

Statistics and Data Analysis (HEP at LHC)

□ Statistical basics for physics

- Random processes
- Probability distributions

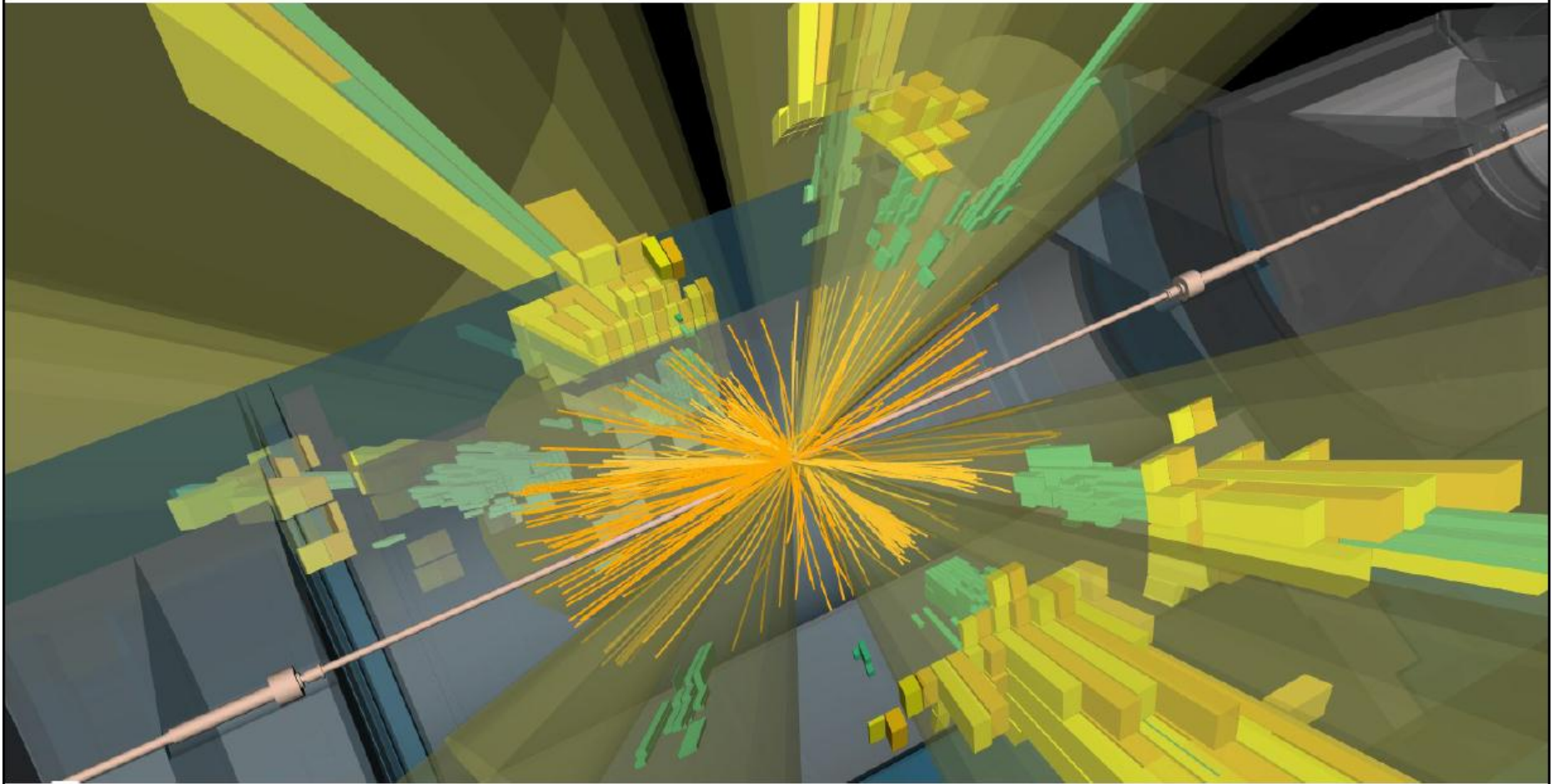
□ Describing physics measurements

- Binned and unbinned data
- Model parameters

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

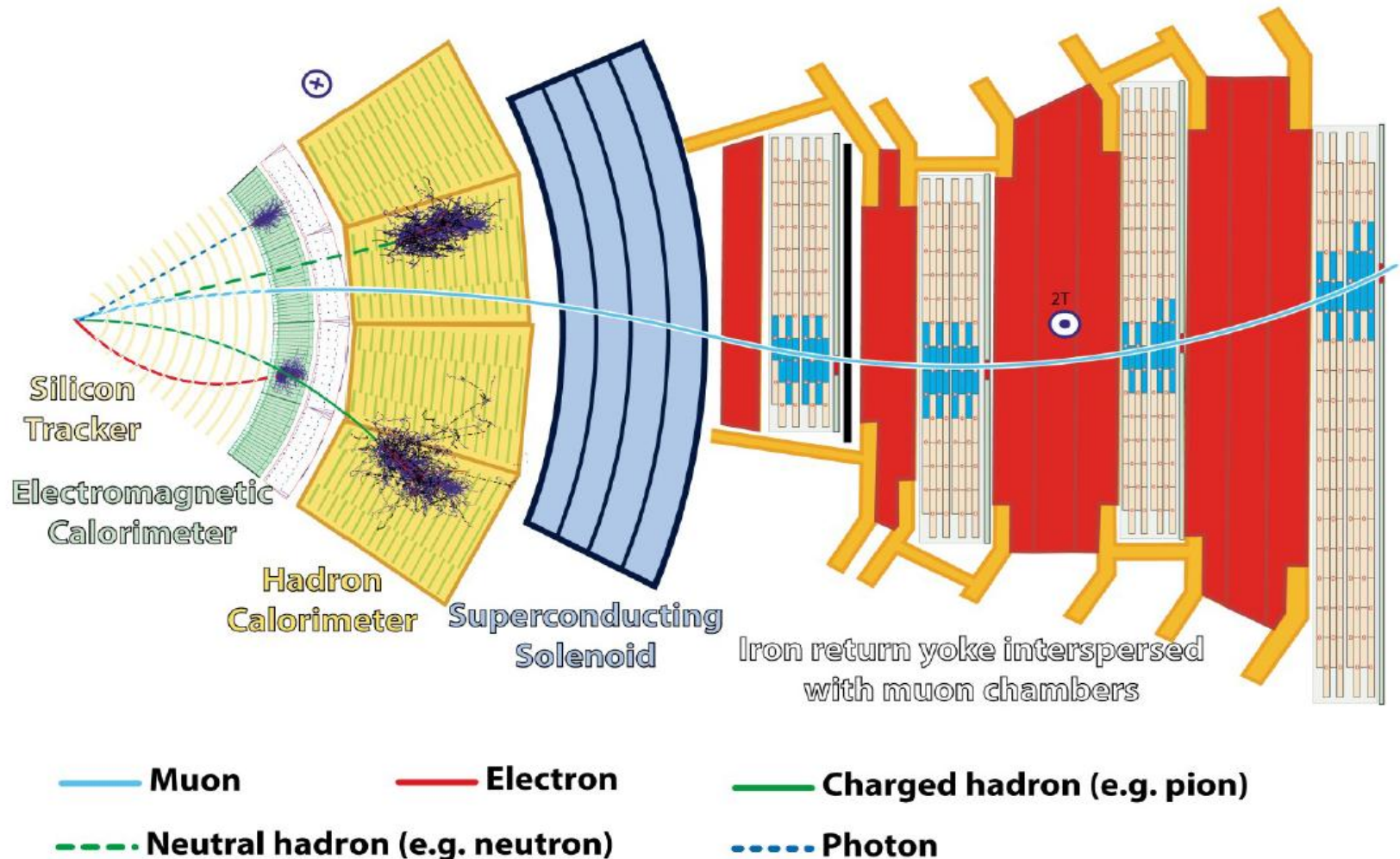
Randomness in High Energy Physics

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

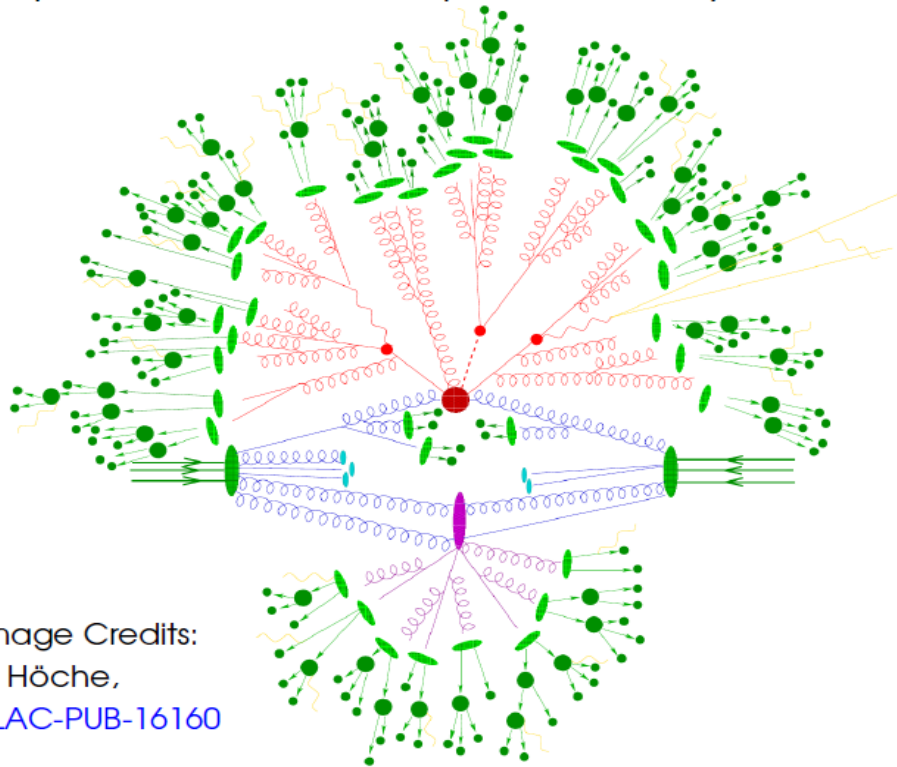


Image Credits:
S. Höche,
SLAC-PUB-16160

Randomness involved in all stages

→ **Classical** randomness: detector response

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

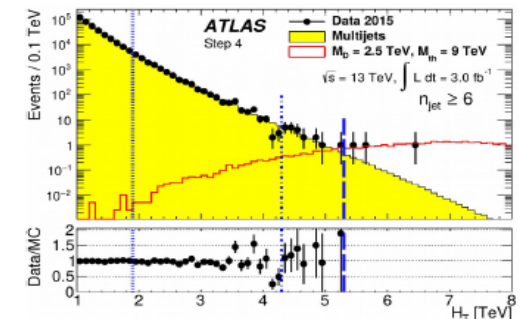
Hard scattering

PDFs, Parton shower, Pileup

Decays

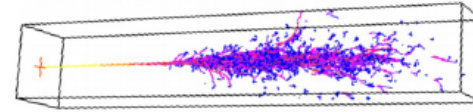
Detector response

Reconstruction

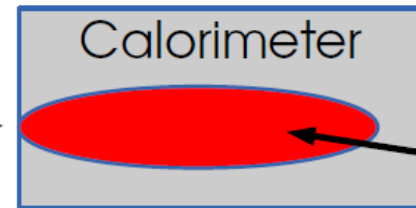


Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

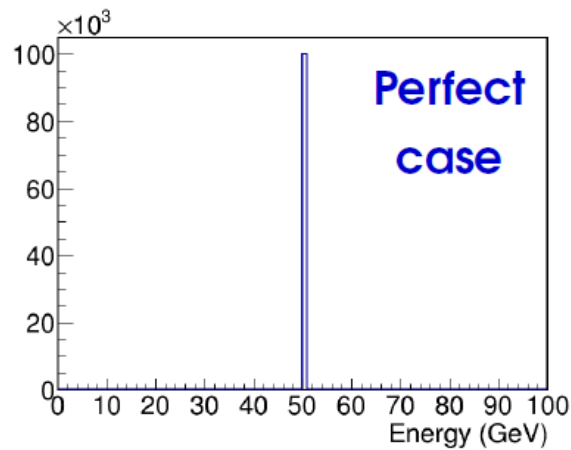


γ



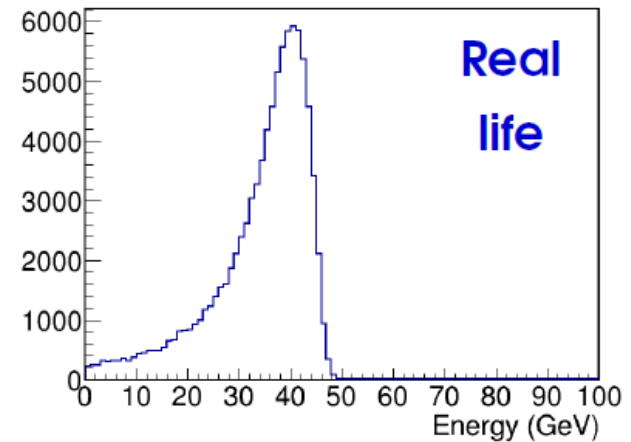
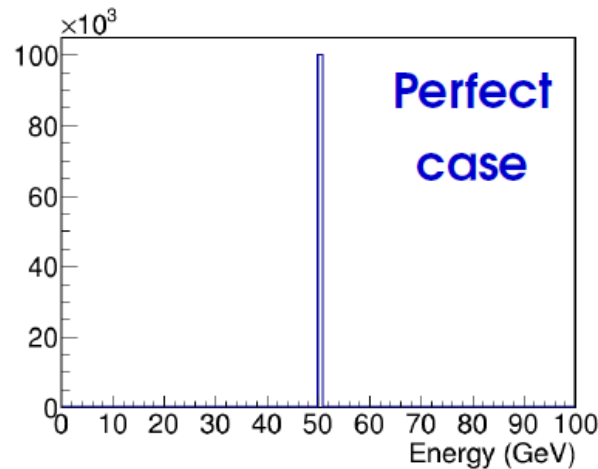
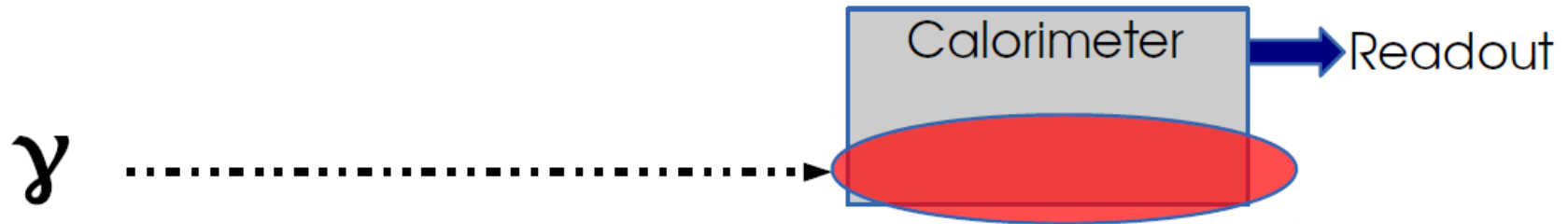
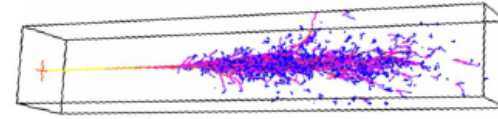
Readout

Energy
deposition



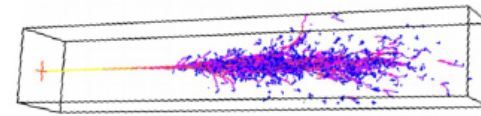
Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter



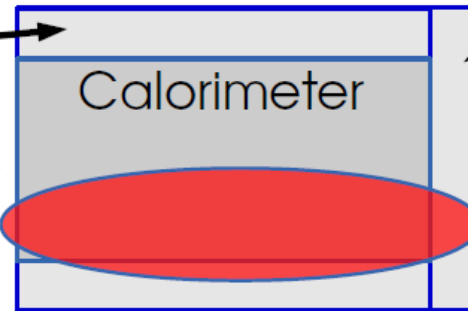
Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

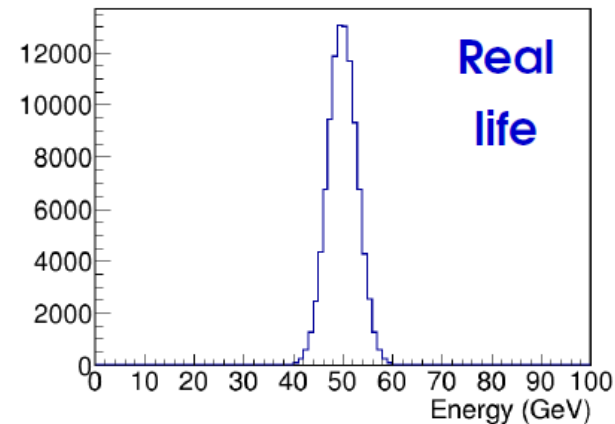
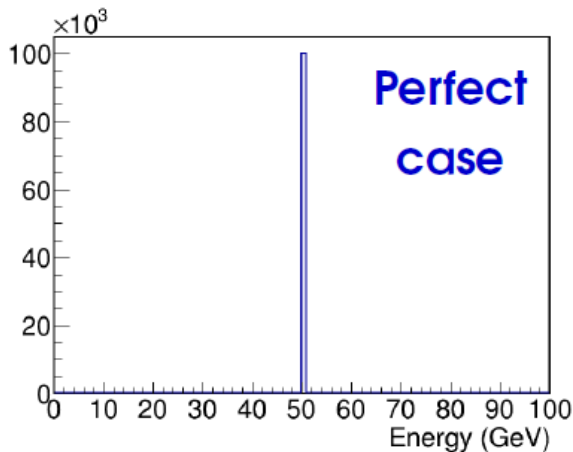


Measure leakage behind calorimeter

Measure leakage
into neighboring cells

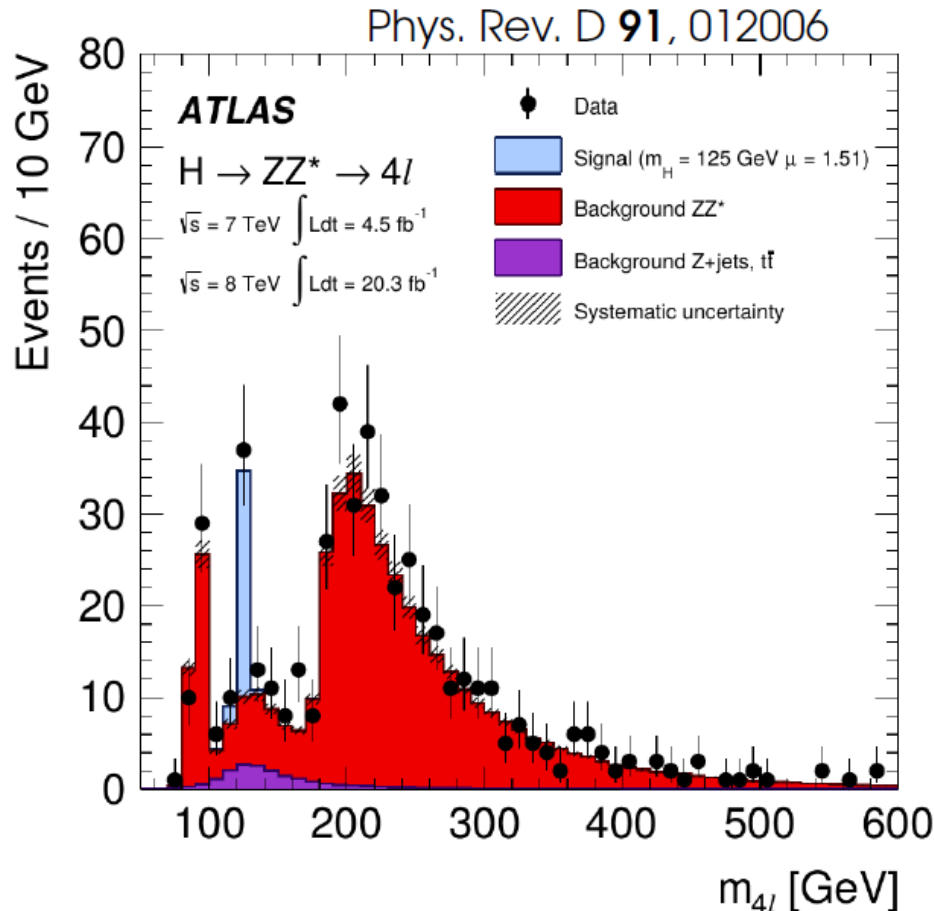


γ



Cannot predict the measured value for a given event
 $\Rightarrow$ **Random process** $\Rightarrow$ **Need a probabilistic description**

Quantum randomness: $H \rightarrow ZZ^* \rightarrow 4l$



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

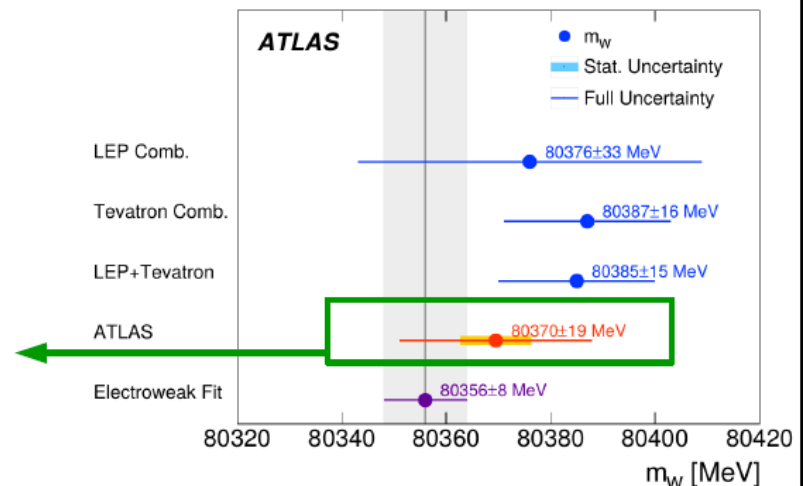
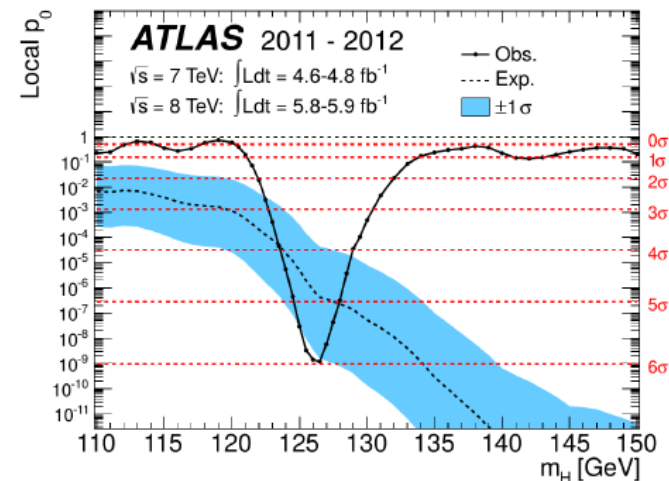
“Will I get an event today ?” → only **probabilistic** answer

Randomness in Physics

Questions with probabilistic answers:

- **Is my Higgs-like excess just a background fluctuation?**
 $\rightarrow$ associated with prob $\sim 10^{-9}$ (by now $\sim 10^{-24}$)
 $\Rightarrow$ 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

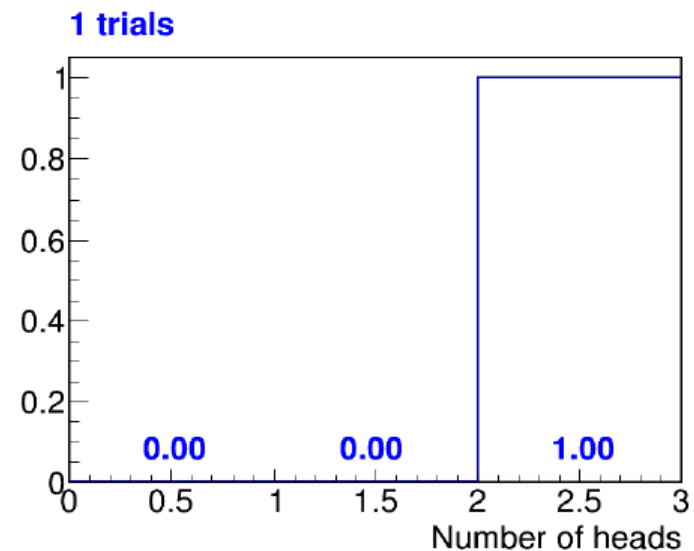
→ Fractions of events in each bin i
converge to a limit p_i

Probability distribution :

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

Properties

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



Probability distributions

Probabilistic treatment of possible outcomes

⇒ **Probability Distribution**

Example: two-coin toss

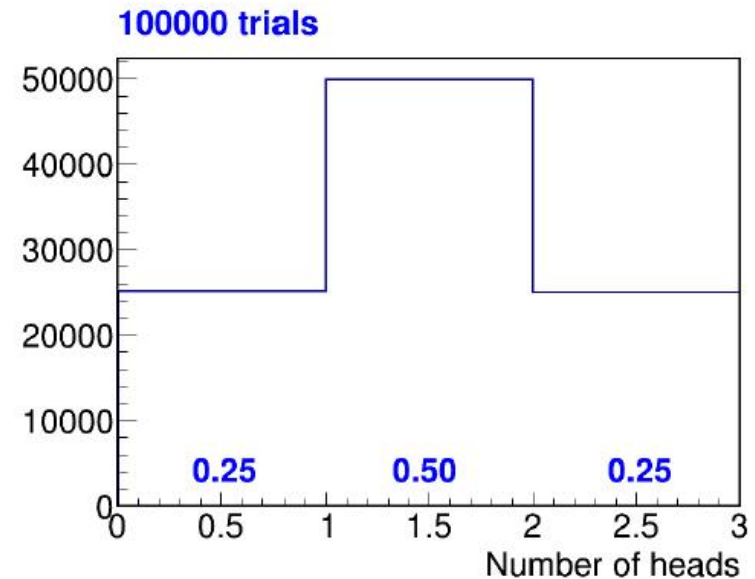
→ Fractions of events in each bin i
converge to a limit p_i

Probability distribution :

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

Properties

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



Probability distributions

Probabilistic treatment of possible outcomes

⇒ **Probability Distribution**

Example: two-coin toss

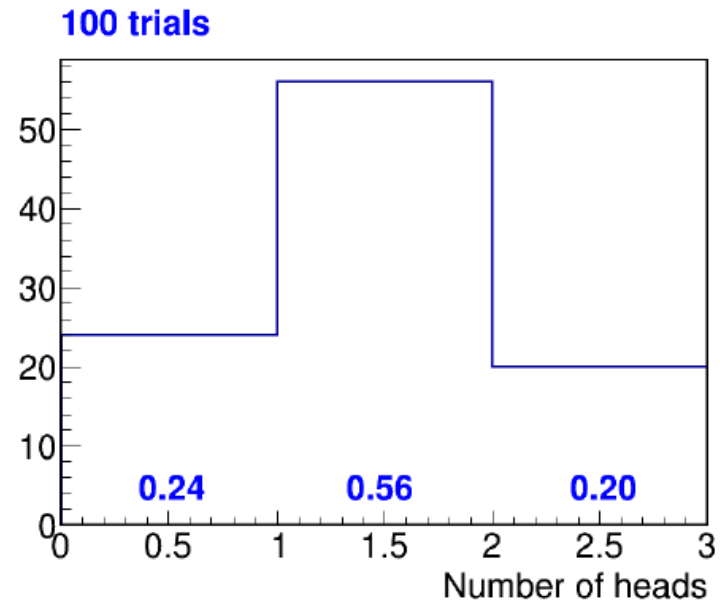
→ Fractions of events in each bin i
converge to a limit p_i

Probability distribution :

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

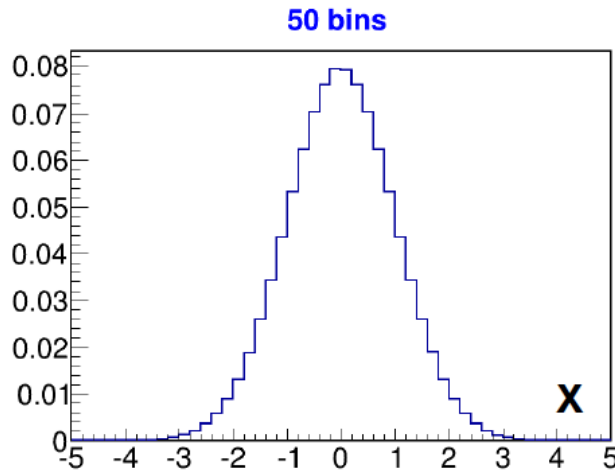
Properties

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



Continuous Variables: PDFs

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



Bin size $\rightarrow 0$:

Probability distribution function $P(x)$

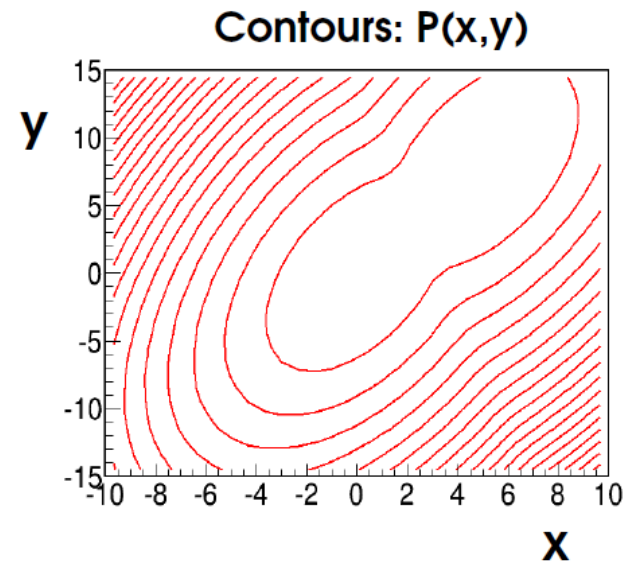
$\rightarrow P(x) > 0, \int P(x) dx = 1$

$\rightarrow$ High values $\Leftrightarrow$ high chance to get
a measurement here

Generalizes to **multiple variables** :

$\rightarrow P(x,y) > 0$

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



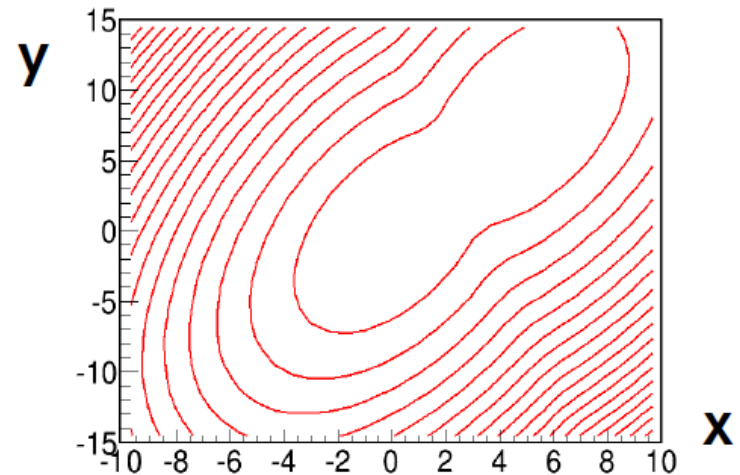
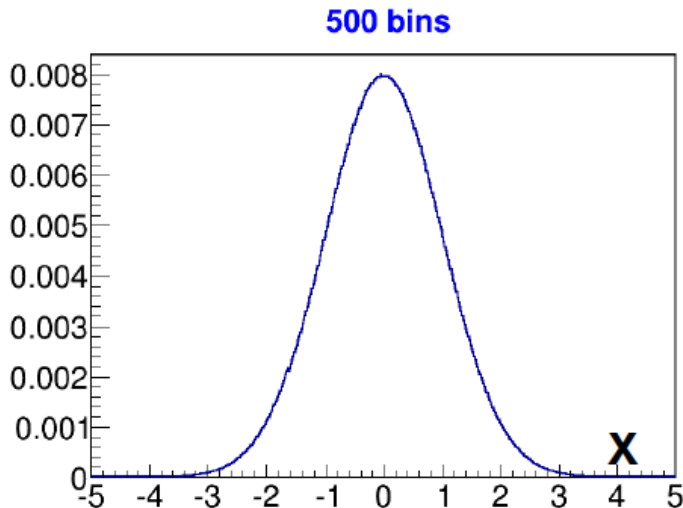
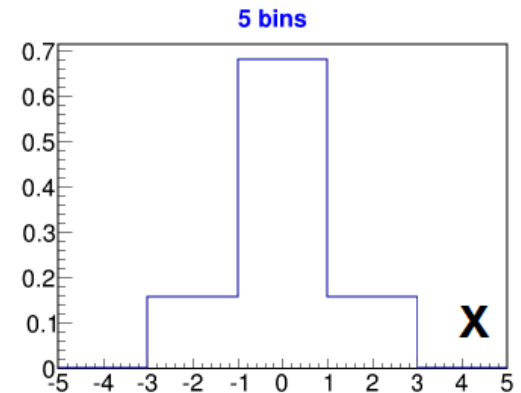
Random Variables

$X, Y, \dots$ are **Random Variables** (continuous or discrete), a.k.a. **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

$E(X) = \langle X \rangle$: **Mean** of X – expected outcome on average over many measurements

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

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

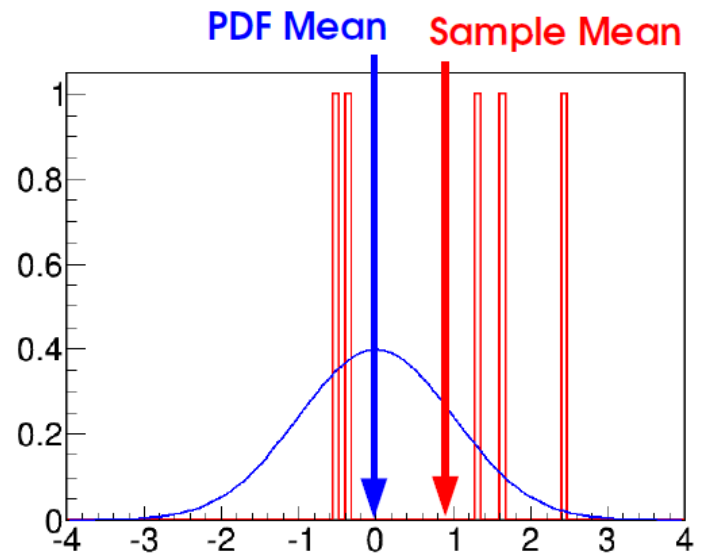
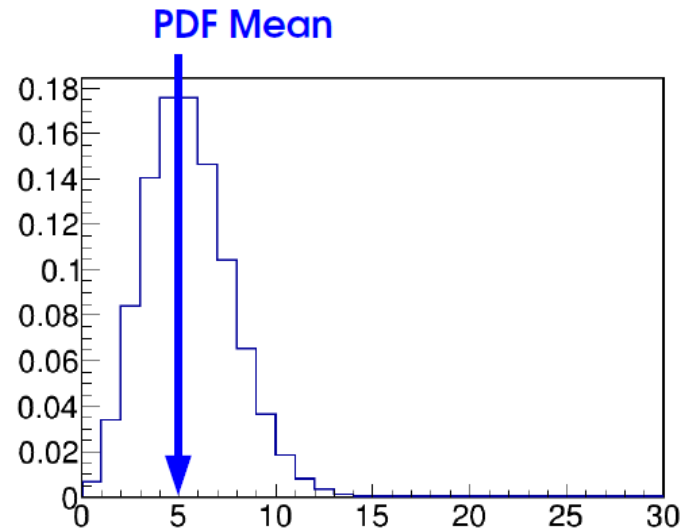
→ Property of the **PDF**

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

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

→ Property of the **sample**

→ approximates the PDF mean.



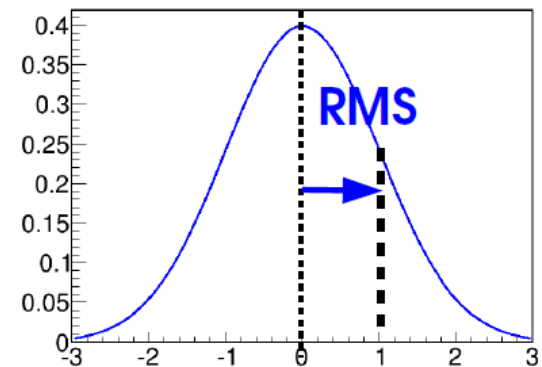
PDF properties: (co)variance

Variance of X:

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

→ Average square of deviation from mean

→ $\text{RMS}(X) = \sqrt{\text{Var}(X)} = \sigma_X$ *standard deviation*

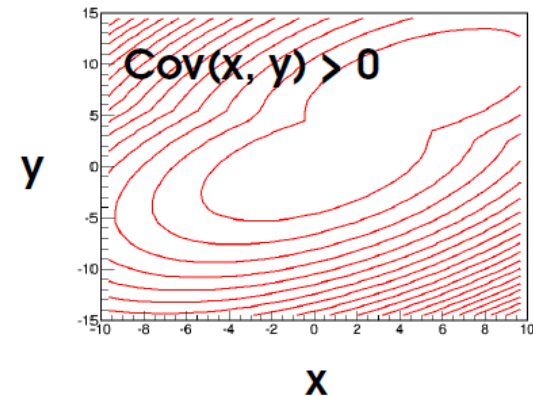


Can be approximated by *sample variance*:

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

Covariance of X and Y:

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

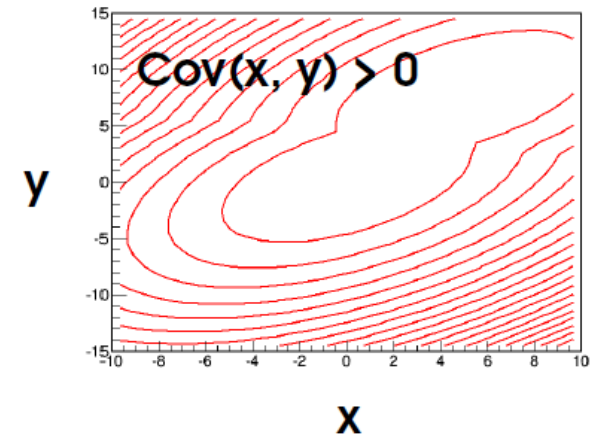
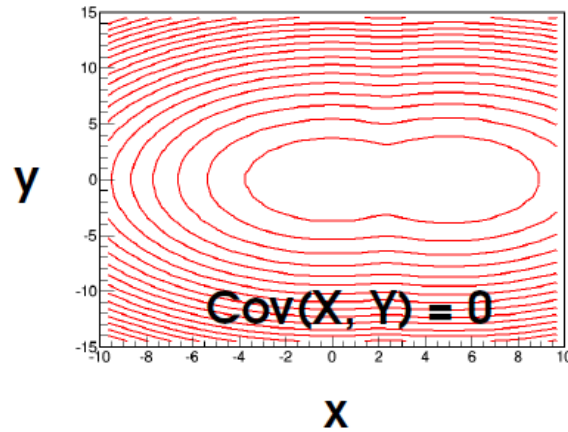
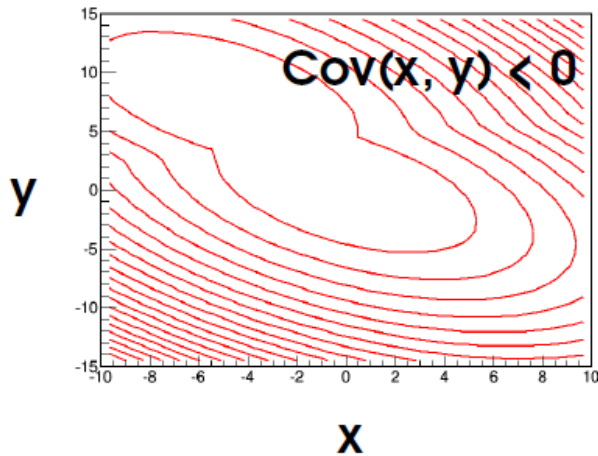


→ Large if variations of X and Y are “synchronized”

Correlation coefficient

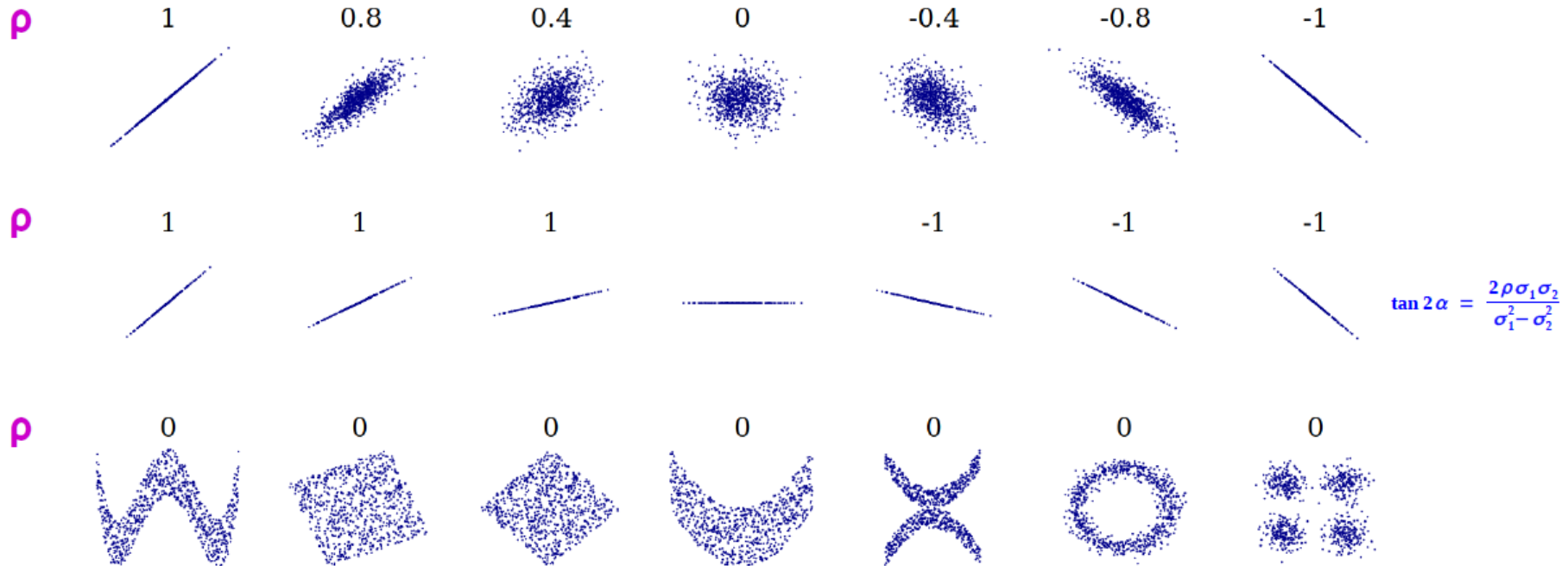
$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} \quad -1 \leq \rho \leq 1$$

PDF properties: (co)variance



„Linear” vs. „non-linear” correlations

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



Source: [Wikipedia](#)

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

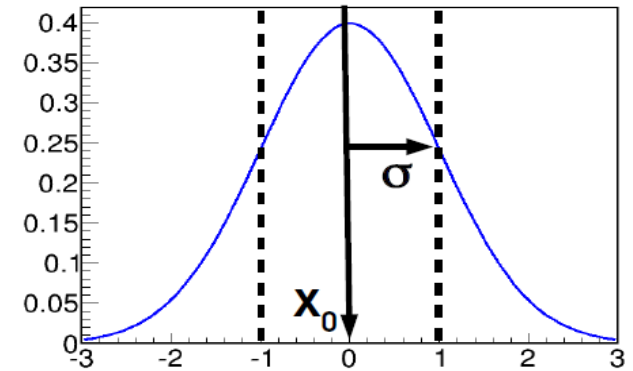
Gaussian PDF

Gaussian distribution:

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

→ Mean : X_0

→ Variance : σ^2 ($\Rightarrow$ RMS = σ)



Generalize to **N** dimensions:

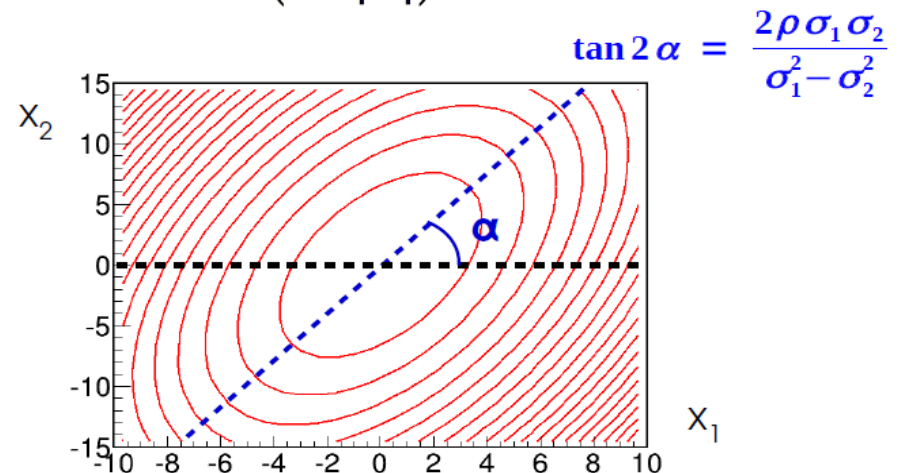
→ Mean : X_0

→ Covariance matrix :

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

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

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



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,

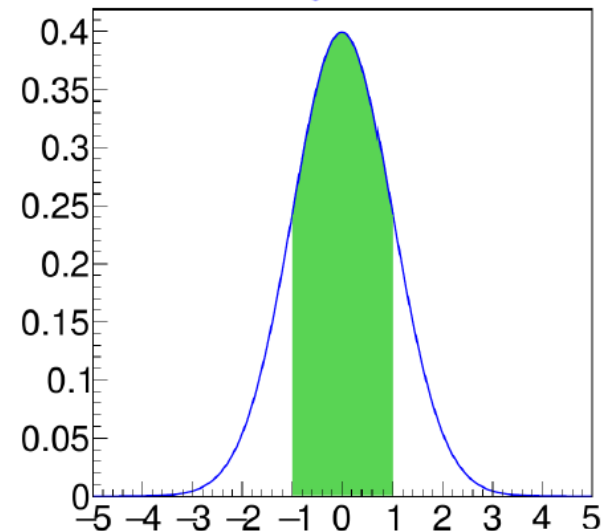
$\Phi(z)$: `ROOT::Math::gaussian_cdf(z)`

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

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



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,

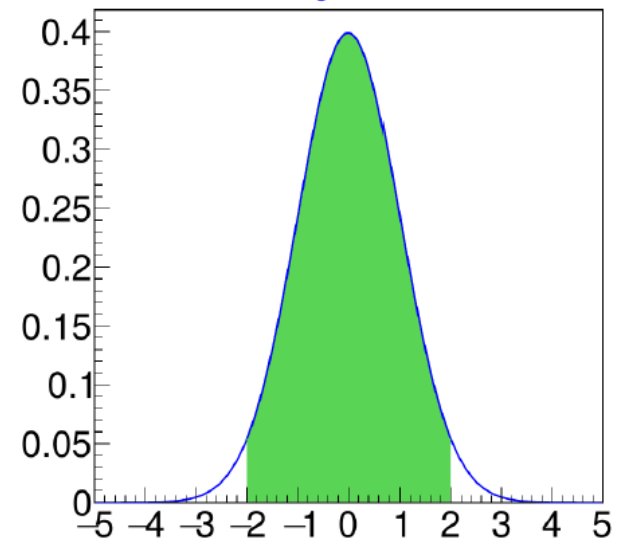
$\Phi(z)$: `ROOT::Math::gaussian_cdf(z)`

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

$P(|x - x_0| < 2\sigma) = 95.4 \%$



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,

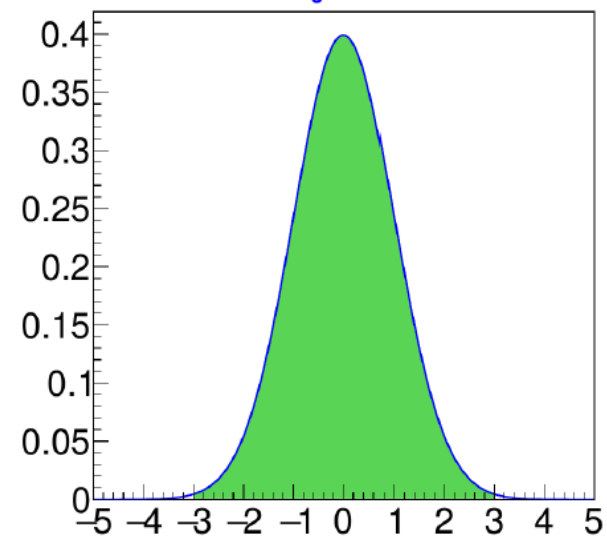
$\Phi(z)$: `ROOT::Math::gaussian_cdf(z)`

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

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

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 $\sqrt{n}$** : smaller fluctuations when averaging over many measurements

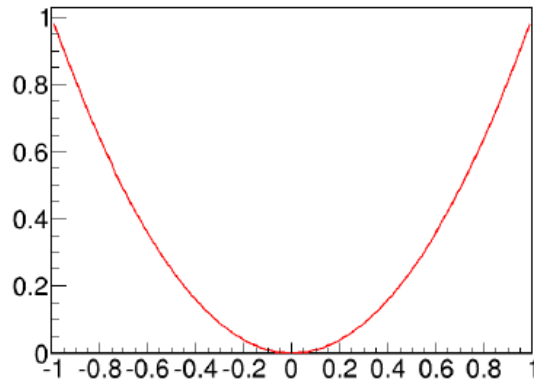
Another version,
for the sum:

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

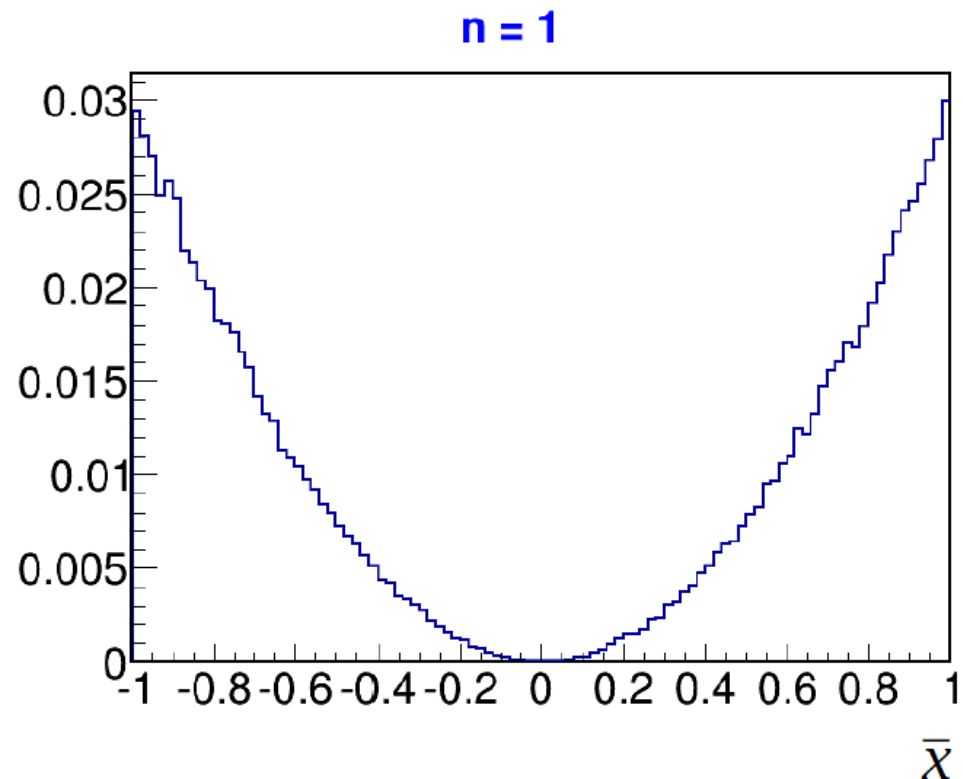
Mean scales like n , but RMS only like $\sqrt{n}$

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay $\cos \theta^*$)



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

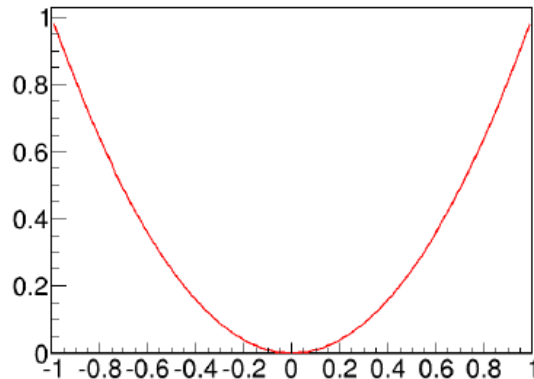


Distribution becomes Gaussian, although very non-Gaussian originally

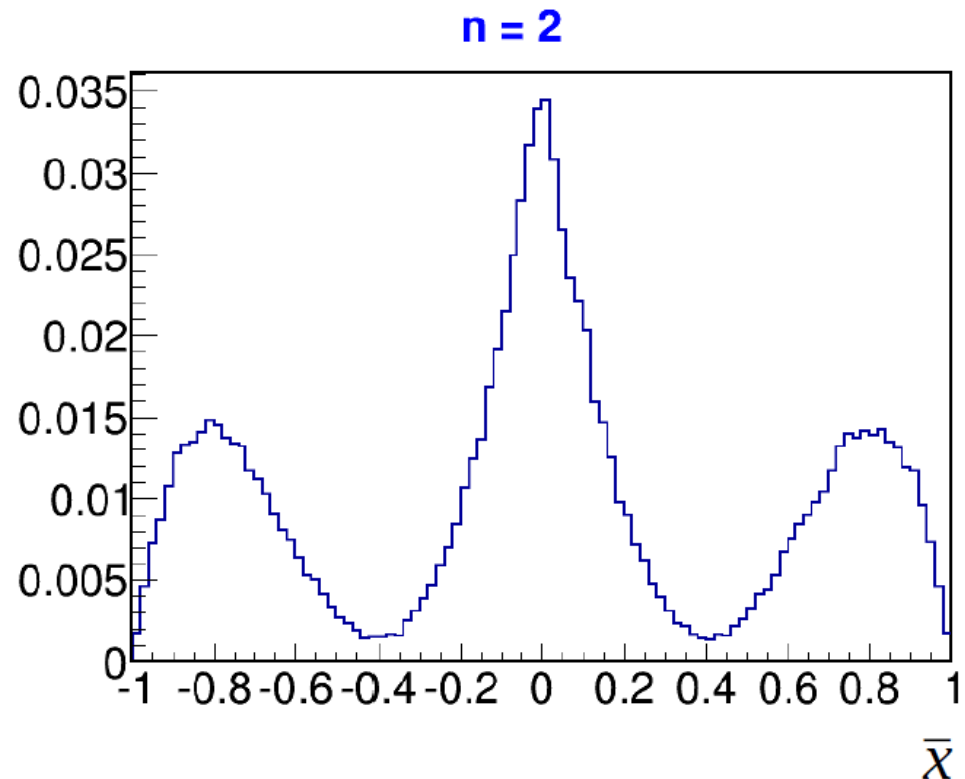
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay $\cos \theta^*$)



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

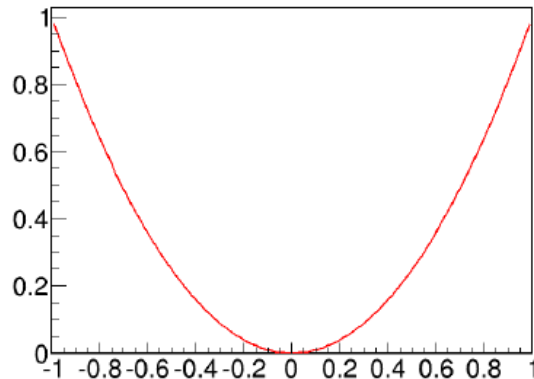


Distribution becomes Gaussian, although very non-Gaussian originally

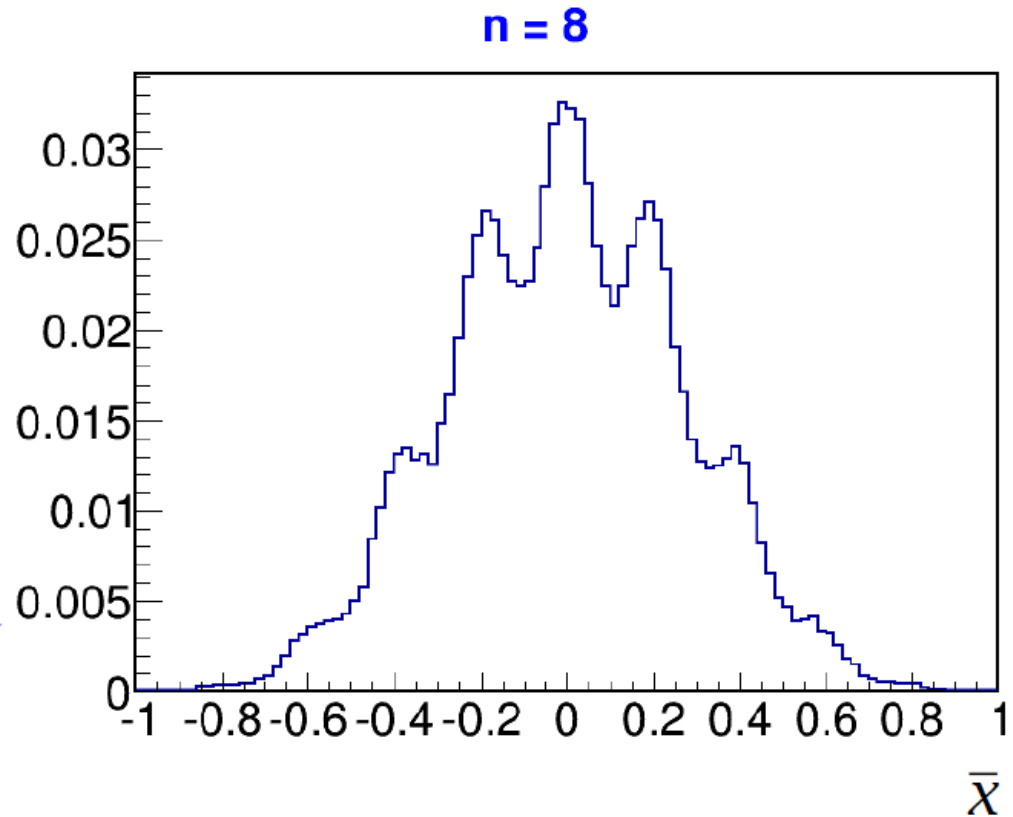
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay $\cos \theta^*$)



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

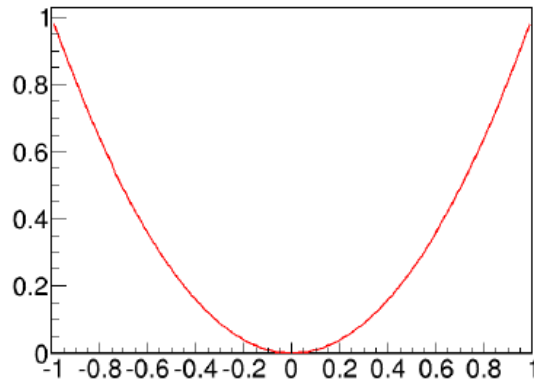


Distribution becomes Gaussian, although very non-Gaussian originally

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

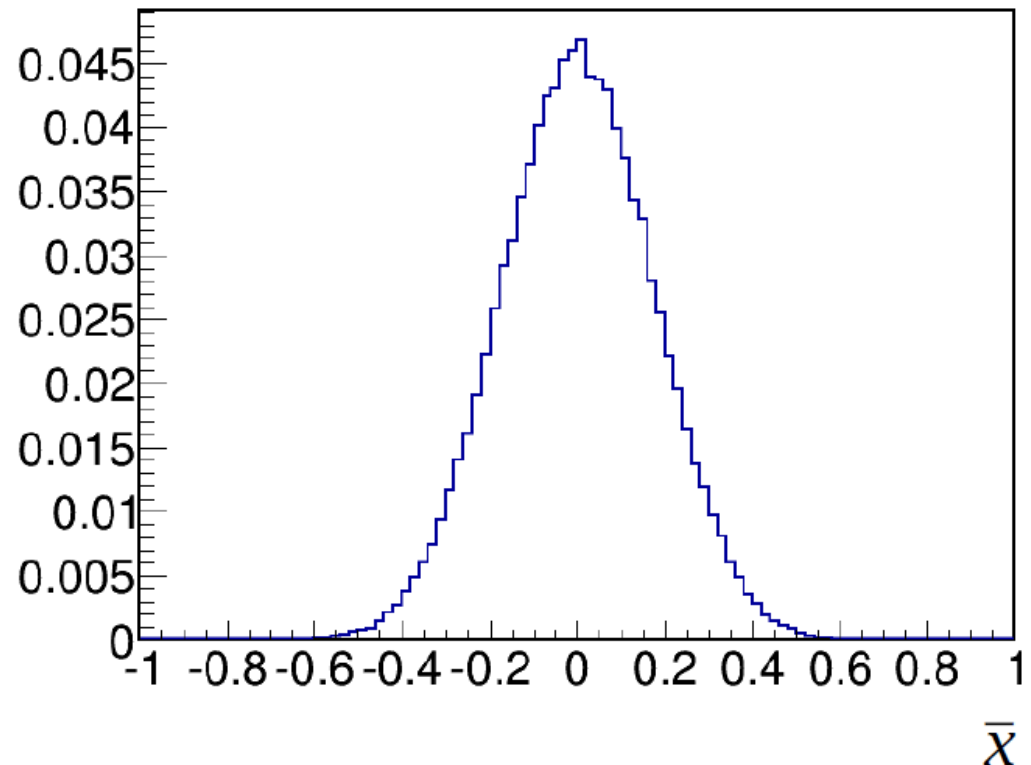
Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay $\cos \theta^*$)



n = 20

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

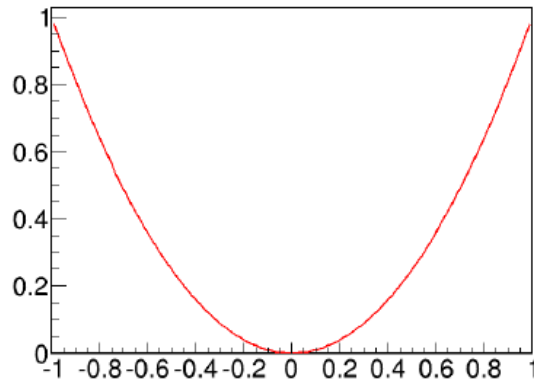


Distribution becomes Gaussian, although very non-Gaussian originally

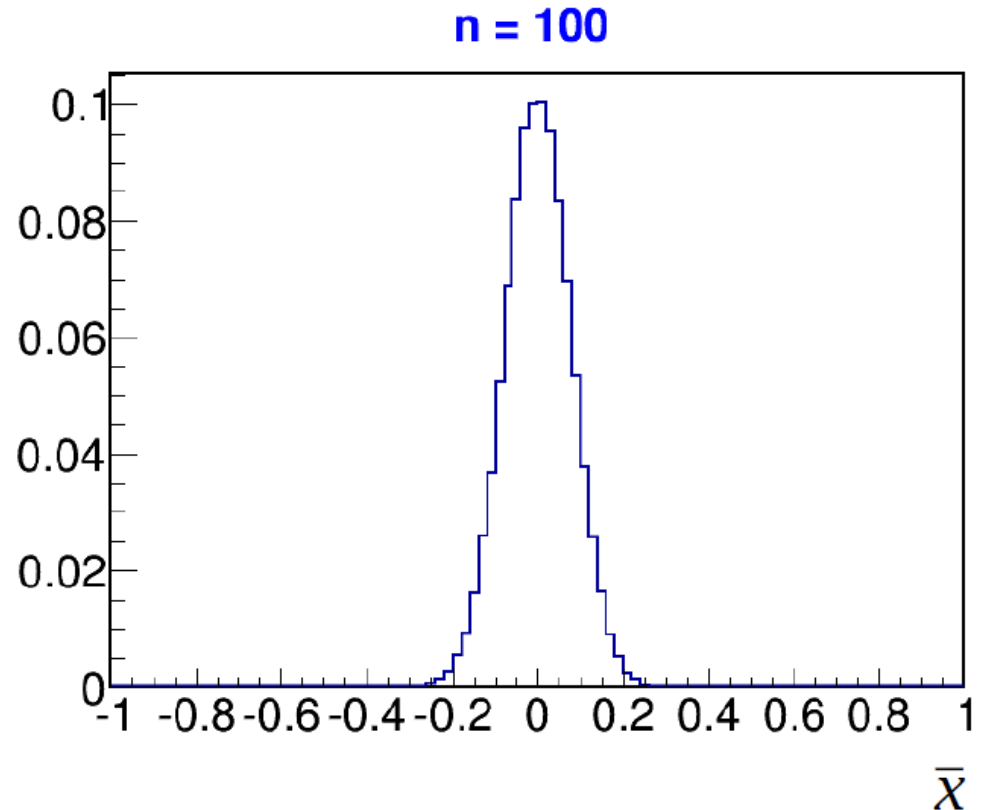
Distribution becomes narrower as expected (as $1/\sqrt{n}$)

Central Limit Theorem in action

Draw events from a parabolic distribution (e.g. decay $\cos \theta^*$)



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



Distribution becomes Gaussian, although very non-Gaussian originally

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

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 $(x_1^0, \dots, x_n^0)$

Distribution depends on n :

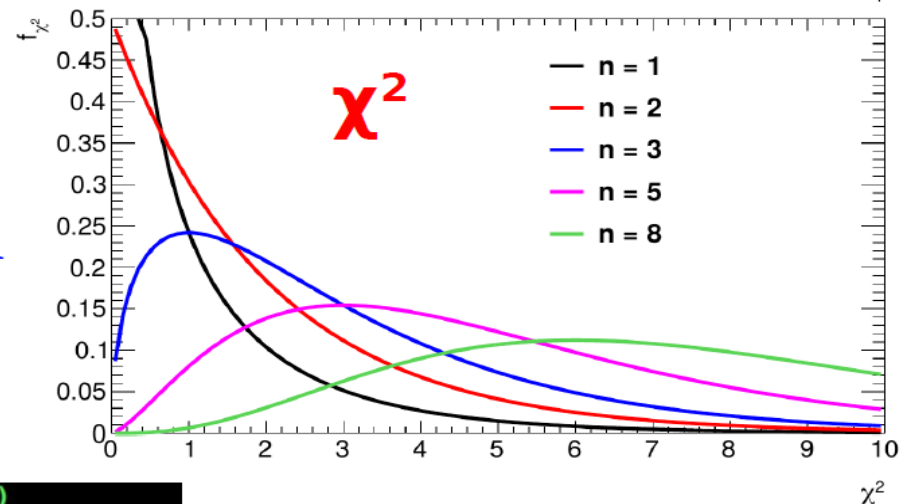
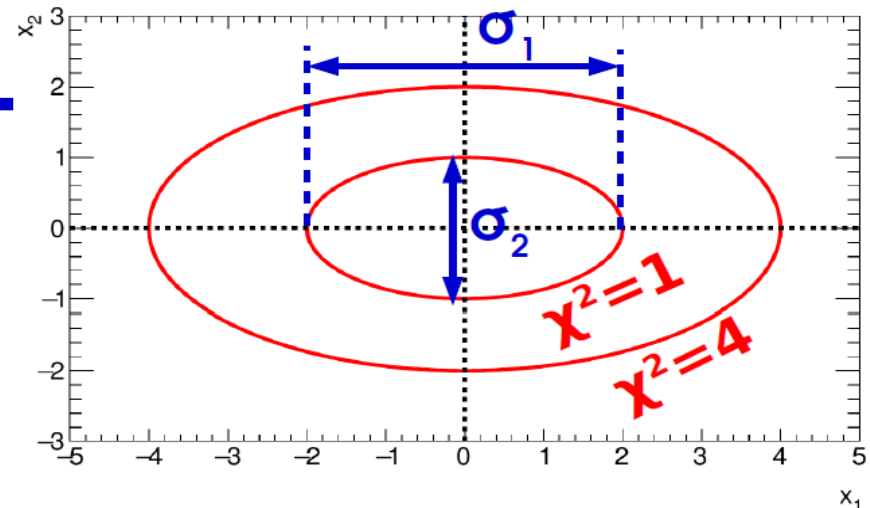
Rule of thumb: χ^2/n should be $\lesssim 1$

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 $(x_1^0, \dots, x_n^0)$

Distribution depends on n :

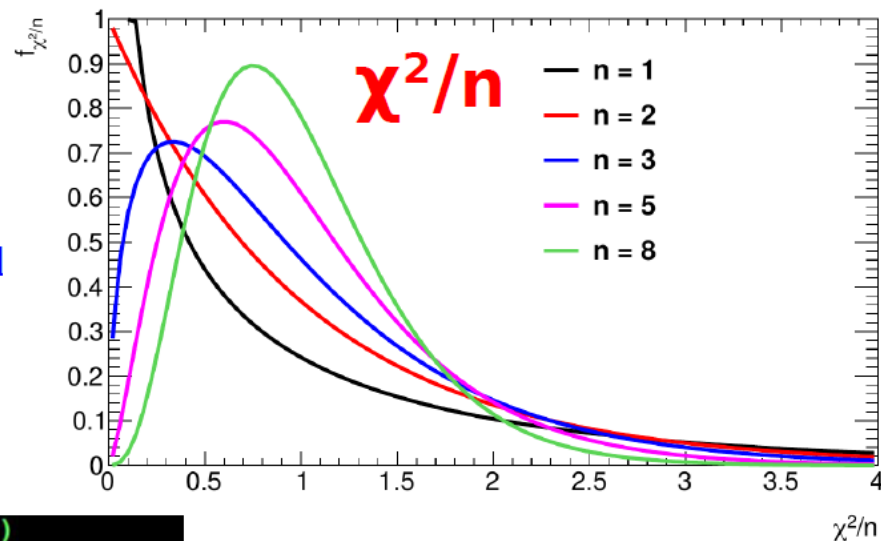
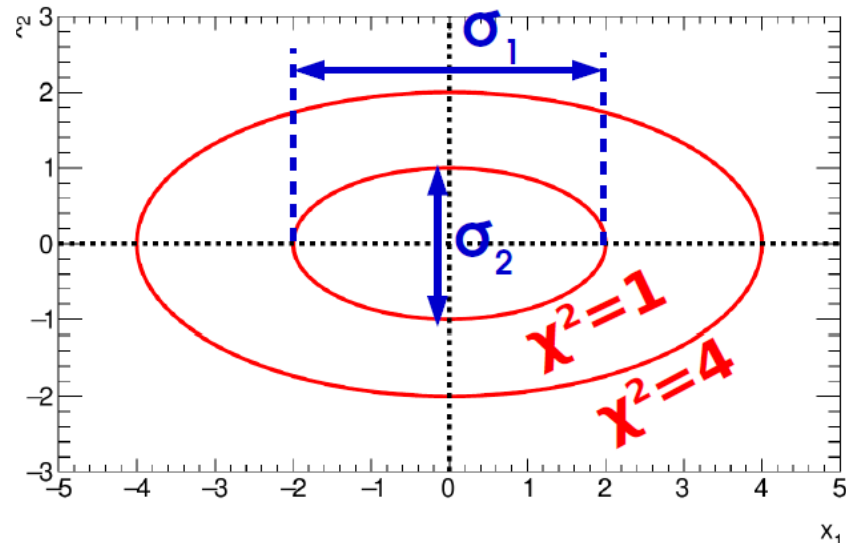
Rule of thumb: χ^2/n should be ≈ 1

Exact distributions in ROOT:

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

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

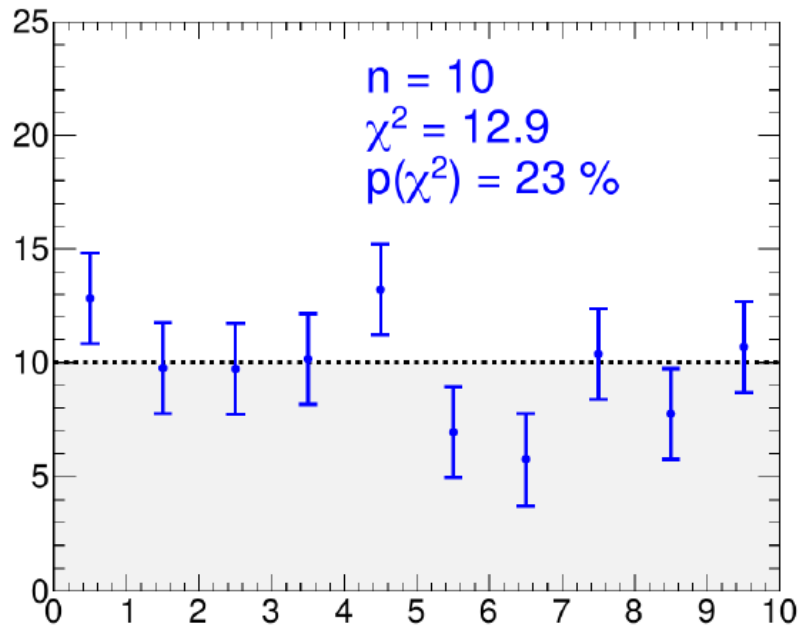
```
root [0] ROOT::Math::chisquared_cdf(1, 1)
(double) 0.68268949
root [1] ROOT::Math::chisquared_cdf(4, 1)
(double) 0.95449974
```



Histogram Chi-squared

Histogram χ^2 with respect to a reference shape:

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



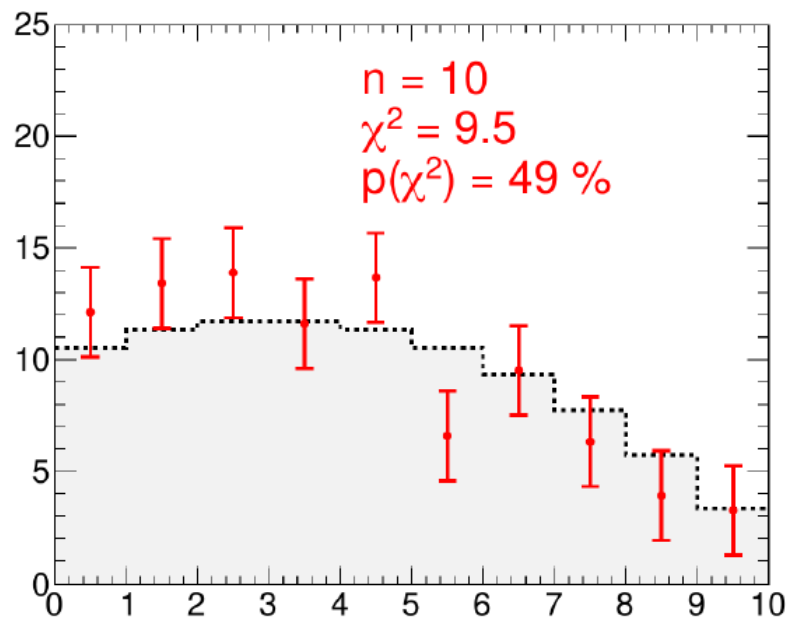
BLUE histogram vs. flat reference

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

Histogram Chi-squared

Histogram χ^2 with respect to a reference shape:

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



BLUE histogram vs. flat reference

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

RED histogram vs. flat reference

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

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* :

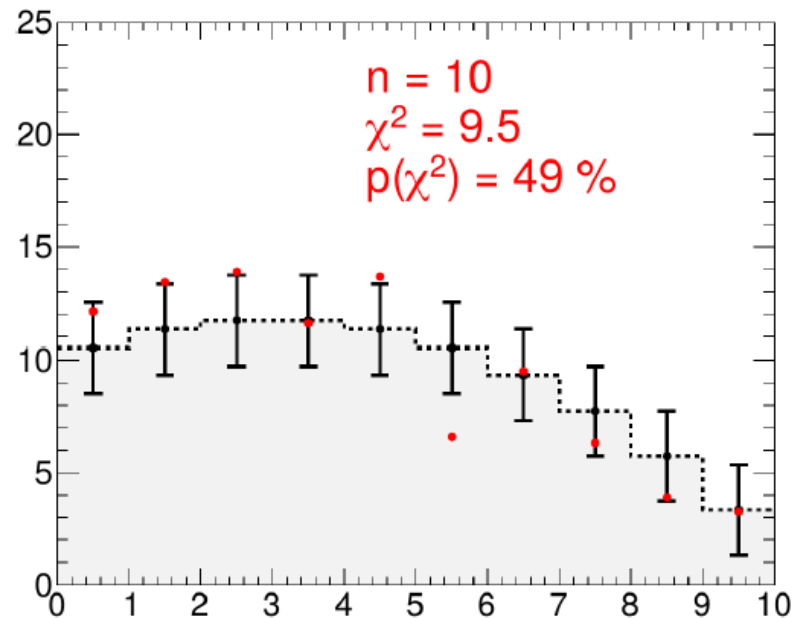
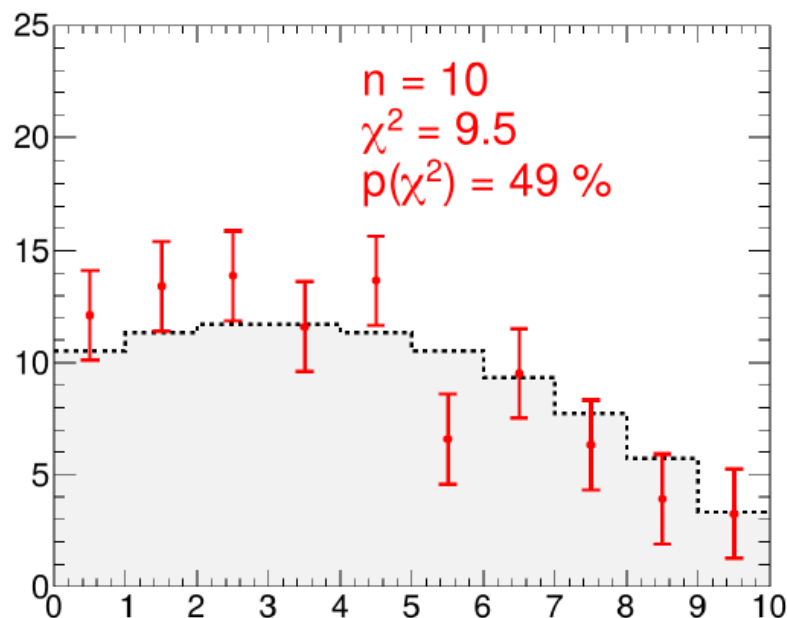
→ **Bin central value** ~ mean of the bin PDF

→ **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: $Z \rightarrow ee$ Inclusive σ^{fid}

Phys. Lett. B 759 (2016) 601

Measurement Principle:

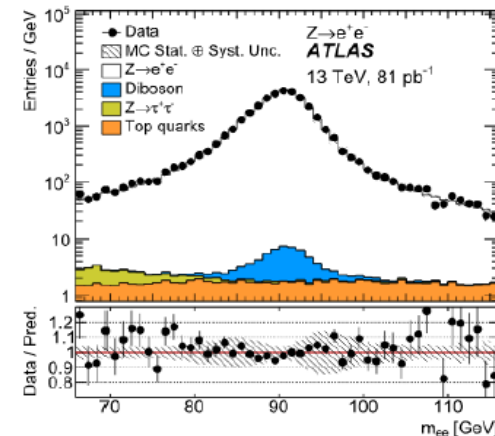
$$\sigma^{\text{fid}} = \frac{n_{\text{data}} - N_{\text{bkg}}}{C_{\text{fid}} L}$$

35000 $\pm$ 187 (points to n_{data})
 175 $\pm$ 8 (points to N_{bkg})
 (81 $\pm$ 2) pb⁻¹ (points to L)
 0.552 $\pm$ 0.006 (points to C_{fid})

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

Simple uncertainty propagation:

$$\sigma^{\text{fid}} = 0.781 \pm 0.004 \text{ (stat)} \pm 0.008 \text{ (syst)} \pm 0.016 \text{ (lumi) nb}$$



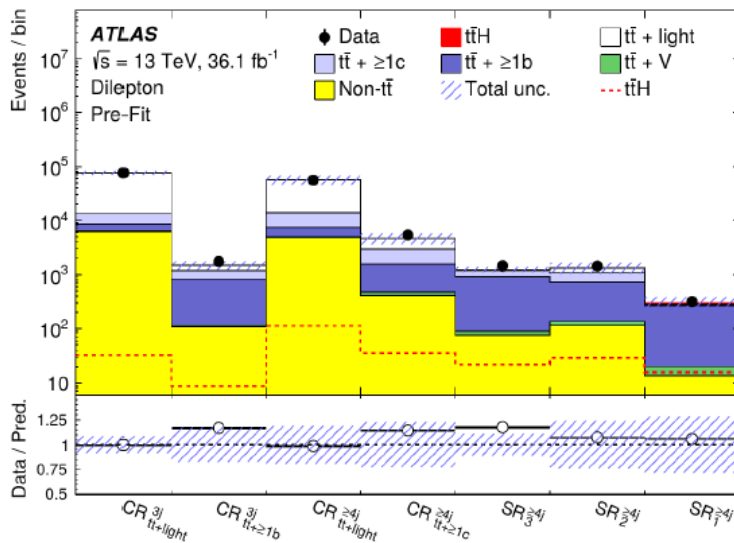
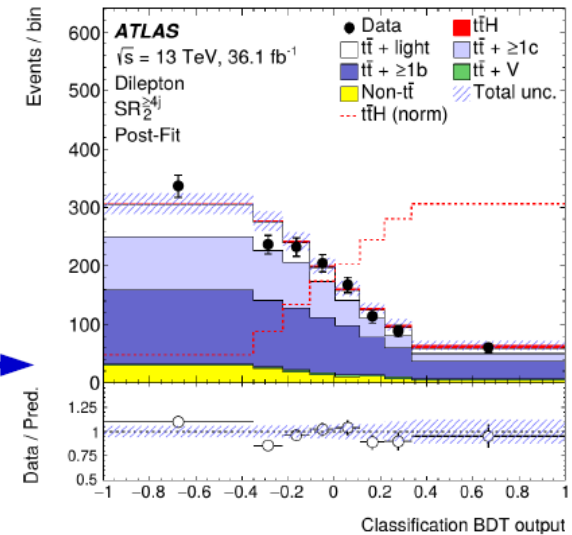
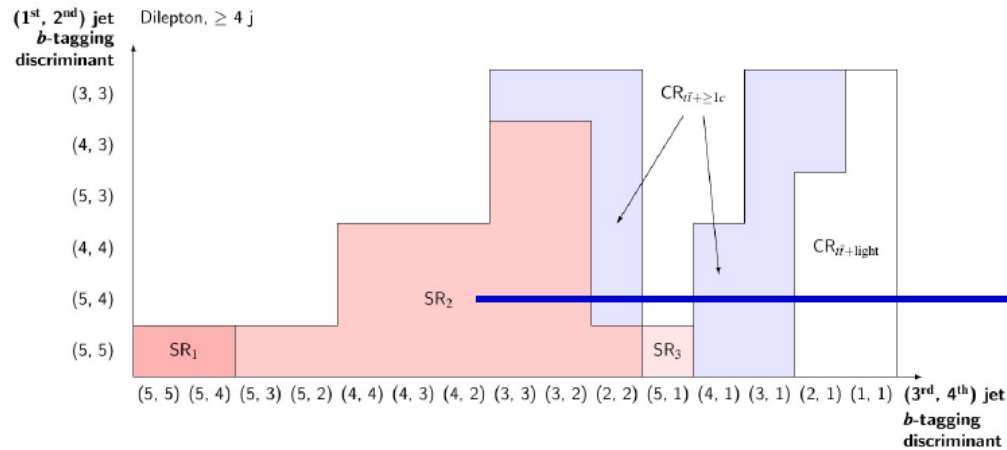
→ Simplest possible example in several ways (from the Statistics point of view!)

→ “Single bin counting”: only data input is n_{data} .

Example analyses

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

arXiv:1712.008895



Event counting in different regions:

Multiple-bin counting

Lots of information available

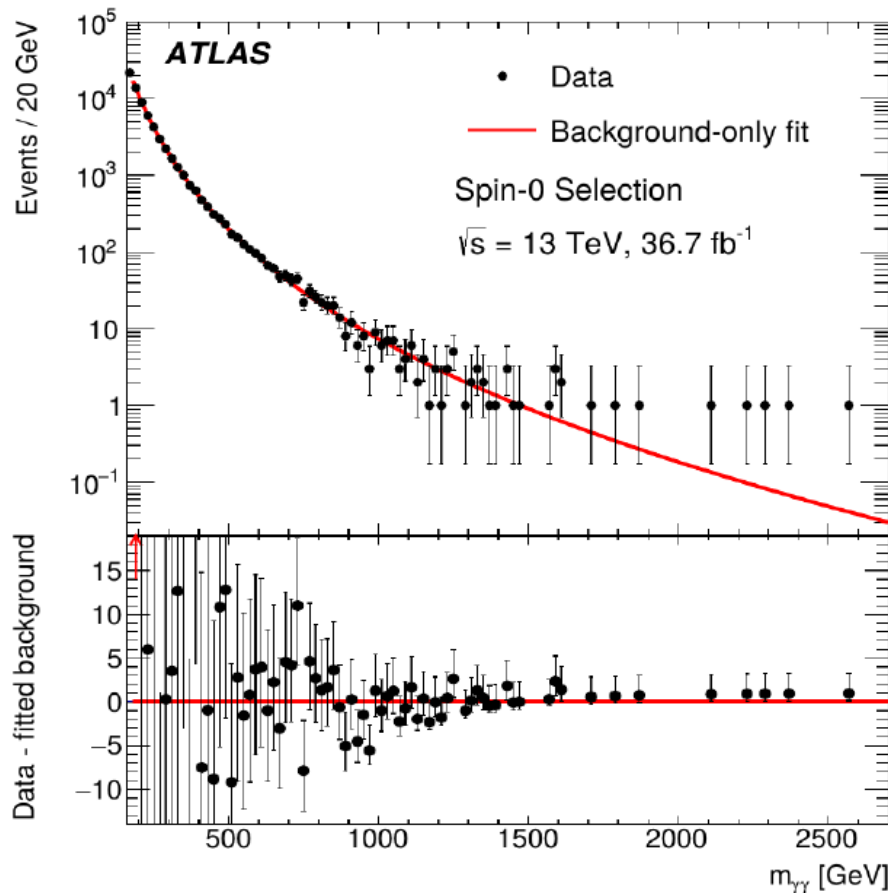
→ Potentially higher sensitivity

→ 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 ?

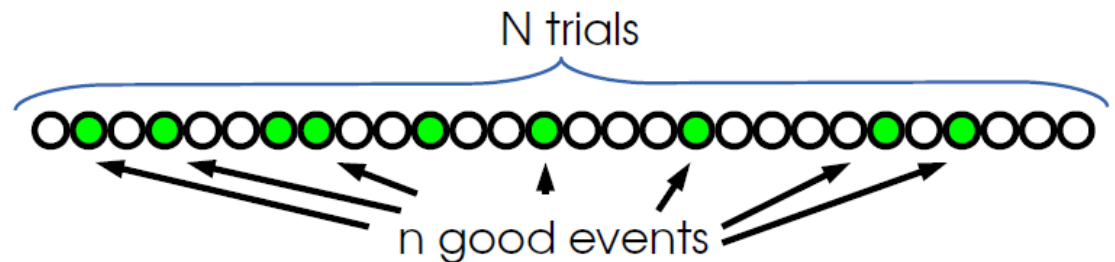
Counting events

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 = $N \cdot p$

Variance = $N \cdot p(1 - p)$



However suppose $p \ll 1$, $N \gg 1$, and let $\lambda = N \cdot p$:

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

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

Mean = λ

Variance = $\lambda \Rightarrow \text{RMS} = \sqrt{\lambda}$

$$(1-p)^{N-n} \stackrel{n \ll N}{\approx} \left(1 - \frac{\lambda}{N}\right)^N \stackrel{N \gg 1}{\approx} e^{-\lambda}$$

For n expected events, the uncertainty is $\sqrt{n}$

Rare processes ?

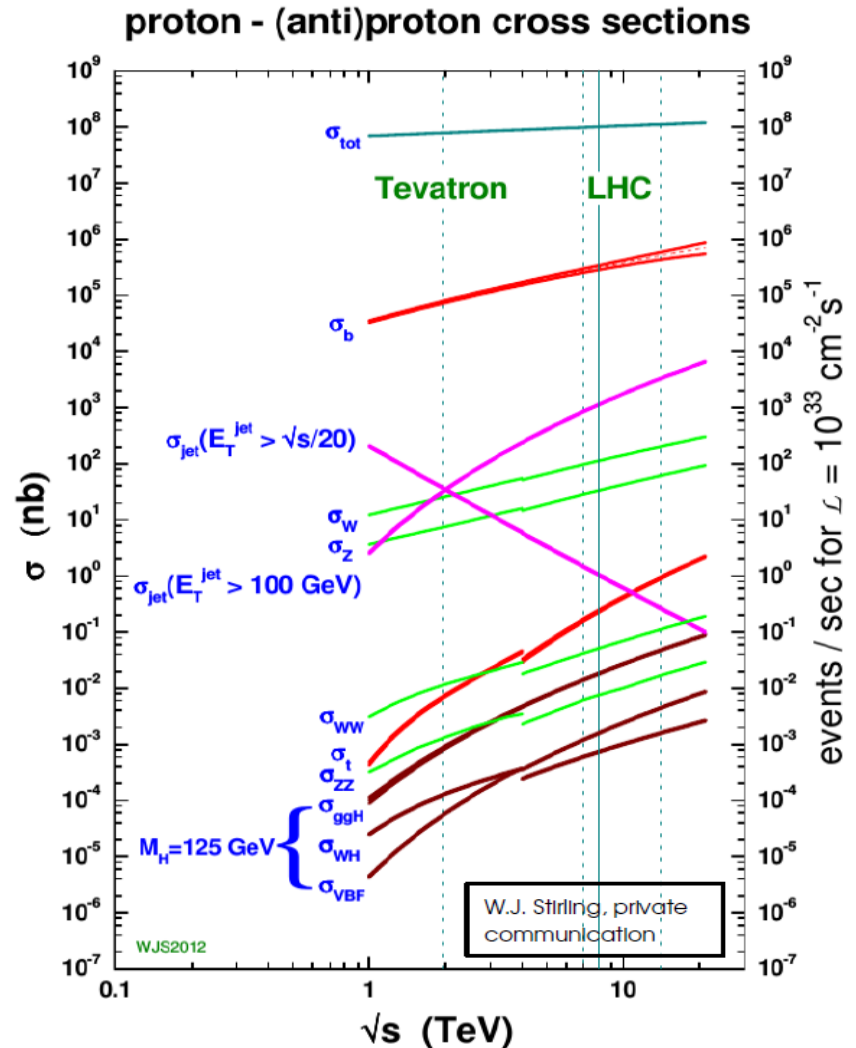
HEP : almost always use Poisson distributions. Why ?

ATLAS :

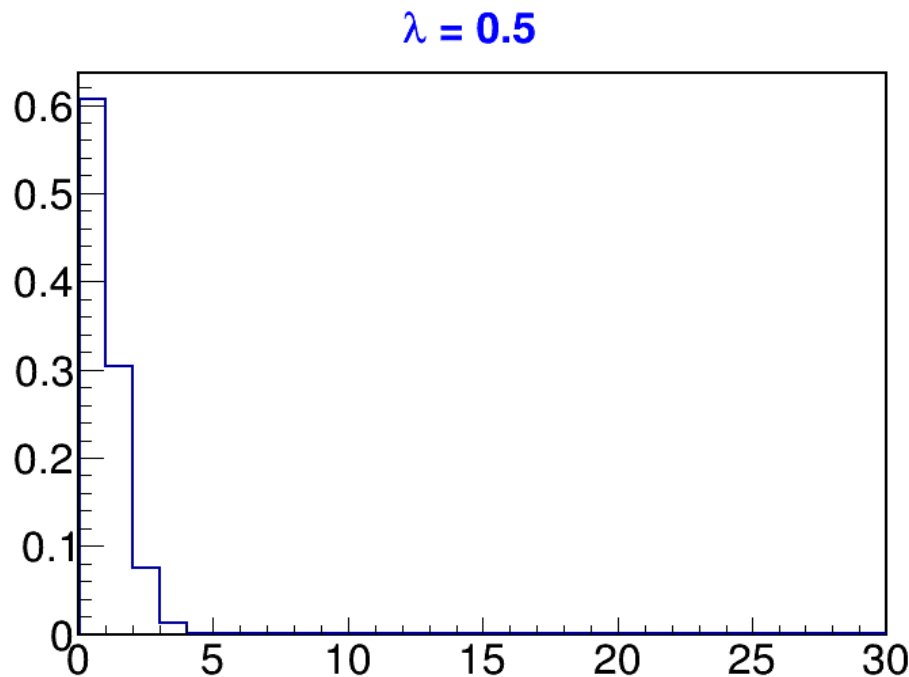
- **Event rate ~ 1 GHz**
($L \sim 10^{34} \text{ cm}^{-2}\text{s}^{-1} \sim 10 \text{ nb}^{-1}/\text{s}$, $\sigma_{\text{tot}} \sim 10^8 \text{ nb}$,)
 - **Trigger rate ~ 1 kHz**
(Higgs rate ~ 0.1 Hz)
- $\Rightarrow p \sim 10^{-6} \ll 1$ ($p_{H \rightarrow \gamma\gamma} \sim 10^{-13}$)

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

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



Poisson distributions



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

λ : expected
number of events

Mean = λ

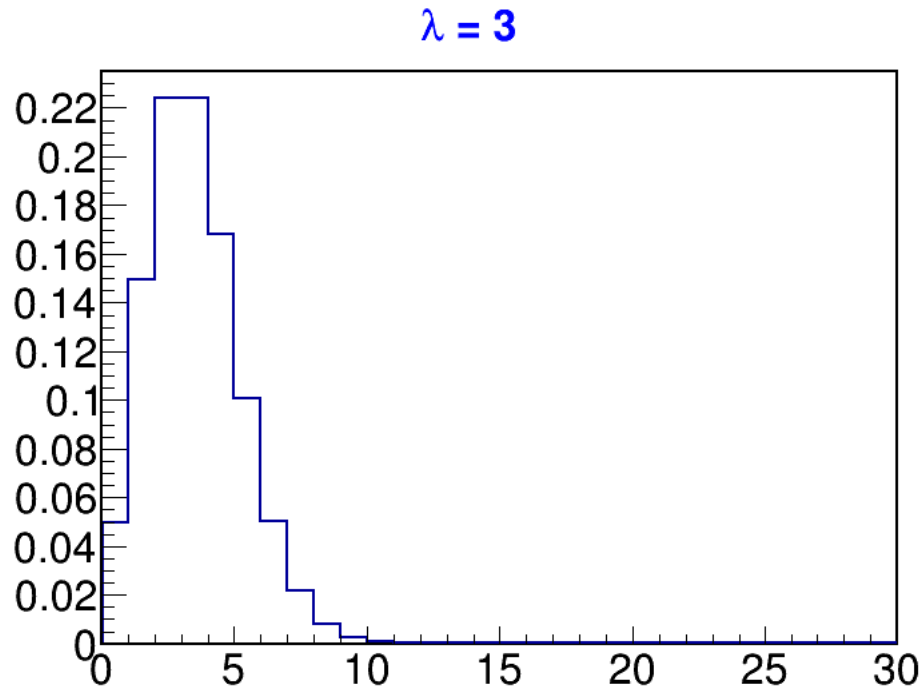
Variance = λ

$\sigma = \sqrt{\lambda}$

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

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

Poisson distributions



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

λ : expected
number of events

Mean = λ

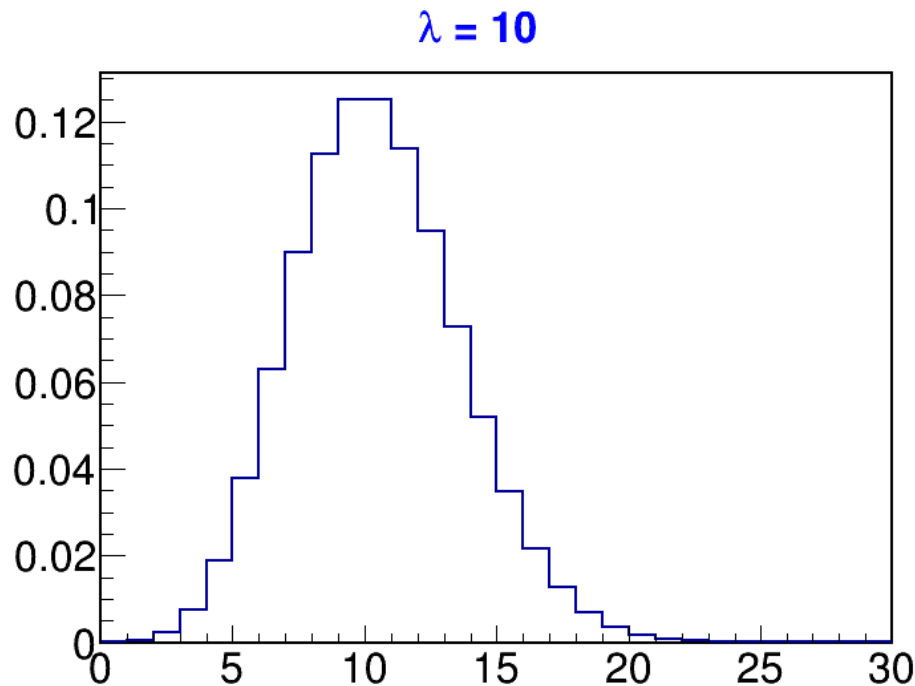
Variance = λ

$\sigma = \sqrt{\lambda}$

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

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

Poisson distributions



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

λ : expected
number of events

Mean = λ

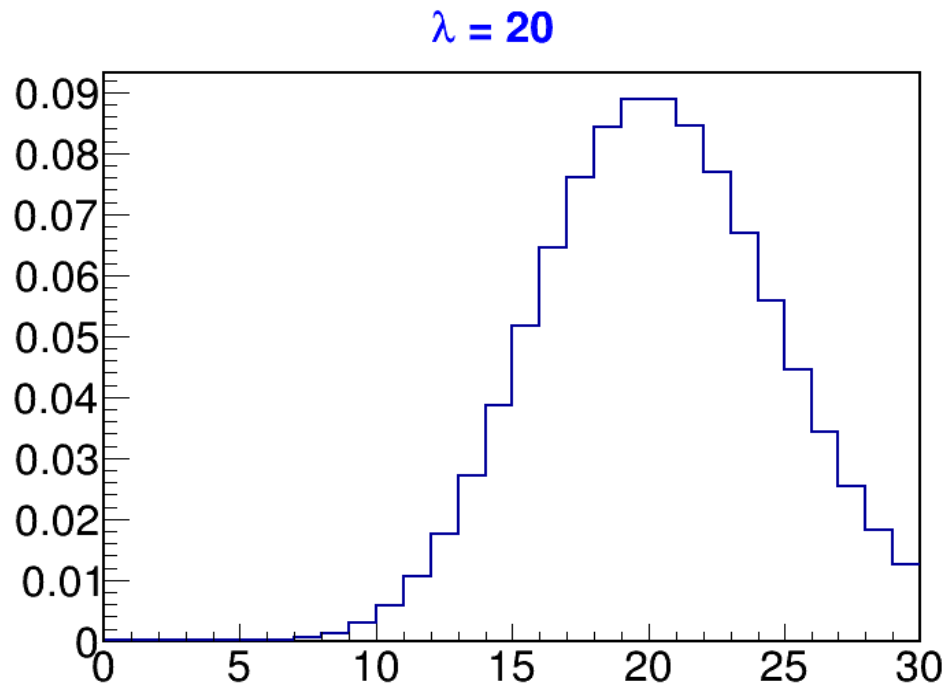
Variance = λ

$\sigma = \sqrt{\lambda}$

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

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

Poisson distributions



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

λ : expected number of events

Mean = λ

Variance = λ

$\sigma = \sqrt{\lambda}$

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

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

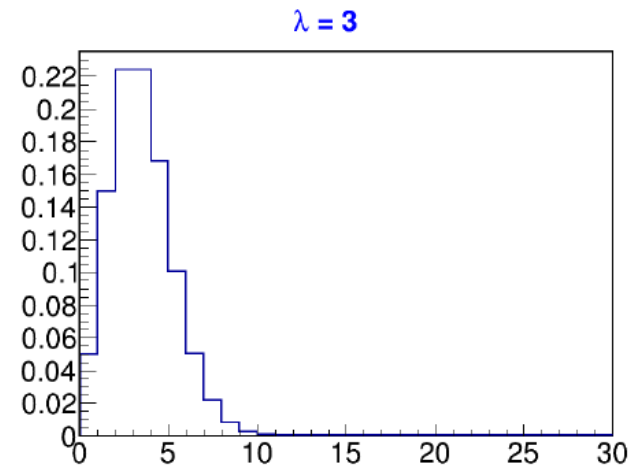
Statistical model for counting

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:

$$35000 \pm (\sqrt{35000} = 187)$$

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

175 ± 8
 $(81 \pm 2) \text{ pb}^{-1}$
 0.552 ± 0.006

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

Phys. Lett. B 759 (2016) 601

Simple uncertainty propagation:

$$\sigma^{\text{fid}} = 0.781 \pm 0.004 \text{ (stat)} \pm 0.008 \text{ (syst)} \pm 0.016 \text{ (lumi) nb}$$

Statistical uncertainty:
 Derived from assumption
 that n_{data} is $\sim \text{Poisson}(S+B)$

Systematics: more on
 this in Lecture 3

Unbinned shape analysis

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

→ Describe shape of the *distribution of m*

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

$H \rightarrow \gamma\gamma$ -inspired example:

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

⇒ Total PDF for a single event: Expected yields : S, B

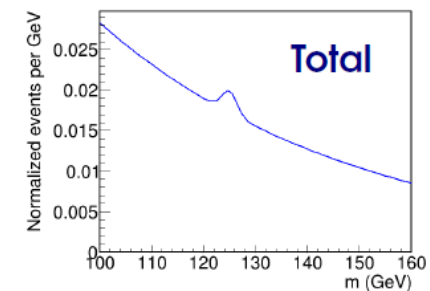
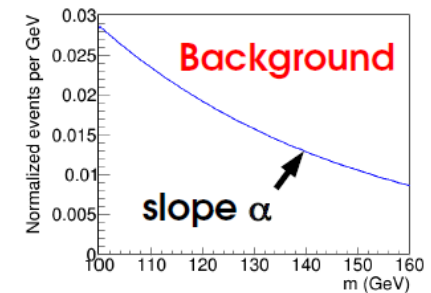
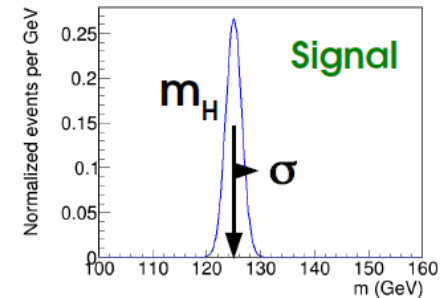
$$P_{\text{total}}(m) = \frac{S}{S+B} G(m; m_H, \sigma) + \frac{B}{S+B} \alpha e^{-\alpha m}$$

⇒ Total PDF for a dataset

Probability to observe n events

Probability to observe the value m_i

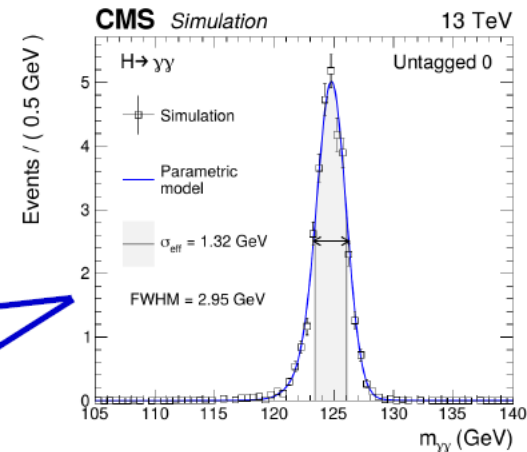
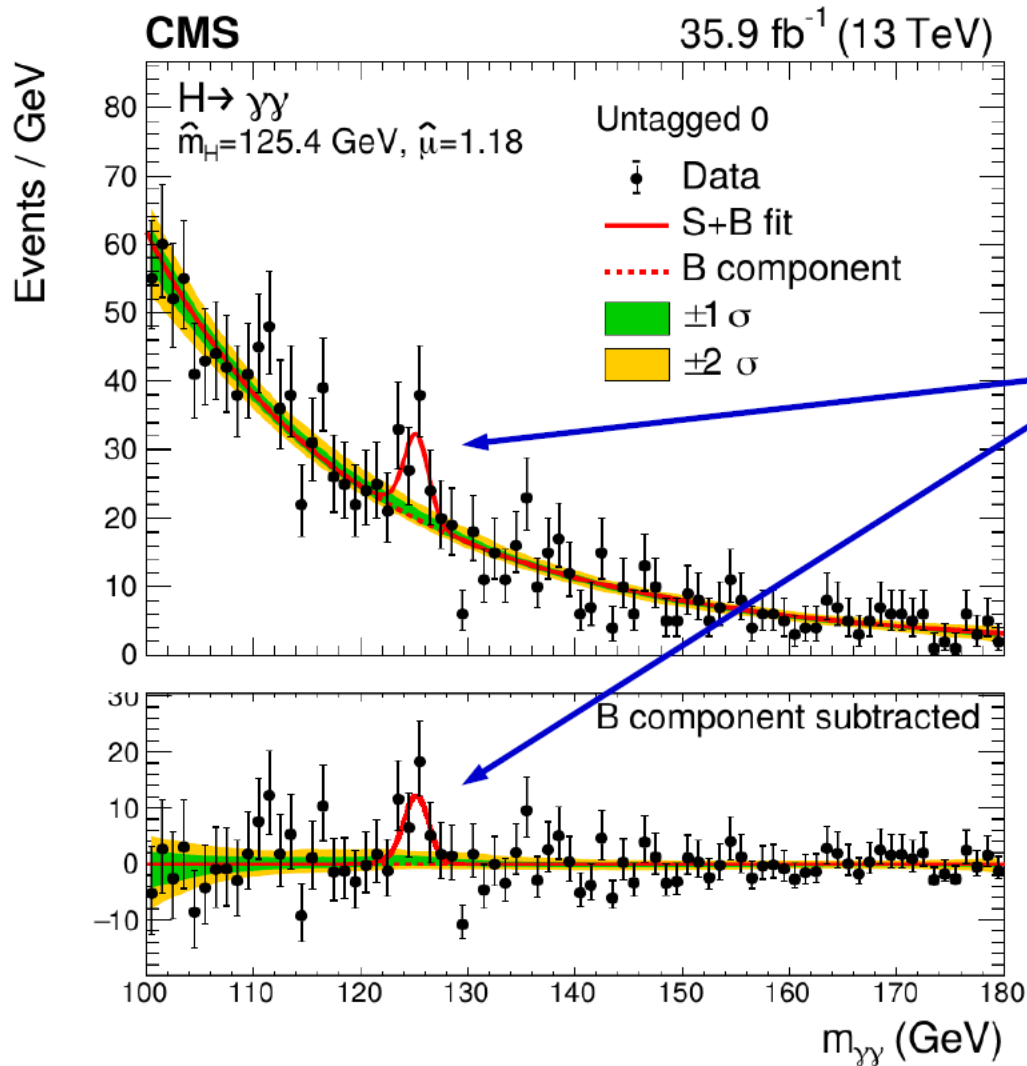
$$P(\{m_i\}_{i=1 \dots n}) = e^{-(S+B)} \frac{(S+B)^n}{n!} \prod_{i=1}^n \left[\frac{S}{S+B} G(m_i; m_H, \sigma) + \frac{B}{S+B} \alpha e^{-\alpha m_i} \right]$$



Unbinned shape analysis

$H \rightarrow \gamma\gamma$

JHEP 11 (2018) 185



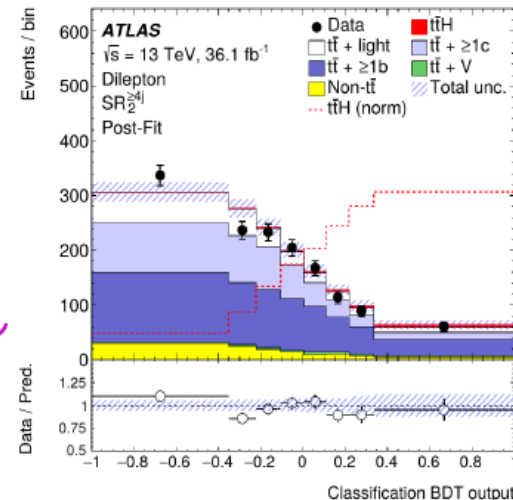
Binned shape analysis

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

→ N_{bins} : number of bins

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

$$P(\{n_i\}; S, B) = \prod_{i=1}^{N_{\text{bins}}} \underbrace{e^{-(Sf_{S,i} + Bf_{B,i})} \frac{(Sf_{S,i} + Bf_{B,i})^{n_i}}{n_i!}}_{\text{Poisson distribution in each bin}}$$



$N_{\text{bins}} = 1$: Counting analysis

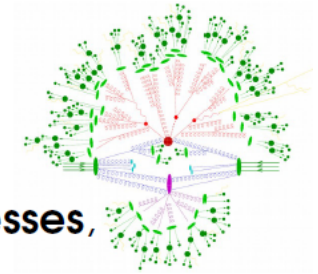
$N_{\text{bins}} \rightarrow \infty$: Unbinned shape analysis (the fractions become PDF values)

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

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

⊖ **MC stat fluctuations** can create artefacts, especially for $S \ll B$.

How to describe data



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

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

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