Statistics and Data Analysis (HEP at LHC)

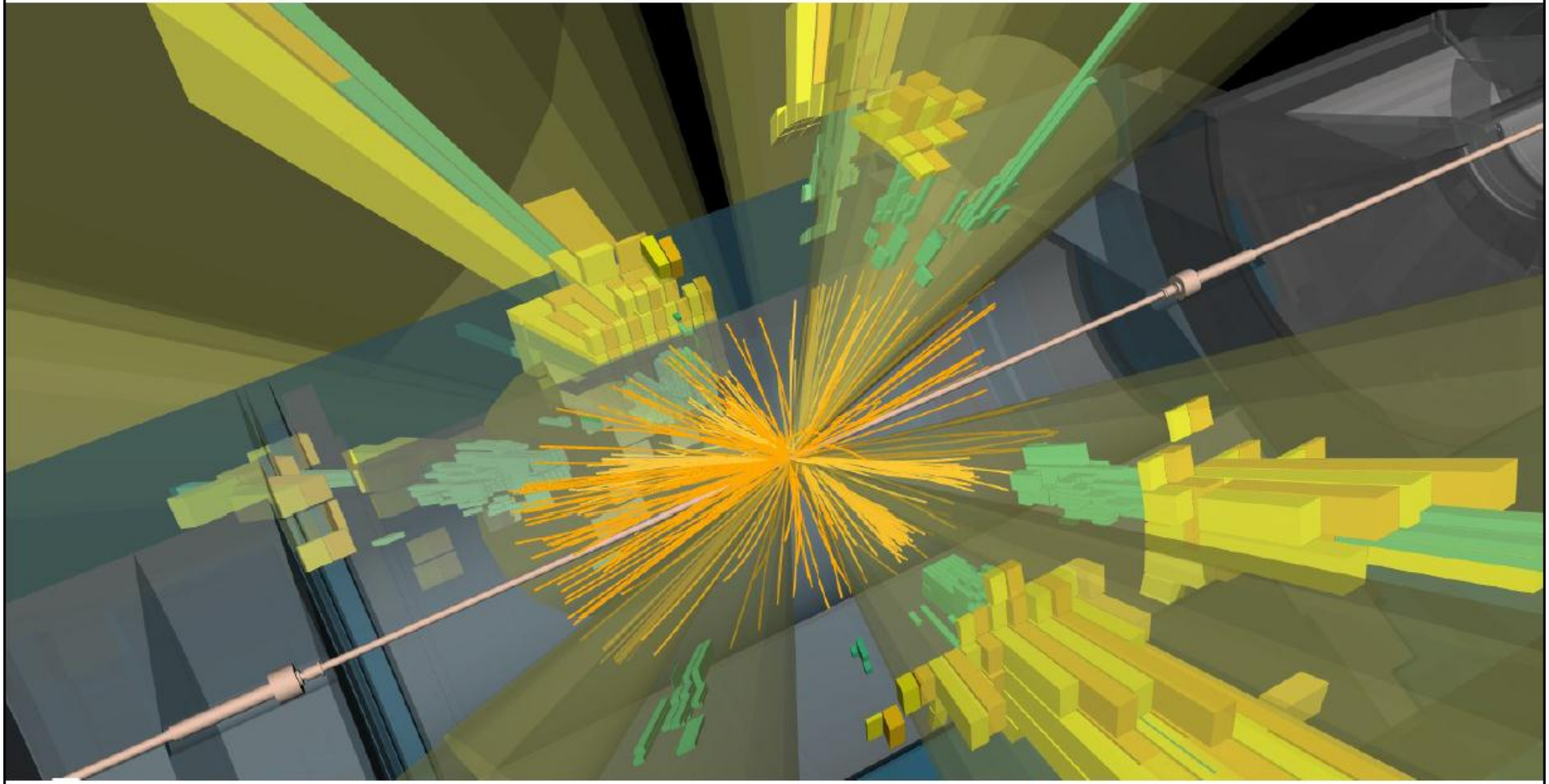
- **Statistical basics for physics**
  - **Random processes**
  - **Probability distributions**
- **Describing physics measurements**
  - **Binned and unbinned data**
  - **Model parameters**

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

Prof. dr hab. Elżbieta Richter-Wąs

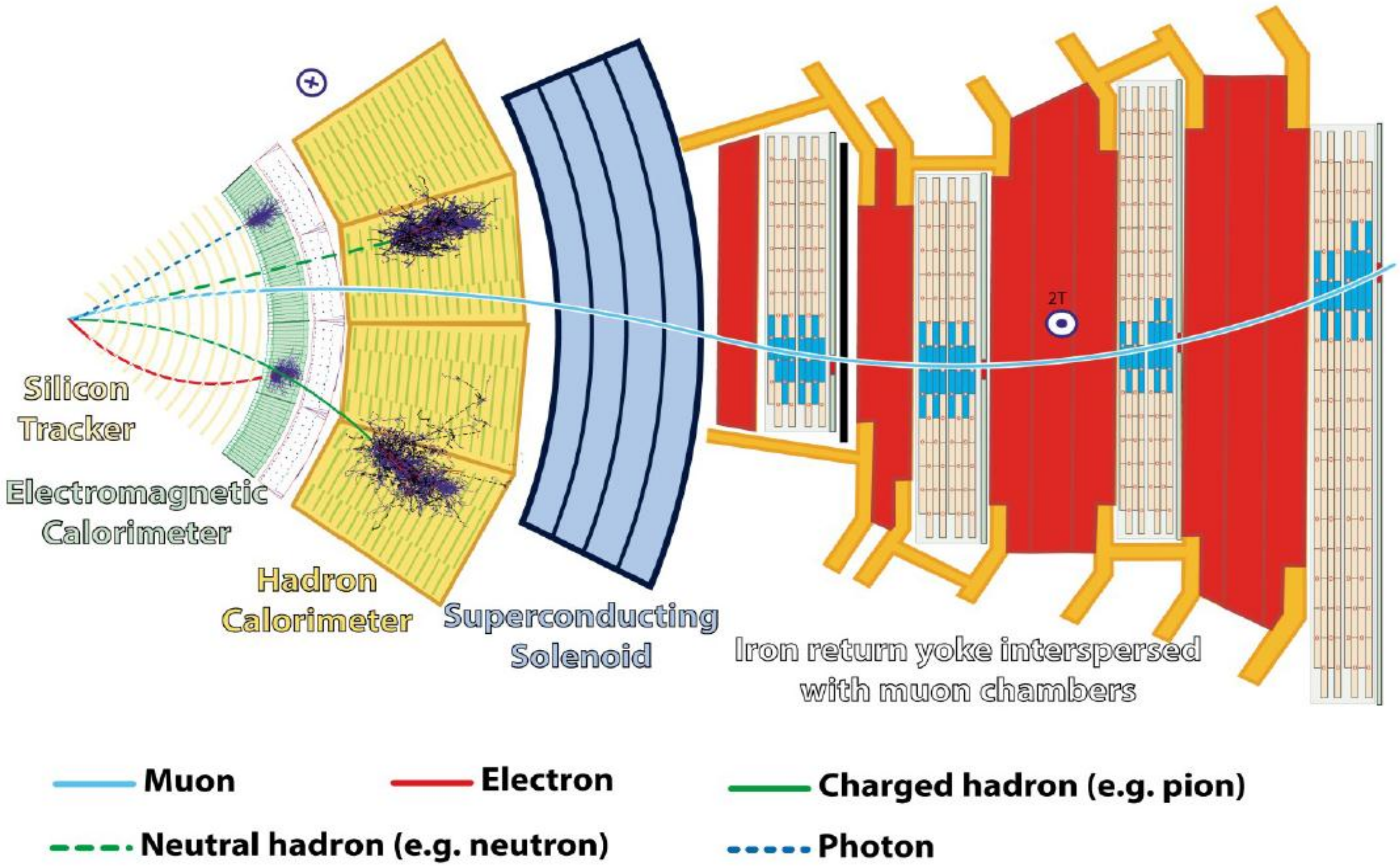
Randomness in High Energy Physics

Experimental data is produced by incredibly complex processes



Randomness in High Energy Physics

Experimental data is produced by incredibly complex processes



Randomness in High Energy Physics

Experimental data is produced by incredibly complex processes

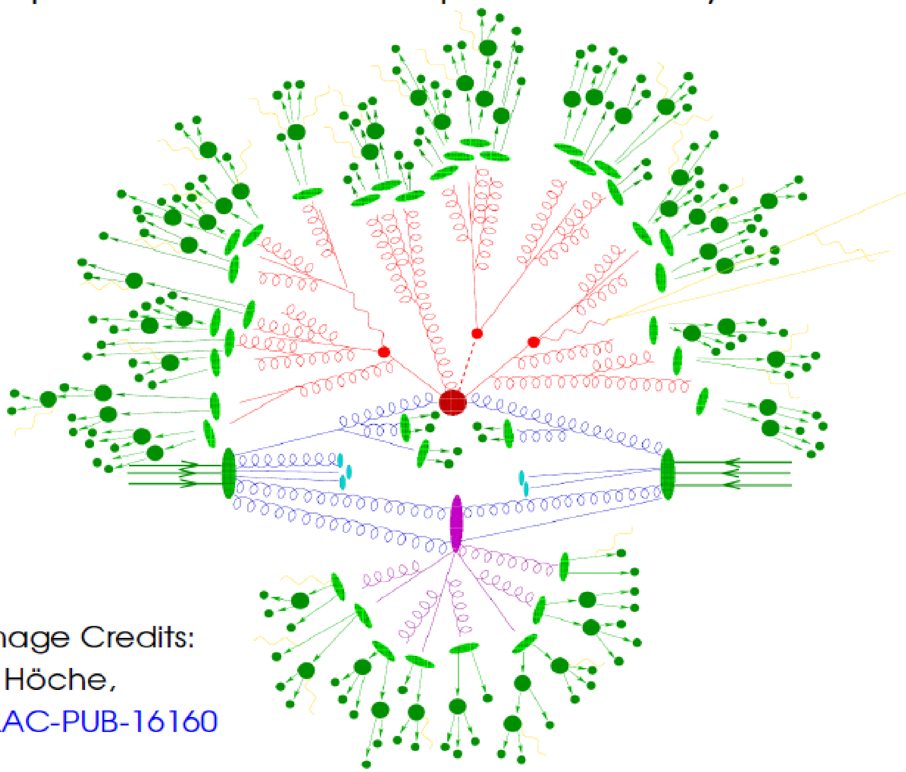


Image Credits:
S. Höche,
SLAC-PUB-16160

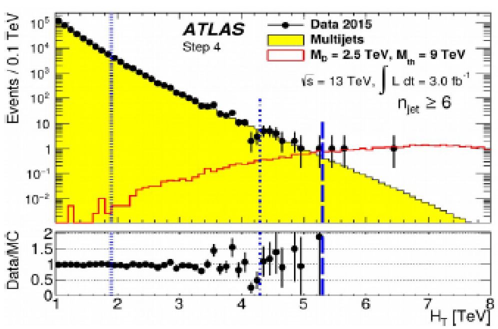
Hard scattering

PDFs, Parton shower, Pileup

Decays

Detector response

Reconstruction



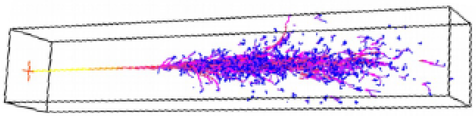
Randomness involved in all stages

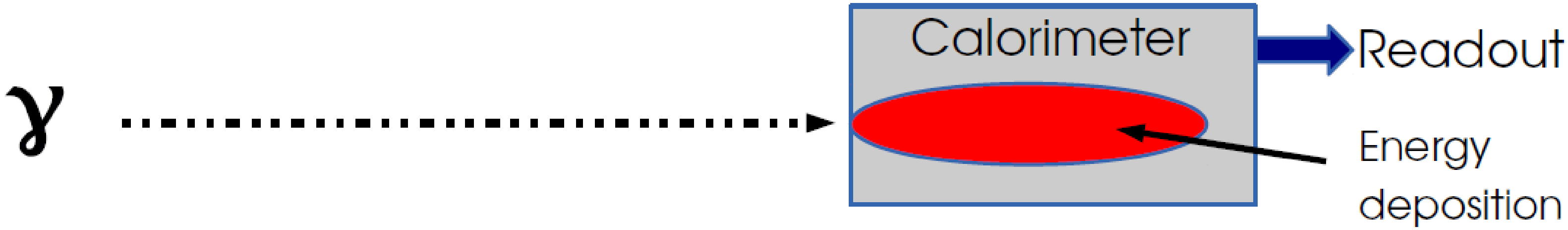
→ **Classical** randomness: detector reponse

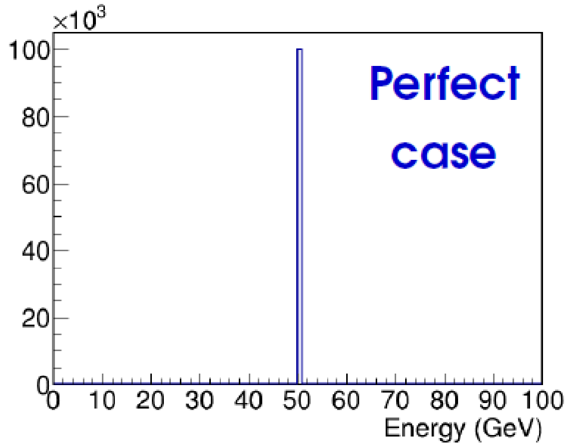
→ **Quantum** effects in particle production, decay

Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

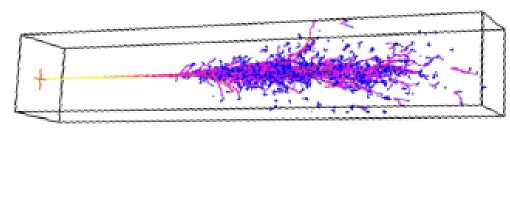


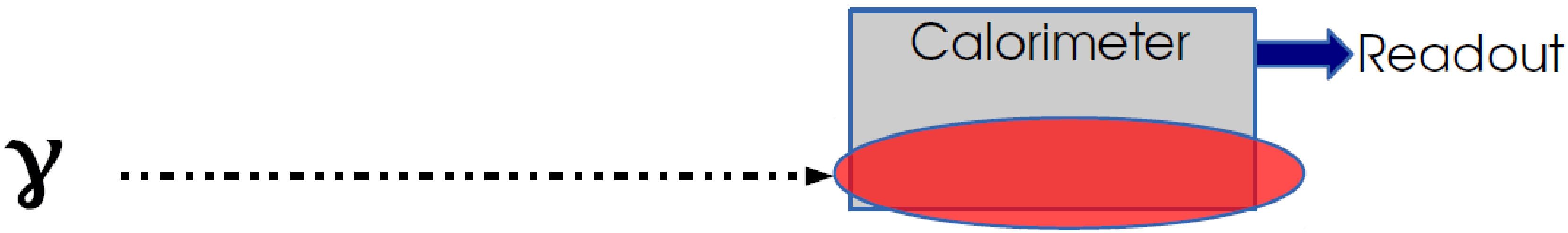


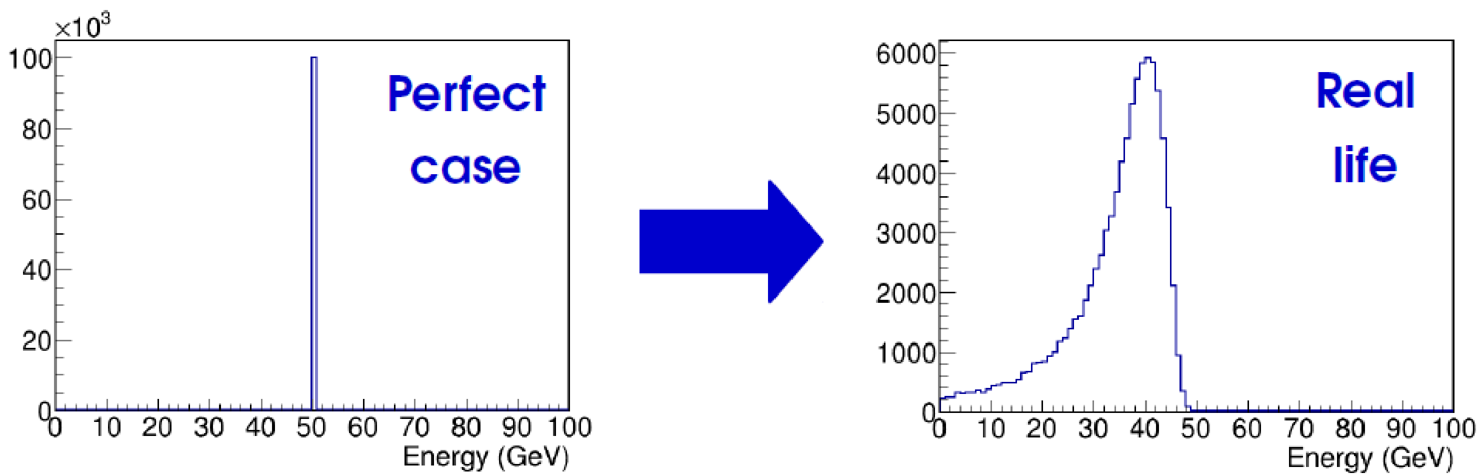


Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

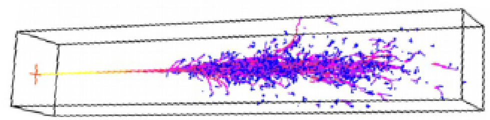


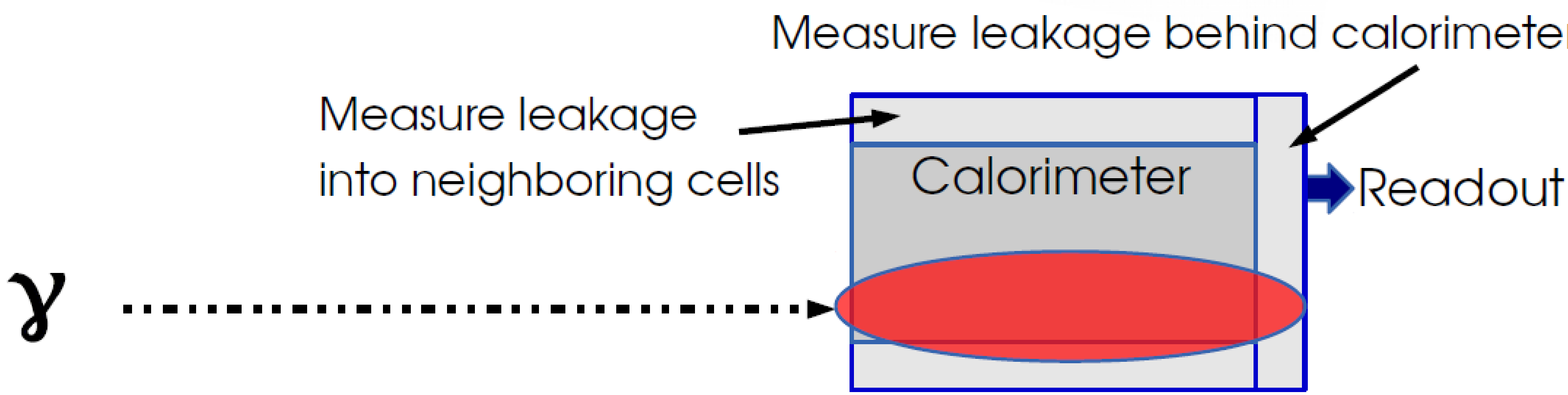


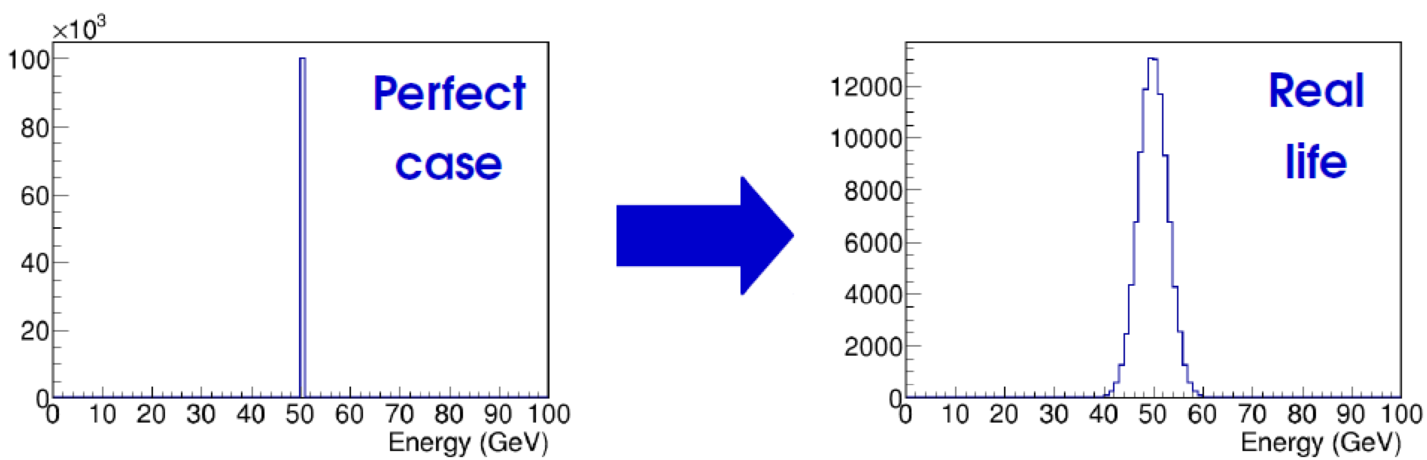


Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter



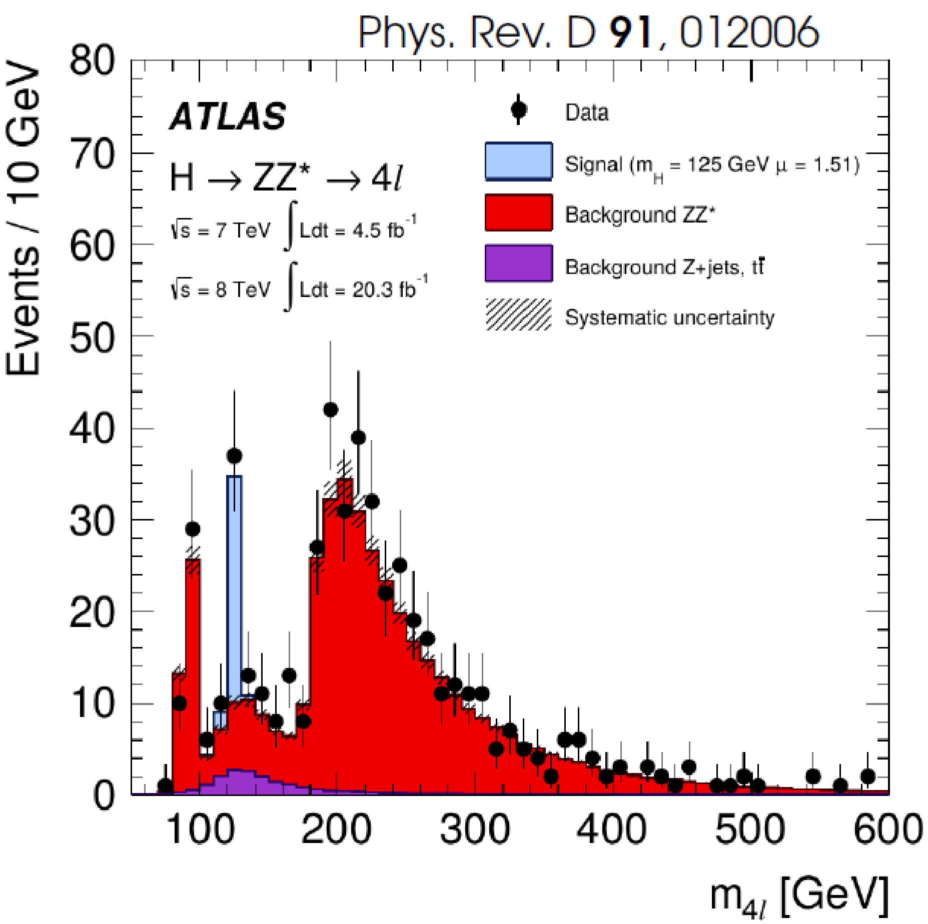




Cannot predict the measured value for a given event

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

Quantum randomness: H->ZZ*-> 4 l



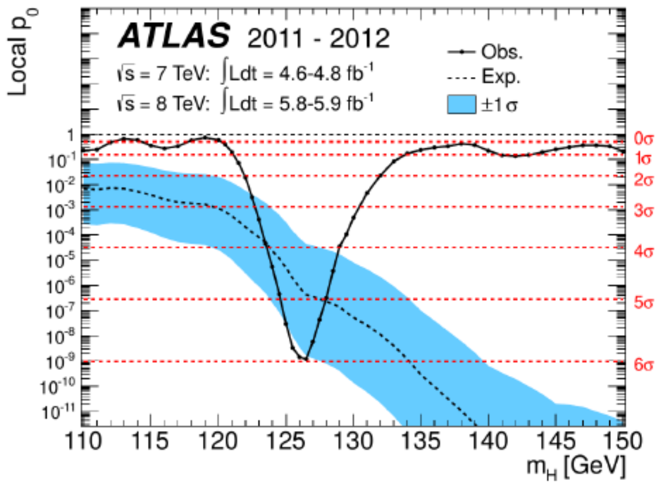
Rare process: Expect 1 signal event every **~6 days**

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

Randomness in Physics

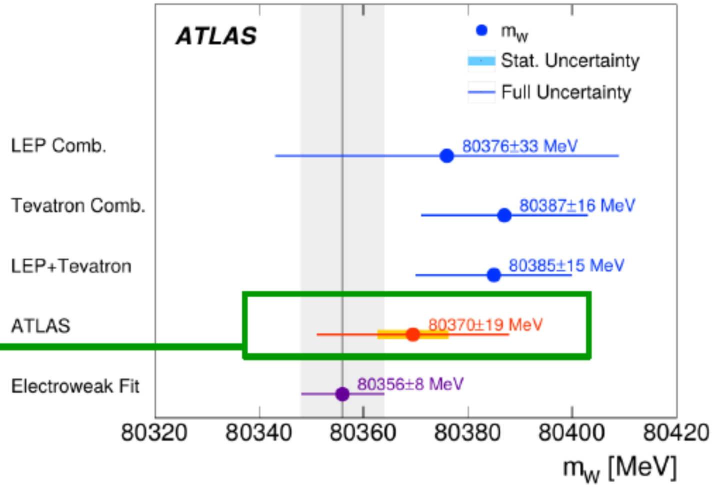
Questions with probabilistic answers:

- **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



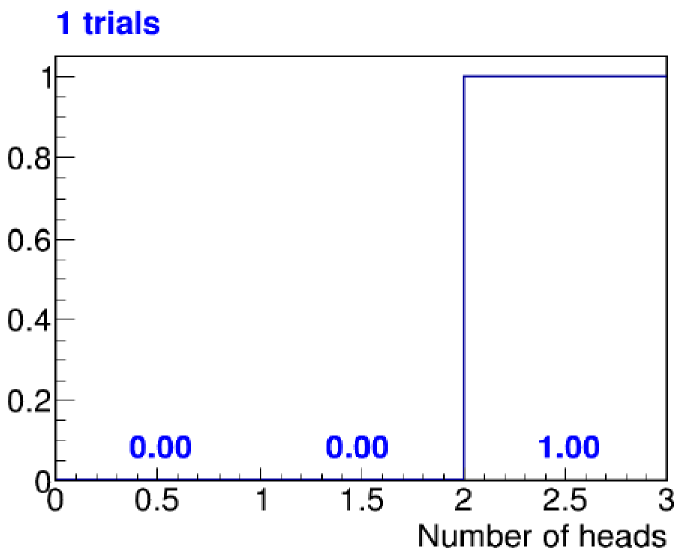
Probability distributions

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

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

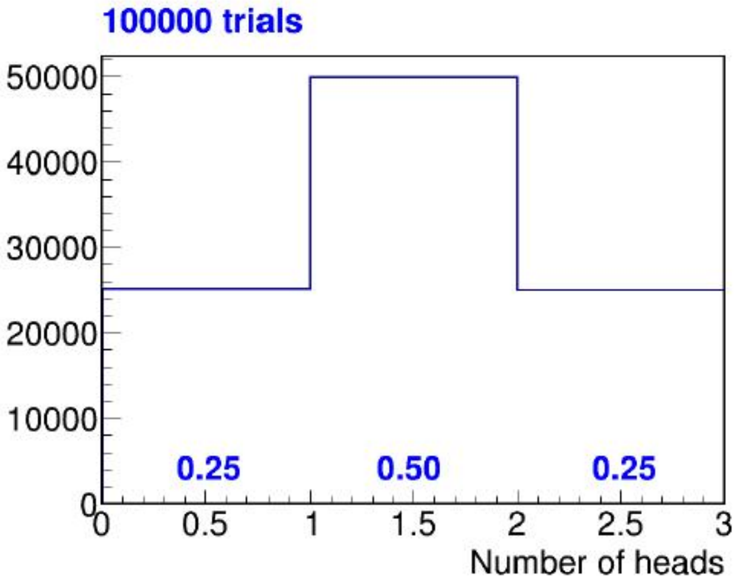
Probability distributions

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

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

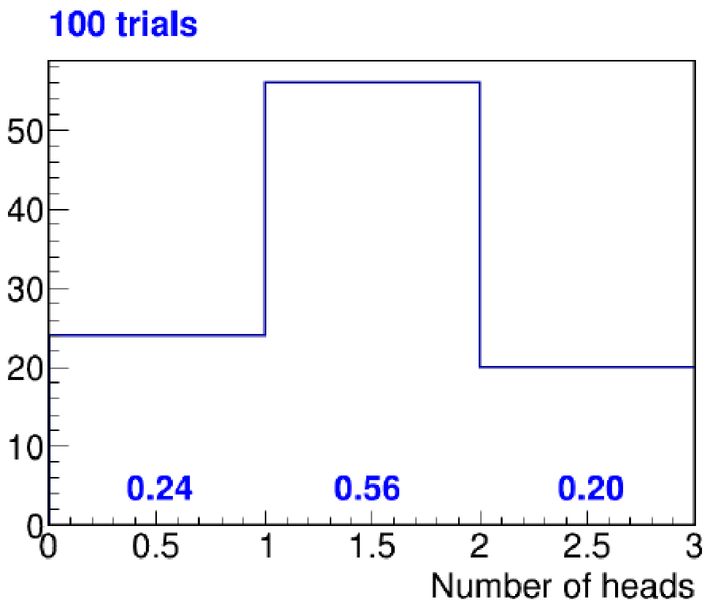
Probability distributions

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 :

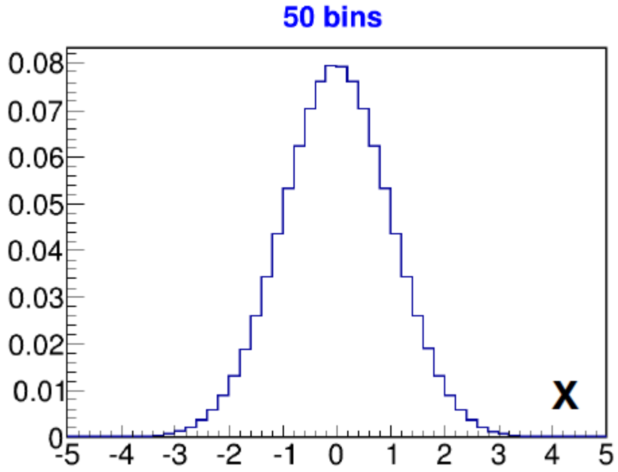
$\{ P_i \}$ for i = 0, 1, 2

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)

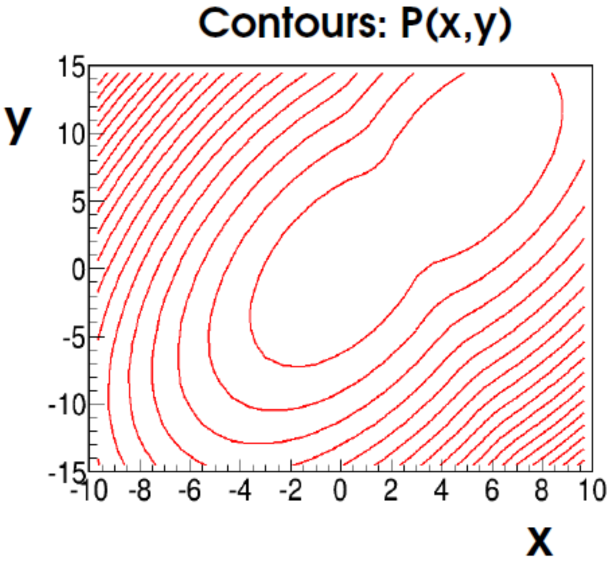
→ $P(x) > 0$, $\int P(x)\,dx = 1$

→ High values ⇔ high chance to get a measurement here

Generalizes to **multiple variables** :

→ $P(x,y) > 0$

→ $\int P(x,y)\,dx\,dy = 1$

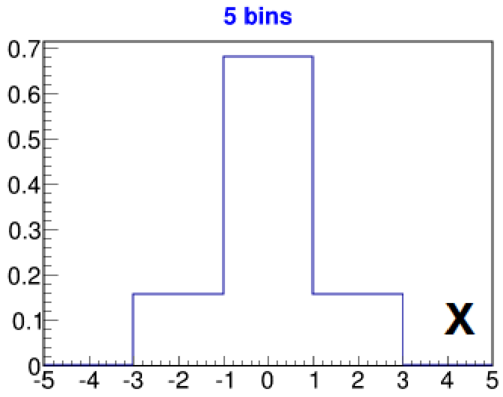


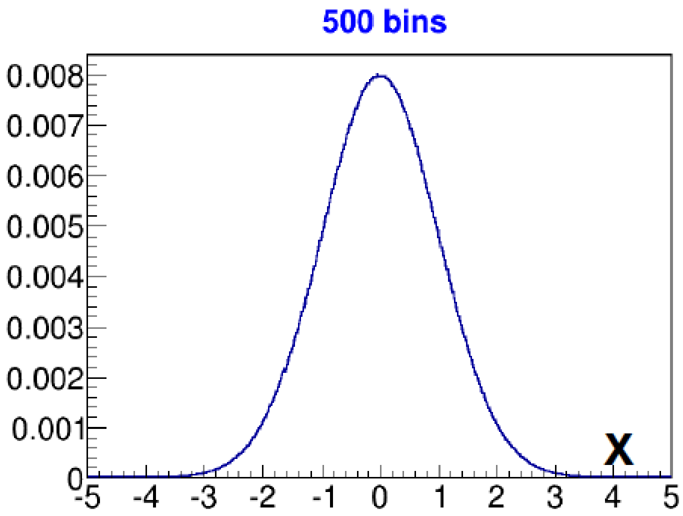
Random Variables

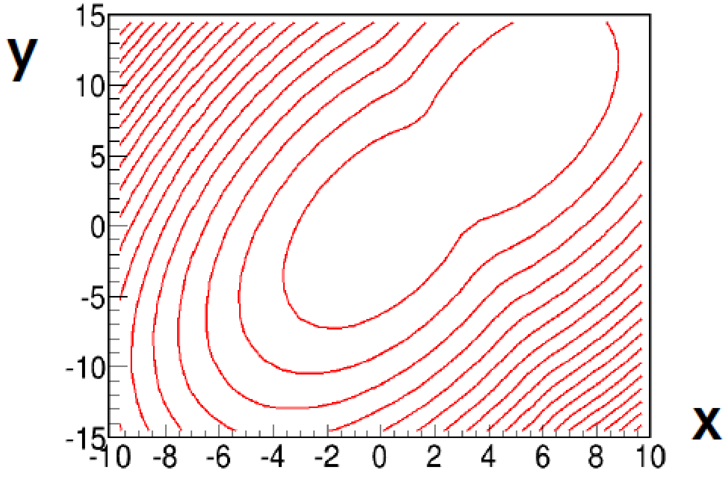
X, Y... are **Random Variables** (continuous or discrete), a.ka. **observables** :
→ X can take any value x, with probability **P(X=x).**

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

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







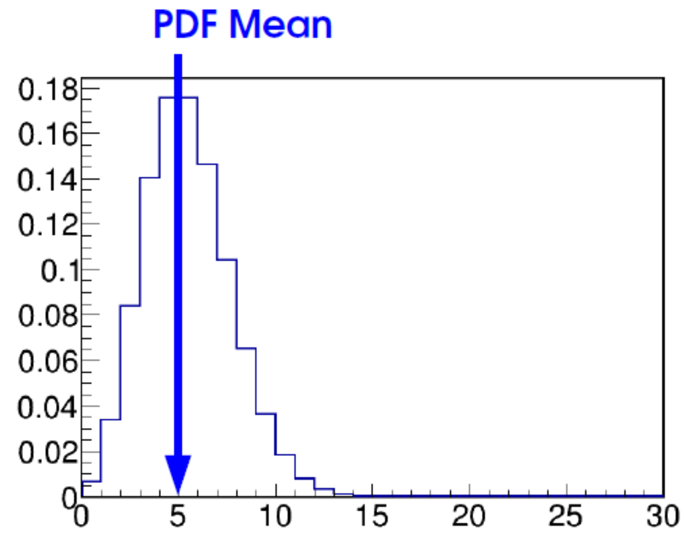
PDF properties: mean

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

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

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

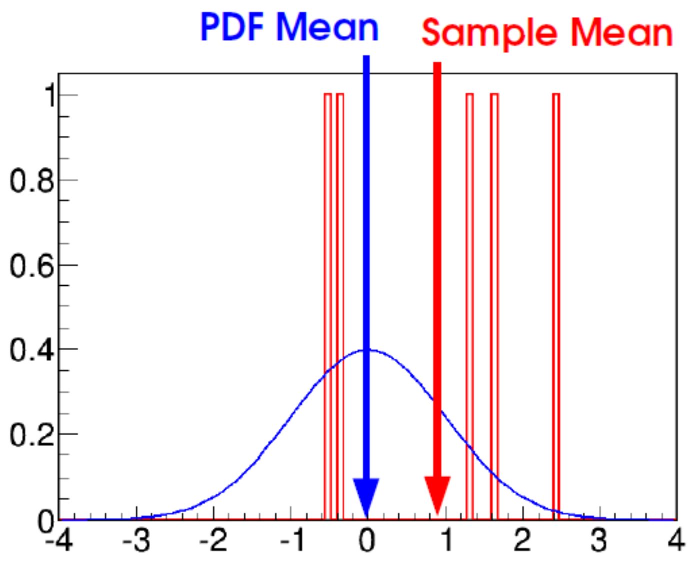
→ Property of the **PDF**



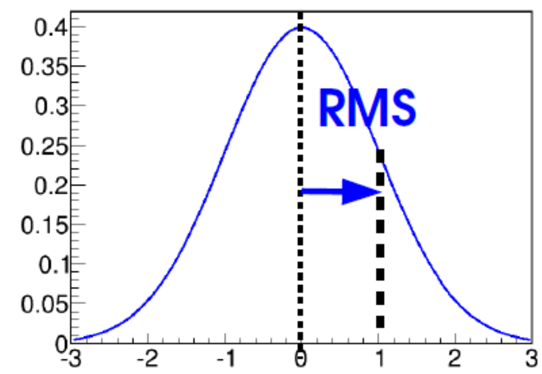
For measurements $x_1 \ldots x_n$,
then can compute the **Sample mean**:

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

→ Property of the **sample**
→ approximates the PDF mean.



PDF properties: (co)variance



Variance of X:

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

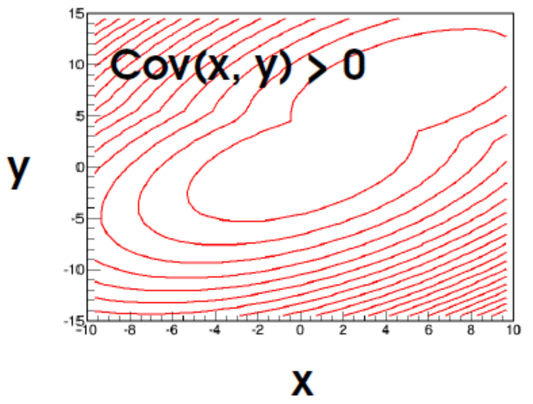
→ Average square of deviation from mean

→ **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:

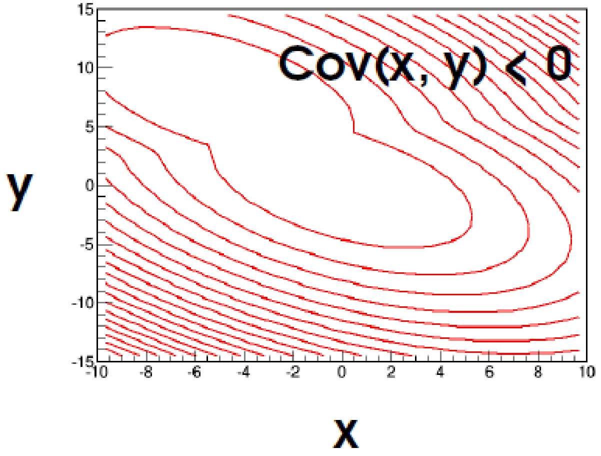


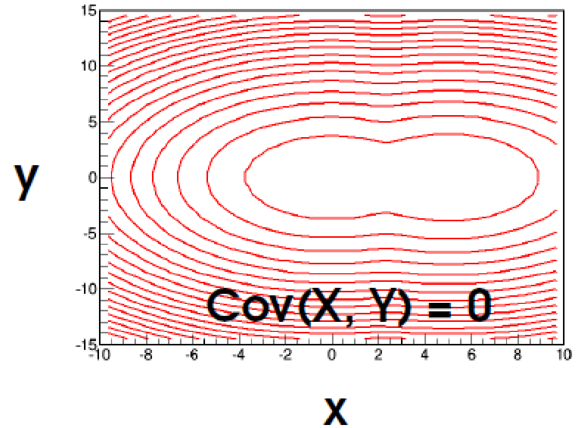
$$\mathrm{Cov}(X,Y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$$

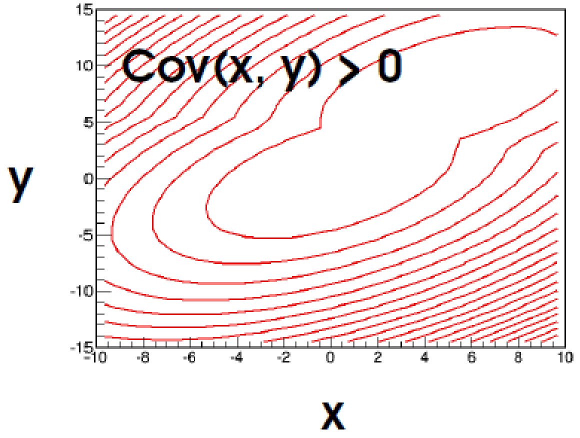
→ Large if variations of X and Y are "synchronized"

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

PDF properties: (co)variance

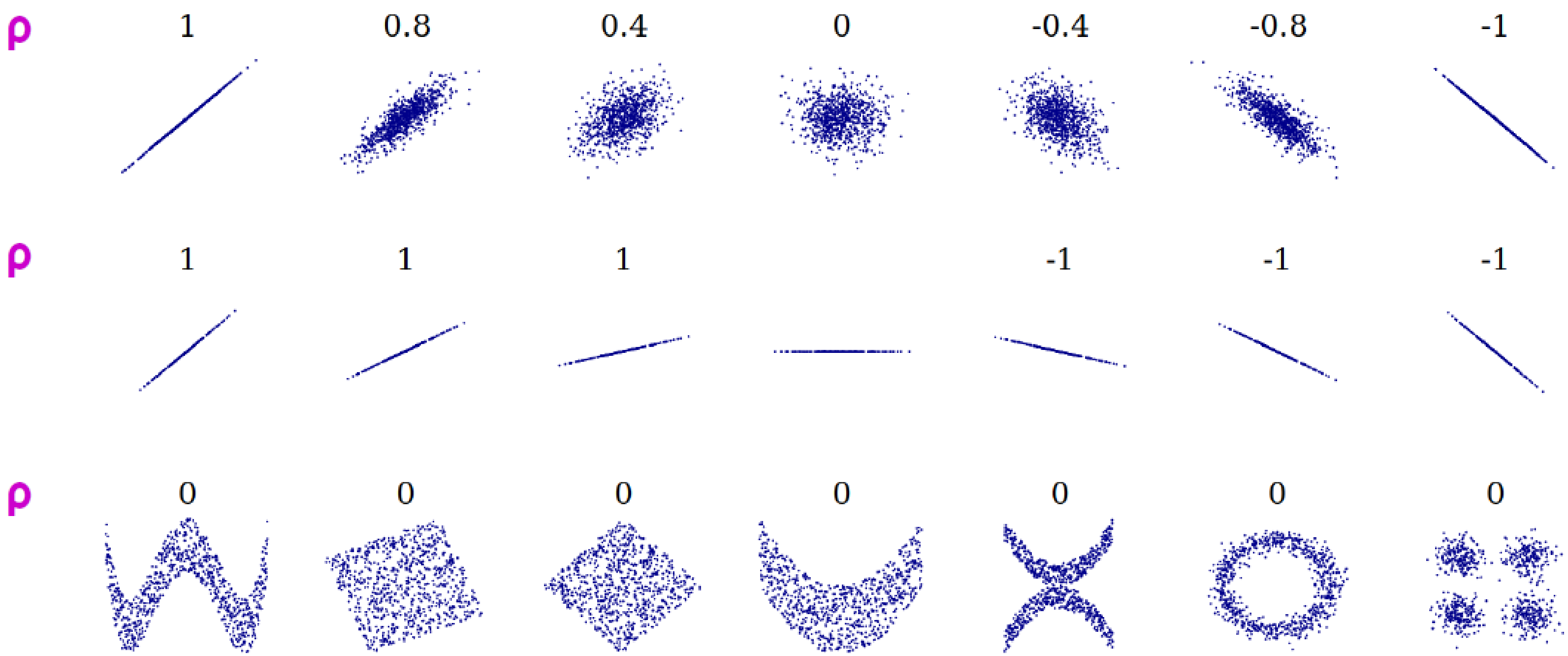






„Linear” vs. „non-linear” correlations

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



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

Source: Wikipedia

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

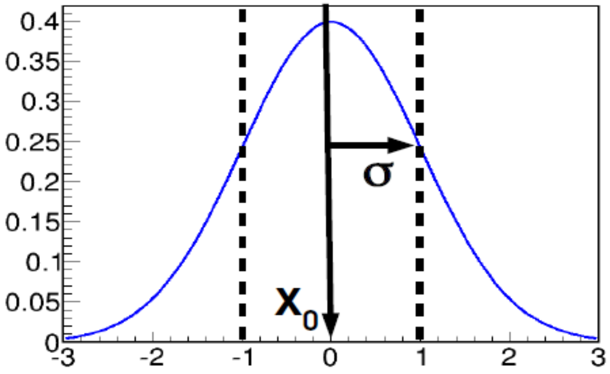
Gaussian PDF

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 ($\Rightarrow$ RMS = σ)



Generalize to **N dimensions**:

→ **Mean** : X_0

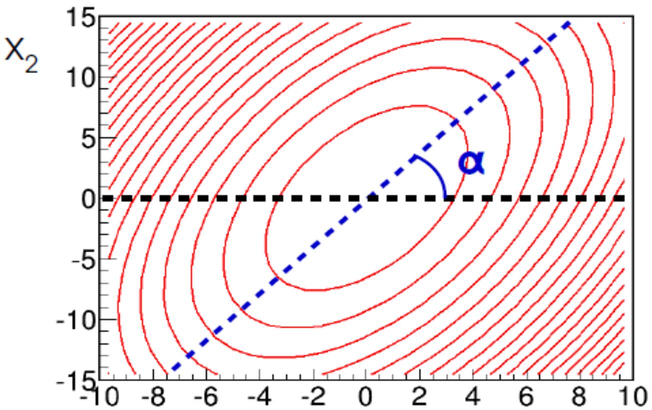
→ **Covariance matrix** :

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

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

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

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



Gaussian quantiles

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

$G(x; x_0, \sigma)$ depends only on $z \sim G(z; 0,1)$

Probability $P(|x - x_0| > Z\sigma)$ to be away from the mean:

Z	$P(\|x - x_0\| > Z\sigma)$
1	**0.317**
2	**0.045**
3	**0.003**
4	**3×10^{-5}**
5	**6×10^{-7}**

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

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

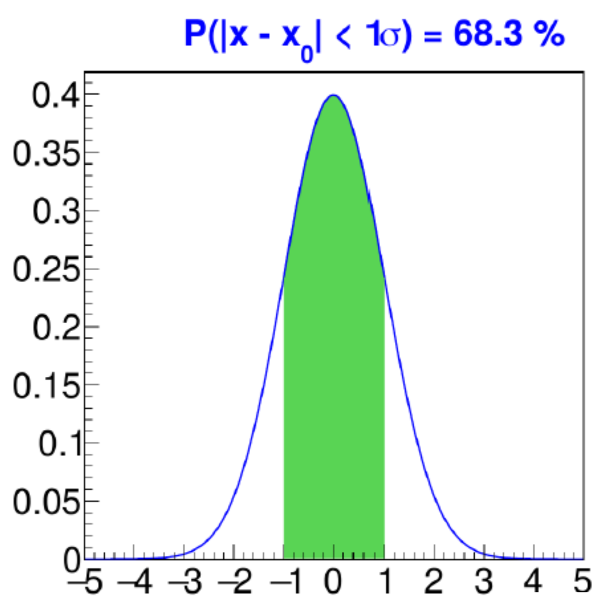
In ROOT,

$\Phi(z)$: `ROOT::Math::gaussian_cdf(z)`

$\Phi^{-1}(p)$: `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

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

$G(x;x_0,\sigma)$ depends only on $z \sim G(z;0,1)$

Probability $P(|x - x_0| > Z\sigma)$ to be away from the mean:

Z	$P(\|x - x_0\| > Z\sigma)$
1	**0.317**
2	**0.045**
3	**0.003**
4	**3×10^{-5}**
5	**6×10^{-7}**

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

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

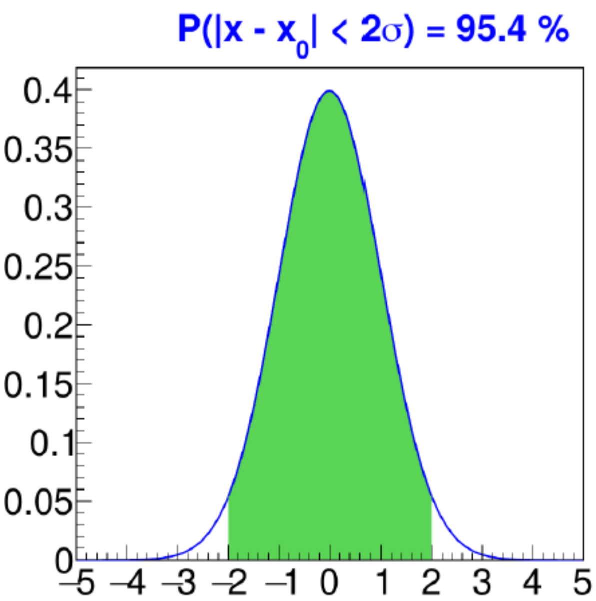
In ROOT,

$\Phi(z)$: `ROOT::Math::gaussian_cdf(z)`

$\Phi^{-1}(p)$: `ROOT::Math::gaussian_quantile(p,1)`

and add "_c" to use $1-\Phi$ 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

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

$G(x; x_0, \sigma)$ depends only on $z \sim G(z; 0,1)$

Probability $P(|x - x_0| > Z\sigma)$ to be away from the mean:

Z	$P(\|x - x_0\| > Z\sigma)$
1	0.317
2	0.045
3	0.003
4	3×10^{-5}
5	6×10^{-7}

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

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

In ROOT,

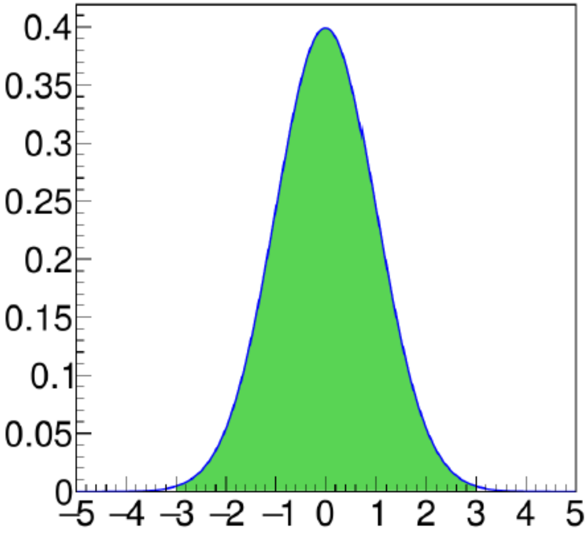
$\Phi(z)$: `ROOT::Math::gaussian_cdf(z)`

$\Phi^{-1}(p)$: `ROOT::Math::gaussian_quantile(p,1)`

and add "_c" to use $1-\Phi$ 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
```

$P(|x - x_0| < 3\sigma) = 99.7\ \%$



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

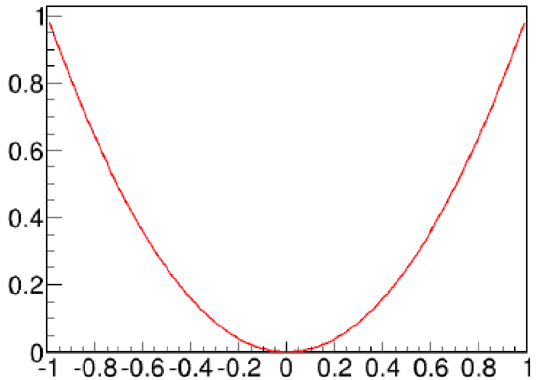
Another version, for the sum:

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

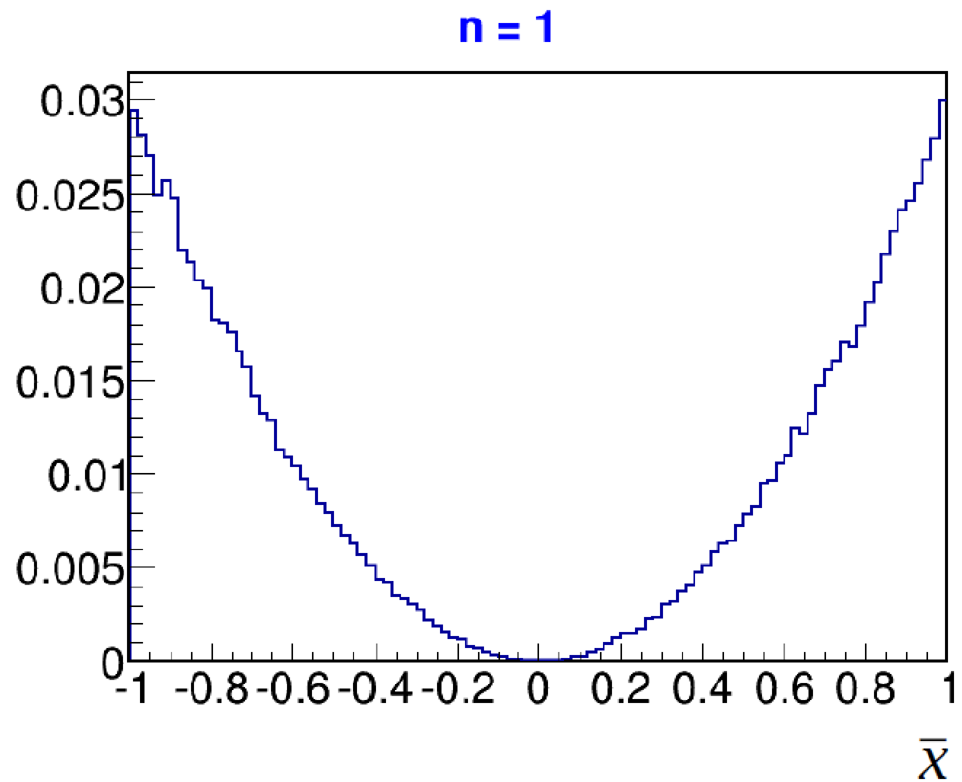
Mean scales like n, but RMS only like √n

Central Limit Theorem in action

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



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

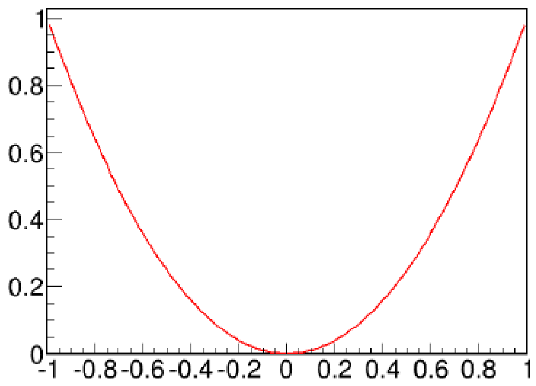


Distribution becomes Gaussian, although very non-Gaussian originally

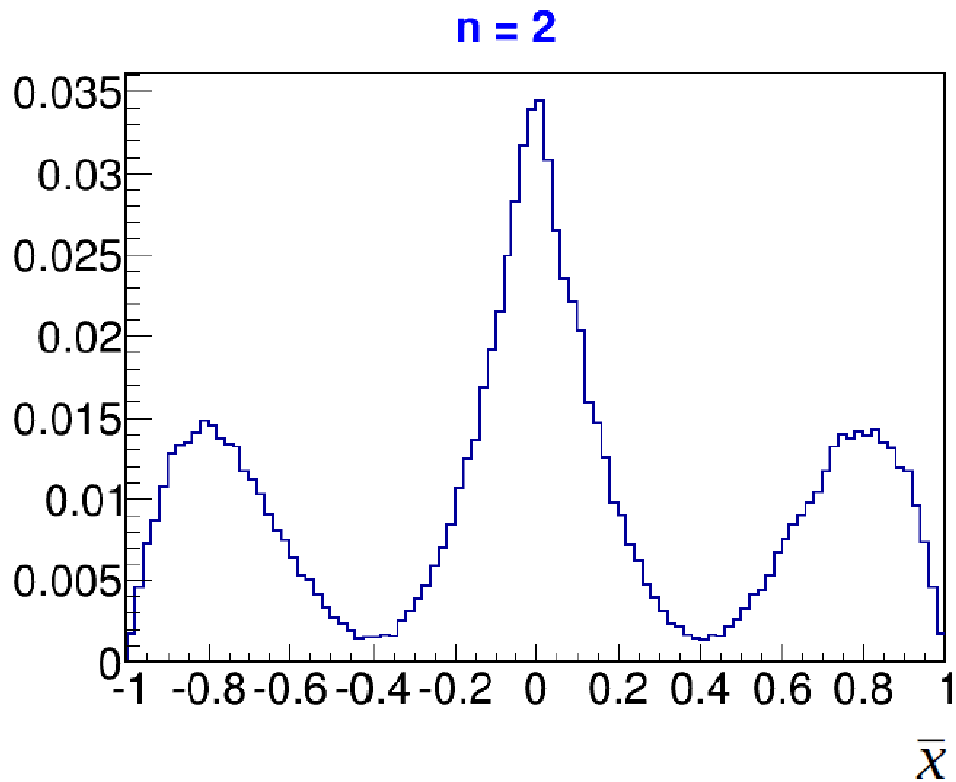
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

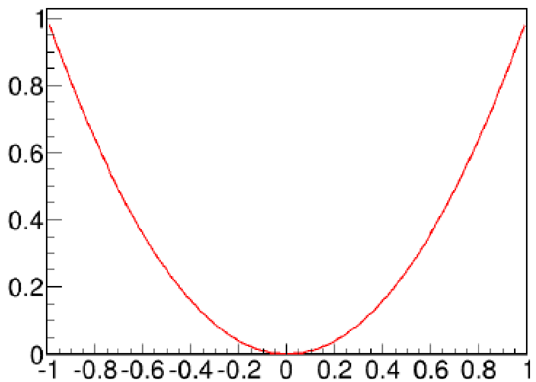


Distribution becomes Gaussian, although very non-Gaussian originally

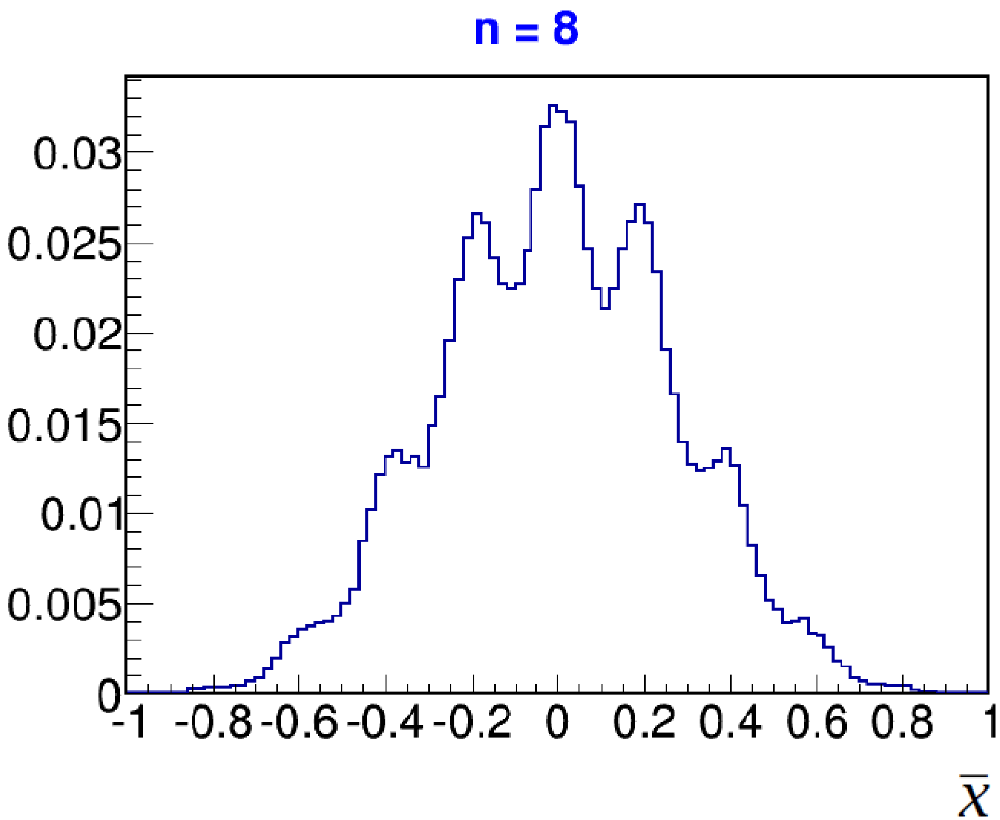
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

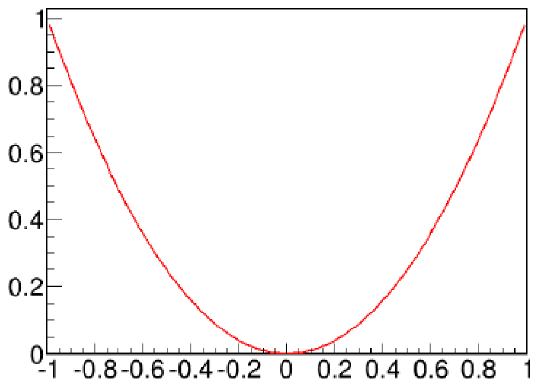


Distribution becomes Gaussian, although very non-Gaussian originally

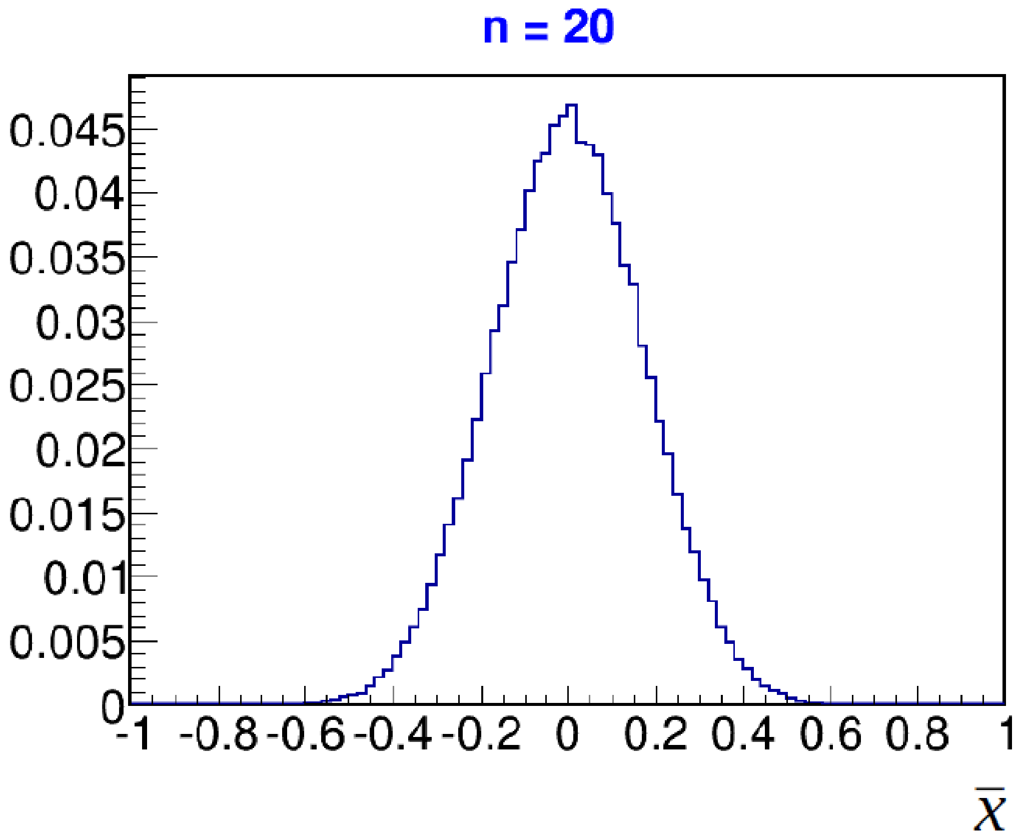
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

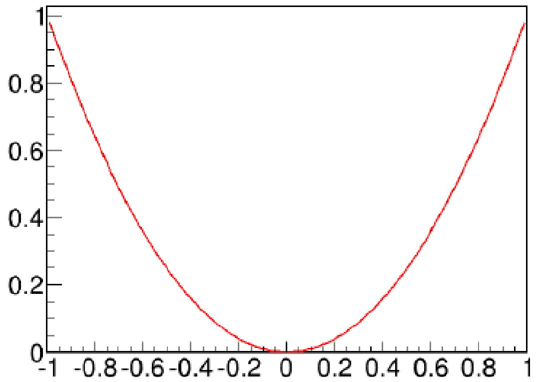


Distribution becomes Gaussian, although very non-Gaussian originally

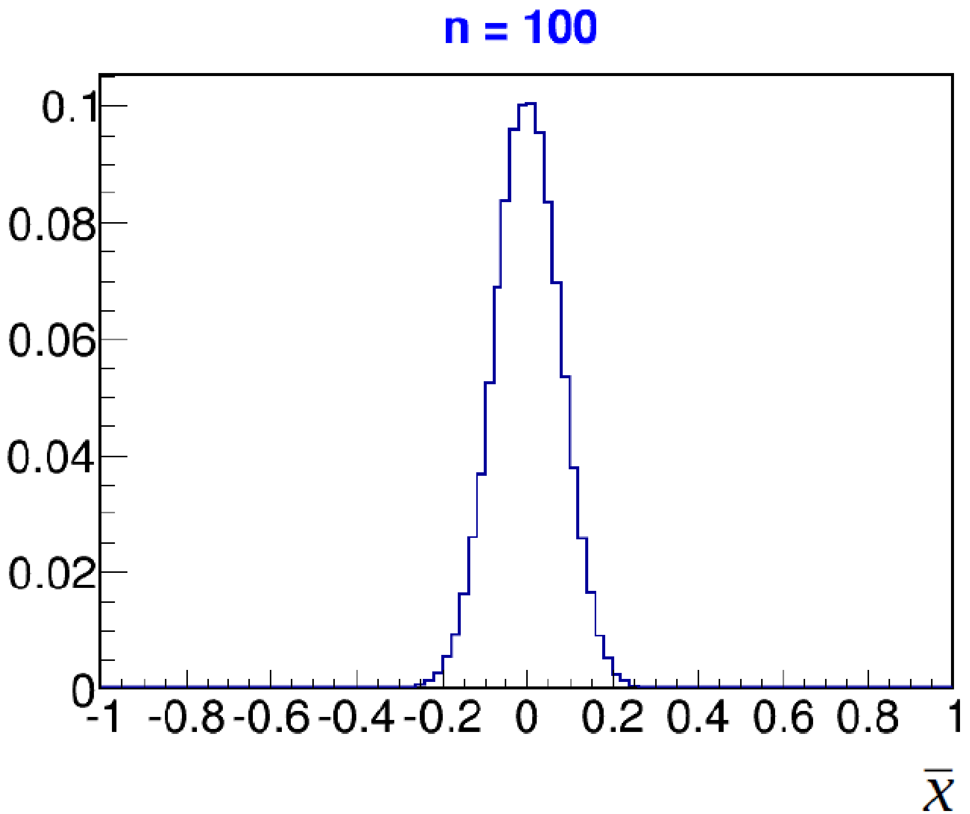
Distribution becomes narrower as expected (as 1/√n)

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)

Chi-squared

Chi-squared

Multiple Independent Gaussian variables x_i: Define

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

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

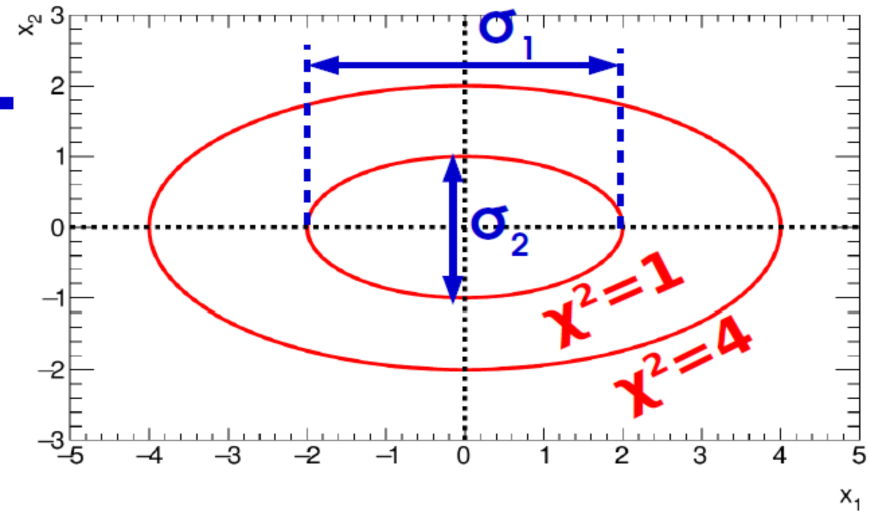
Distribution depends on n :

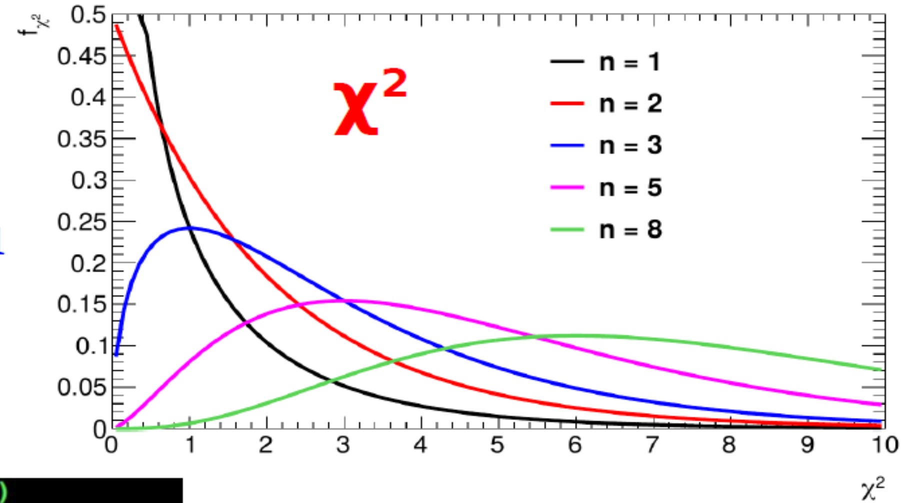
Rule of thumb: χ^2/n **should be** $\lesssim 1$

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 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 $(x_1^0 \ldots . x_n^0)$

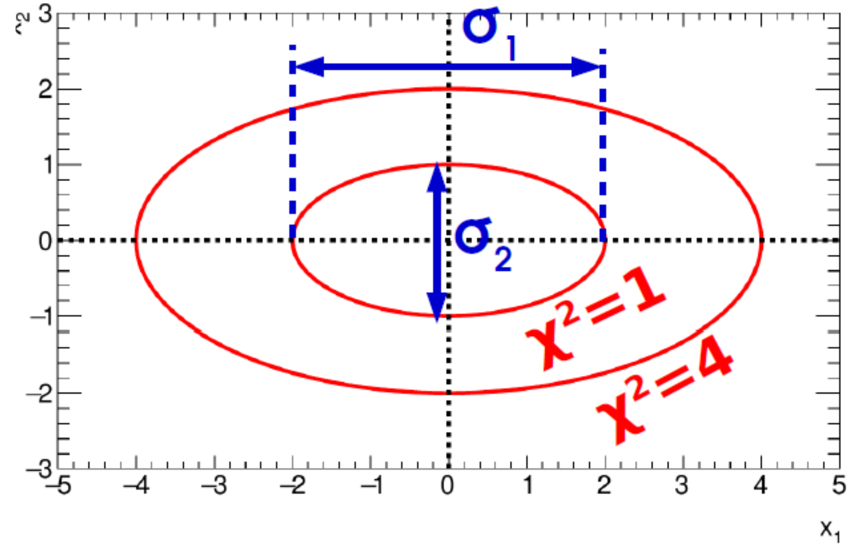
Distribution depends on n :

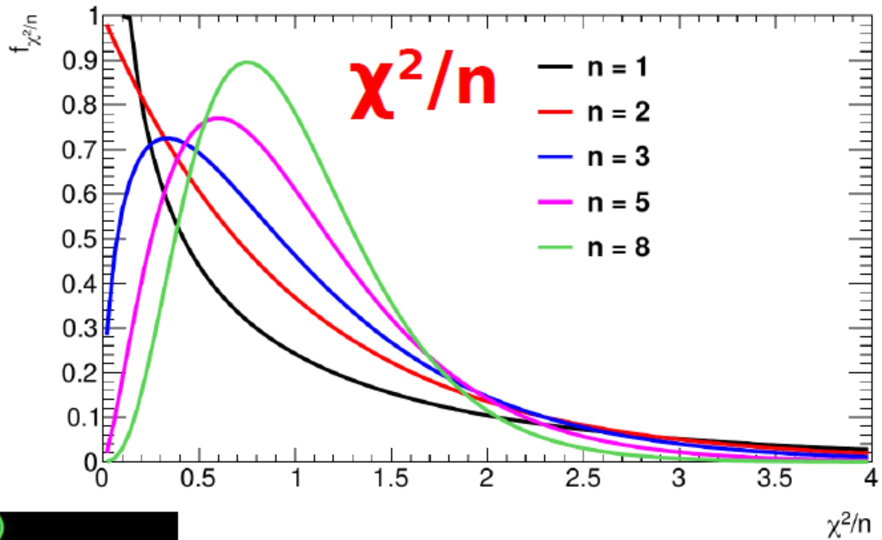
Rule of thumb: χ^2/n **should be** $\lesssim 1$

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
```

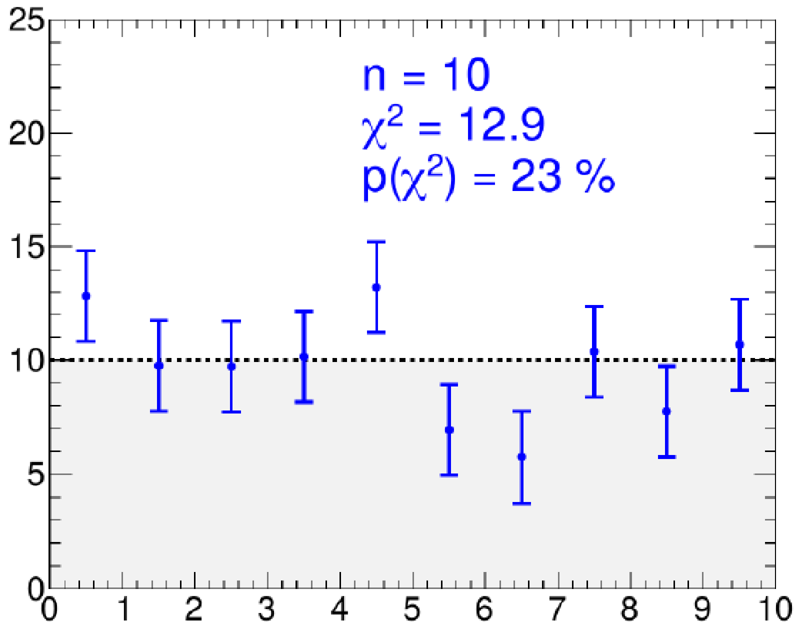




Histogram Chi-squared

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)



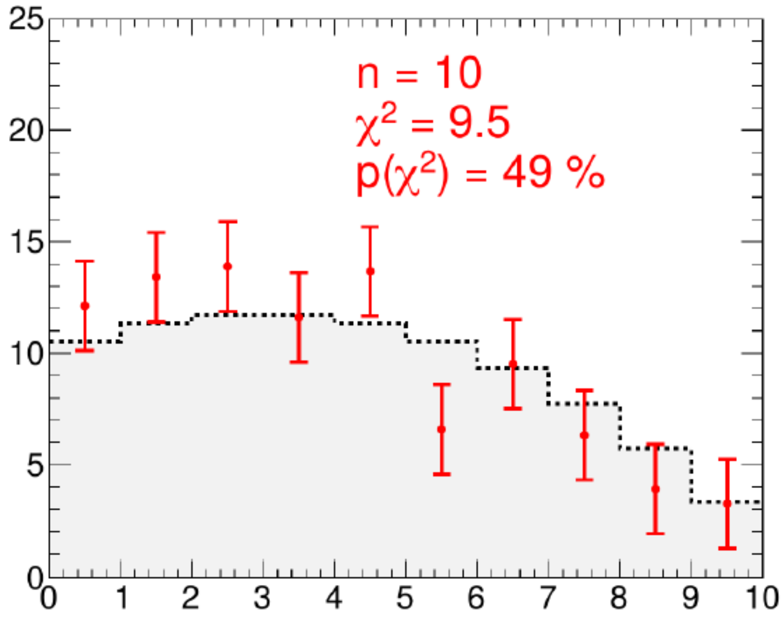
BLUE histogram vs. flat reference

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

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 vs. flat reference

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

RED histogram vs. flat reference

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

RED histogram vs. correct reference

$\chi^2 = 9.5$, $p(\chi^2=9.5, n=10) =$ **49%** ✔

ROOT commands:

```
root [0] ROOT::Math::chisquared_cdf_c(12.9, 10)
(double) 0.22931681
root [1] ROOT::Math::chisquared_cdf_c(38.8, 10)
(double) 2.7519383e-05
```

Error Bars

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

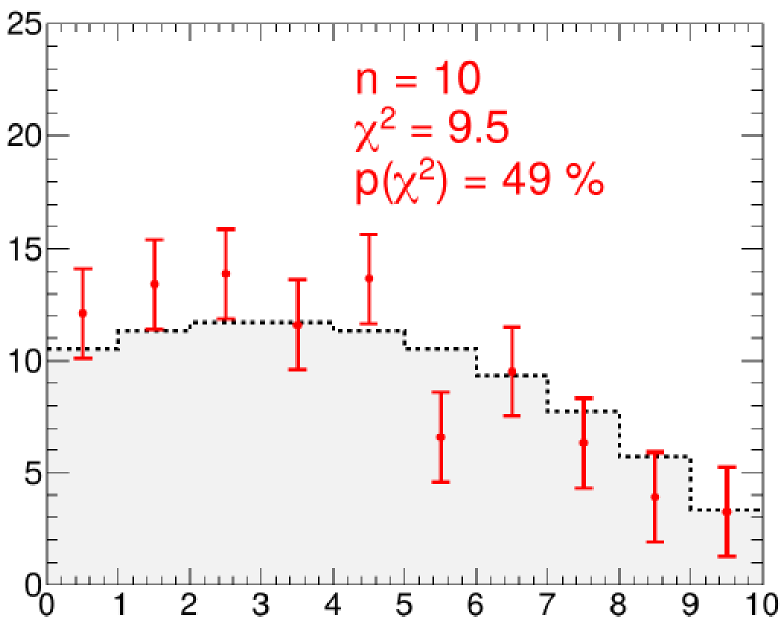
→ ***Bin central value*** ~ *mean of the bin PDF*

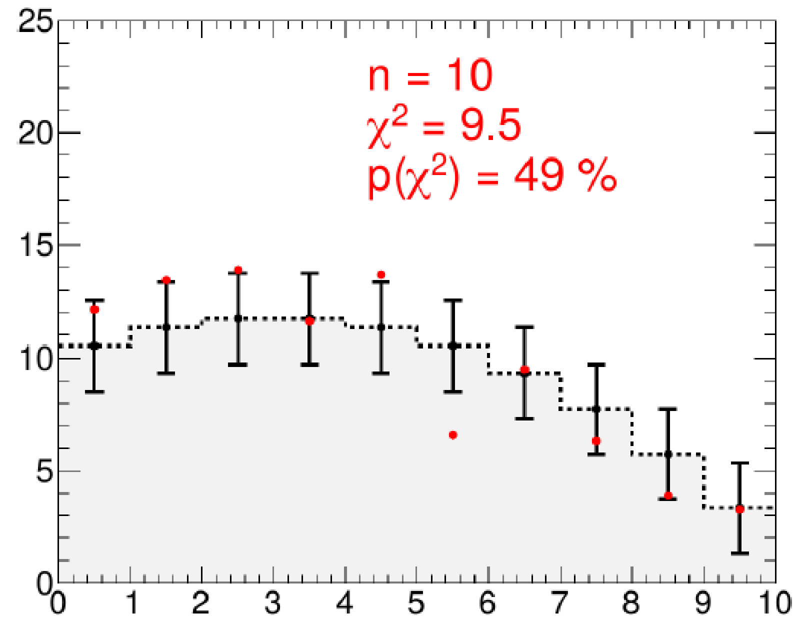
→ ***Bin uncertainty*** ~ *RMS of the bin PDF*

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**.





Example analyses

Example 1: $Z \to ee$ Inclusive σ^{fid}

Phys. Lett. B 759 (2016) 601

Measurement Principle:

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

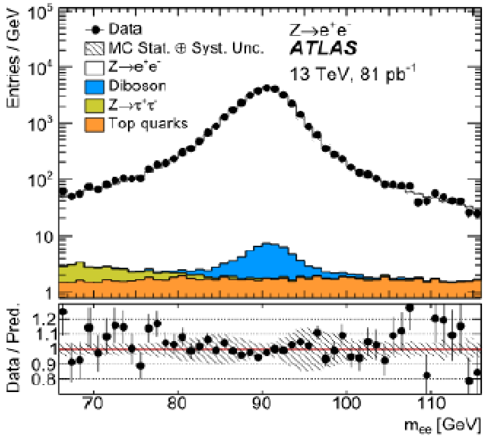
35000 ± 187

175 ± 8

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

(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

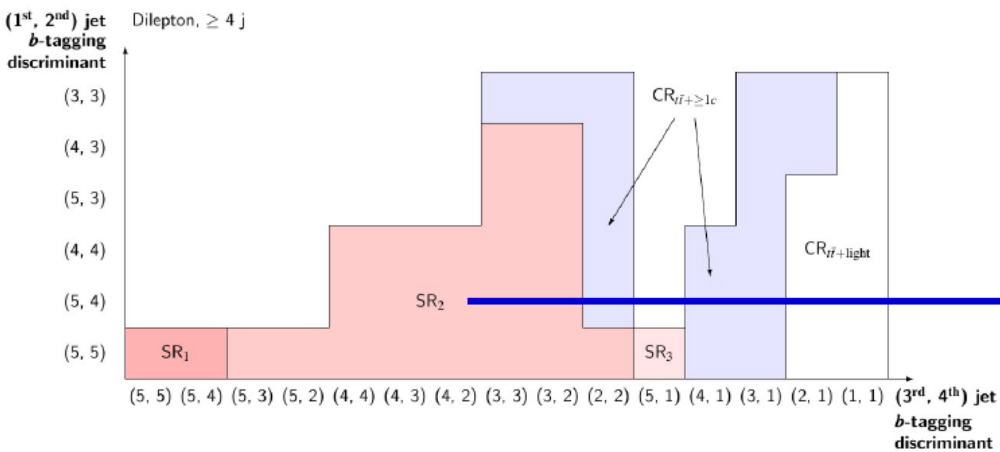
→ **Simplest possible example** in several ways (from the Statistics point of view!)

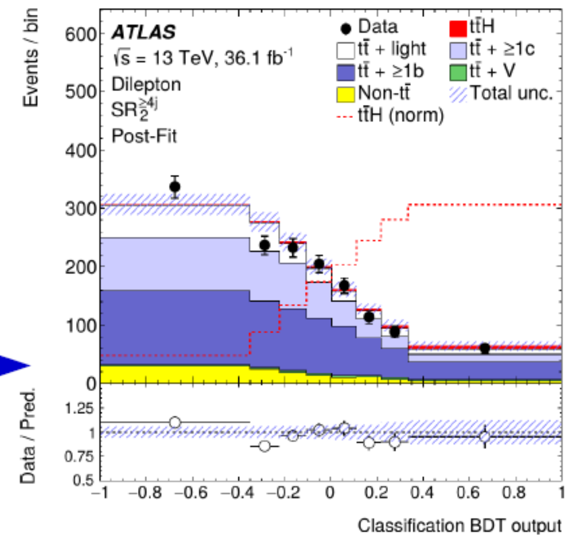
→ **"Single bin counting"** : only data input is $\mathbf{n}_{data}$.

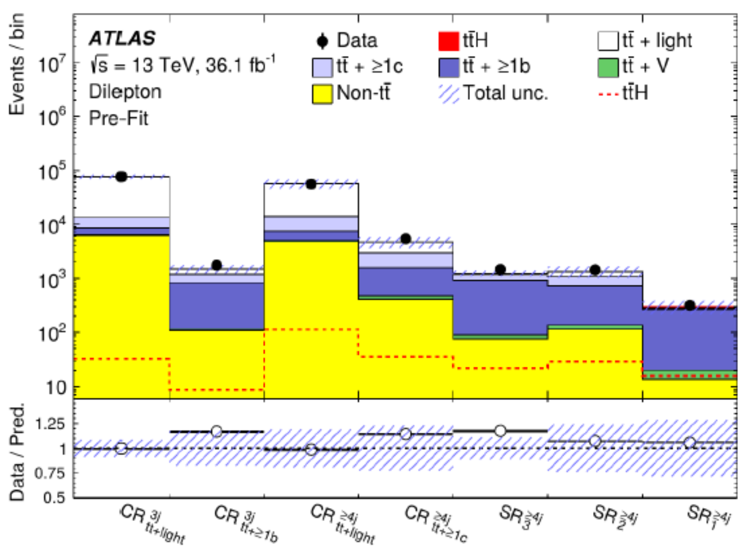
Example analyses

Example 2: ttH→bb

arXiv:1712.008895







Event counting in different regions:

Multiple-bin counting

Lots of information available

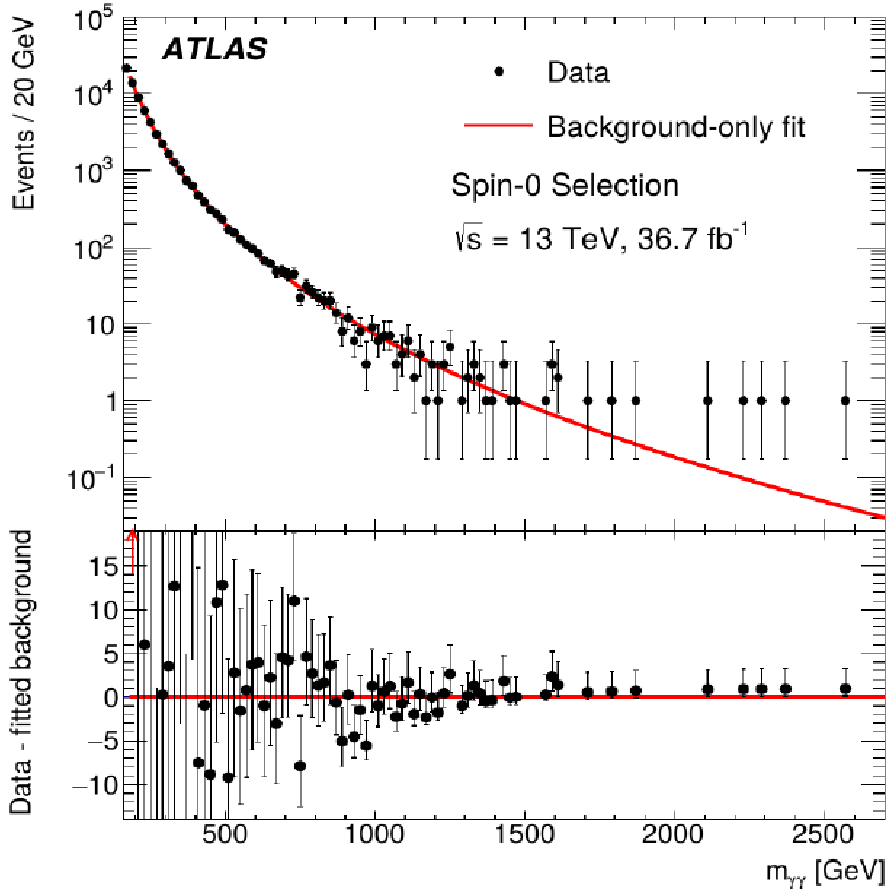
→ Potentially higher sensitivity

→ How to make optimal use of it ?

Example analyses

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

→ No binning effects
→ Use all available information

→ How to describe the shapes ?

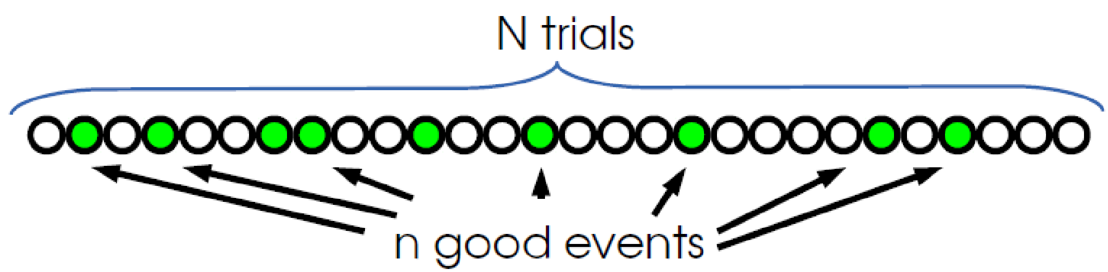
Counting events

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

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

Mean = **λ**

Variance = **λ** ⇒ **RMS = √λ**

For n expected events, the uncertainty is √n

Rare processes ?

HEP : almost always use Poisson distributions. Why ?

ATLAS :

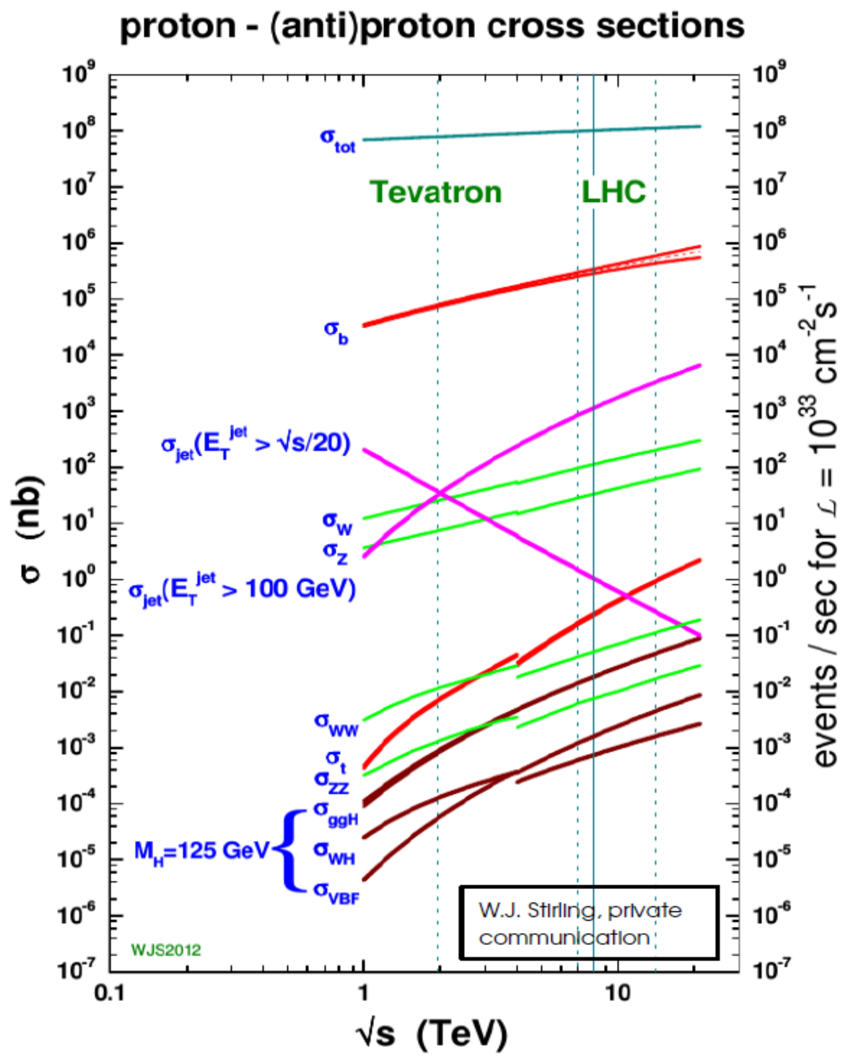
- **Event rate ~ 1 GHz**
  ($L \sim 10^{34}\ cm^{-2}s^{-1} \sim 10\ nb^{-1}/s$, $\sigma_{tot} \sim 10^{8}$ nb,)
- **Trigger rate ~ 1 kHz**
  (Higgs rate ~ **0.1 Hz**)

$\Rightarrow p \sim 10^{-6} \ll 1$ ($p_{H\to\gamma\gamma} \sim 10^{-13}$)

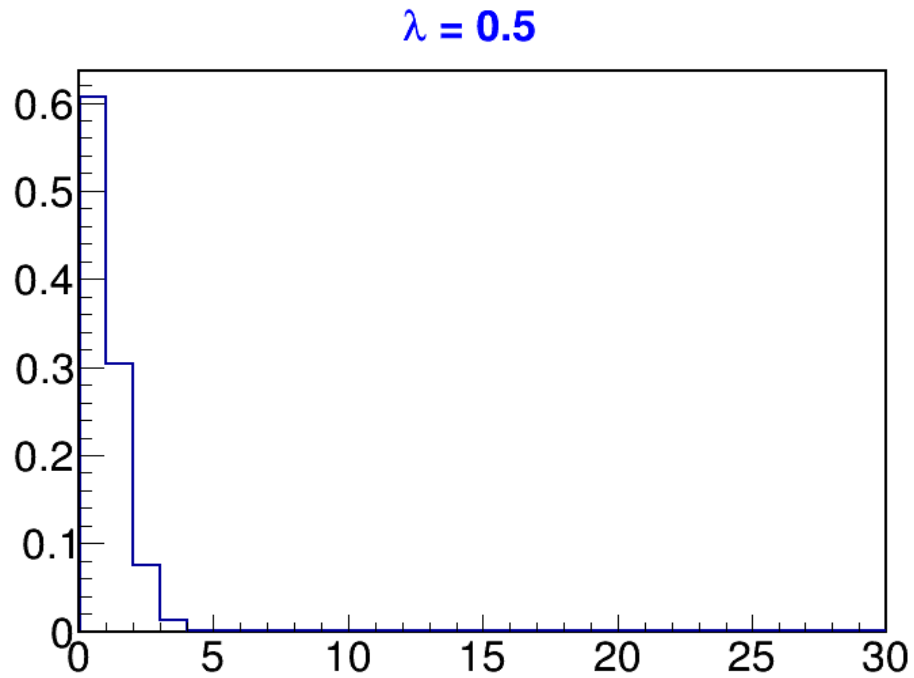
A day of data: $N \sim 10^{14} \gg 1$

⇒ **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

Mean = λ

Variance = λ

σ = √λ

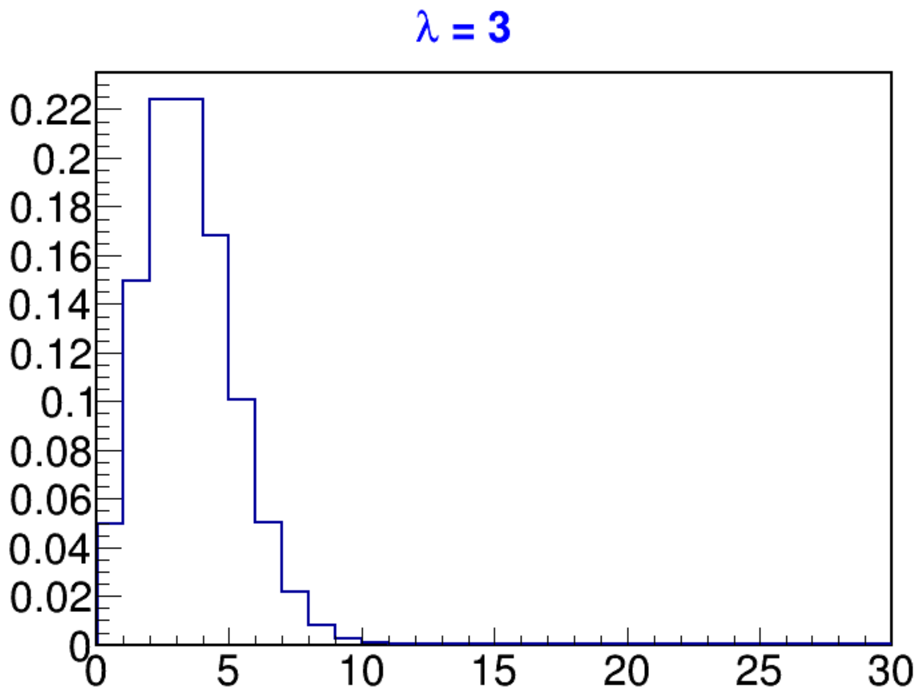
- **Discrete distribution** (positive integers only), **asymmetric for small λ**
- Typical variation (RMS) of n events is **√n**
- Central limit theorem : becomes **Gaussian for large λ** :

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

Poisson distributions

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

λ : expected number of events



Mean = λ

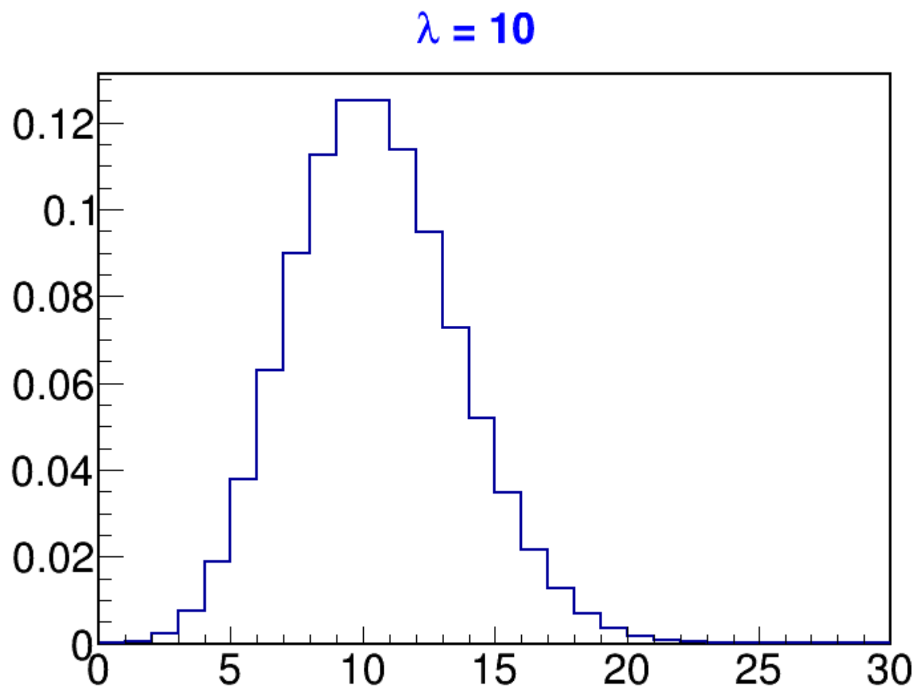
Variance = λ

σ = √λ

- **Discrete distribution** (positive 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

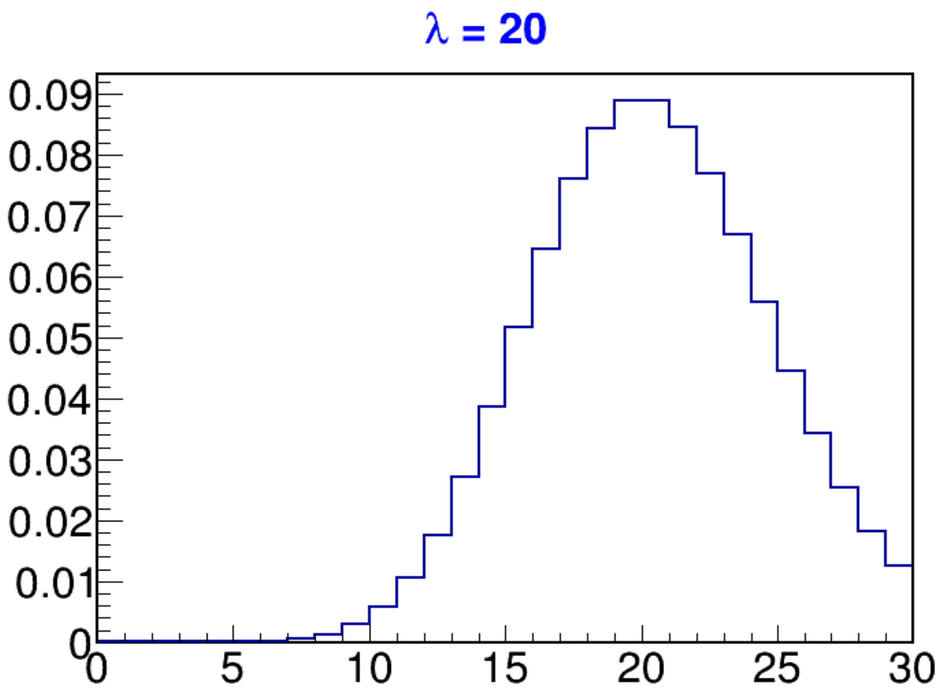
- **Discrete distribution** (positive integers only), **asymmetric for small λ**
- Typical variation (RMS) of n events is **√n**
- Central limit theorem : becomes **Gaussian for large λ** :

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

Poisson distributions

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

λ : expected number of events



Mean = λ

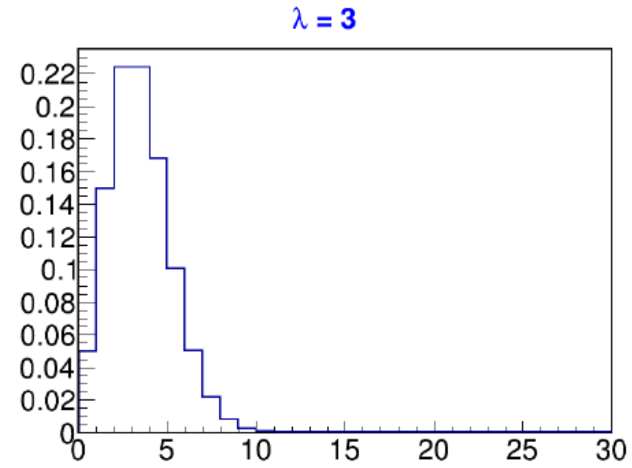
Variance = λ

σ = √λ

- **Discrete distribution** (positive 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: 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->ee inclusive σ^{fid}

Measurement Principle:

35000 ± (√35000 = 187)

175 ± 8

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

(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

Statistical uncertainty: Derived from assumption that n_{data} is ~ Poisson(S+B)

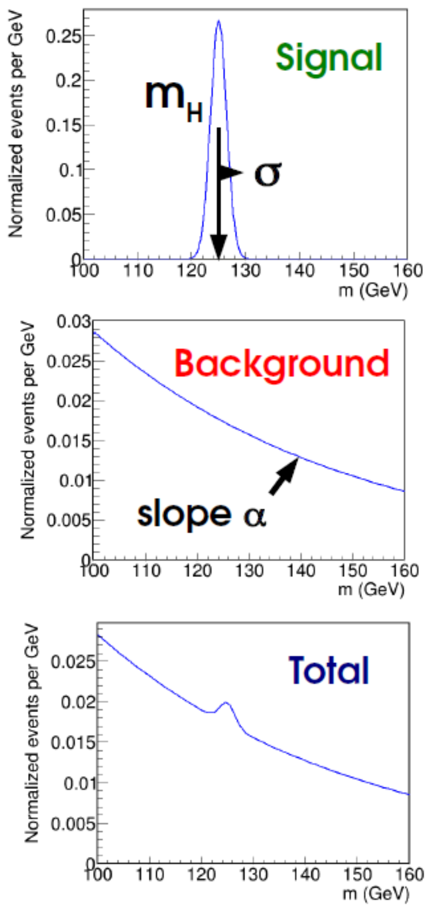
Systematics: more on this in Lecture 3

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

⇒ **Total PDF for a single event**:

Expected yields : **S, B**

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

⇒ **Total PDF for a dataset**

Probability to observe n events

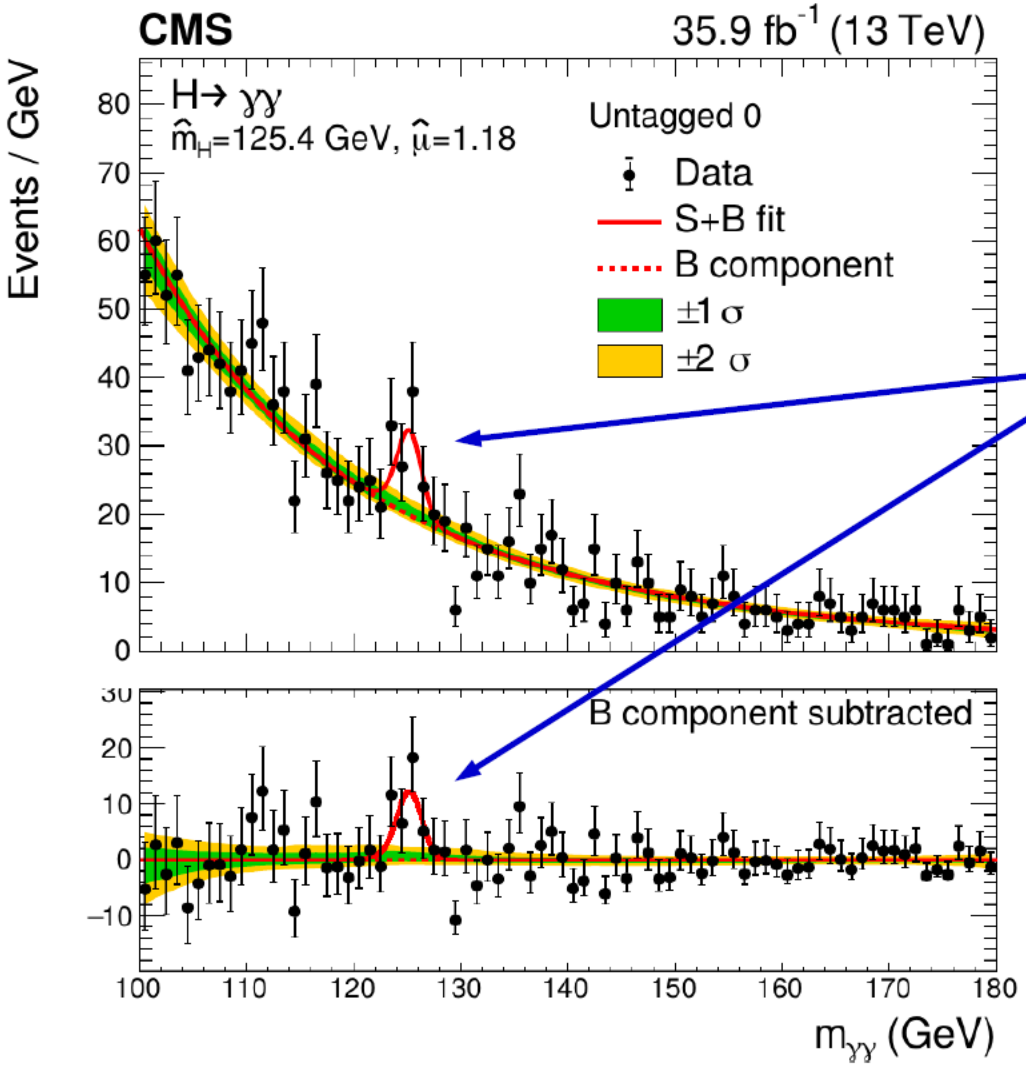
Probability to observe the value m_i

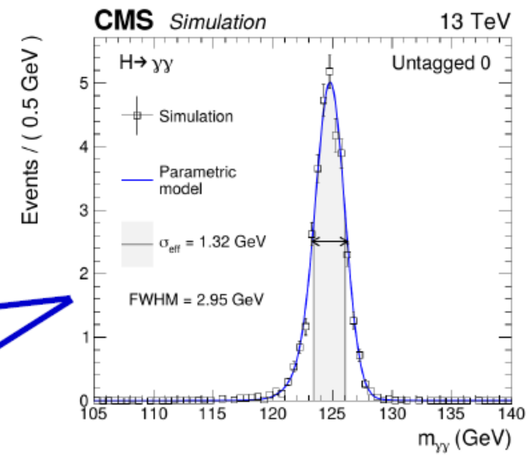
$$P(\{m_i\}_{i=1\ldots n}) = e^{-(S+B)} \frac{(S+B)^n}{n!} \prod_{i=1}^{n} \frac{S}{S+B} G(m_i; m_H, \sigma) + \frac{B}{S+B} \alpha\, e^{-\alpha m_i}$$

Unbinned shape analysis

H→γγ

JHEP 11 (2018) 185





Binned shape analysis

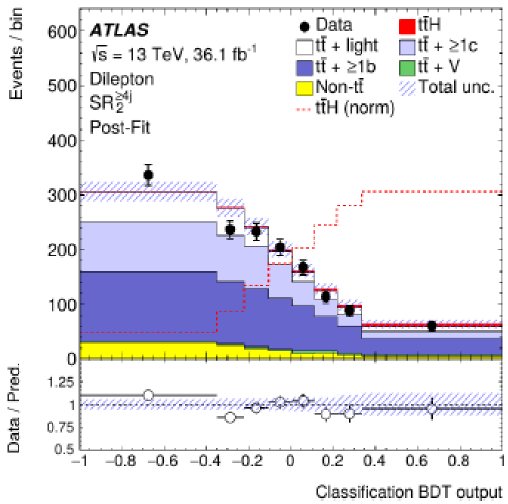
Instead of using $m_1 \ldots m_n$ directly, can build a ***histogram*** $n_1 \ldots n_N$.

→ N_{bins} : number of bins

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

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

Poisson distribution in each bin



N_{bins}=1: Counting analysis

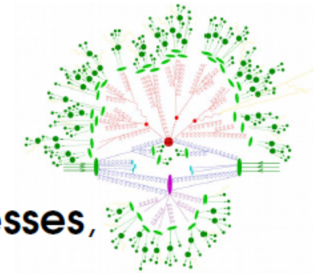
$N_{bins}\rightarrow\infty$: Unbinned shape analysis (the fractions become PDF values)

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.

How to describe data



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

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

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