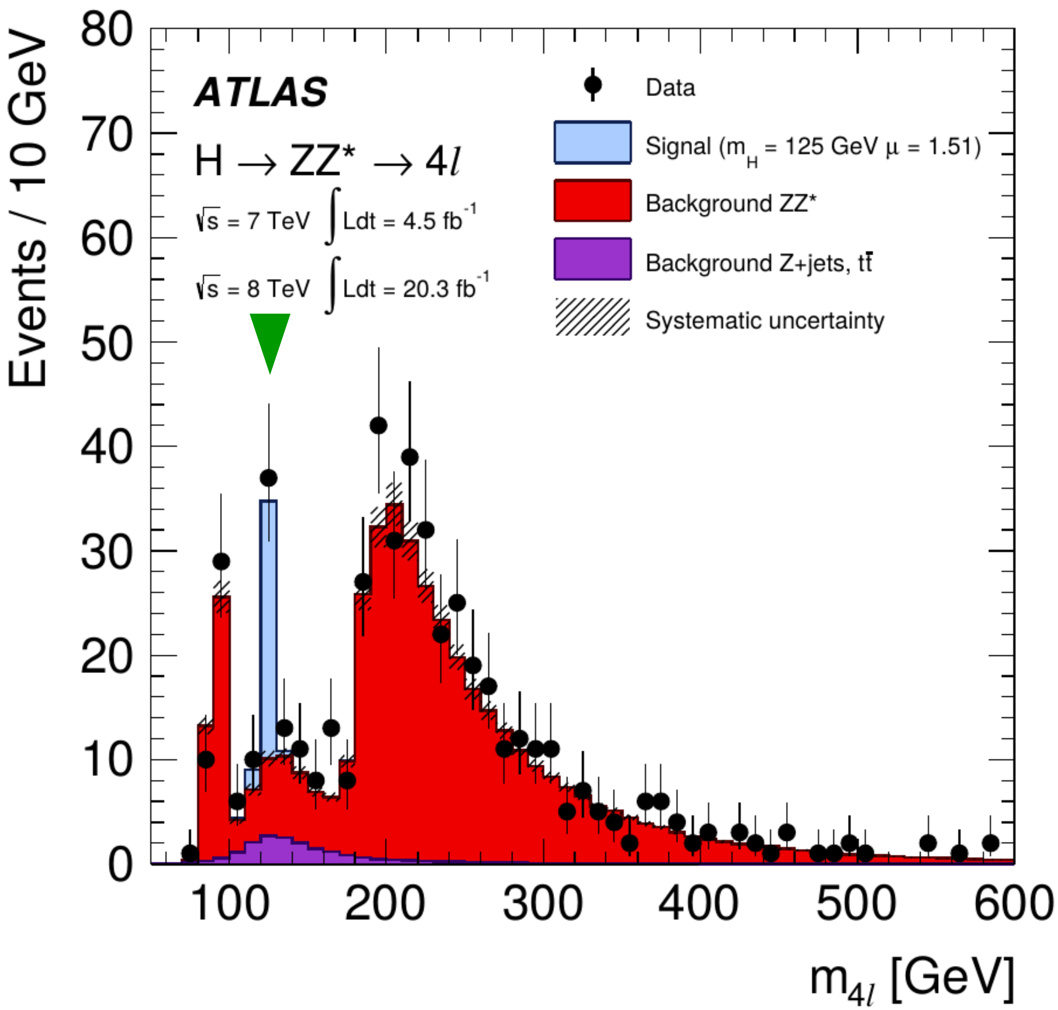


http://www.phdcomics.com/comics/archive.php?comicid=1489

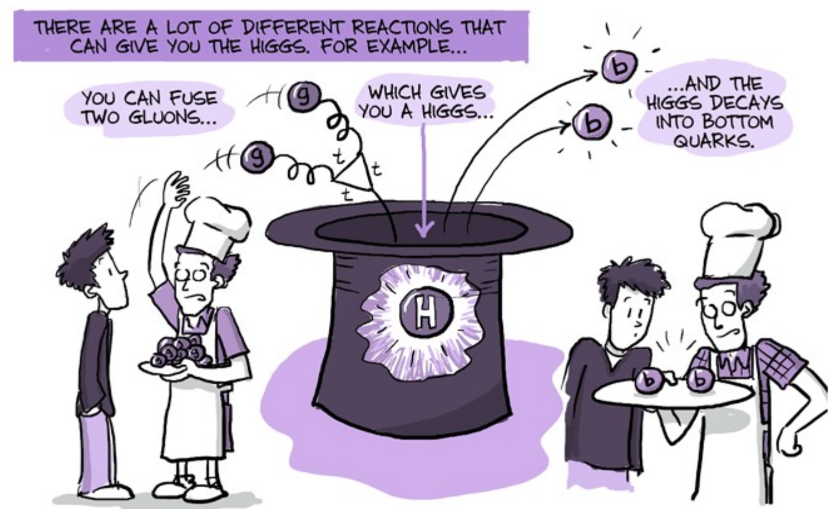
View online

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

Statistical Modeling

Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

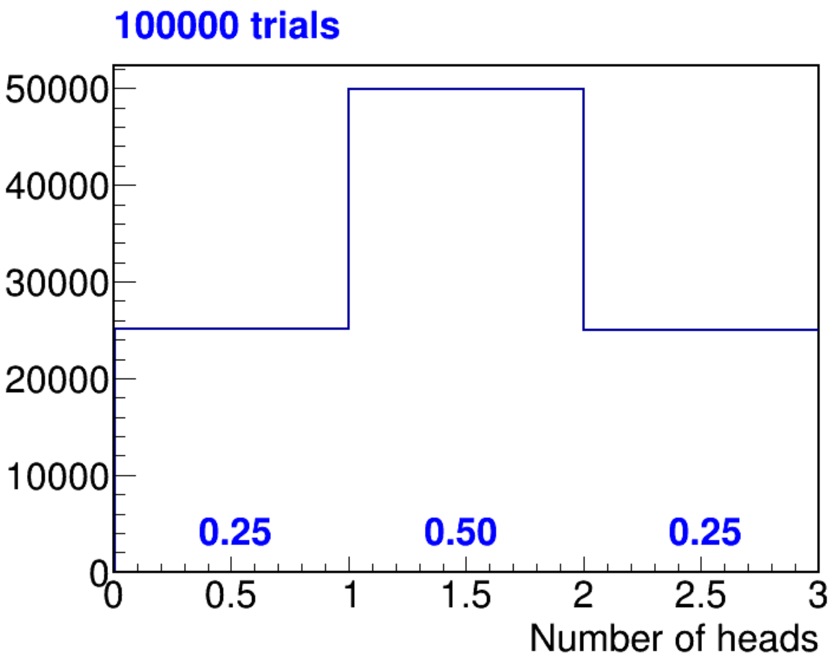
→ Fractions of events in each bin i converge to a limit p_i

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

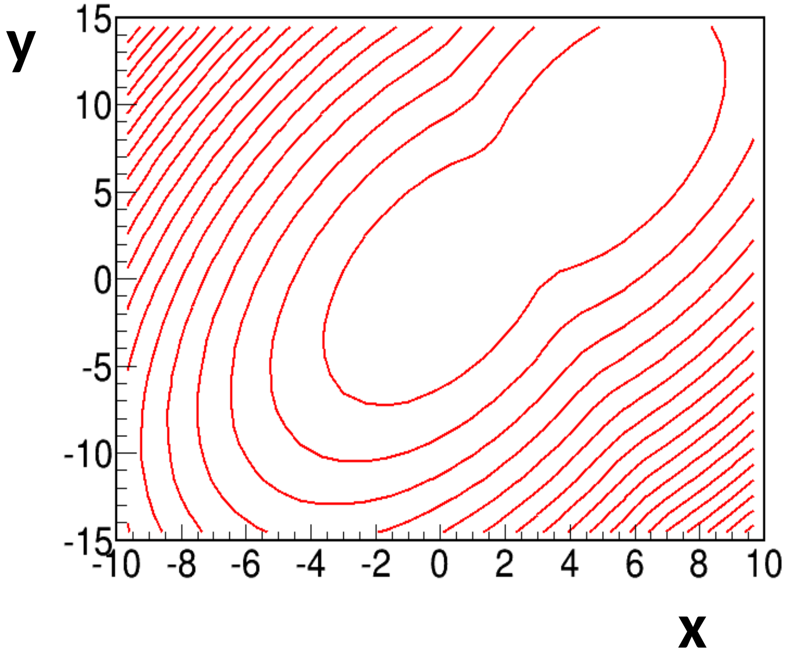
x

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

Generalizes to **multiple variables** :

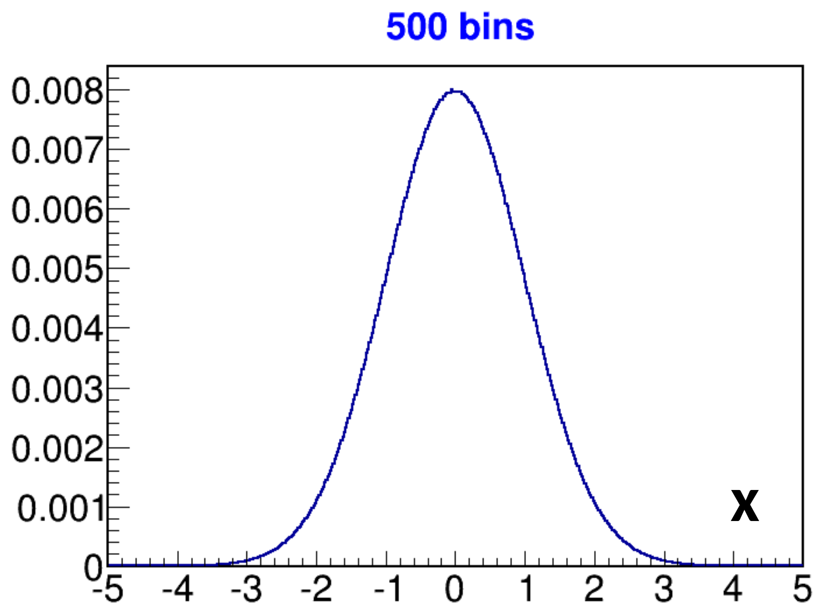
$P(x,y) > 0, \int P(x,y)\,dx\,dy = 1$

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

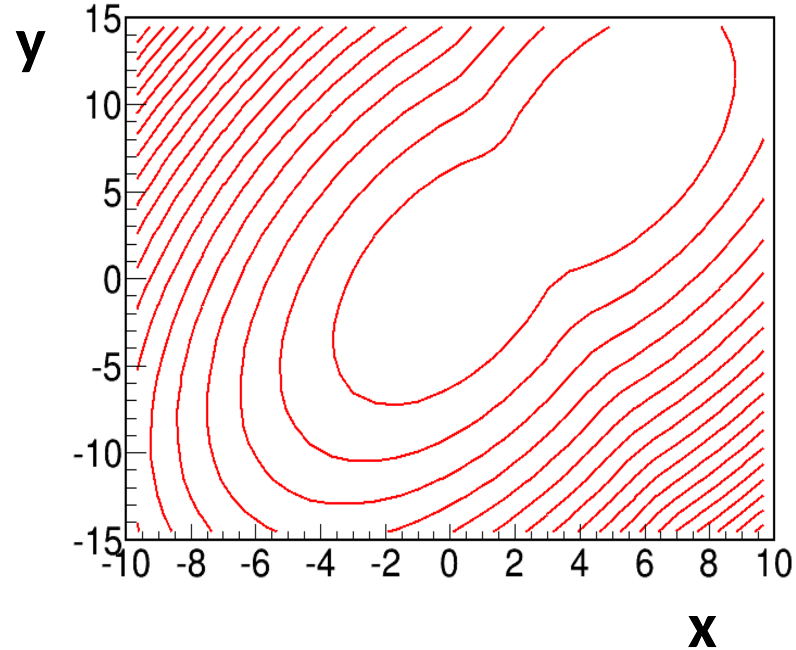
⇒ High chance to get a measurement here

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

Generalizes to **multiple variables** :

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

Contours: P(x,y)



PDF Properties: Mean

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

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

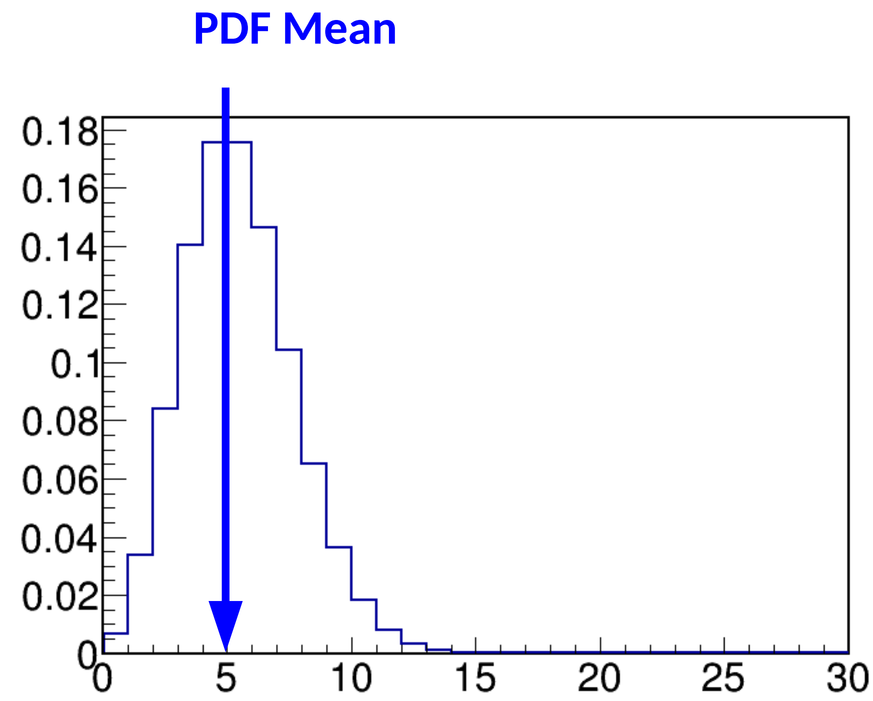
$$\langle X \rangle = \int x \, P(x) \, dx$$

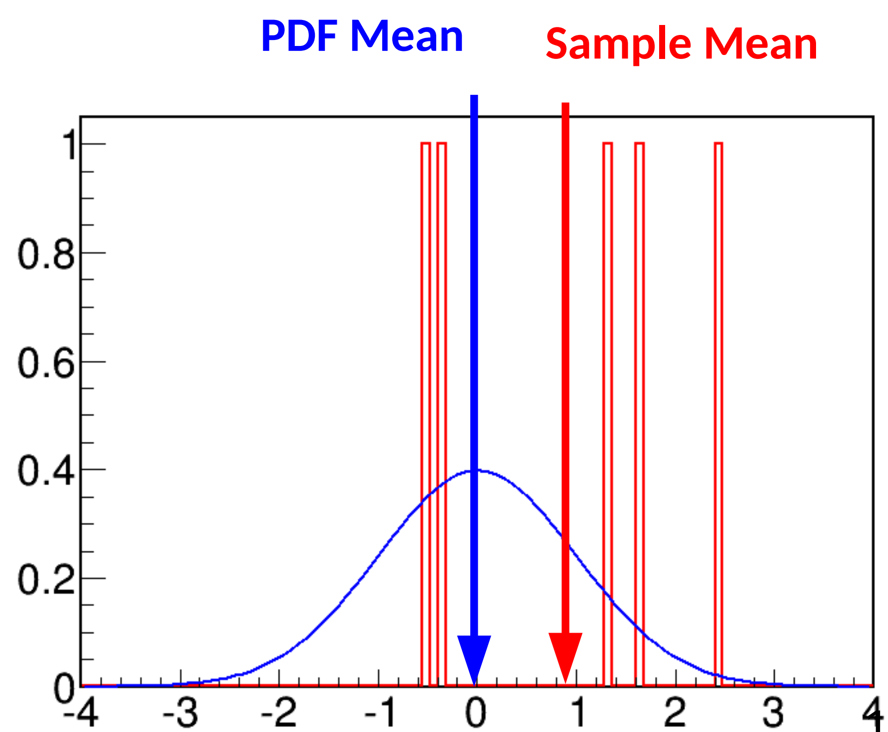
→ Property of the **PDF**

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:

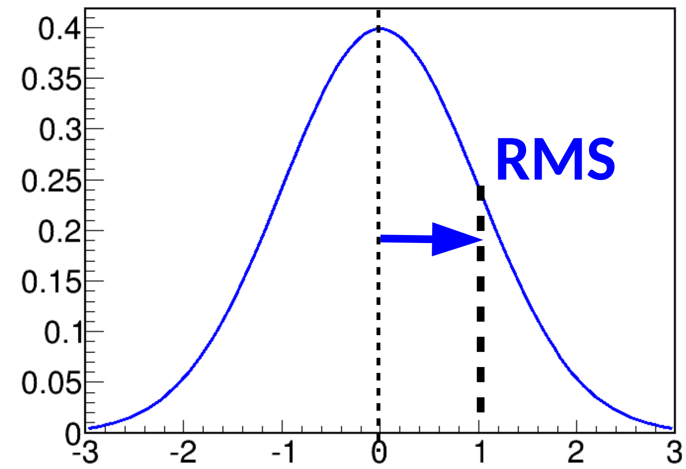
$$\mathrm{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$$

→ Average square of deviation from mean

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

Can be approximated by **sample variance**:

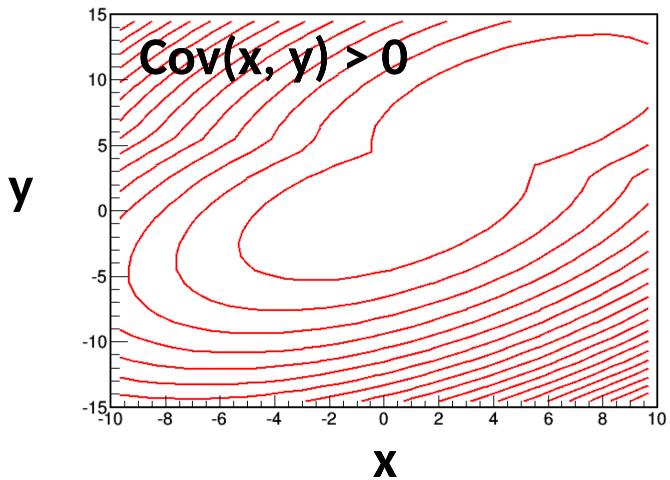
$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$



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

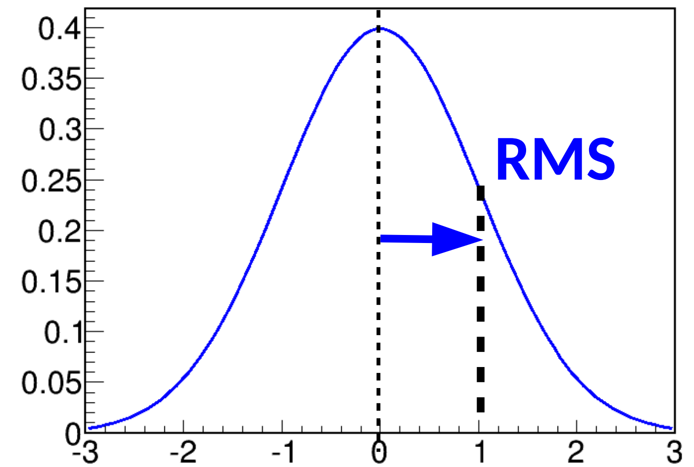
$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

$-1 \le \rho \le 1$

PDF Properties: (Co)variance

Variance of X:

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



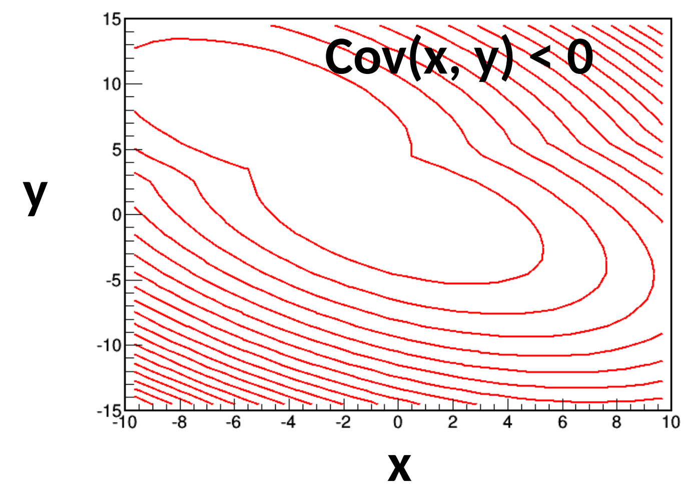
→ Average square of deviation from mean

→ **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 $\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$ $-1 \leq \rho \leq 1$

PDF Properties: (Co)variance

Variance of X:

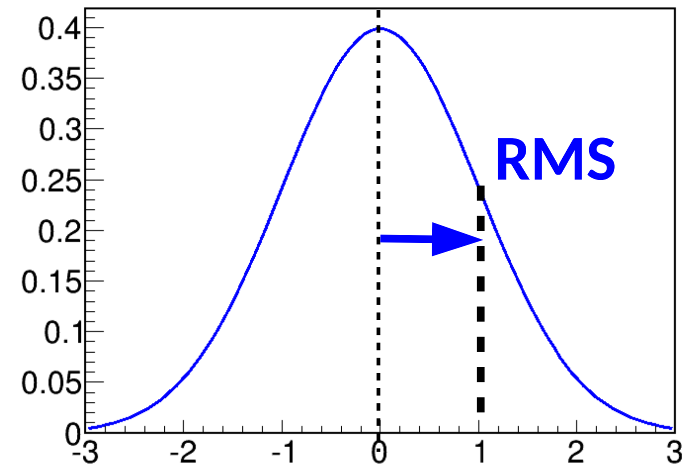
$$\mathrm{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$$

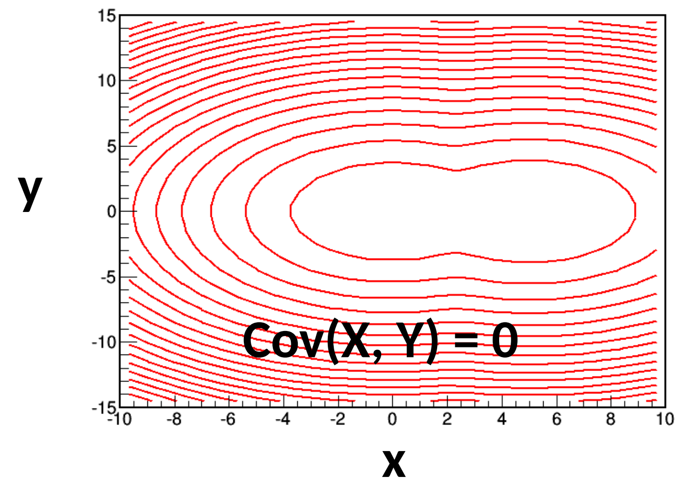
→ Average square of deviation from mean

→ **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

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

$-1 \leq \rho \leq 1$

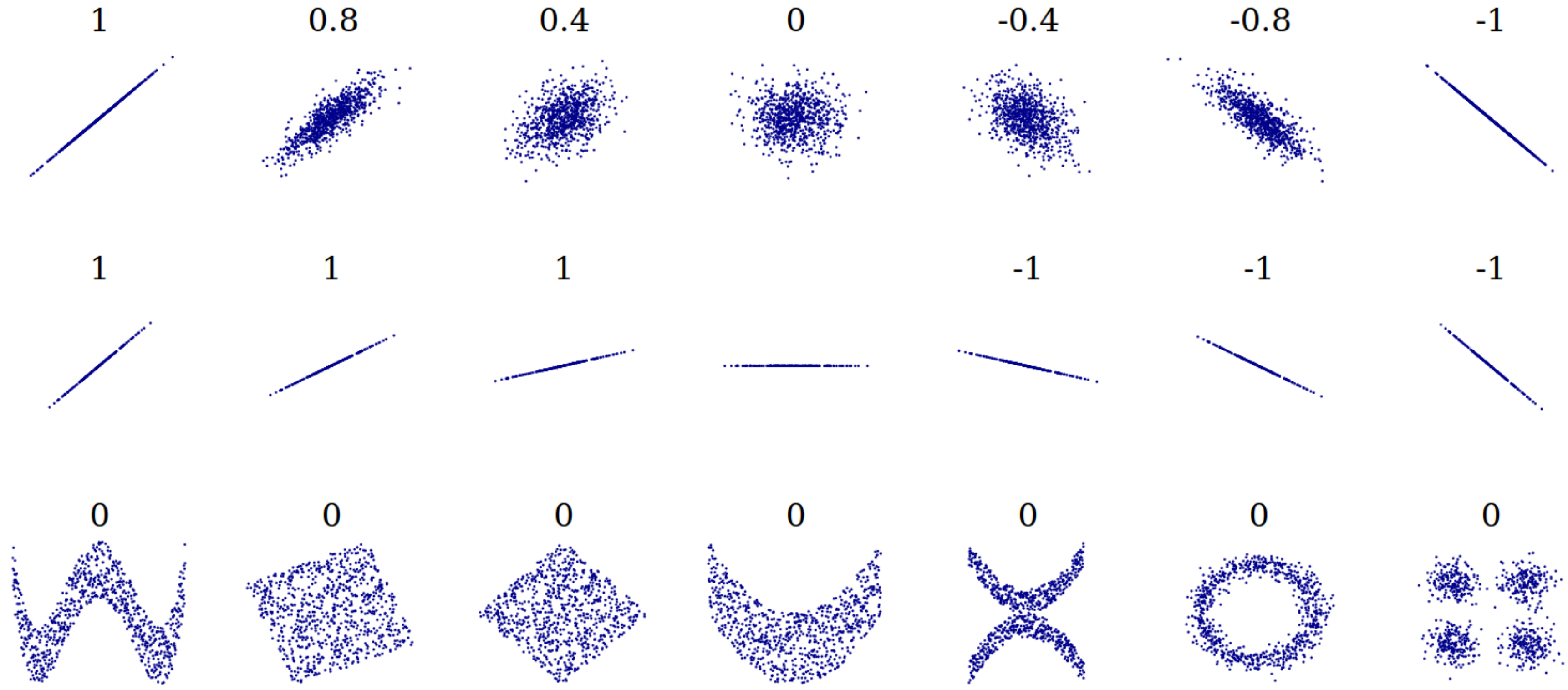
“Linear” vs. “non-linear” correlations

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

ρ

ρ

ρ



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

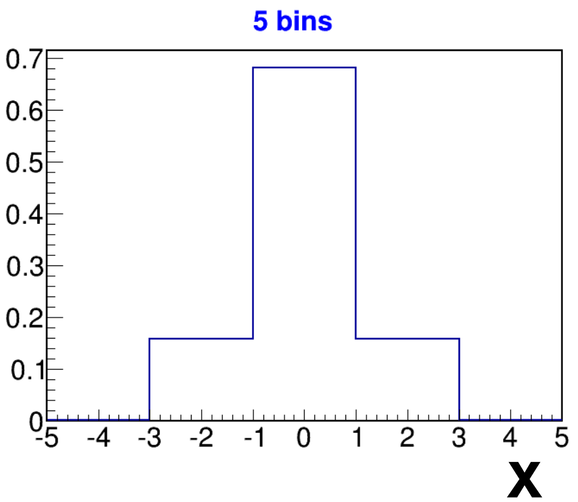
Source: Wikipedia

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

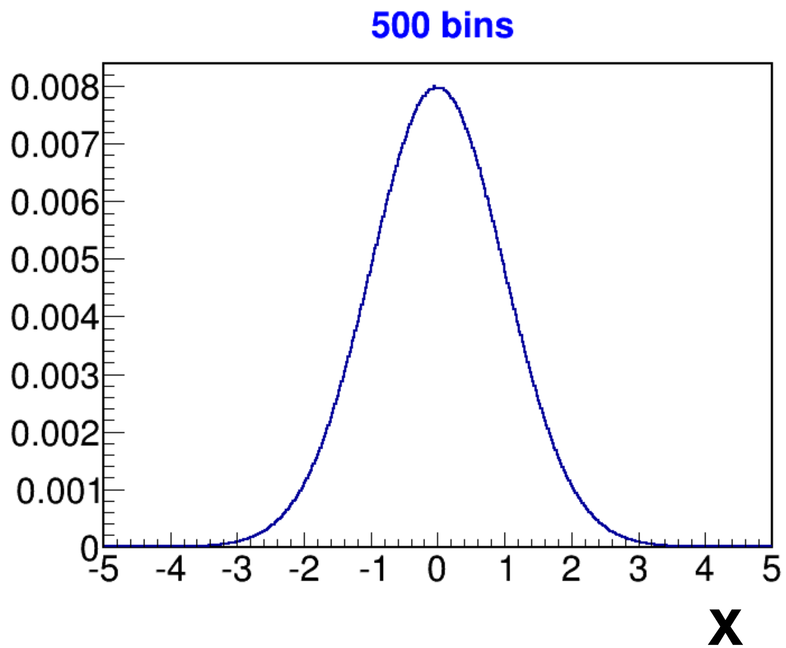
Some vocabulary...

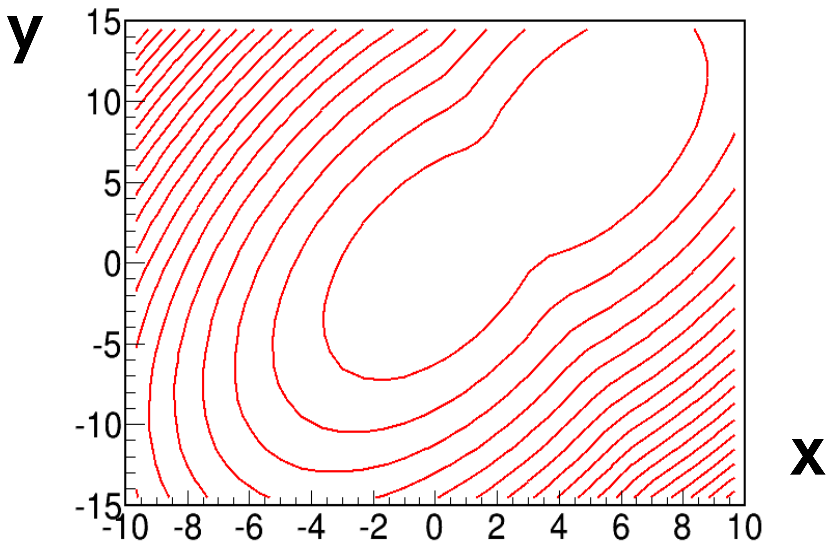
X, Y... are **Random Variables** (continuous or discrete), a.ka. **observables** :
→ X can take any value x, with probability **P(X=x).**

→ P(X=x) 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).

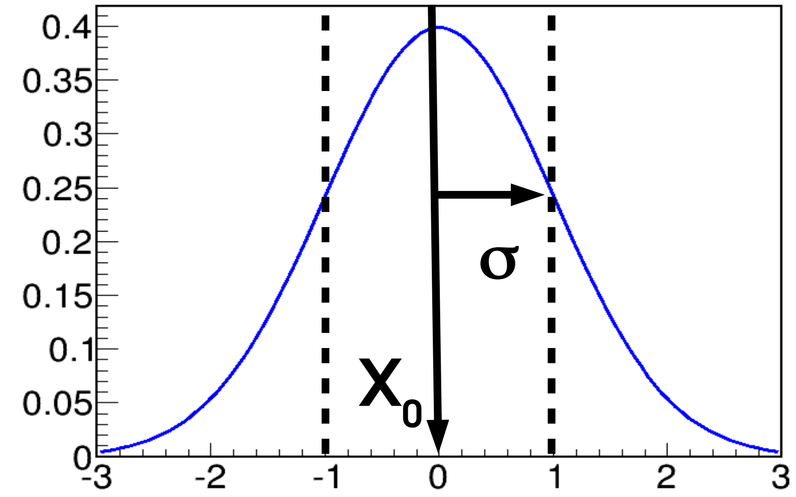




Gaussian PDF

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**:

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

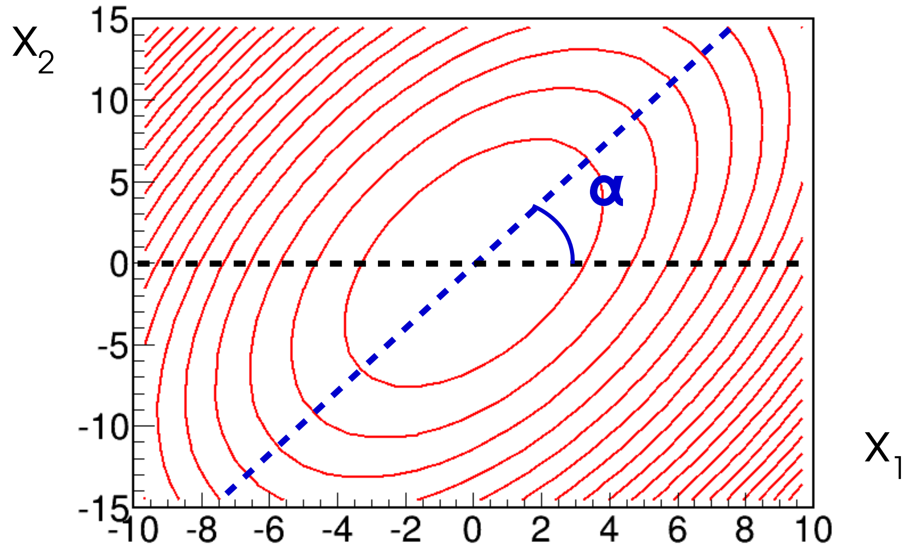
→ **Mean** : X_0

→ **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

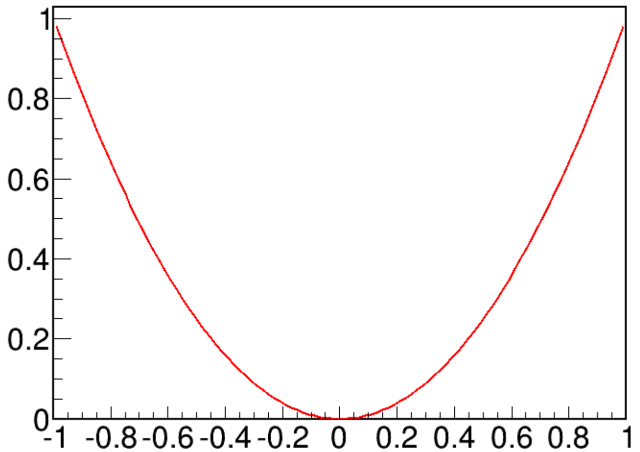
Another version:

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

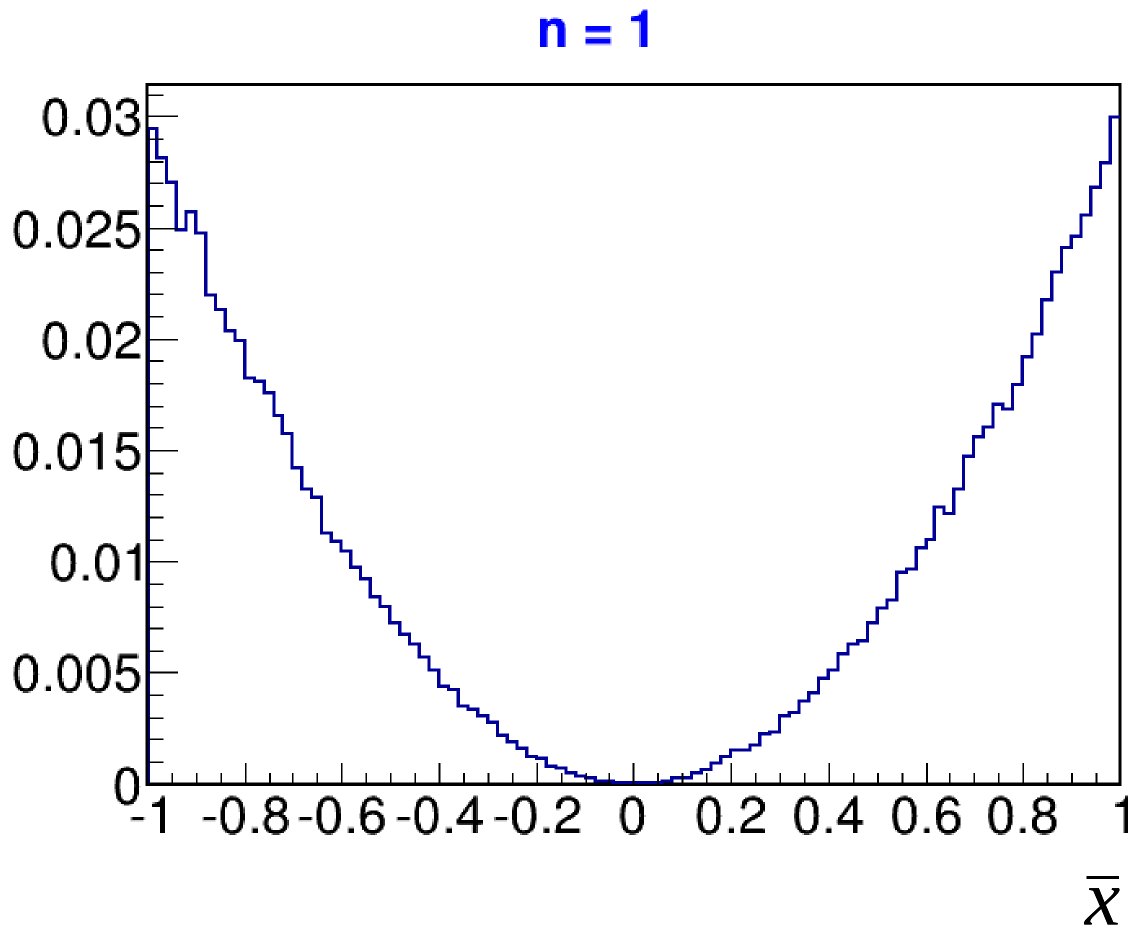
Mean scales like n, but RMS only like √n

Central Limit Theorem in action

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



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

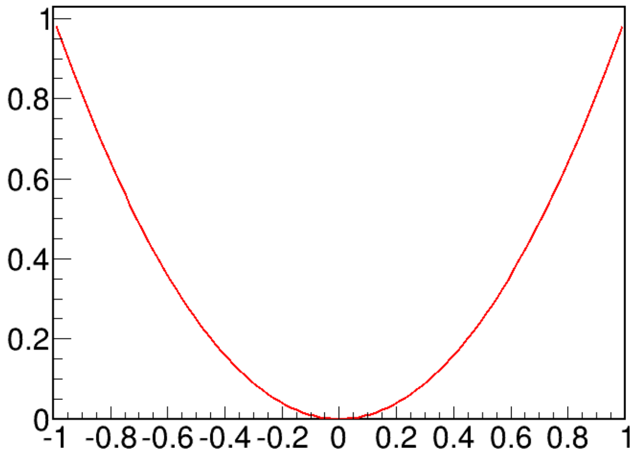


Distribution becomes Gaussian, although very non-Gaussian originally

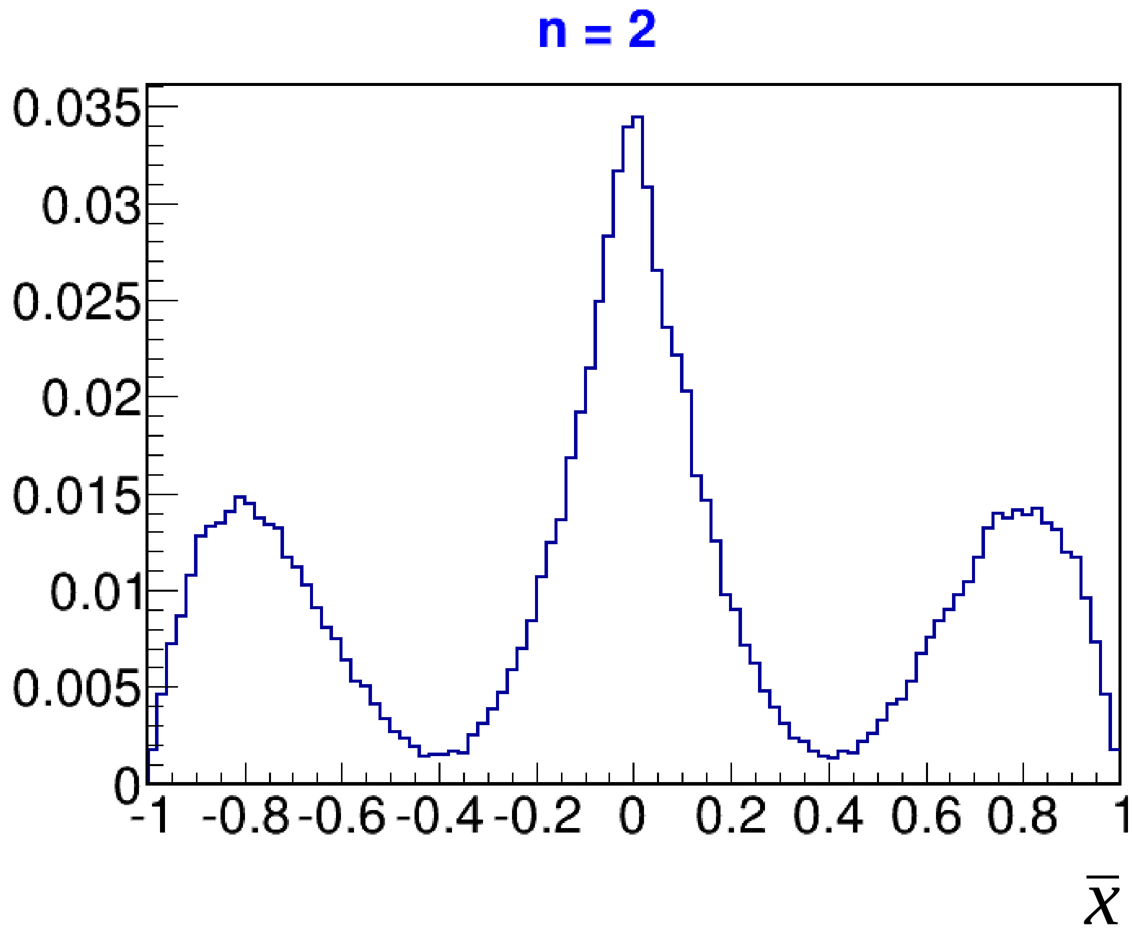
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

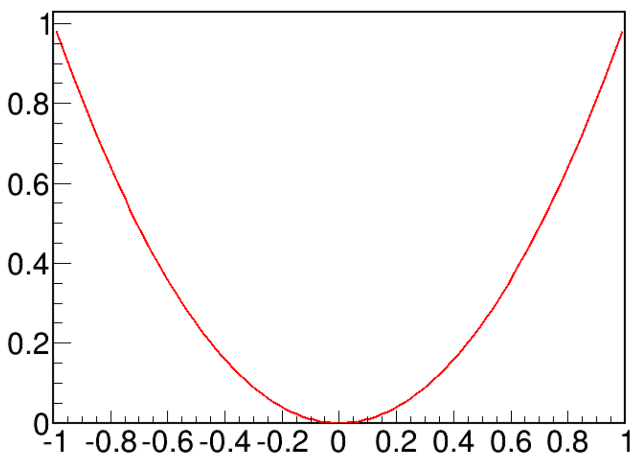


Distribution becomes Gaussian, although very non-Gaussian originally

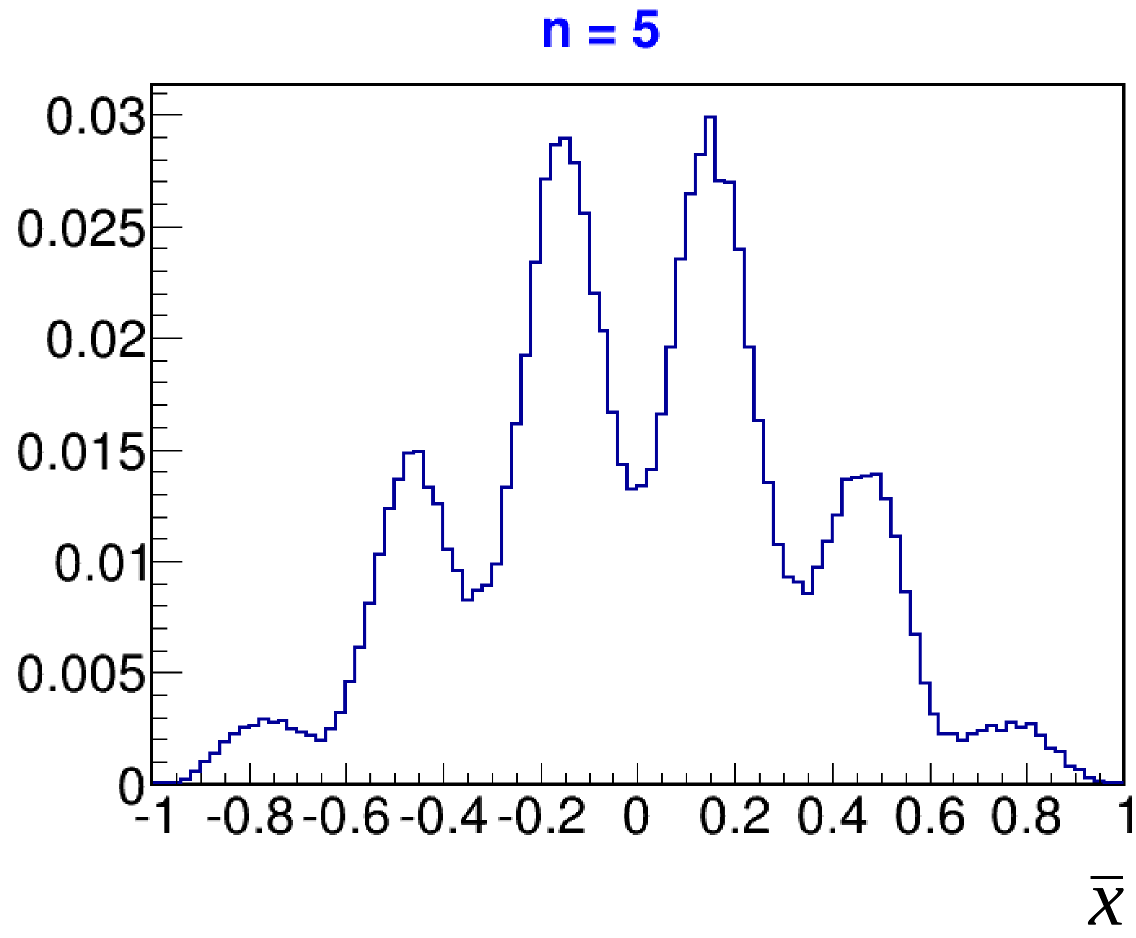
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

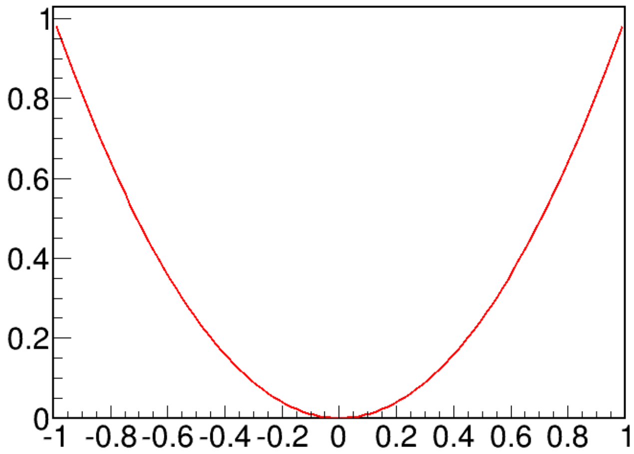


Distribution becomes Gaussian, although very non-Gaussian originally

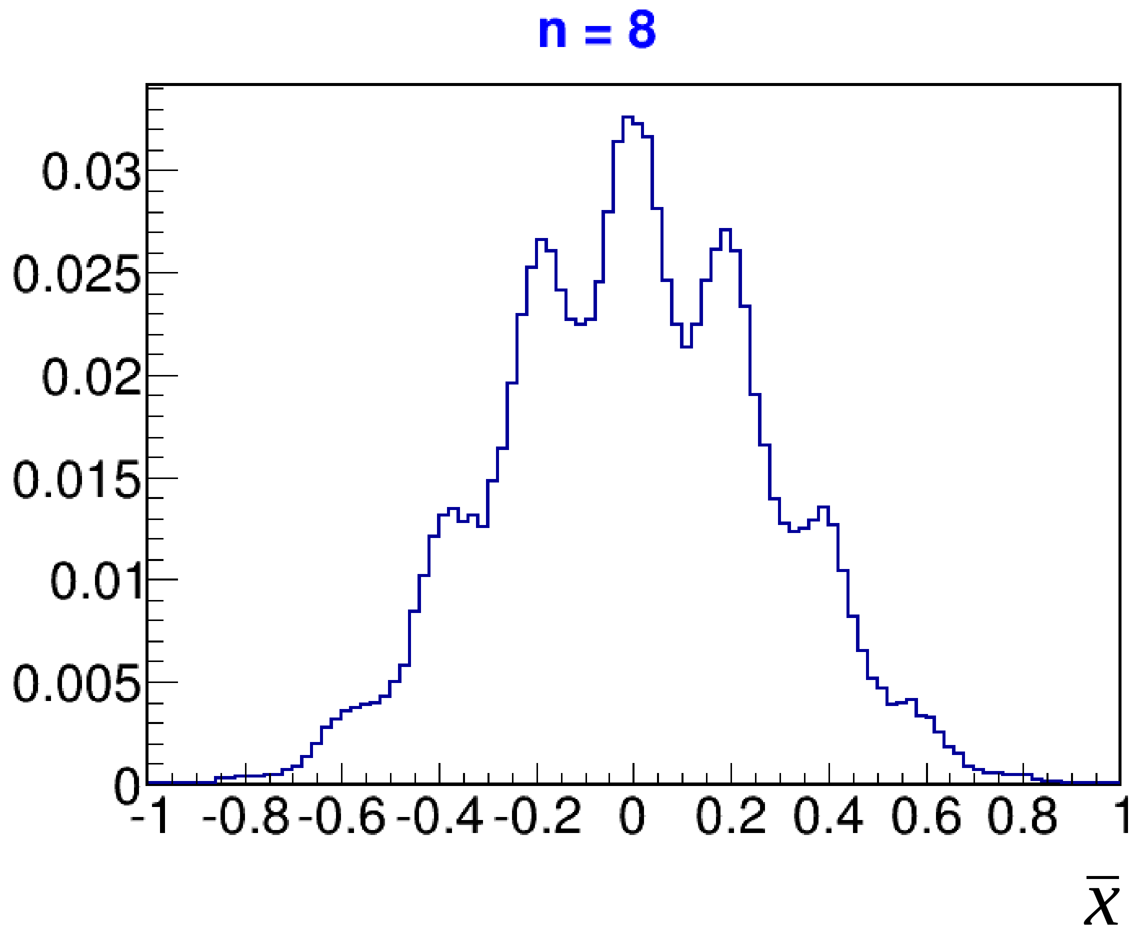
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

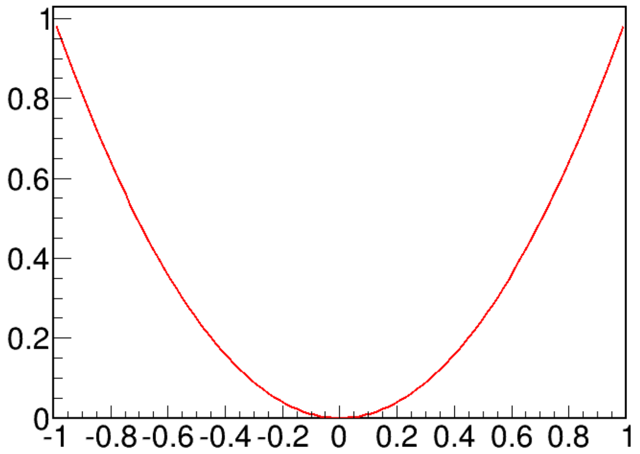


Distribution becomes Gaussian, although very non-Gaussian originally

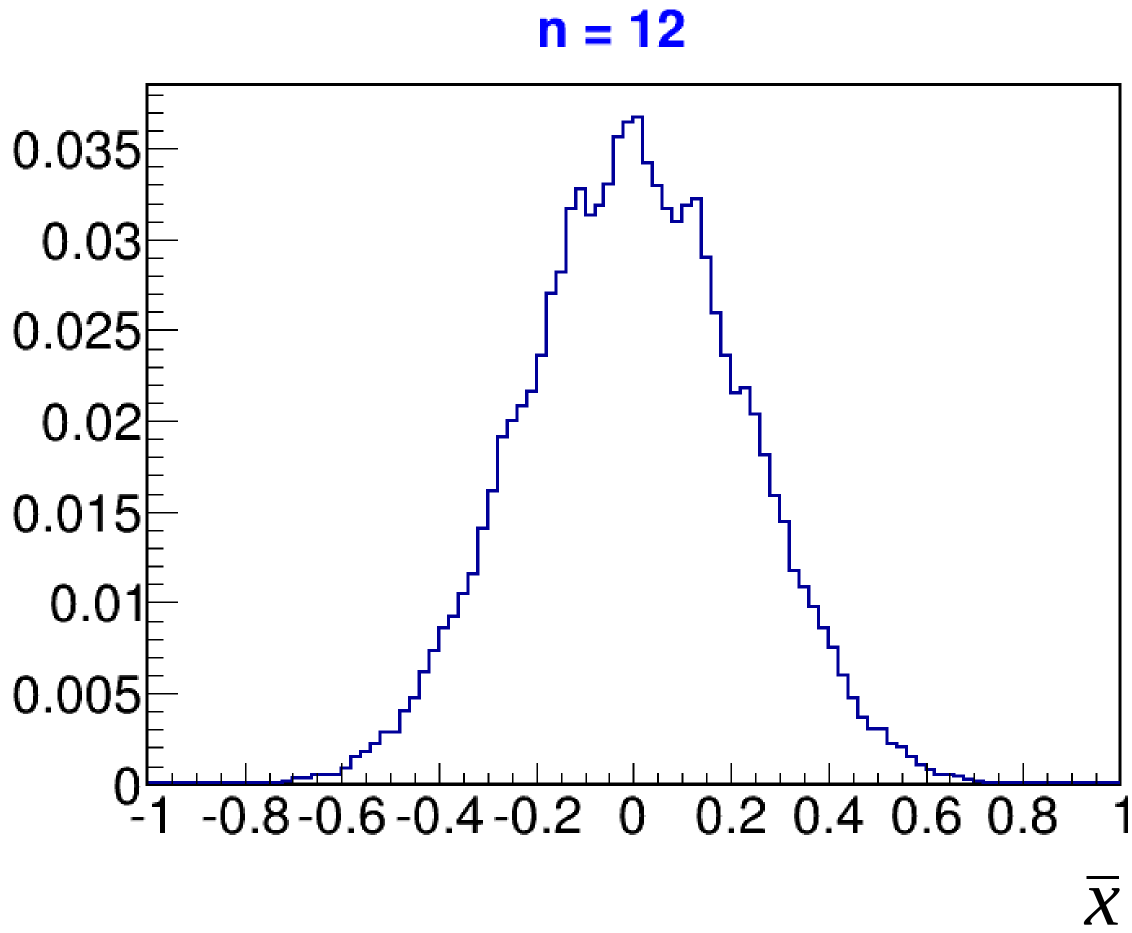
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

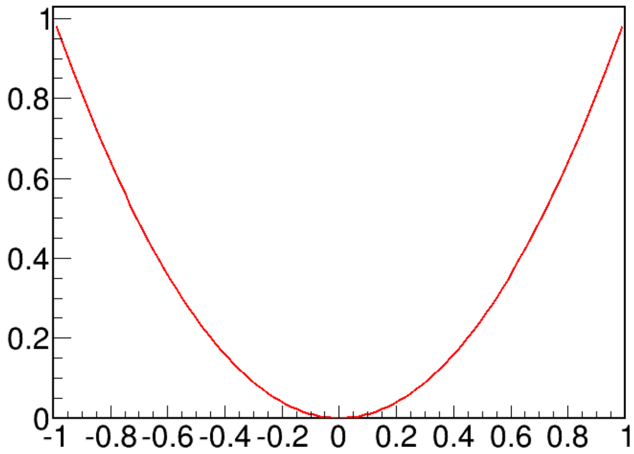


Distribution becomes Gaussian, although very non-Gaussian originally

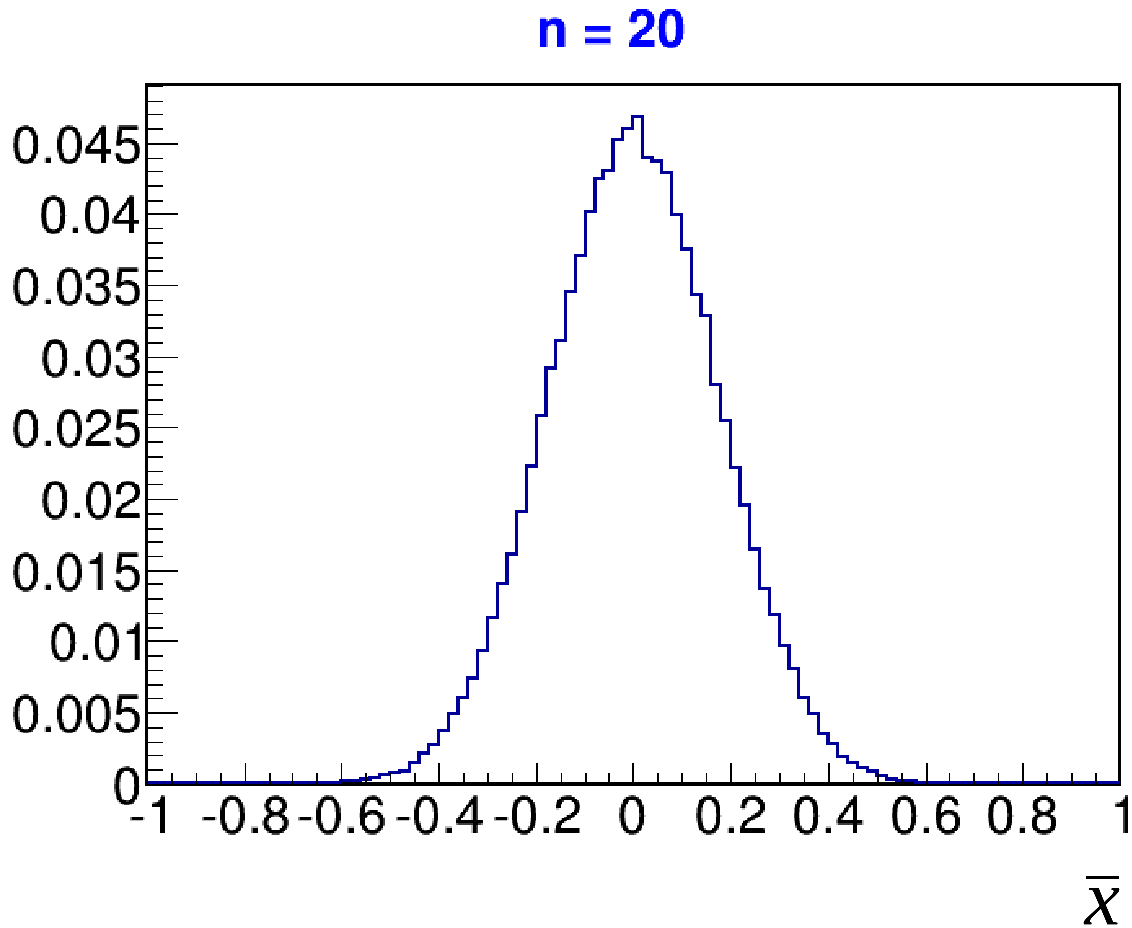
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

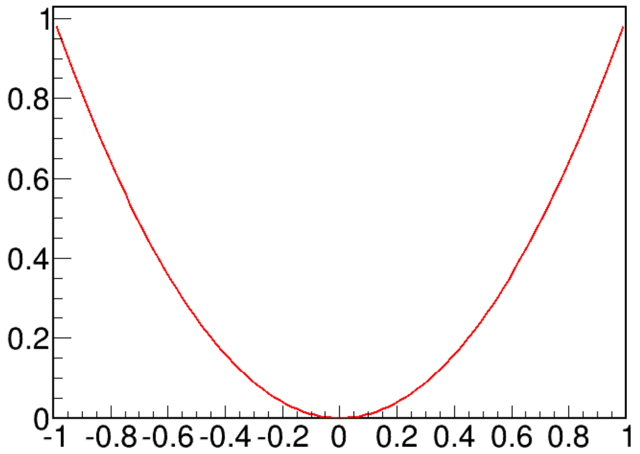


Distribution becomes Gaussian, although very non-Gaussian originally

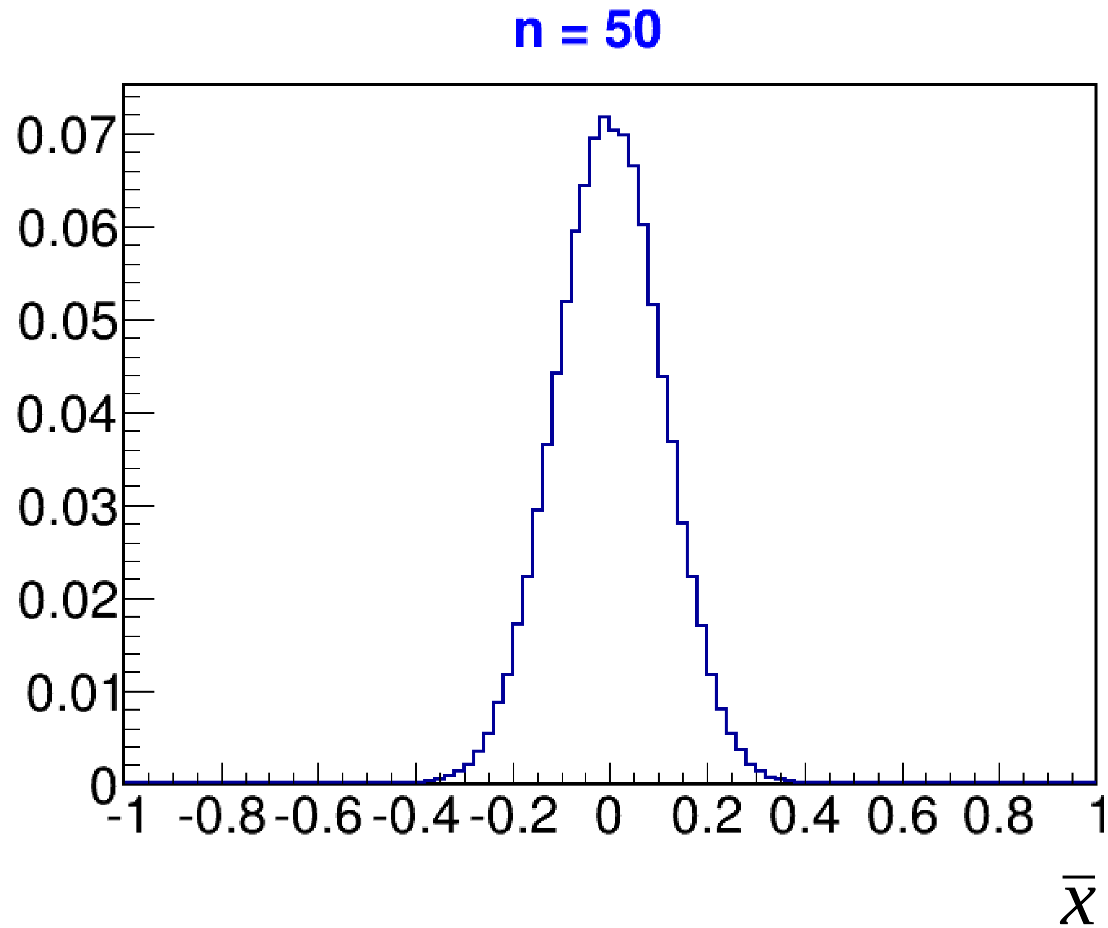
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

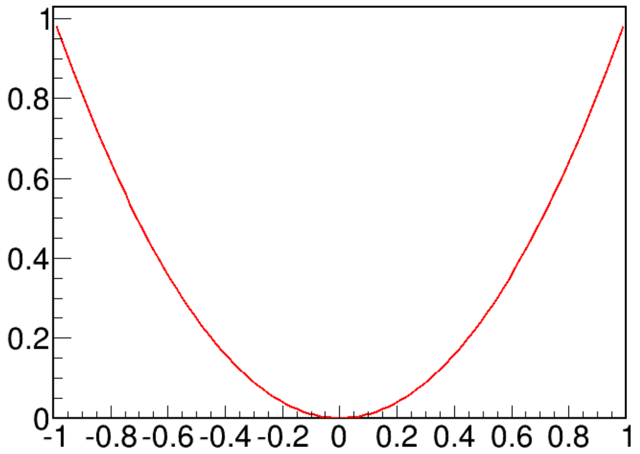


Distribution becomes Gaussian, although very non-Gaussian originally

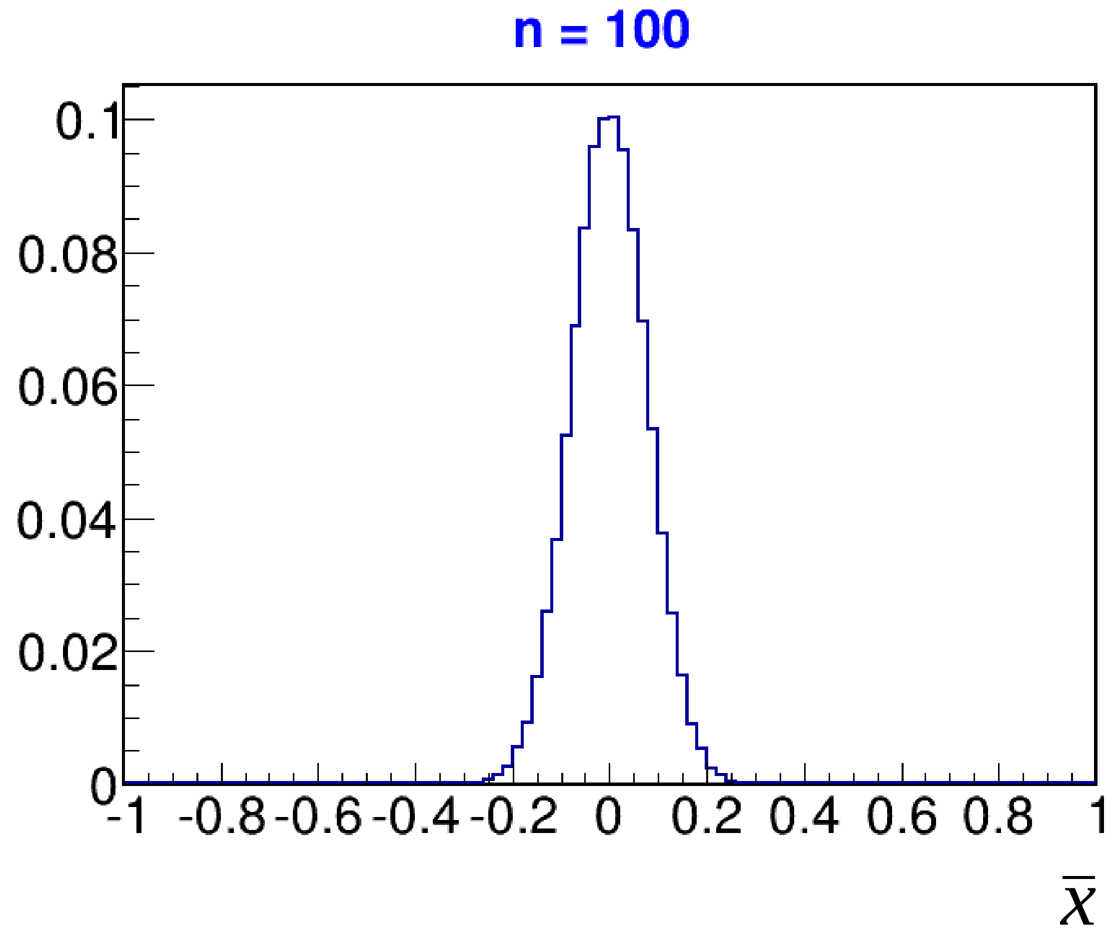
Distribution becomes narrower as expected (as 1/√n)

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

G(x; X_0,σ) depends only on z ~ G(z; 0,1)

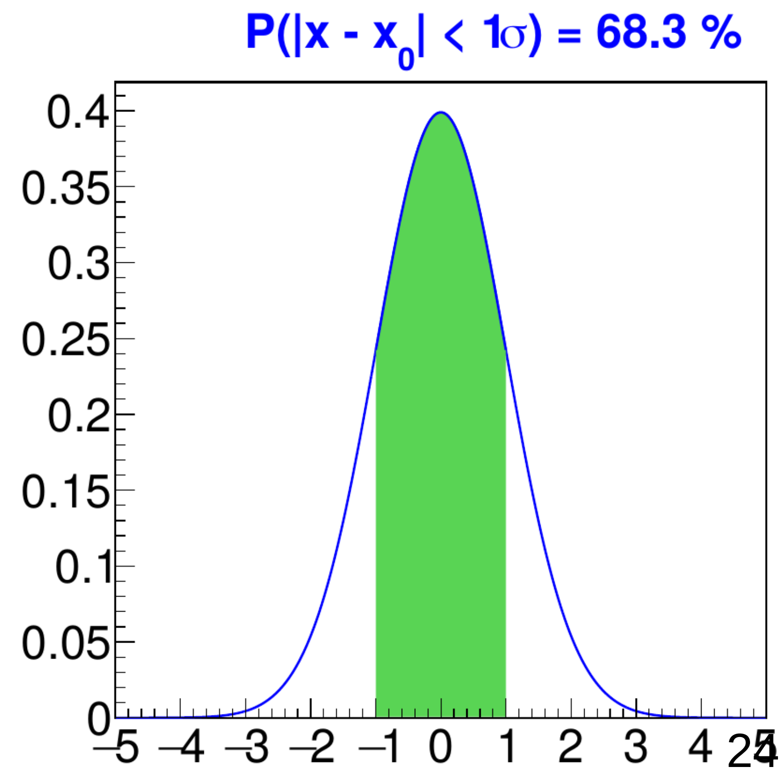
Probability P(|x – X_0| > Zσ) to be away from the mean:

Z	P(|x – X_0| > Zσ)
1	**0.317**
2	**0.045**
3	**0.003**
4	**3 x 10^{-5}**
5	**6 x 10^{-7}**

Cumulative Distribution Function (CDF)

of the Gaussian :

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



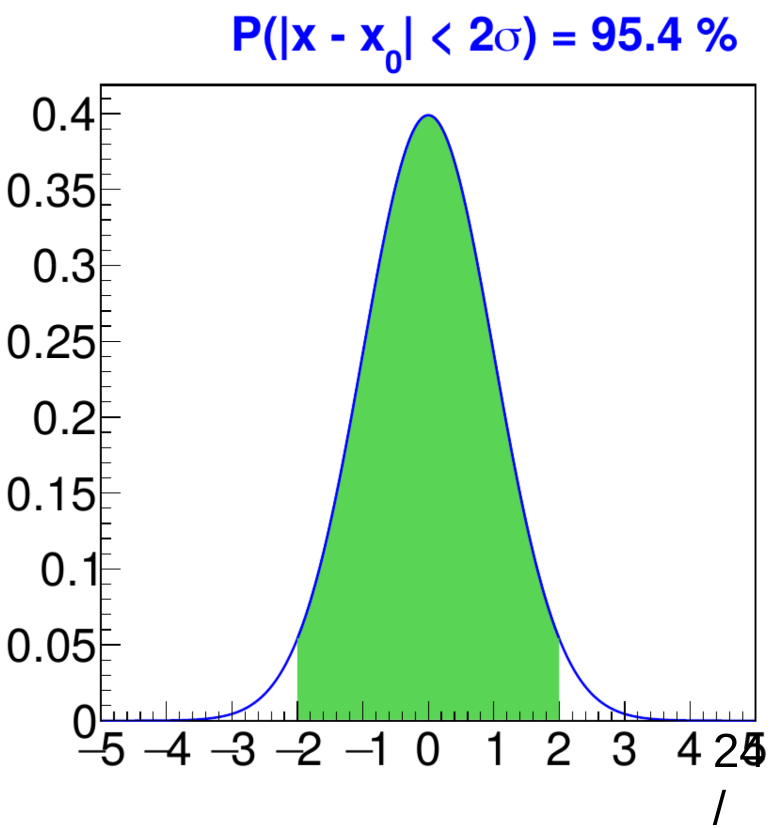
Gaussian Quantiles

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

G(x; X_0,σ) depends only on z ~ G(z; 0,1)

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

Z	P(|x – X_0| > Zσ)
1	**0.317**
2	**0.045**
3	**0.003**
4	**3 x 10^{-5}**
5	**6 x 10^{-7}**



Cumulative Distribution Function (CDF)

of the Gaussian :

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

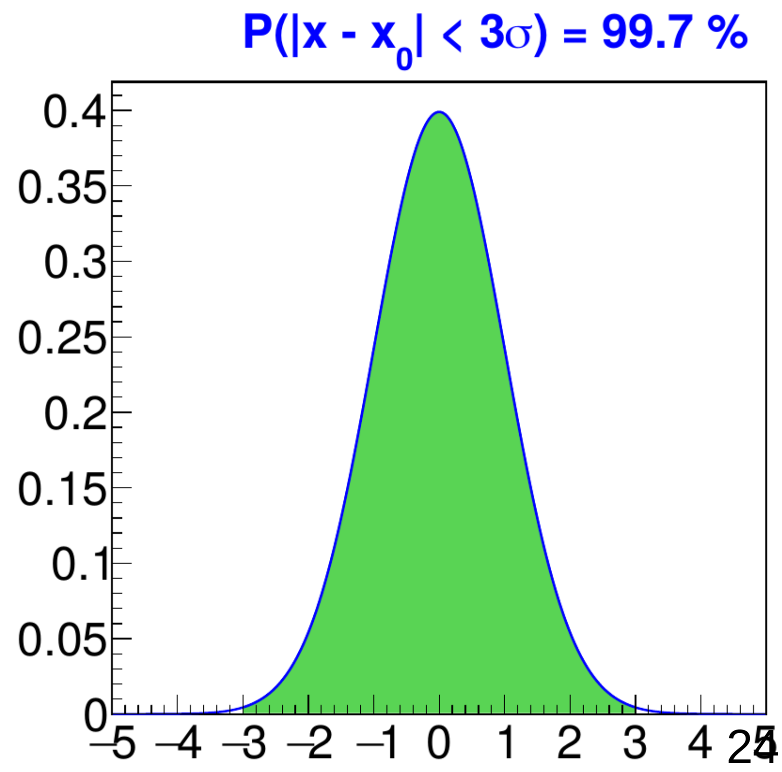
Gaussian Quantiles

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

G(x; X_0,σ) depends only on z ~ G(z; 0,1)

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

Z	P(|x – X_0| > Zσ)
1	**0.317**
2	**0.045**
3	**0.003**
4	**3 x 10^{-5}**
5	**6 x 10^{-7}**



Cumulative Distribution Function (CDF)

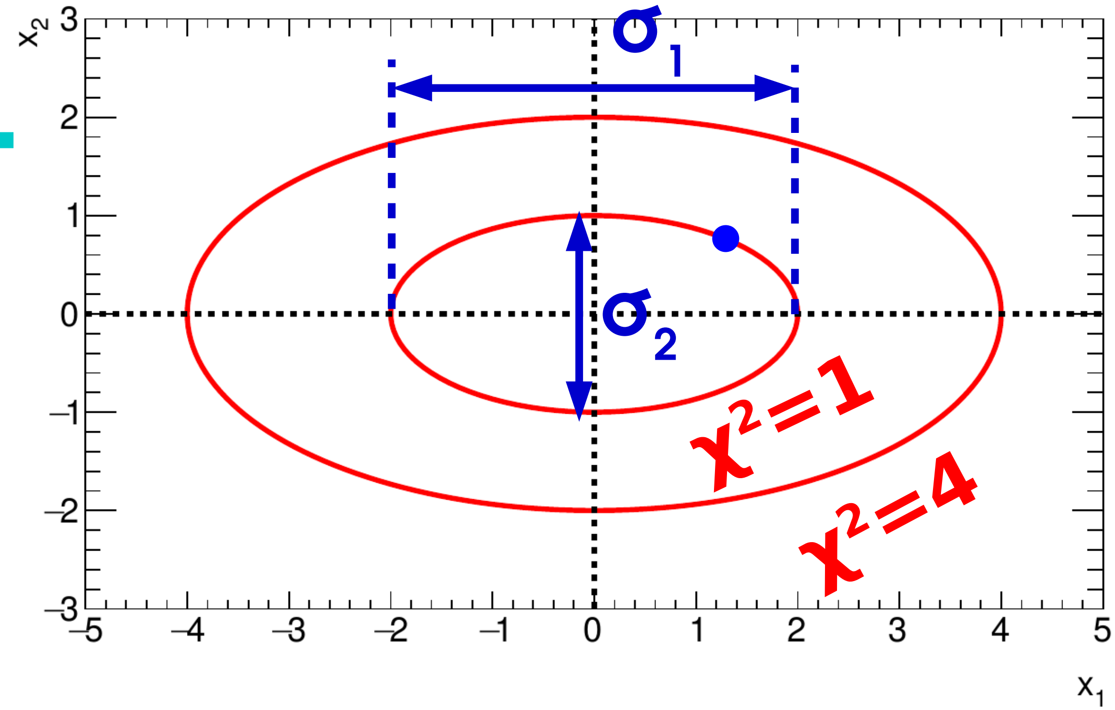
of the Gaussian :

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

Chi-squared

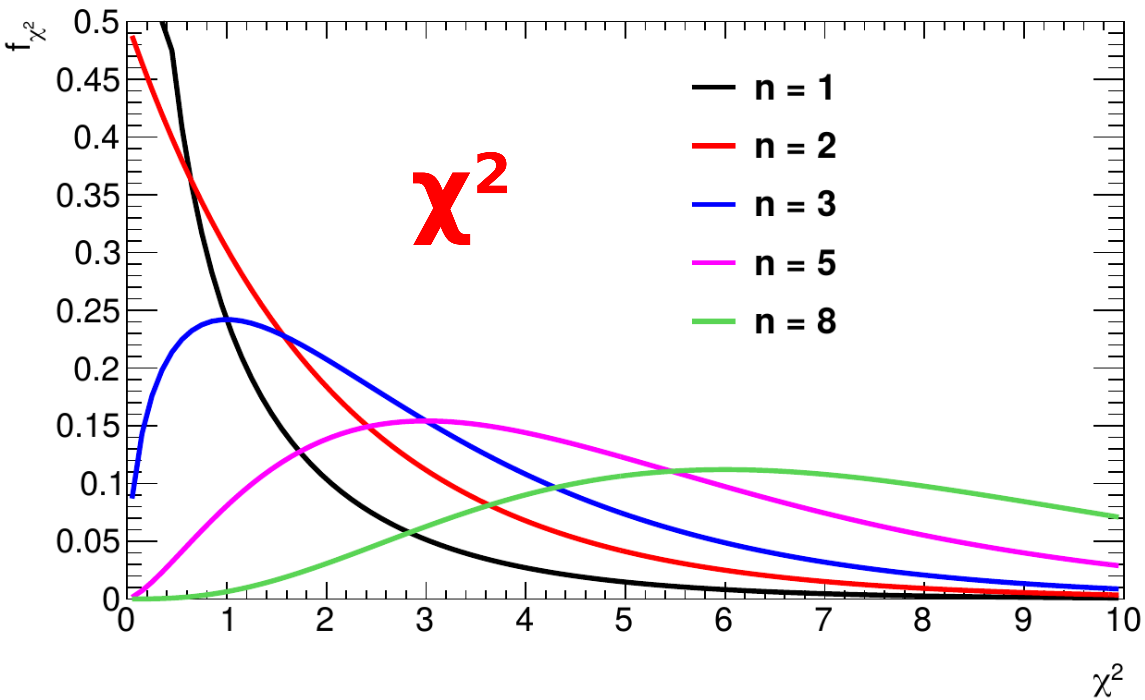
Multiple Independent Gaussian variables x_i: Define

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



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

Distribution depends on n :



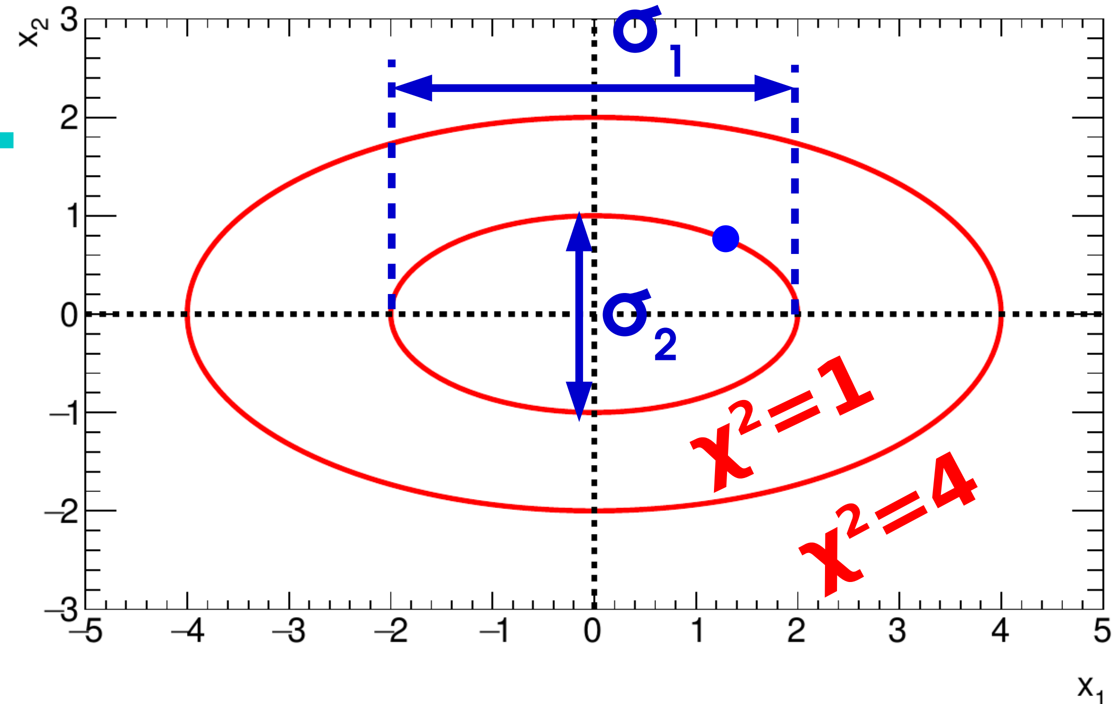
Rule of thumb:

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

Chi-squared

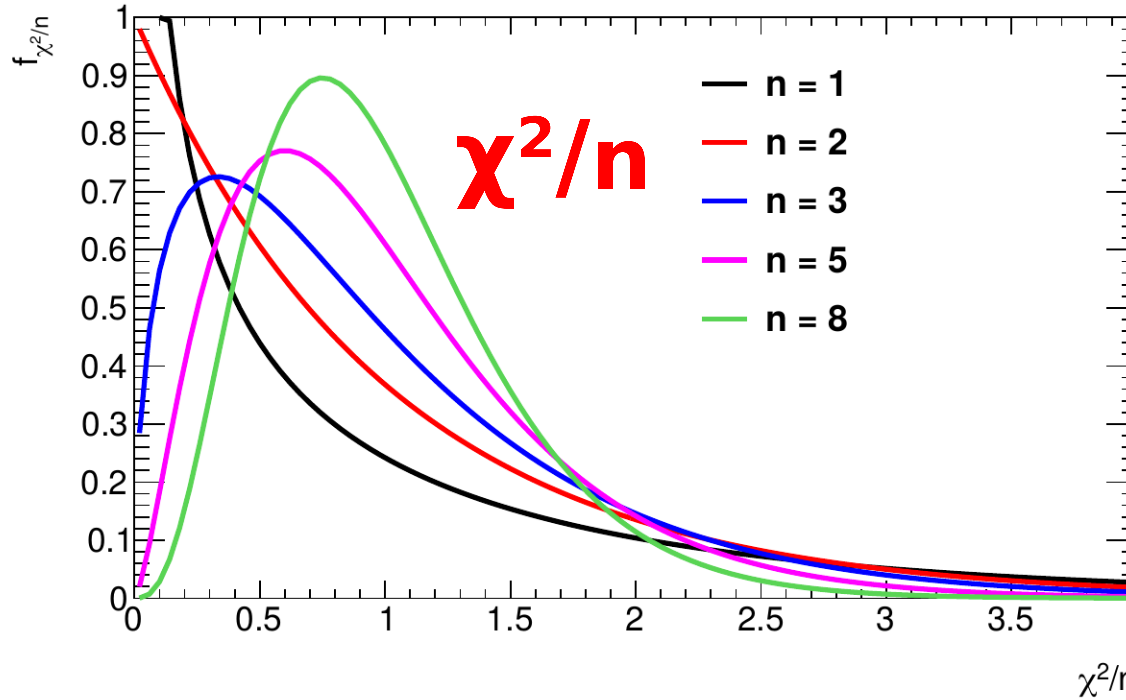
Multiple Independent Gaussian variables x_i : Define

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



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

Distribution depends on n :



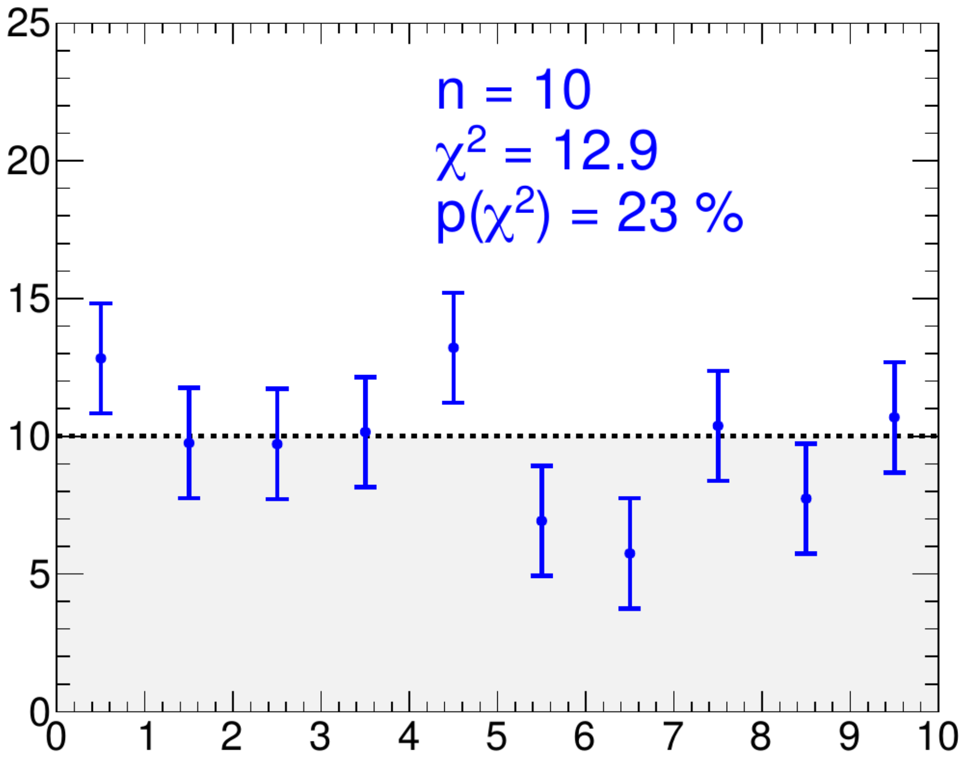
Rule of thumb:

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

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)



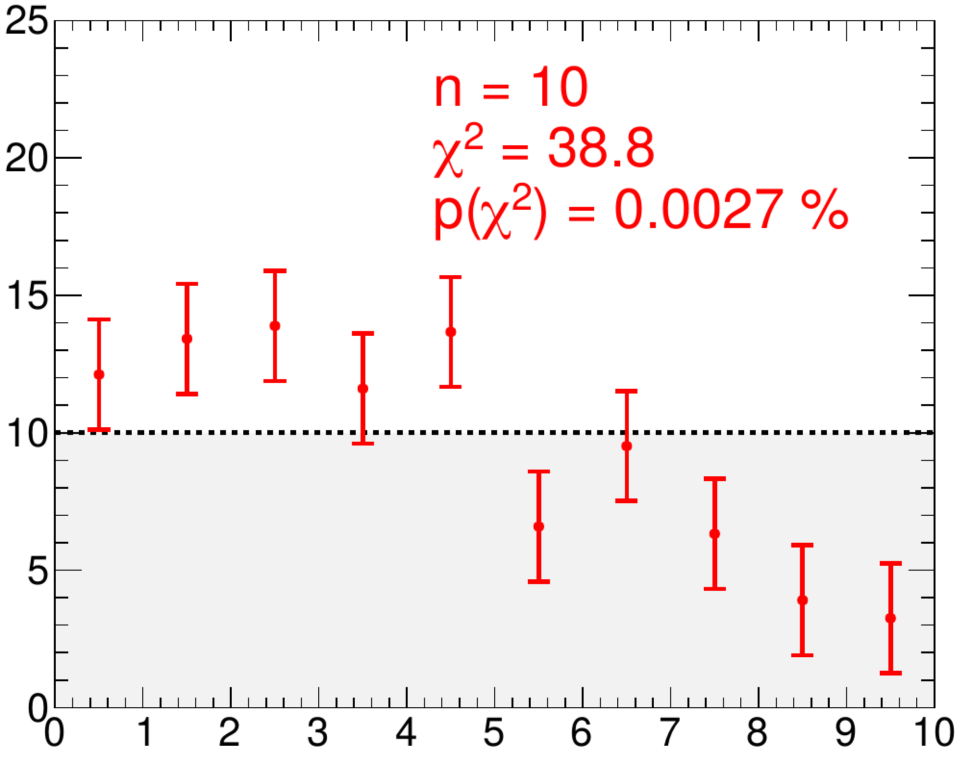
BLUE histogram vs. flat reference

χ^2 = 12.9,   p(χ^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

χ^2 = 12.9, p(χ^2=12.9, n=10) = **23%**

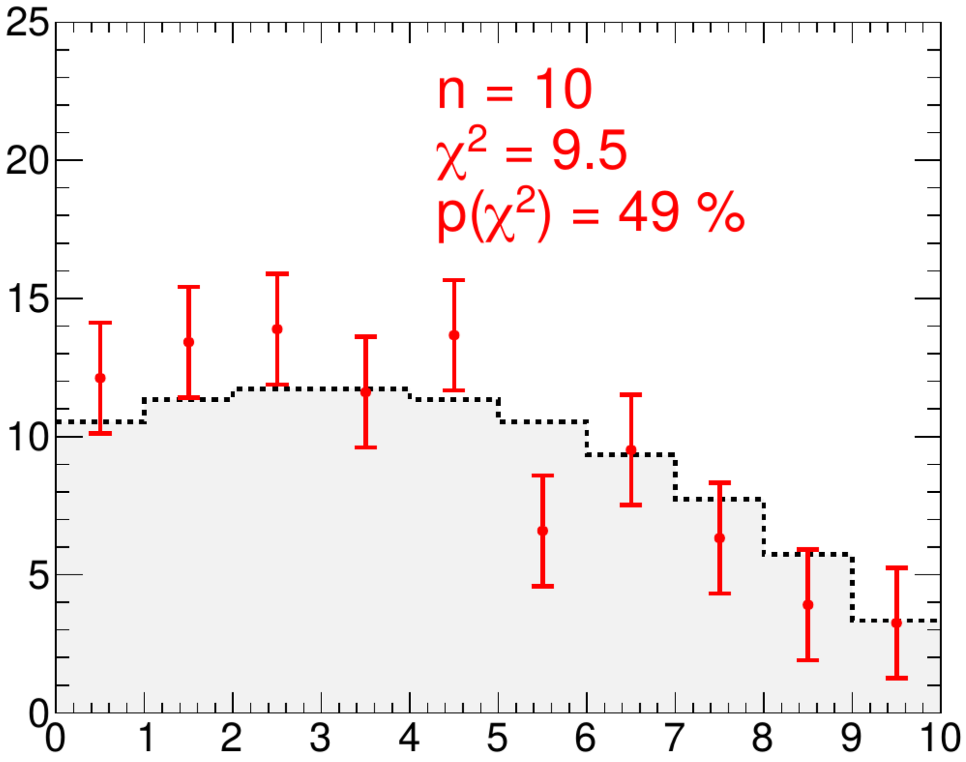
RED histogram vs. flat reference

χ^2 = 38.8, p(χ^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

χ^2 = 12.9,   p(χ^2=12.9, n=10) = **23%**

RED histogram vs. flat reference

χ^2 = 38.8, p(χ^2=38.8, n=10) = **0.003%** ✘

RED histogram vs. correct reference

χ2 = 9.5, p(χ2=9.5, n=10) = 49%

Statistical Modeling

Example 1: Z counting

Phys. Lett. B 759 (2016) 601

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

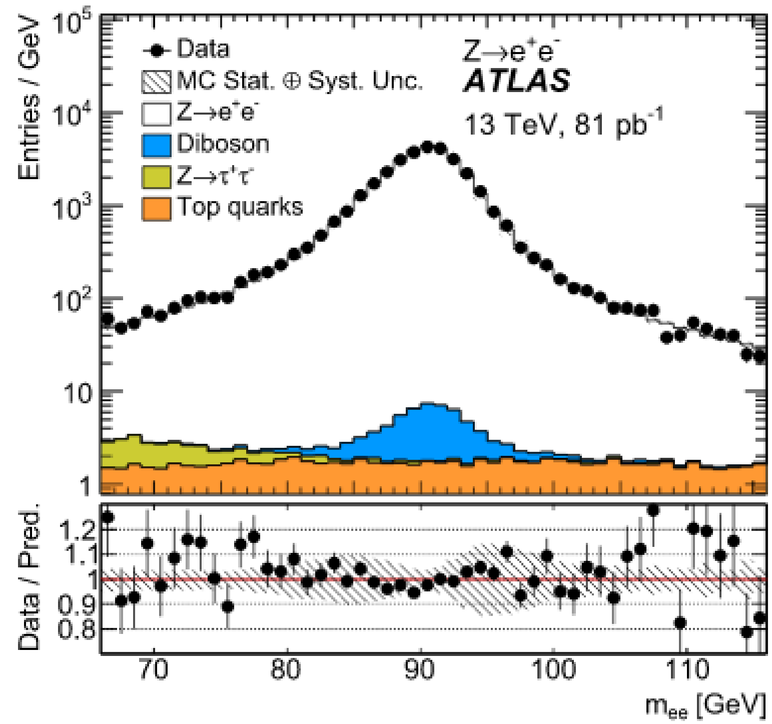
35000 ± 187

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

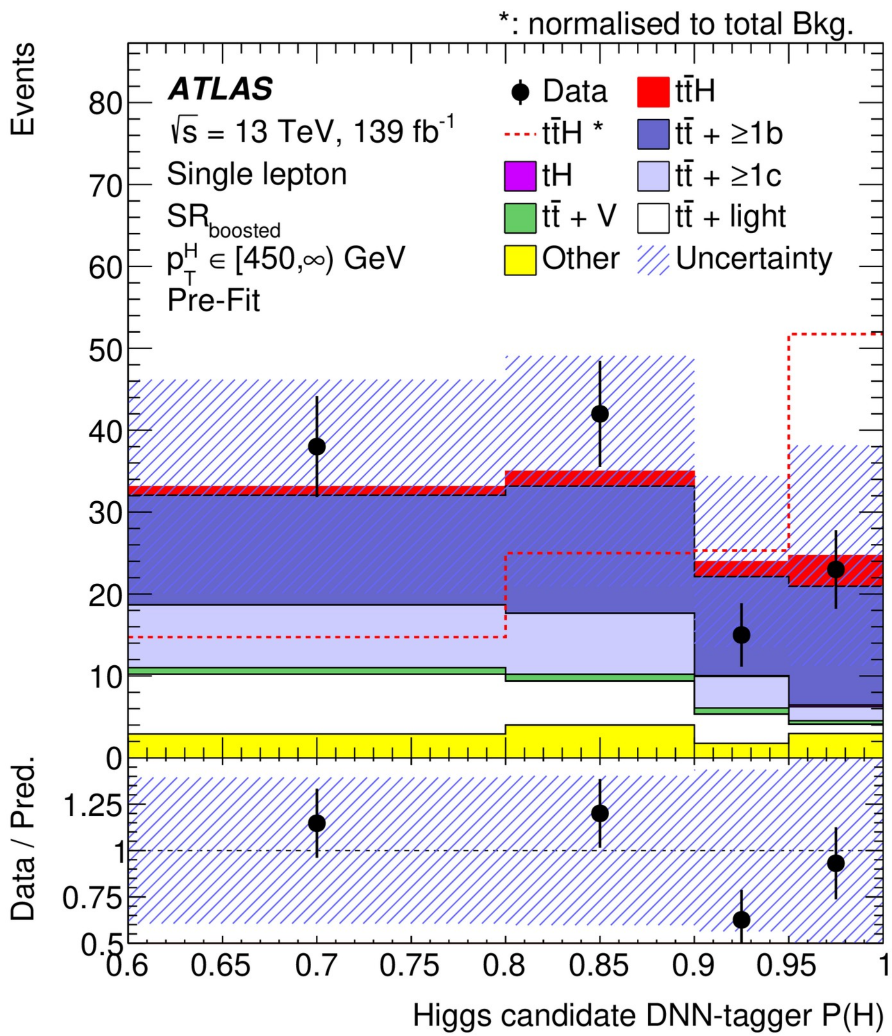
Fluctuations in the data counts

Other uncertainties (assumptions, parameter values)

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

Example 2: ttH→bb

arXiv:2111.06712



Event counting in different regions:

Multiple-bin counting

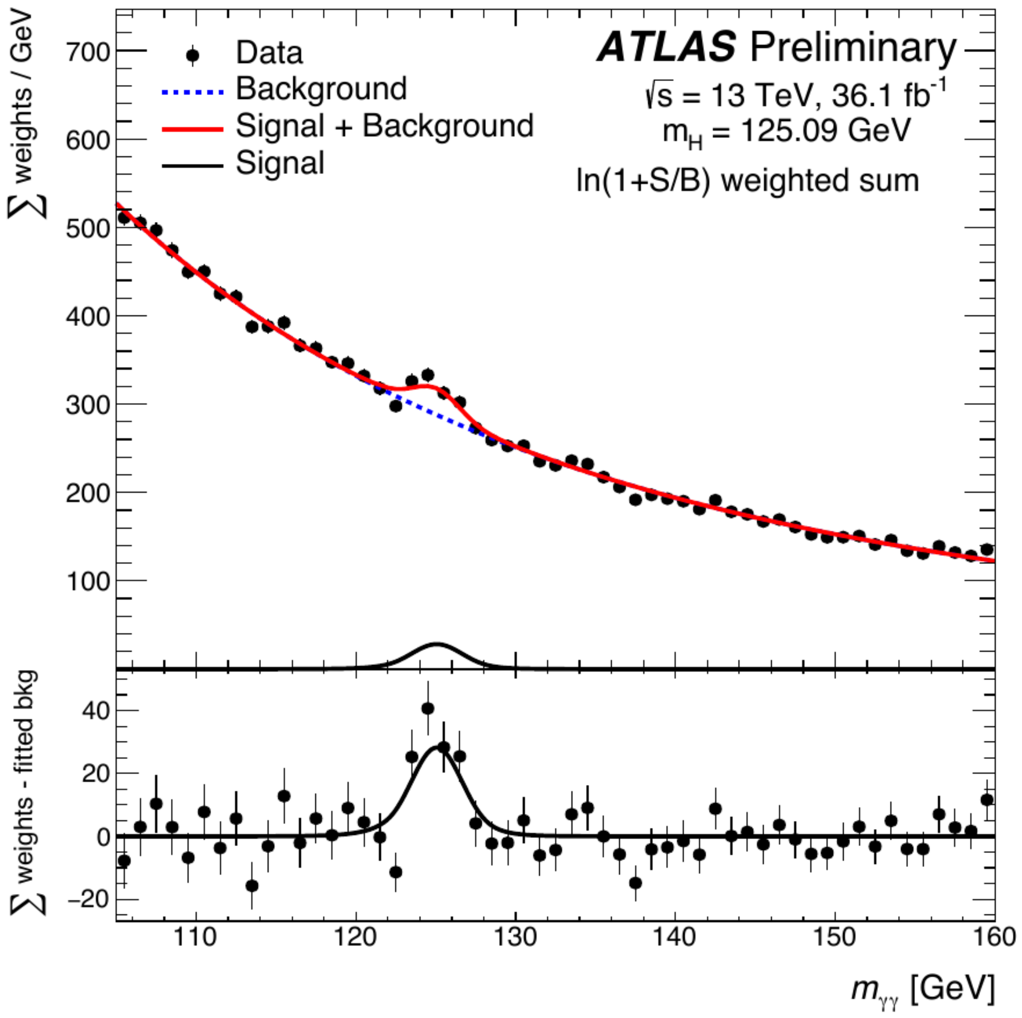
Lots of information available

→ Potentially higher sensitivity

→ How to make optimal use of it ?

Example 3: unbinned modeling

ATLAS-CONF-2017-045



All modeling done using continuous distributions:

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

How to count

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

→ In principle, binomial process

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

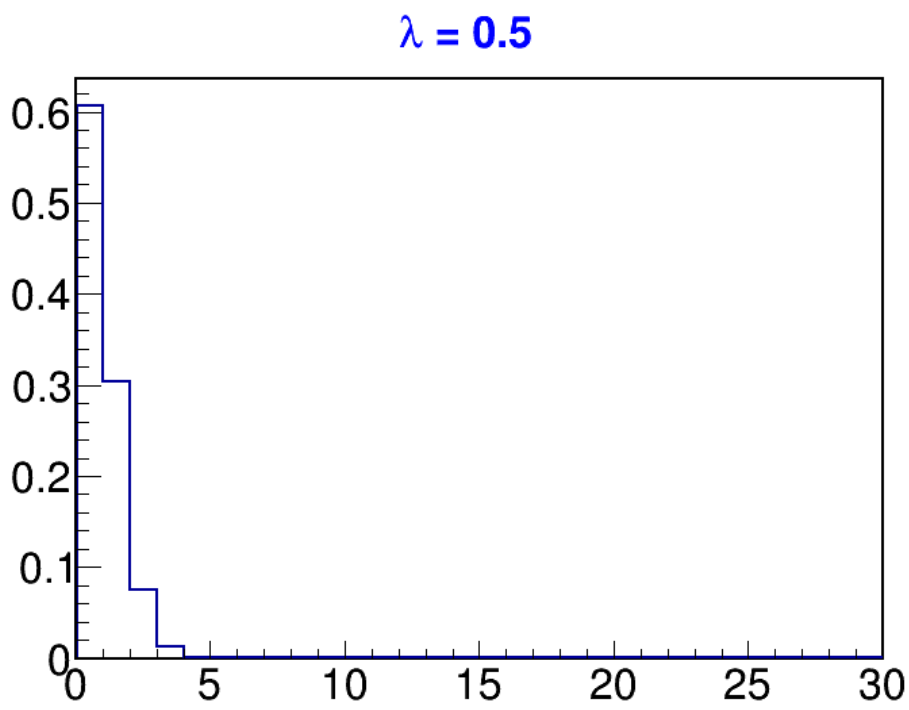
→ *i.e.* very rare process, but very many trials so still expect to see good events

Poisson distribution

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

λ = NP

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √N

Central limit theorem :

becomes Gaussian for large λ :

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

How to count

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

→ In principle, binomial process

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

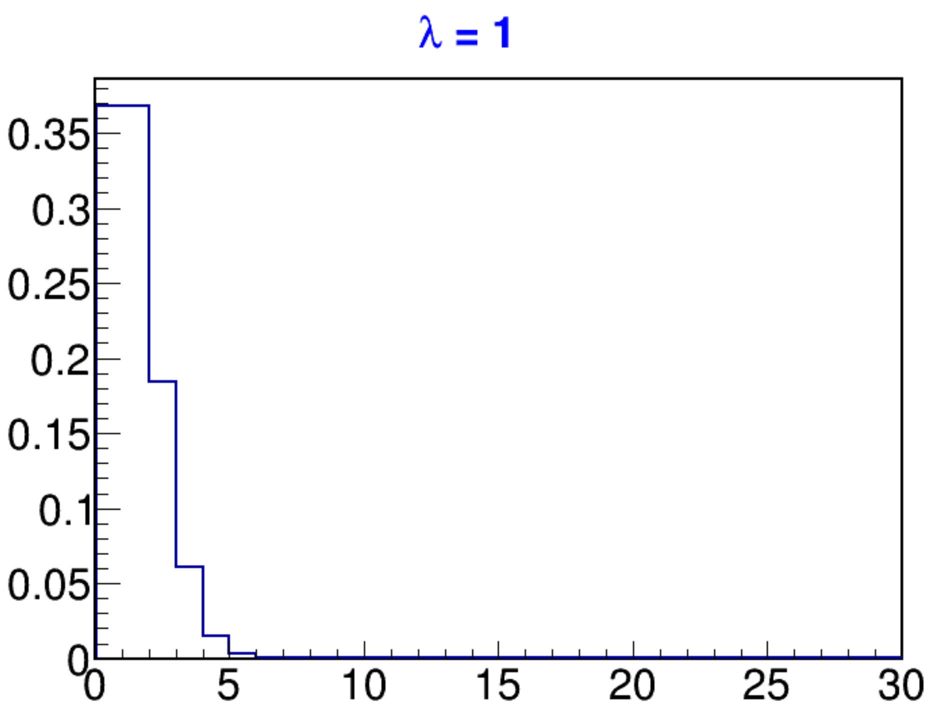
→ *i.e.* very rare process, but very many trials so still expect to see good events

Poisson distribution

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

λ = NP

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √N

Central limit theorem :

becomes Gaussian for large λ :

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

How to count

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

→ In principle, binomial process

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

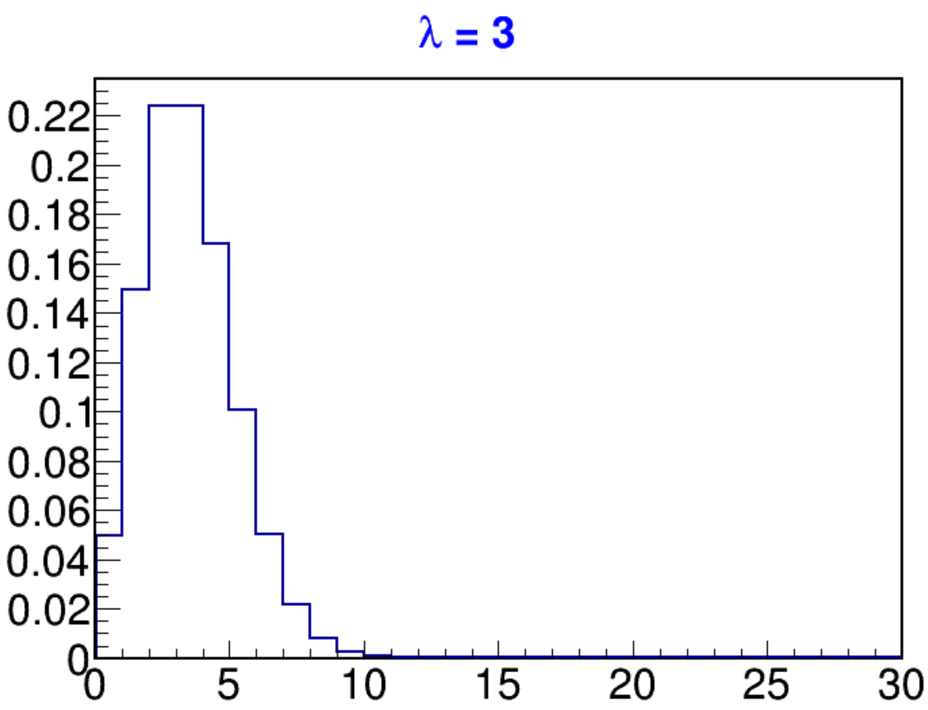
→ *i.e.* very rare process, but very many trials so still expect to see good events

Poisson distribution

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

λ = NP

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √N

Central limit theorem :

becomes Gaussian for large λ :

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

How to count

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

→ In principle, binomial process

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

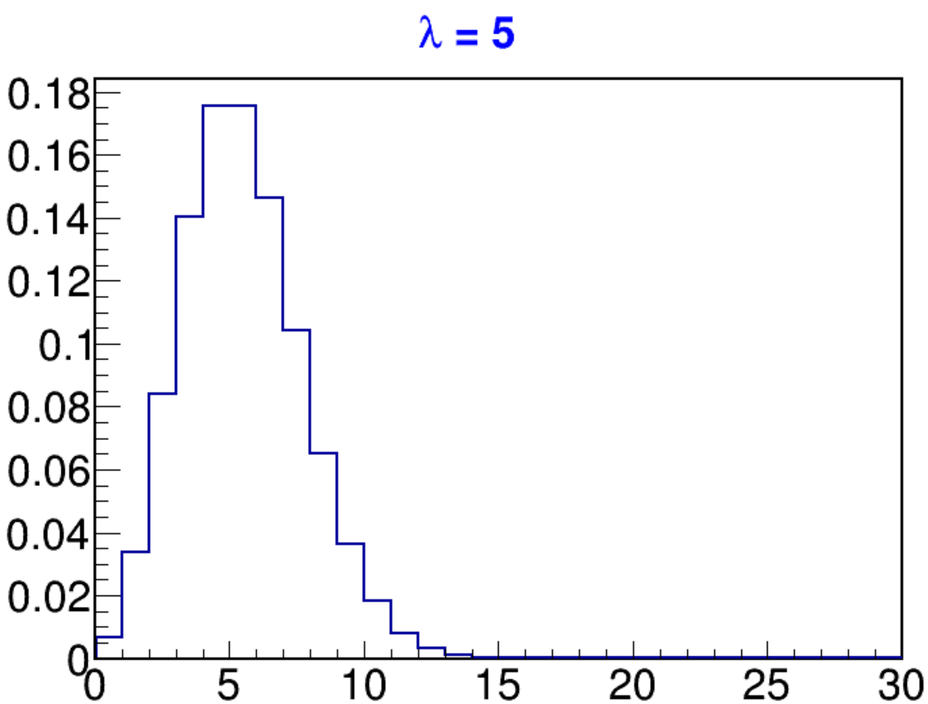
→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution

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

λ = NP

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √N

Central limit theorem :

becomes **Gaussian for large λ** :

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

How to count

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

→ In principle, binomial process

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

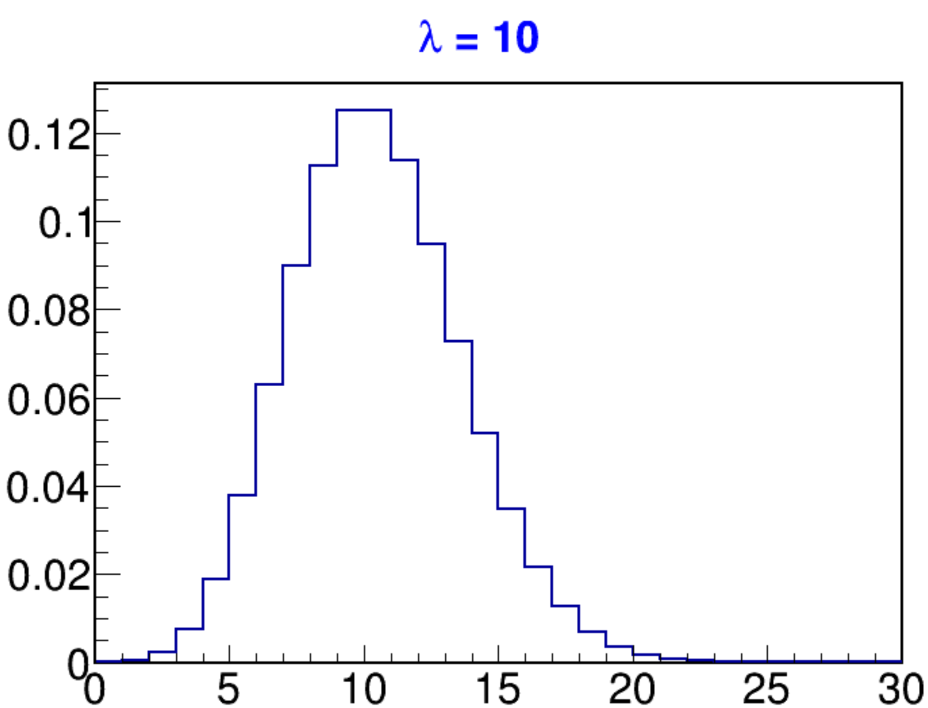
→ *i.e.* very rare process, but very many trials so still expect to see good events

Poisson distribution

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

λ = NP

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √N

Central limit theorem :

becomes Gaussian for large λ :

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

How to count

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

→ In principle, binomial process

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

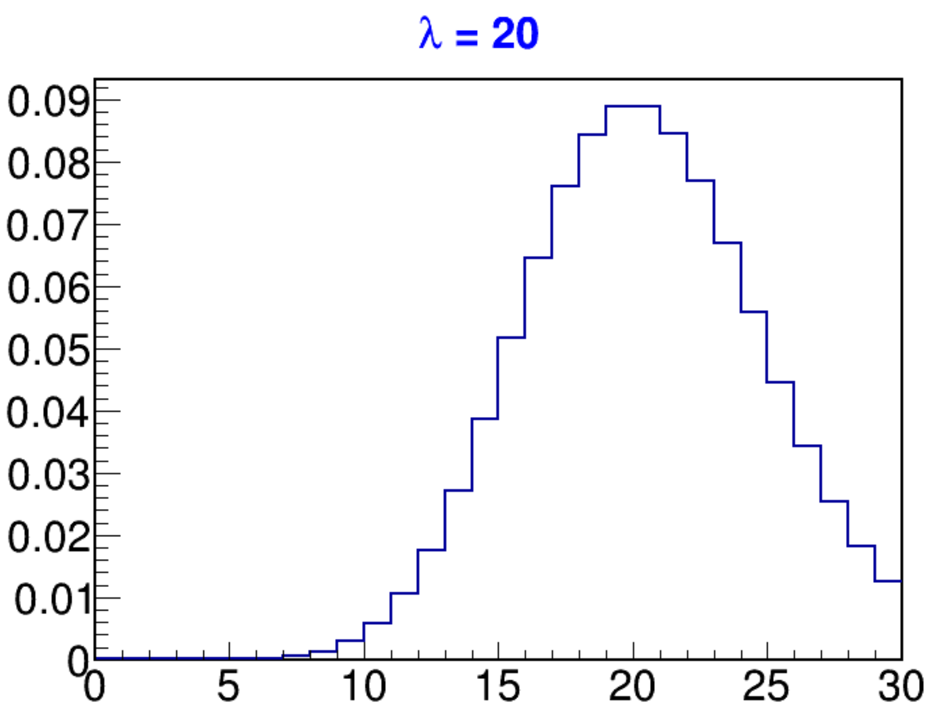
→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution

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

$\lambda = NP$

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



Mean = λ

Variance = λ

σ = √λ

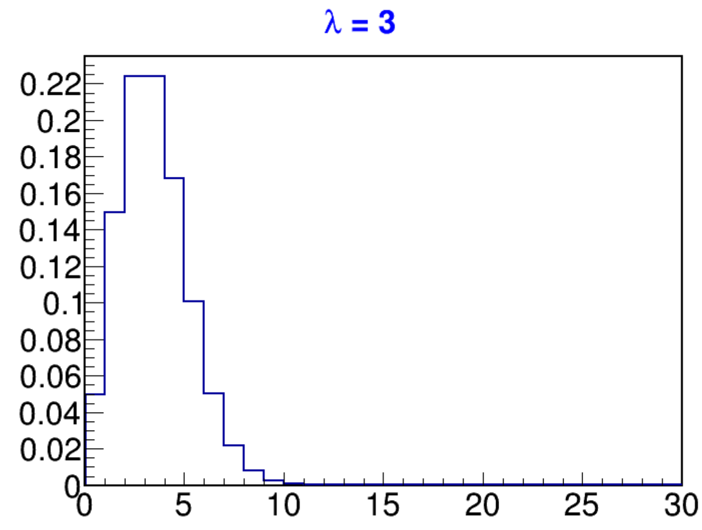
For a counting measurement, RMS = √N

Central limit theorem :

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.**

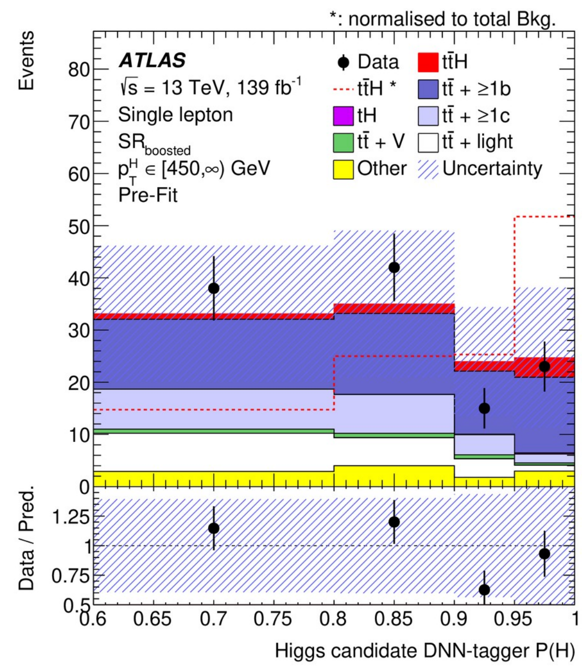
Multiple counting bins

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

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

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

Poisson distribution in each bin



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

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

Also not always possible to generate sufficiently large MC samples
MC stat fluctuations can create artefacts, especially for $S \ll B$.

Model Parameters

Model typically includes:

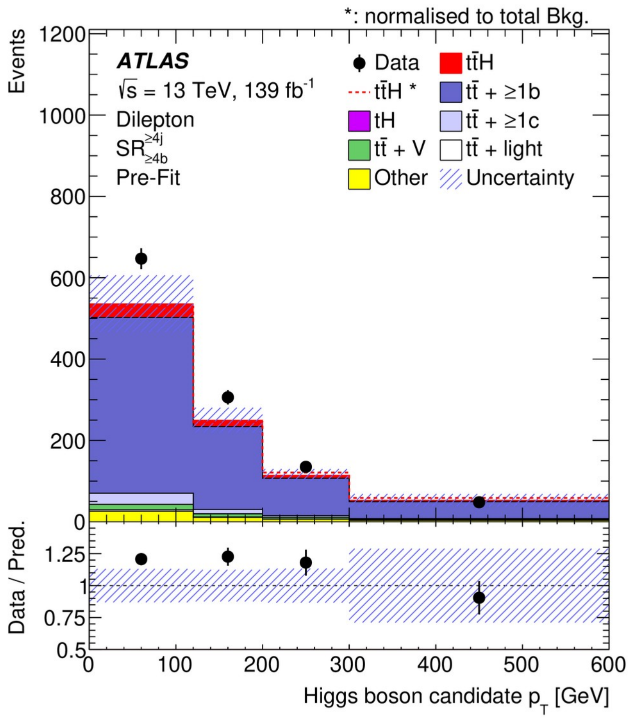
- **Parameters of interest** (POIs) : what we want to measure

  → **S**, $\mathbf{m_W}$, ...

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

  → Background levels (**B**)

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

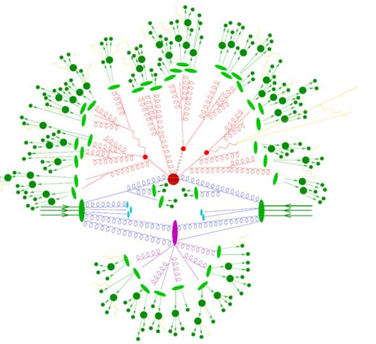


NPs must be either:

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

→ **Constrained by the data**

Takeaways



Random data must be described using a statistical model:

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

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

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

Maximum Likelihood 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

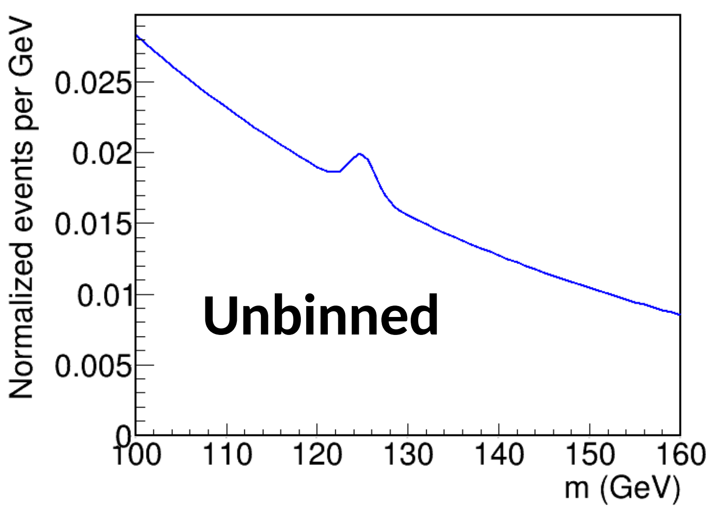
Can draw random events according to PDF : **generate *pseudo-data***

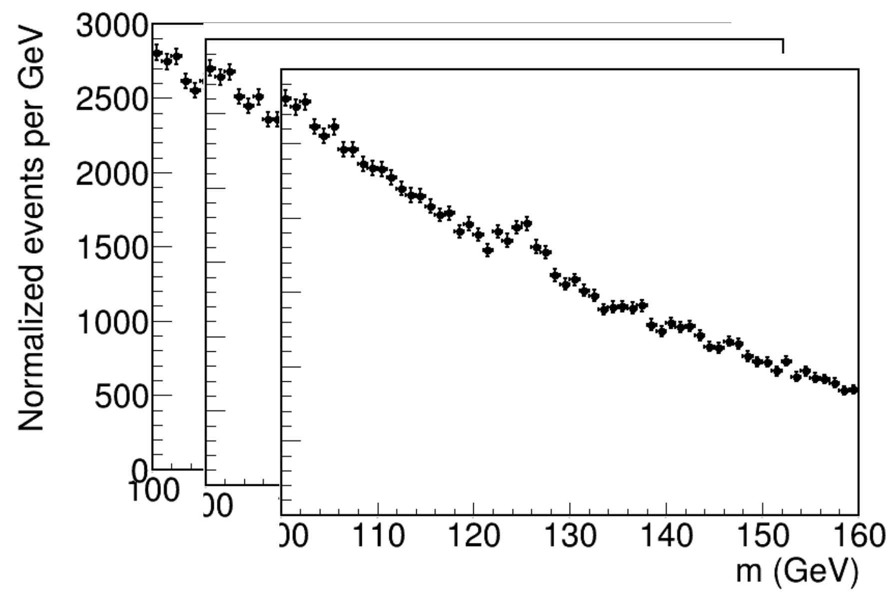
$$P(\lambda = 5)$$

2, 5, 3, 7, 4, 9, ….

Each entry = separate “experiment”

Generate





What a PDF is also for: Likelihood

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

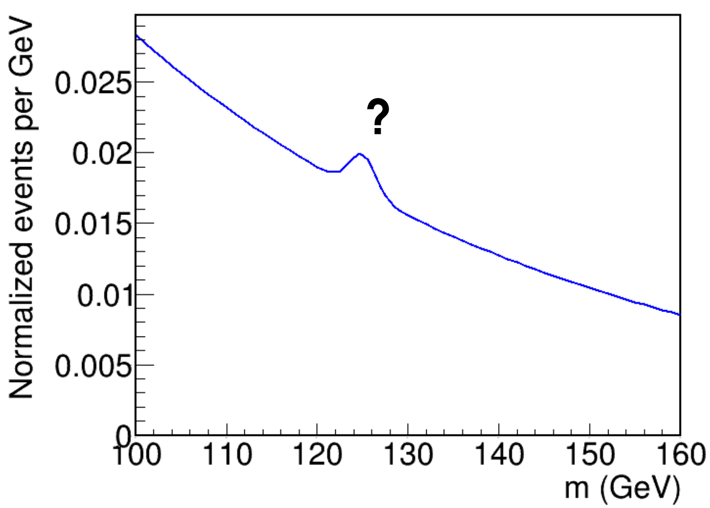
⇒ Possible outcomes of the experiment, for given parameter values

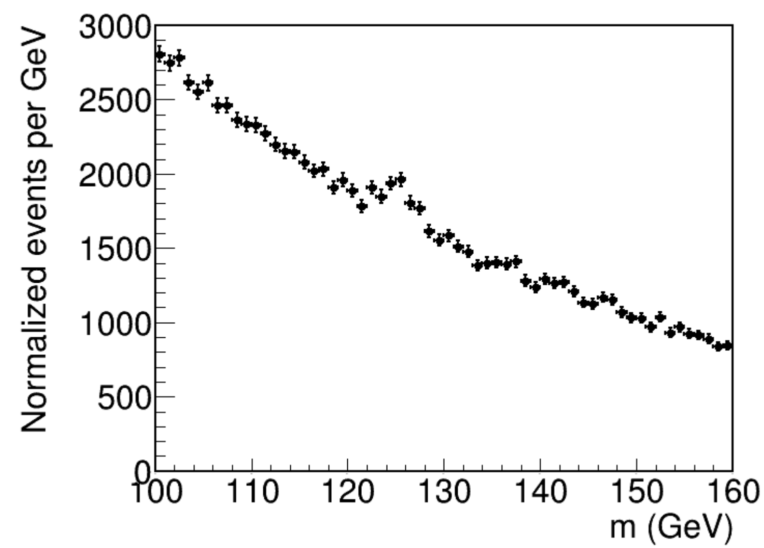
We want the other direction: **use data to get information on parameters**

$$P(\lambda = ?)$$

2

Estimate





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

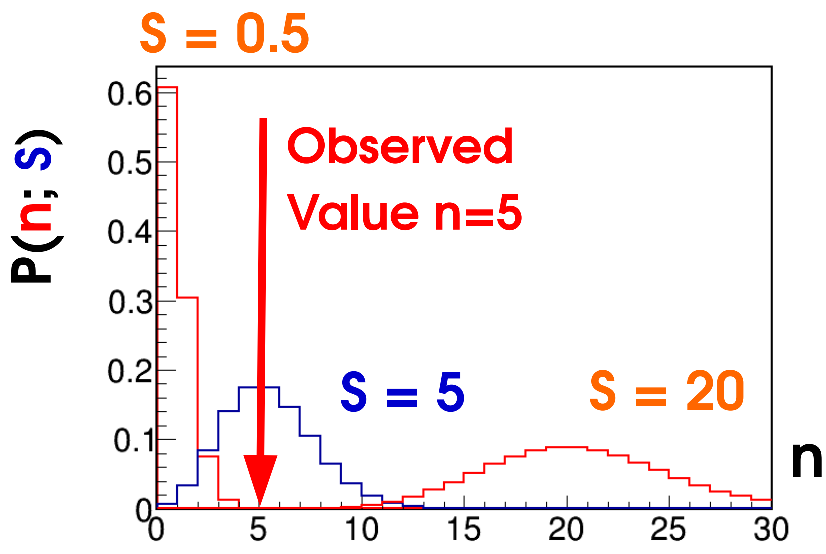
→ **same as the PDF,** but seen as function of the parameters

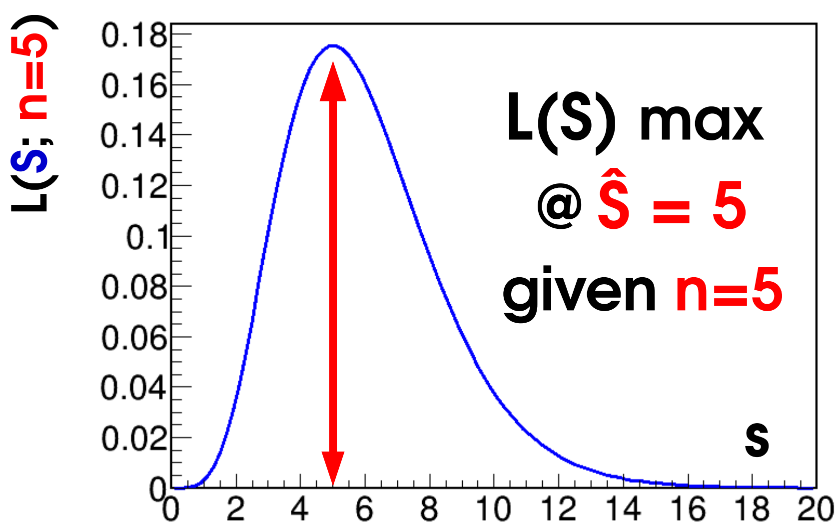
Maximum Likelihood Estimation

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

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

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



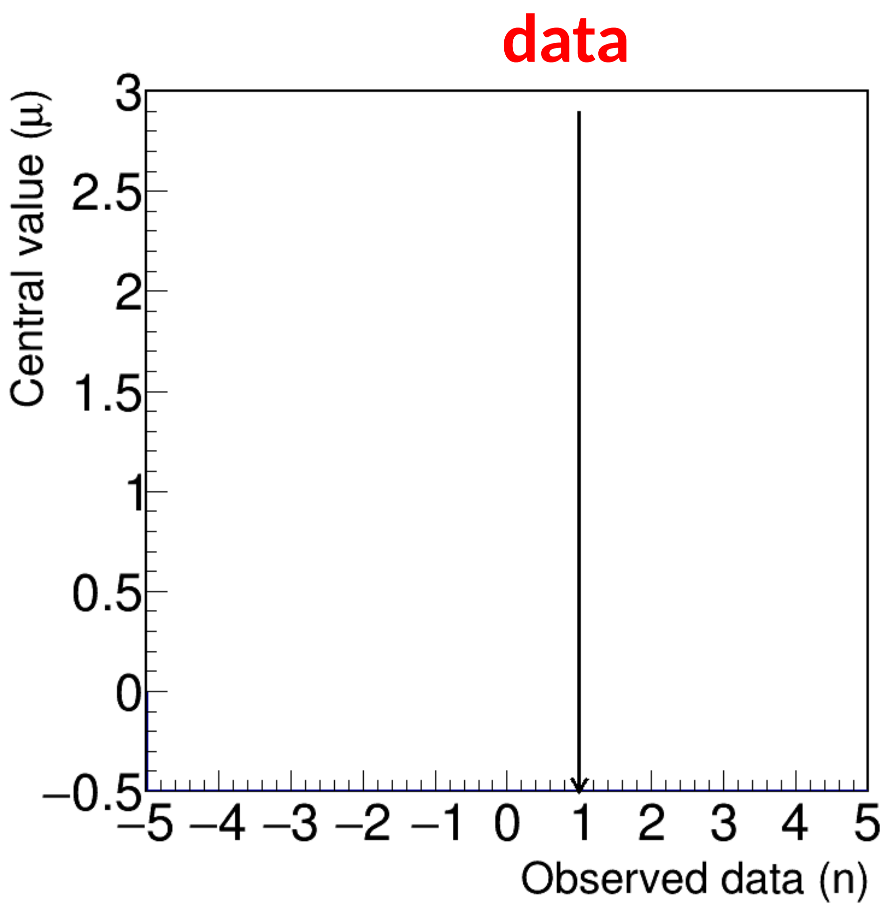


MLE: the value of μ for which this data was ***most likely to occur***

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

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

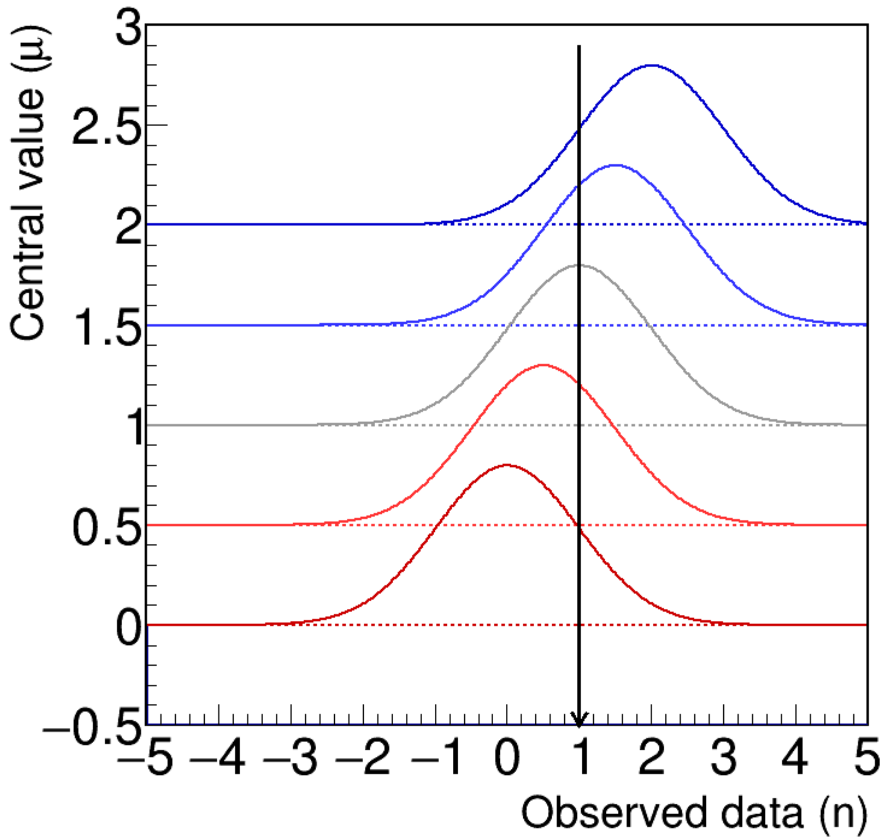
Gaussian case



Best-fit of Gaussian PDF mean to observed data

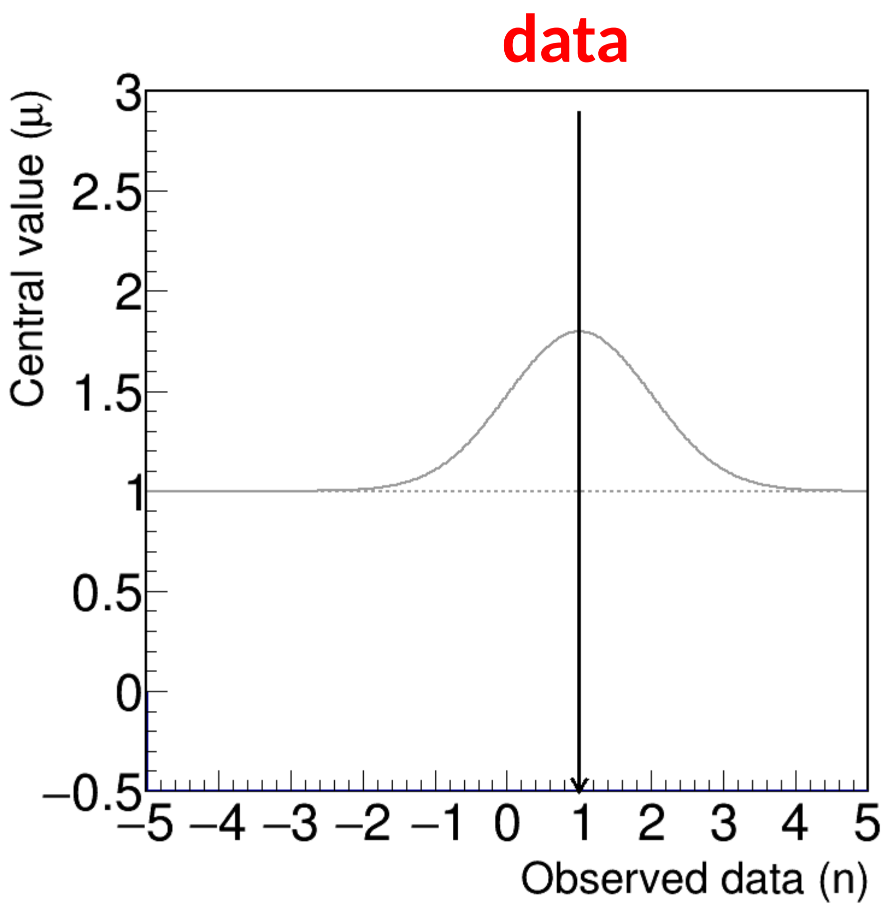
Gaussian case

data



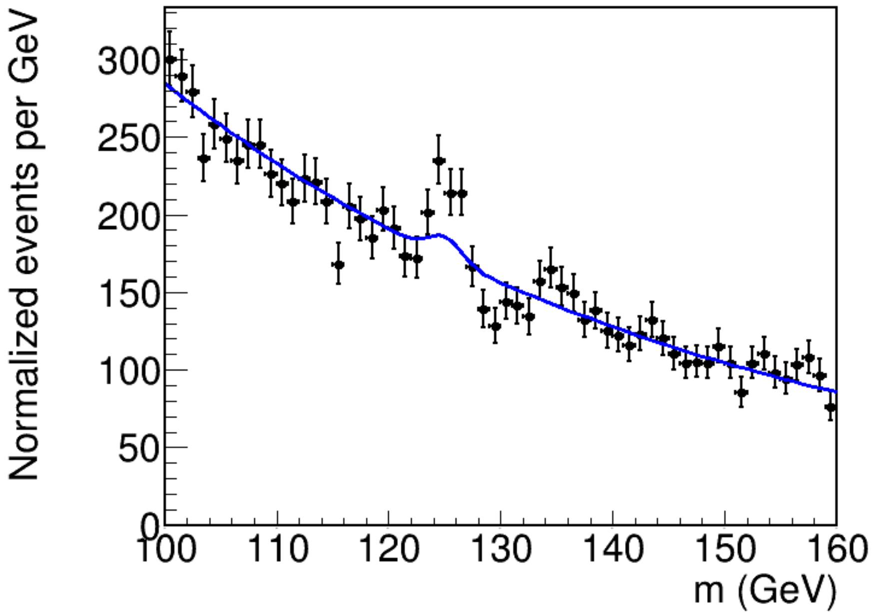
Best-fit of Gaussian PDF mean to observed data

Gaussian case



Best-fit of Gaussian PDF mean to observed data

Multiple Gaussian bins



-2 log Likelihood:

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

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

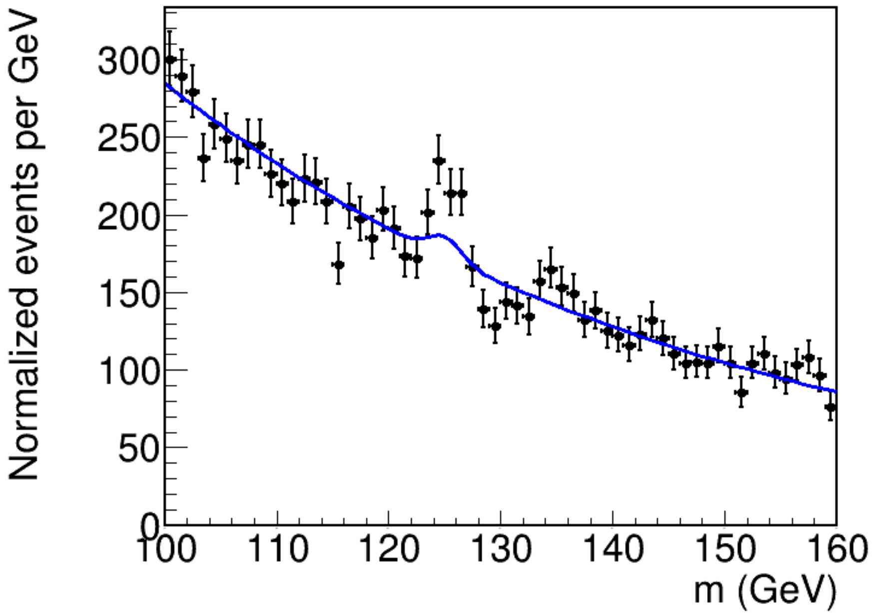
However typically need to perform non-linear minimization in other cases.

HEP practice:

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

  → Many algorithms – gradient-based, etc.

Multiple Gaussian bins



-2 log Likelihood:

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

Maximum likelihood ⇔ Minimum χ^2

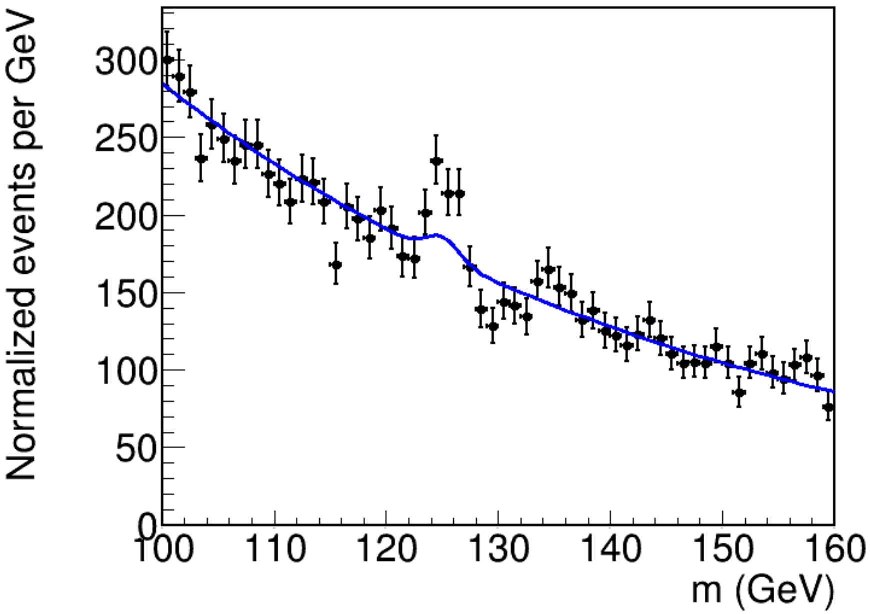
⇔ Least-squares minimization

However typically need to perform non-linear minimization in other cases.

HEP practice:

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

Multiple Gaussian bins



-2 log Likelihood:

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

Maximum likelihood ⇔ Minimum χ^2

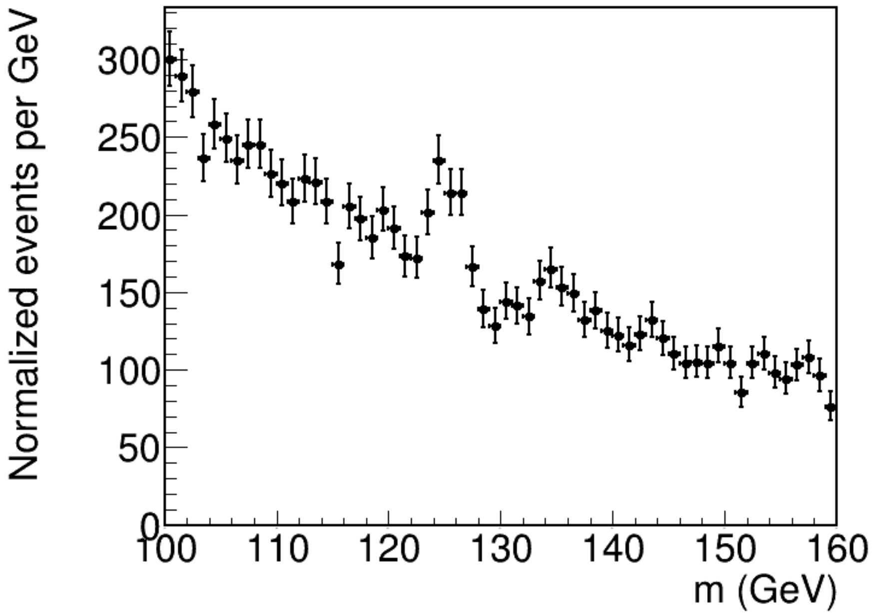
⇔ Least-squares minimization

However typically need to perform non-linear minimization in other cases.

HEP practice:

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

Multiple Gaussian bins



-2 log Likelihood:

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

Maximum likelihood ⇔ Minimum χ^2

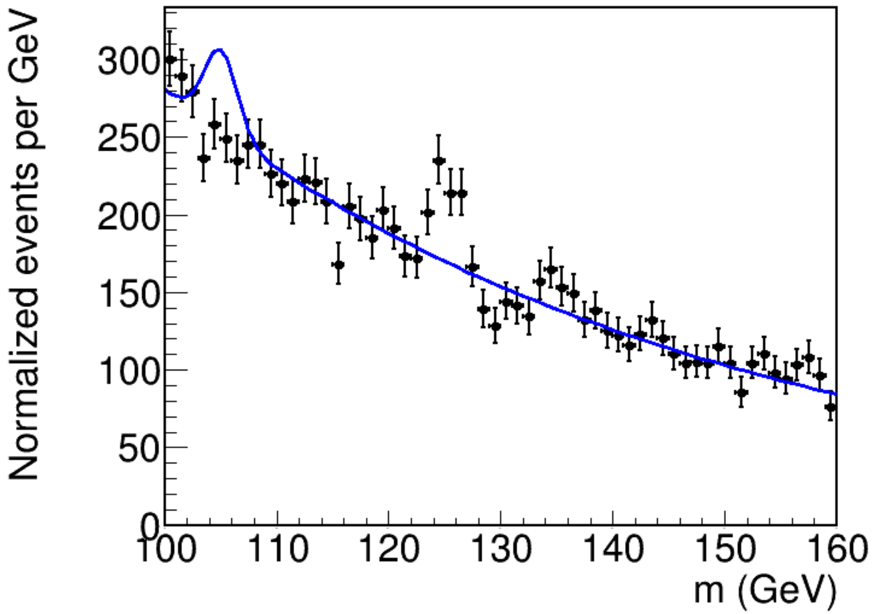
⇔ Least-squares minimization

However typically need to perform non-linear minimization in other cases.

HEP practice:

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

Multiple Gaussian bins



-2 log Likelihood:

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

Maximum likelihood ⇔ Minimum χ^2

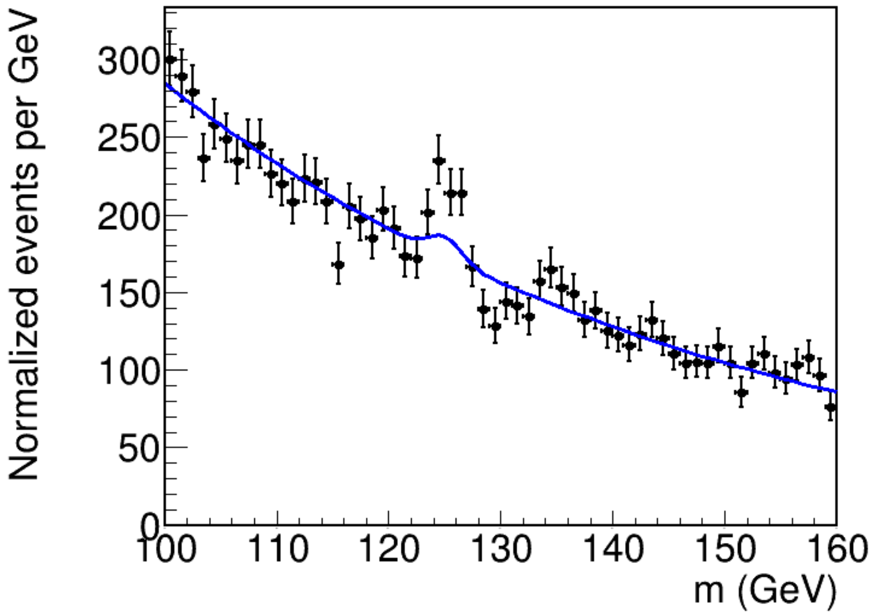
⇔ Least-squares minimization

However typically need to perform non-linear minimization in other cases.

HEP practice:

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

Multiple Gaussian bins



-2 log Likelihood:

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

Maximum likelihood ⇔ Minimum χ^2

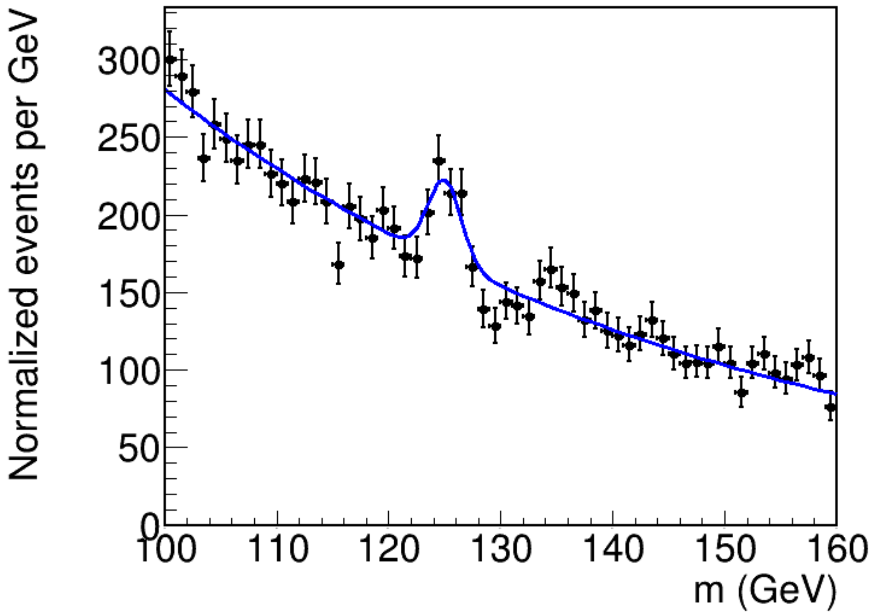
⇔ Least-squares minimization

However typically need to perform non-linear minimization in other cases.

HEP practice:

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

Multiple Gaussian bins



-2 log Likelihood:

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

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

However typically need to perform non-linear minimization in other cases.

HEP practice:

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

  → Many algorithms – gradient-based, etc.

Hands-ons

Each lecture statistics lecture comes with “hands-on” exercises.

The hands-on session will be based on **Jupyter notebooks** built using the **`numpy/scipy/pyplot`** stack.

If you have a computer, **please install** anaconda before the start of the class.

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.

No hands-on today, but have a look after the course.

Please be prepared to run the hands-ons during lectures 2 and 3 !

Links to resources

The hands-on resources for each lecture are listed below:

Lecture 1	notebook [solutions]	binder [solutions]	**Today**
Lecture 2	notebook	binder	
Lecture 3	notebook	binder	

- **Use the notebook links if you have a local install**: save the notebook locally and open it with your JupyterLab installation.
- **Use the binder links to use public servers**: the links will open the notebooks in a remote server sessions in your browser.

Notebooks with solutions to the exercises will be posted after the lectures.

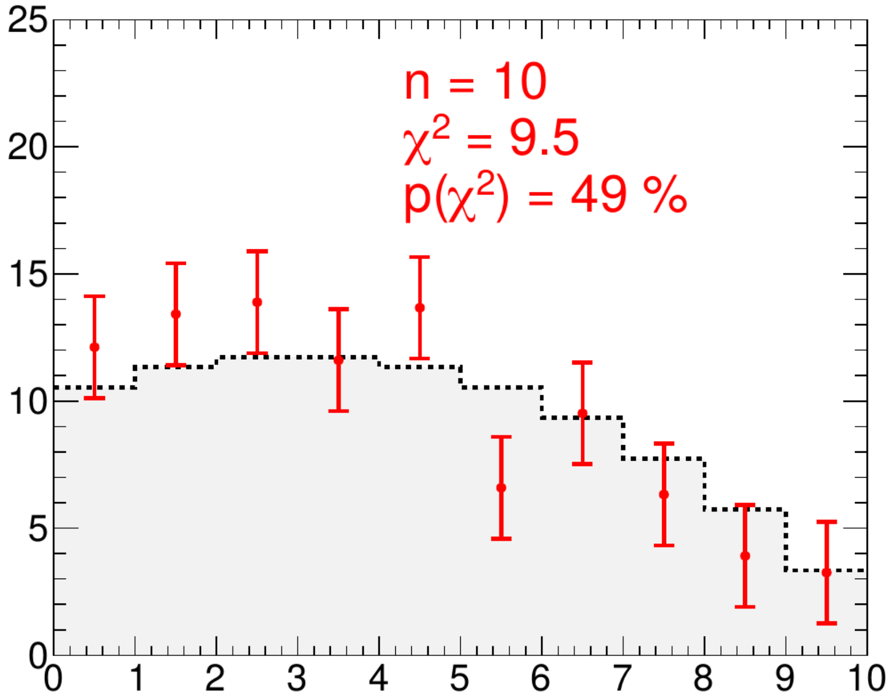
Please let me know in case of technical issues running the notebooks!

Extra Slides

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 :***

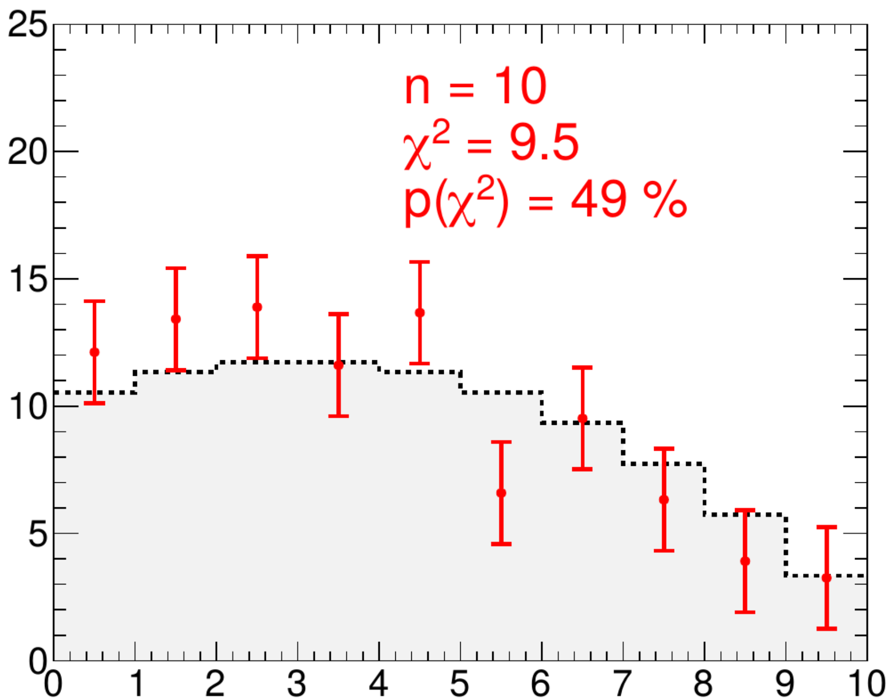
→ ***Bin central value*** *~ mean of the bin PDF*

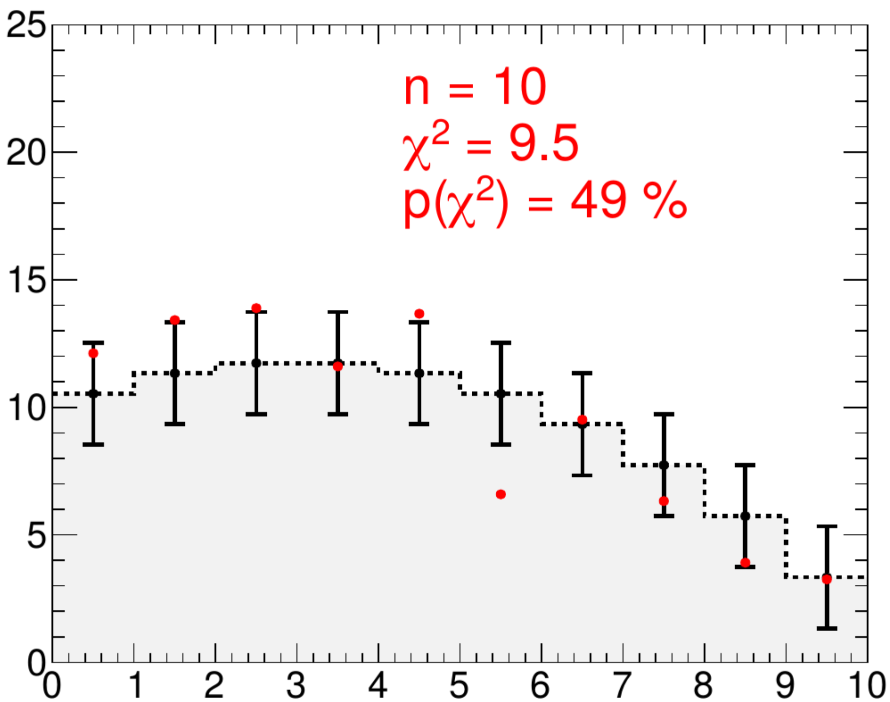
→ ***Bin uncertainty*** *~ RMS of the bin PDF*

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**.





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,)

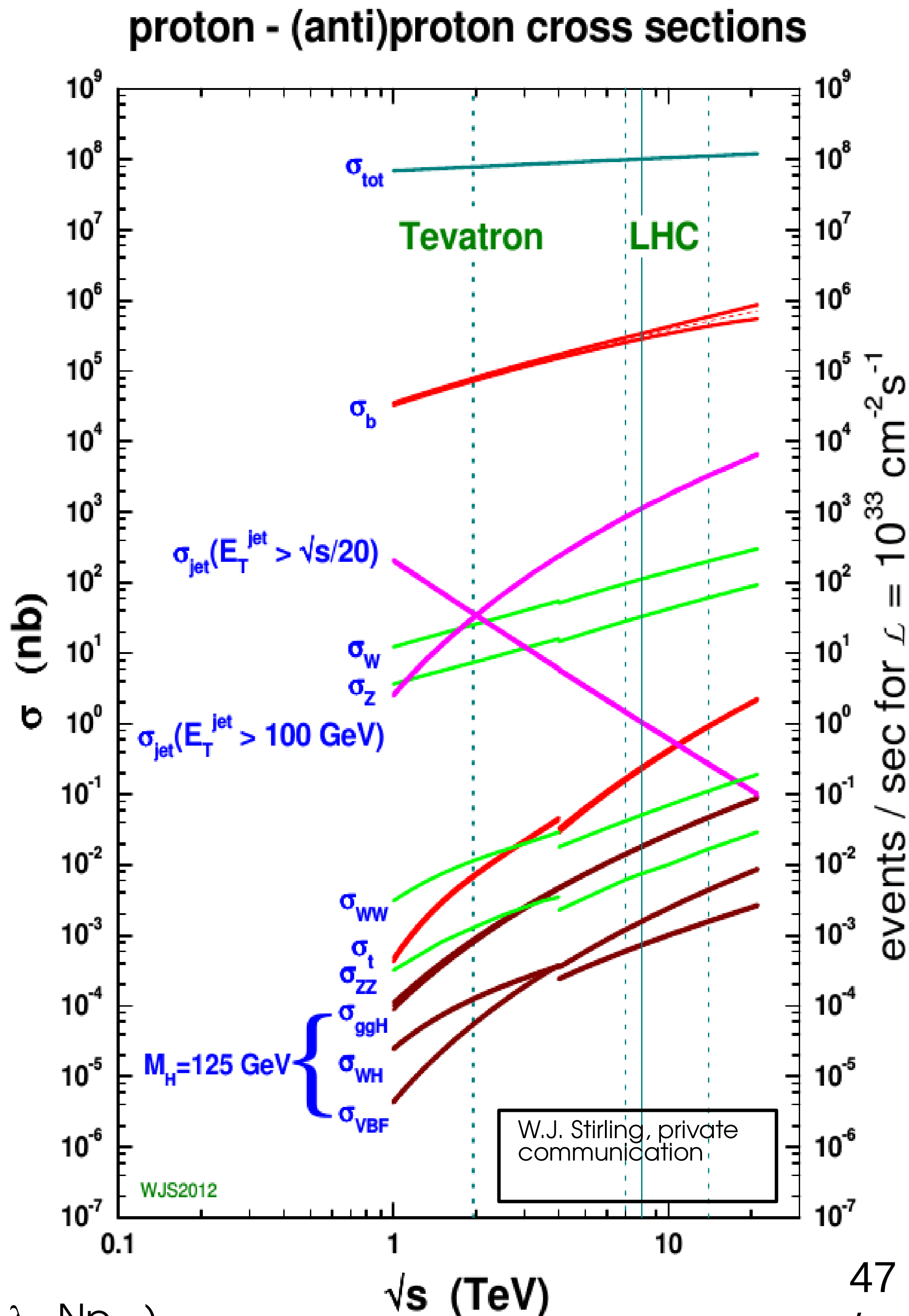
- **Trigger rate ~ 1 kHz**

  (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.



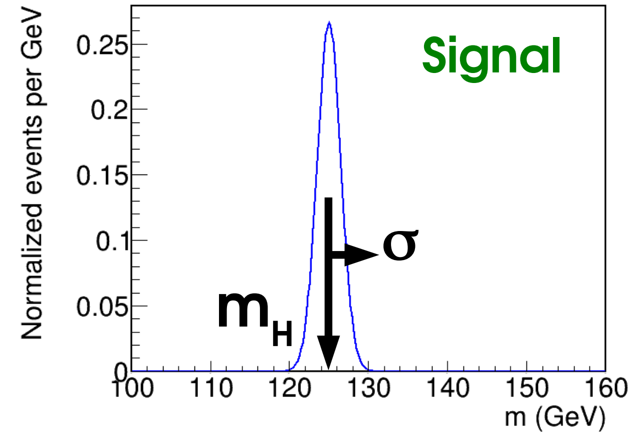
(Large N = design requirement, to get not-too-small λ=Np...)

Unbinned Shape Analysis

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

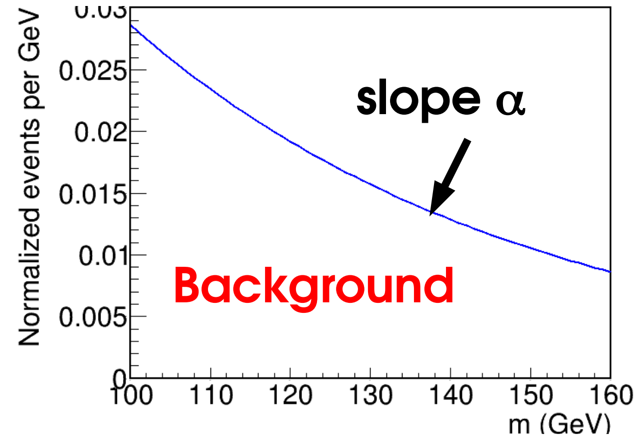
→ Describe shape of the **distribution of m**

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



H→γγ-inspired example:

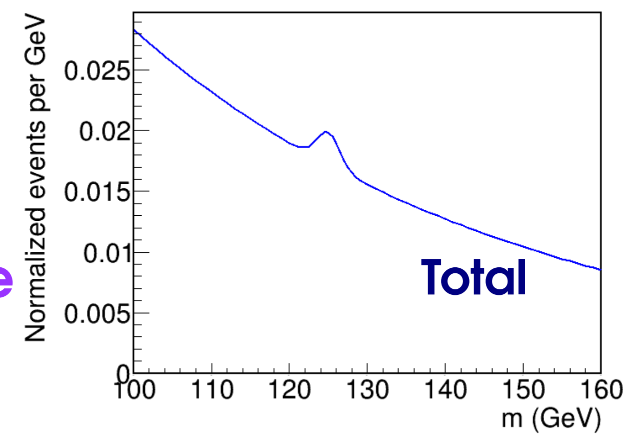
- Gaussian signal $\quad P_{\text{signal}}(m) = G(m; m_H, \sigma)$
- Exponential bkg $\quad P_{\text{bkg}}(m) = \alpha\, e^{-\alpha m}$



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

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

Poisson Example

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

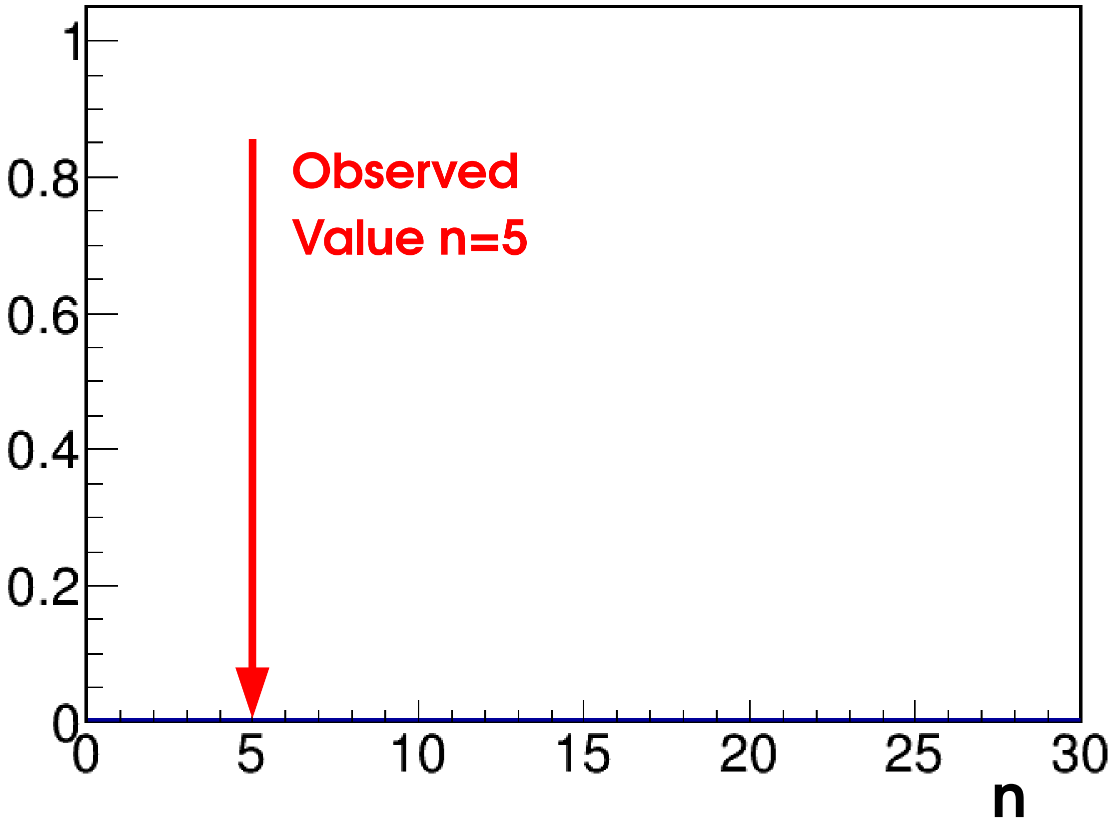
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

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



Poisson Example

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

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

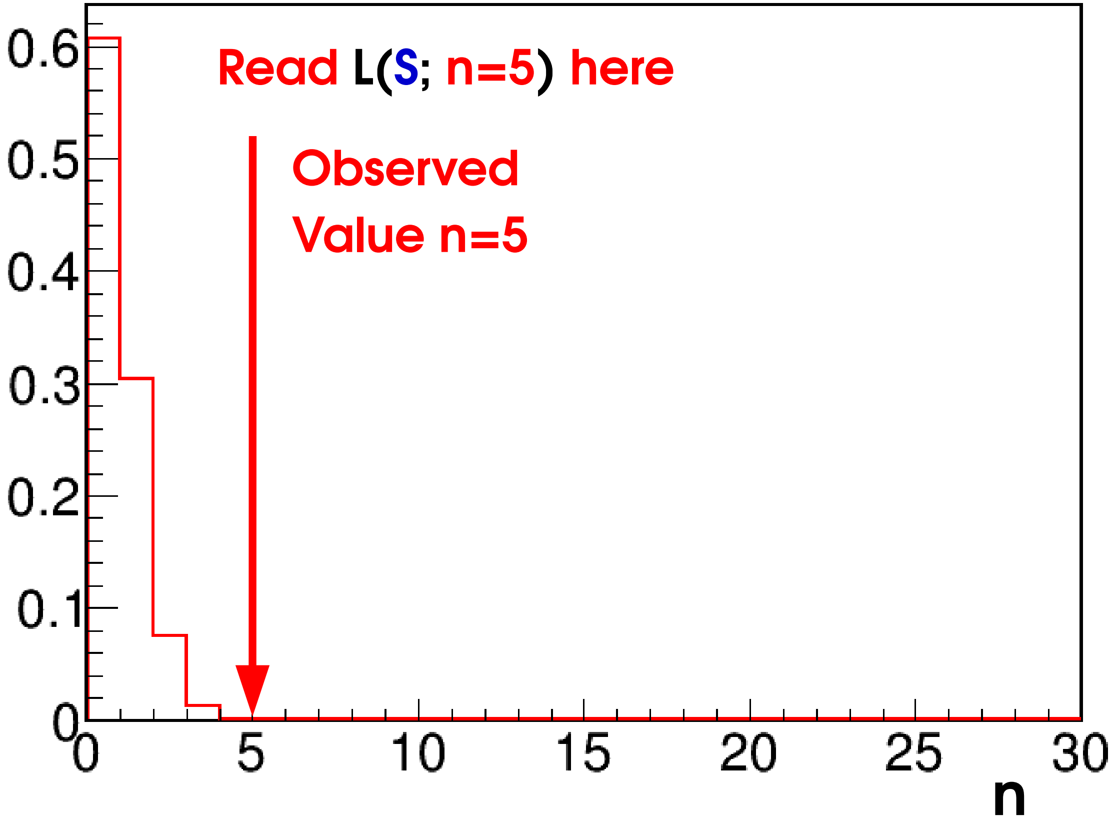
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

P(S = 0.5)
Low
likelihood



Poisson Example

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

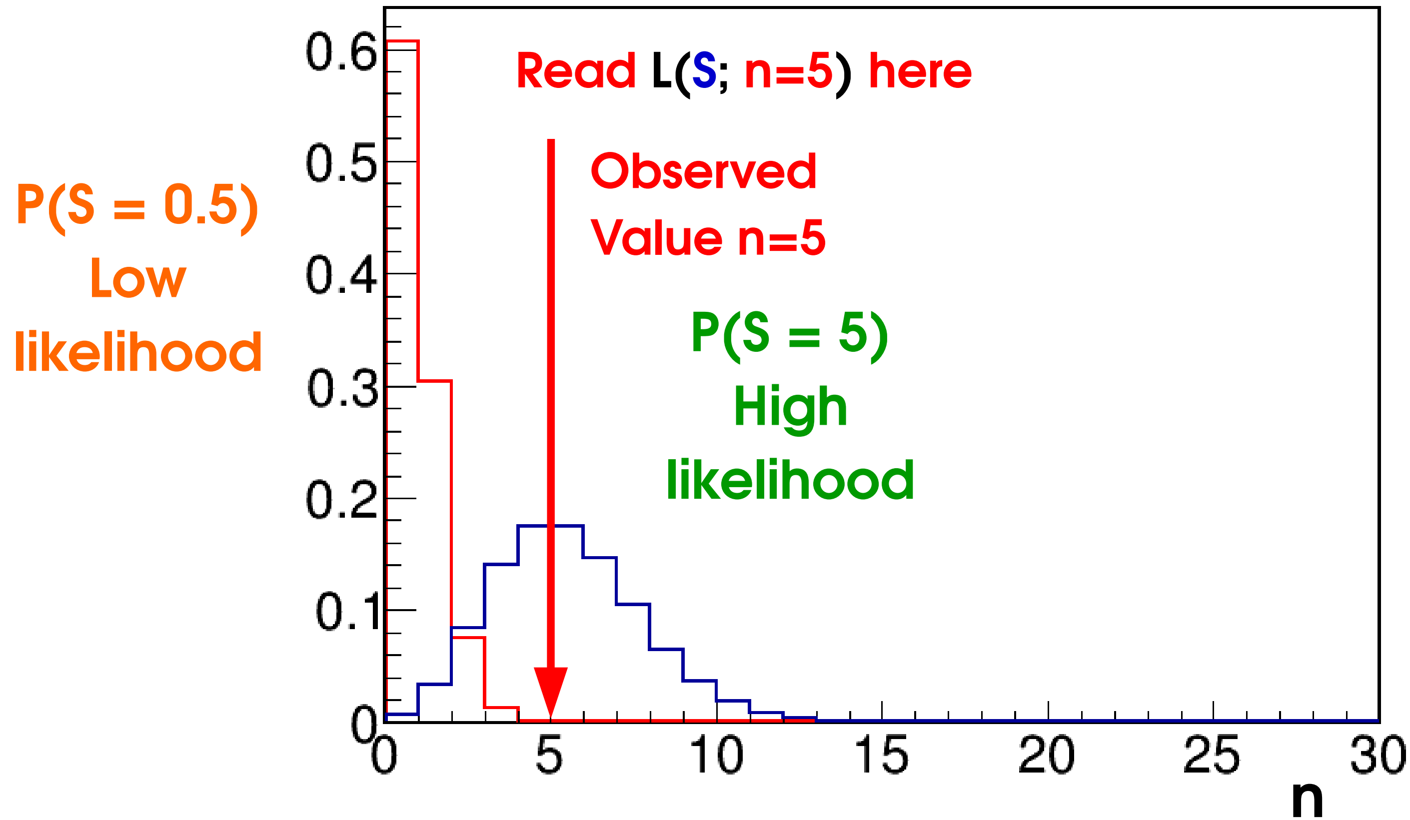
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

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



Poisson Example

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

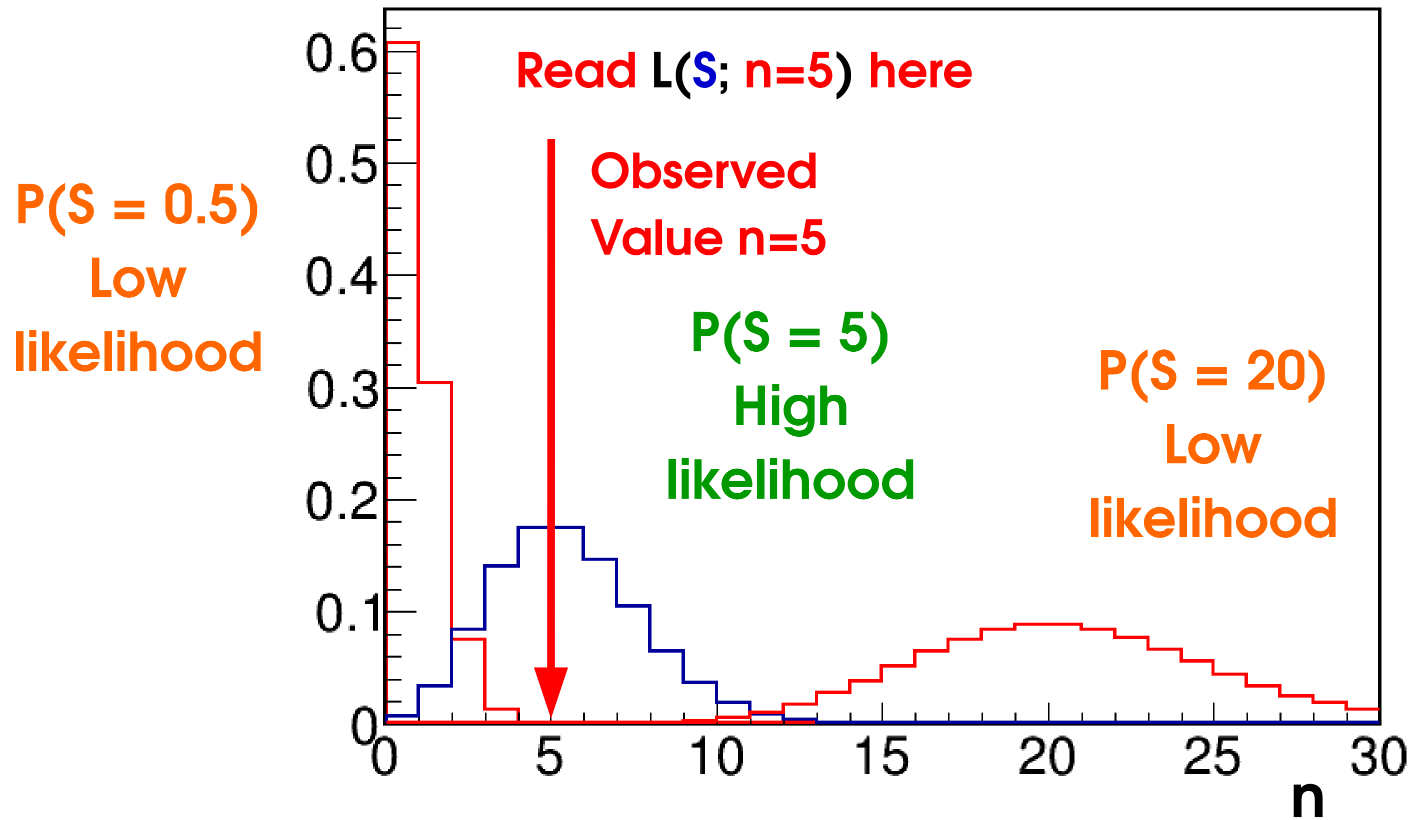
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

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



Poisson Example

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

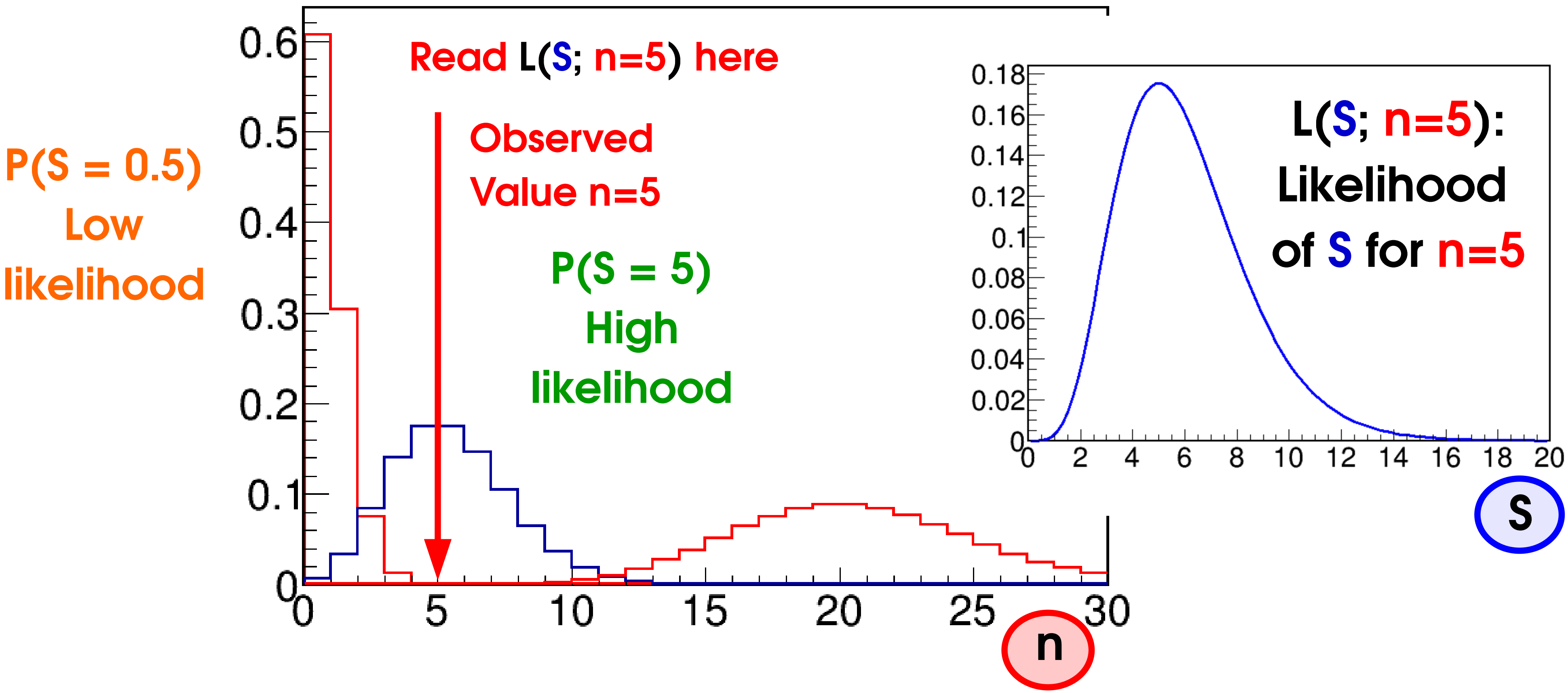
Say we **observe** n=5, want to infer information on the parameter S

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

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

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

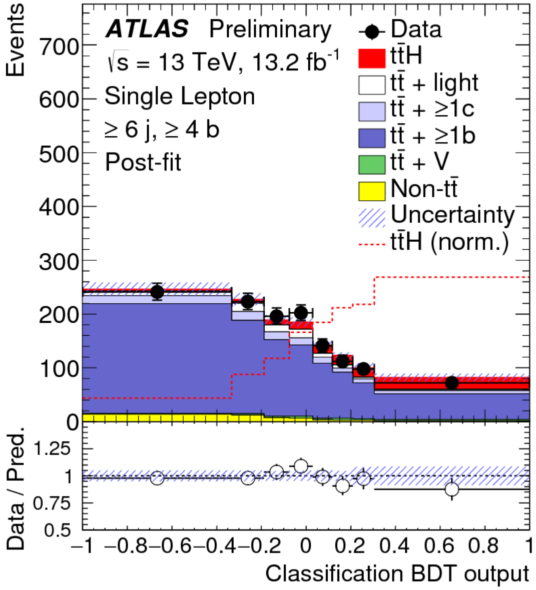
$$L(S;n=5) = e^{-S}\frac{S^5}{5!}$$



MLEs in Shape Analyses

Binned shape analysis:

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



Maximize global L(S) (each bin may prefer a different **S**)
In practice easier to minimize

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

Needs a computer...

In the Gaussian limit

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

χ^2 formula!

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

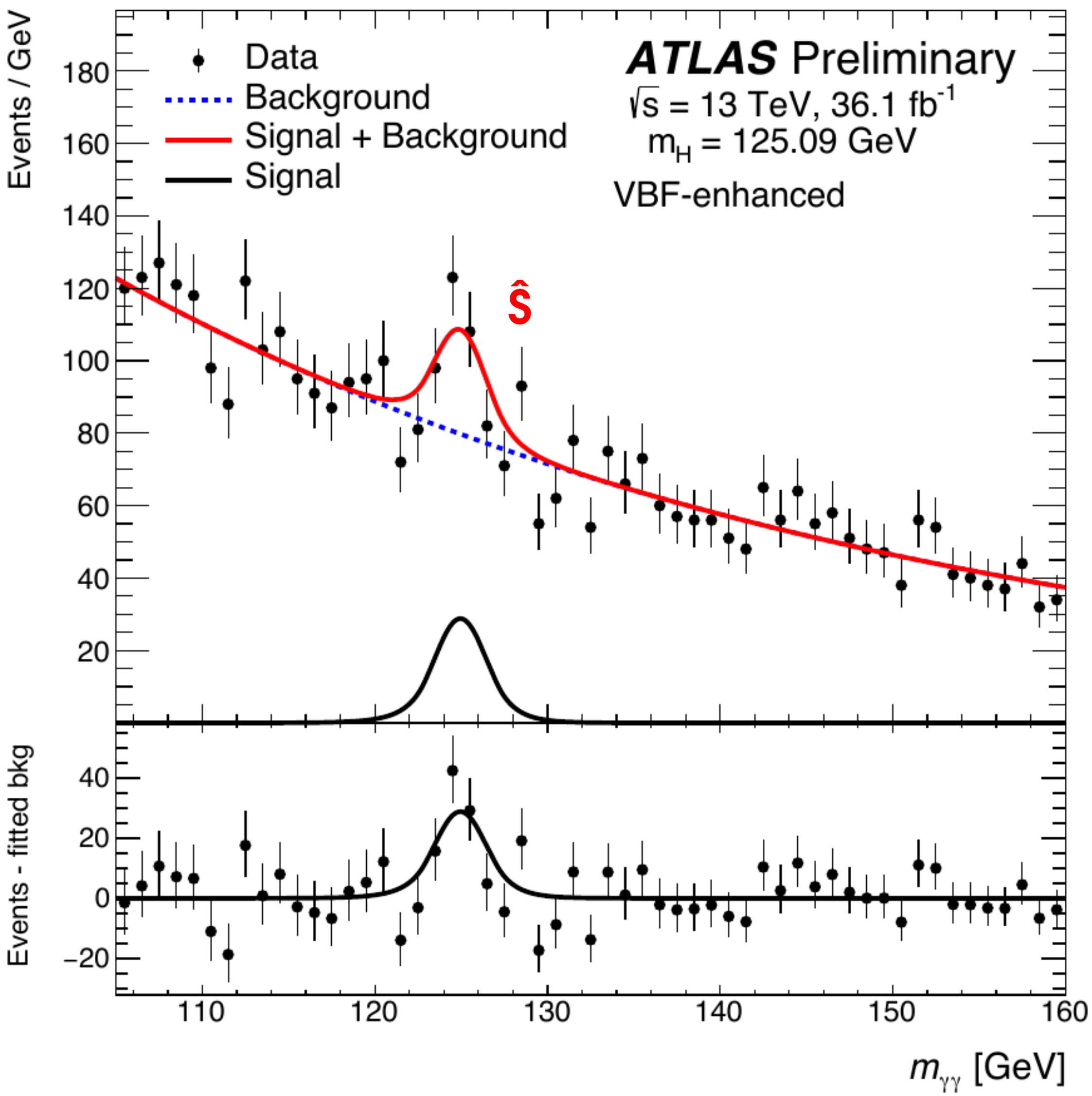
→ **Poisson MLE** (min λ_{Pois}) : ***Best fit value*** in a *likelihood* fit (in `ROOT`, fit option "`L`")

In RooFit, λ_{Pois} ⇒ `RooAbsPdf::fitTo()`, λ_{Gaus} ⇒ `RooAbsPdf::chi2FitTo()`.

In both cases, MLE ⇔ *Best Fit*

H→γγ

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



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

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

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

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

ATLAS-CONF-2017-045

MLE Properties

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

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

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

  for large enough datasets

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

  → MLE captures all the available information in the data

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

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

  → Usually more efficient numerically

  → For Gaussian L, λ is parabolic:

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

Extra: Fisher Information

Fisher Information:

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

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

Gaussian likelihood:

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

→ smaller σ_{Gauss} ⇒ more information.

Cramer-Rao bound:

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

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

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

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

Gaussian case:

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

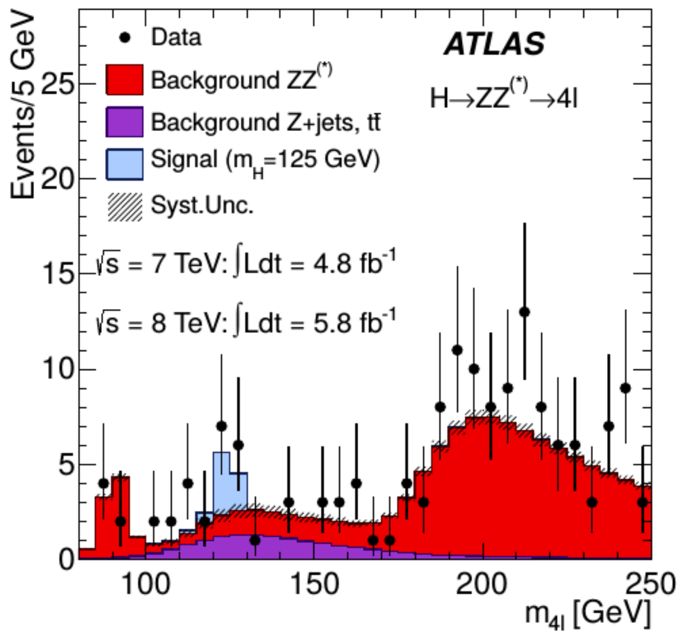
$$P(\tilde{\mu}) \propto \exp\left(-\frac{(\tilde{\mu} - \mu^*)^2}{2\sigma^2_{\tilde{\mu}}} \right)$$

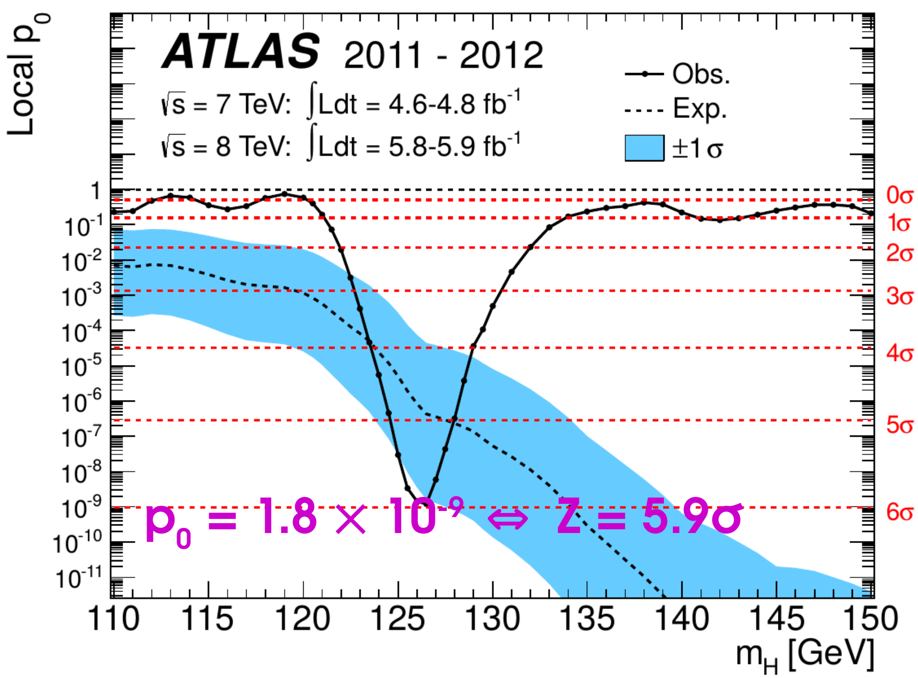
- **MLE**: $\mathrm{Var}(\hat{\mu}) = \sigma_{\hat{\mu}}^2$

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

Some Examples

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





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