PSI Zuoz Summer School 2022

Statistics

Nicolas Berger (LAPP Annecy)

Lecture 1

PSI Zuoz Summer School 2022

Statistics

For Physicists

Nicolas Berger (LAPP Annecy)

Lecture 1

Lecture Plan

Statistics basic concepts (Today)

- **[Basic ingredients** (PDFs, etc.)**]**
- **Parameter estimation** (maximum likelihood, least-squares, …)
- **Model testing** (χ^2 tests, hypothesis testing, p-values, …)

Computing statistical results (Today/Tomorrow)

- **Discovery**
- **Confidence intervals**
- **Upper limits**
- **Systematics and profiling**
- **[Bayesian techniques]**

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

Hands-on sessions

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

If you have a computer with you, **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.

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

The **warmup** item includes material that will not be covered in detail in the class, as well as an introduction to the notebooks. **Please have a look before the beginning of the classes** if you are unfamiliar with this.

Statistics are everywhere

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

Credits: mattbuck / wikimedia

Credits: StatLab

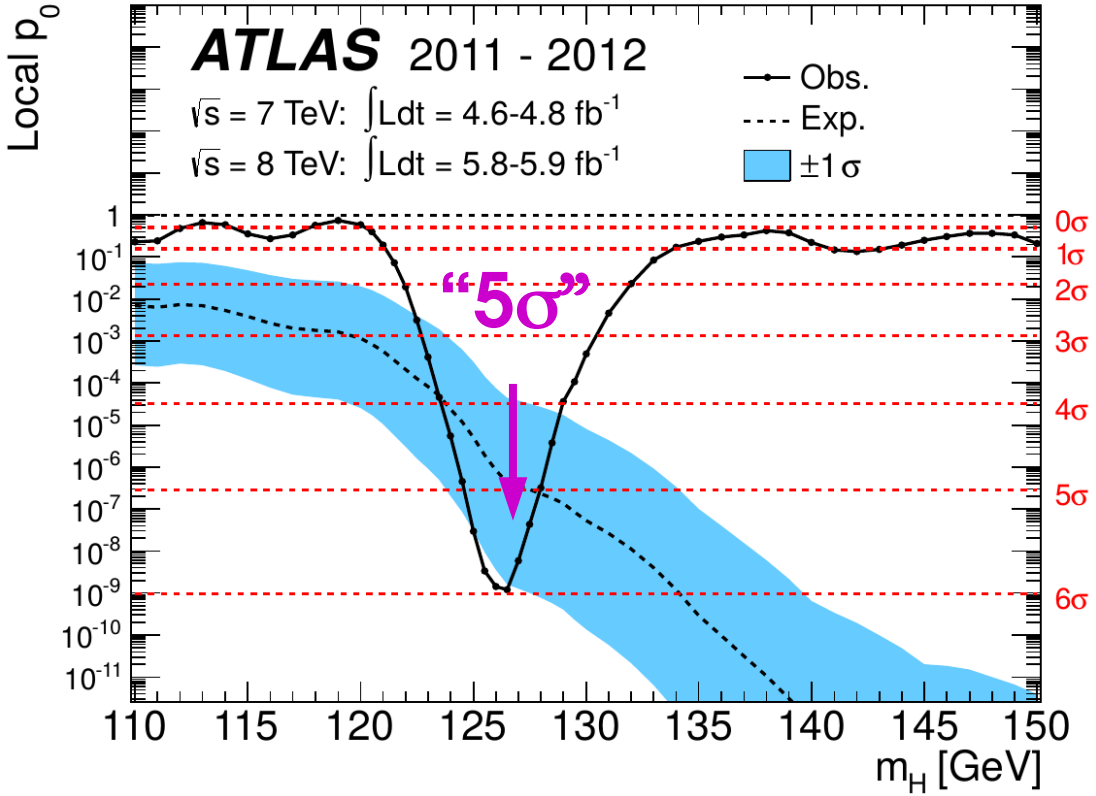
And Physics ?

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

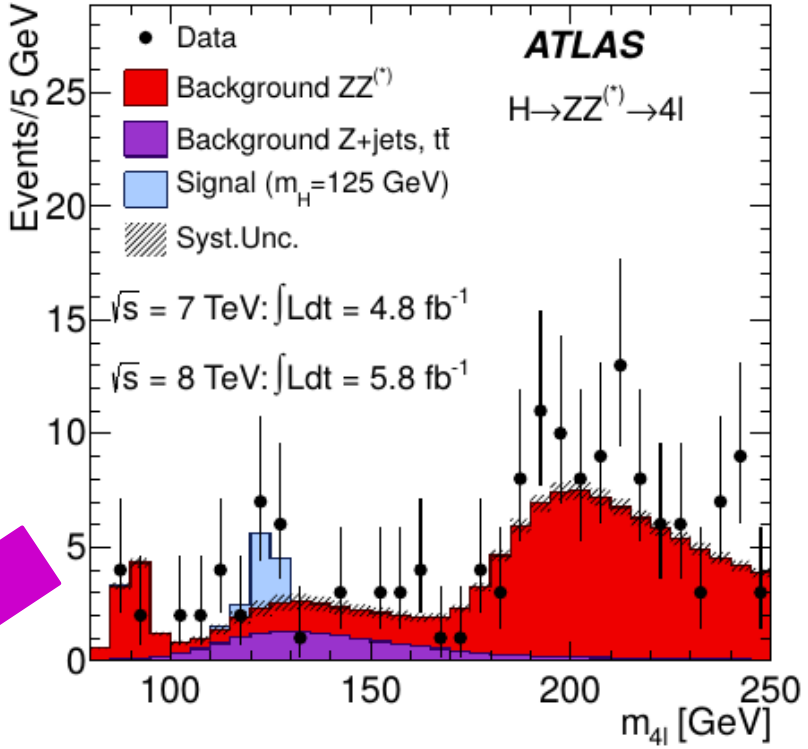
Introduction

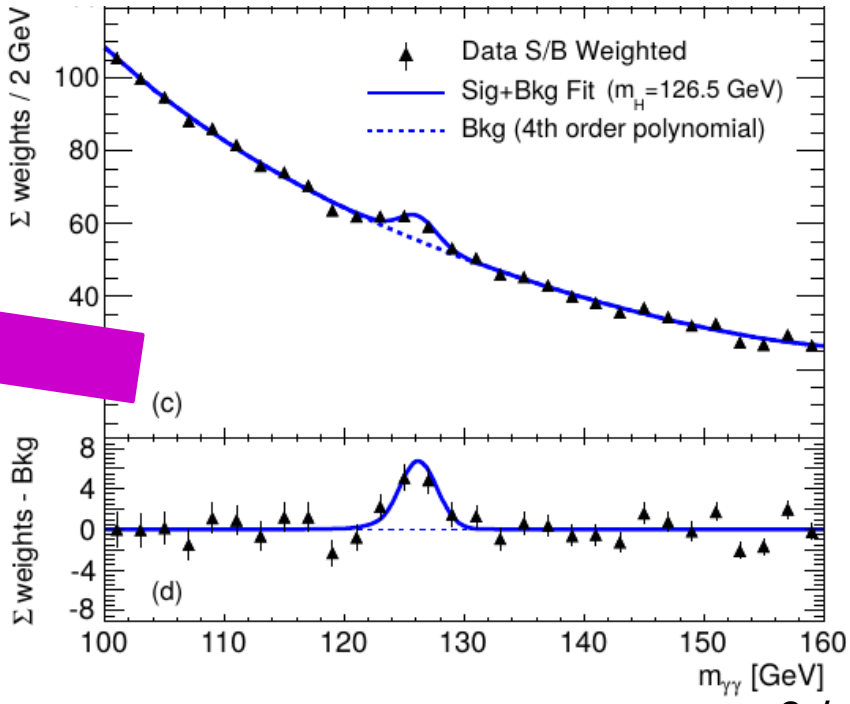
Statistical methods play a critical role in many areas of physics

Higgs discovery : **"We have 5σ" !**



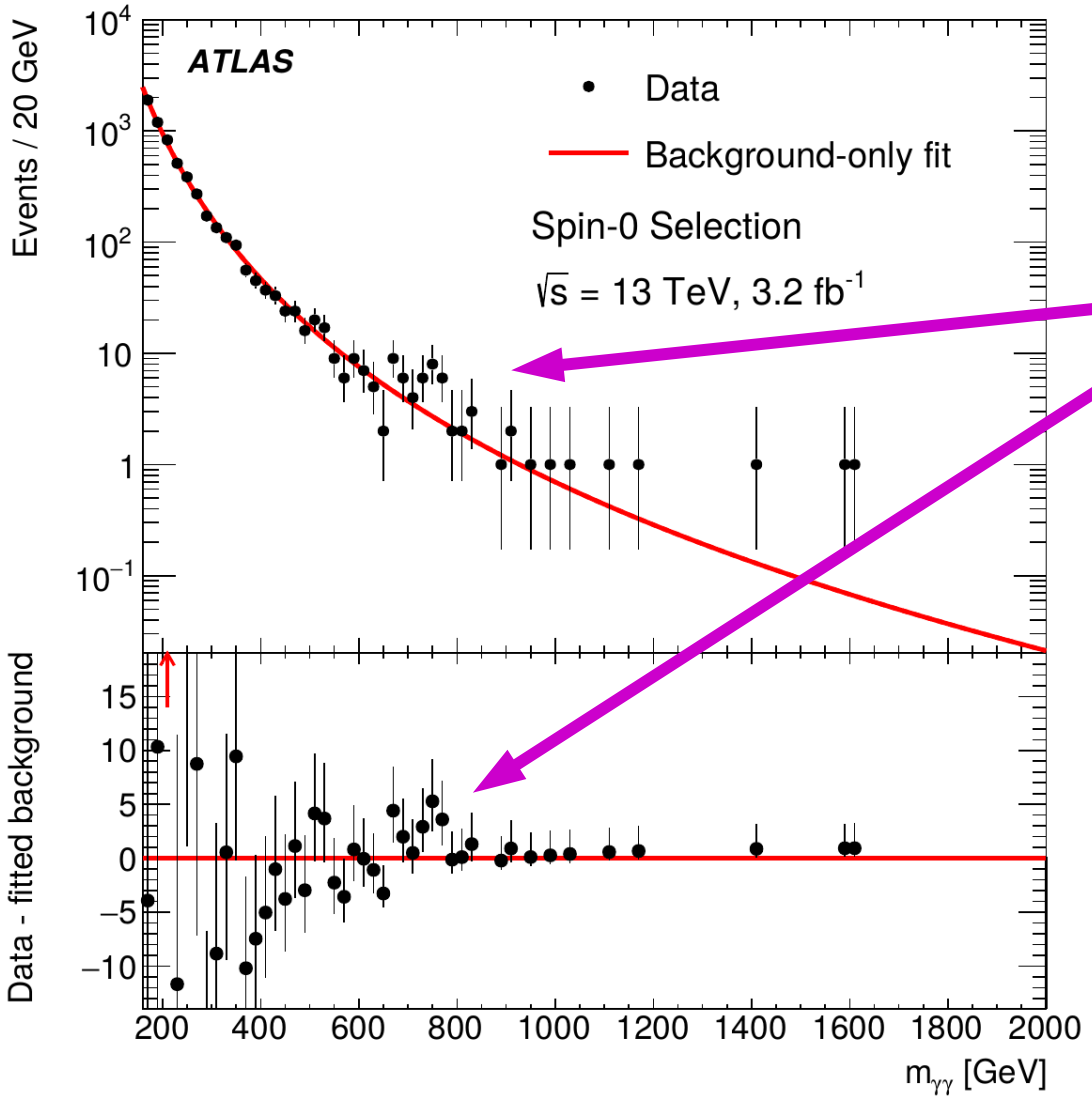
Phys. Lett. B 716 (2012) 1-29





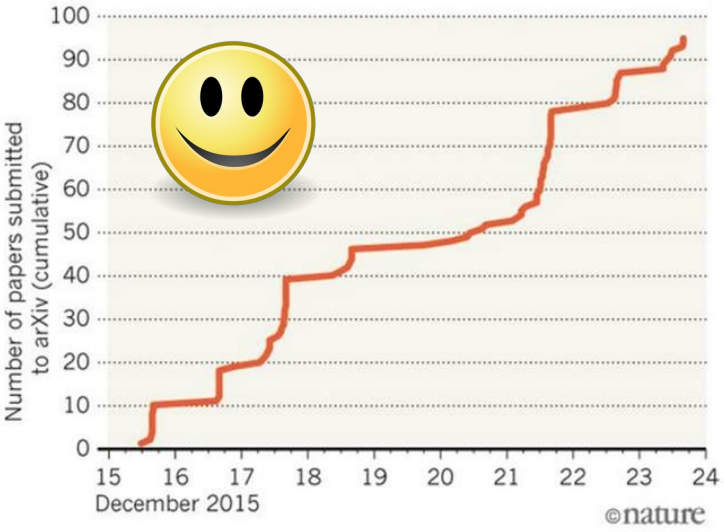
Introduction

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



New Physics ?

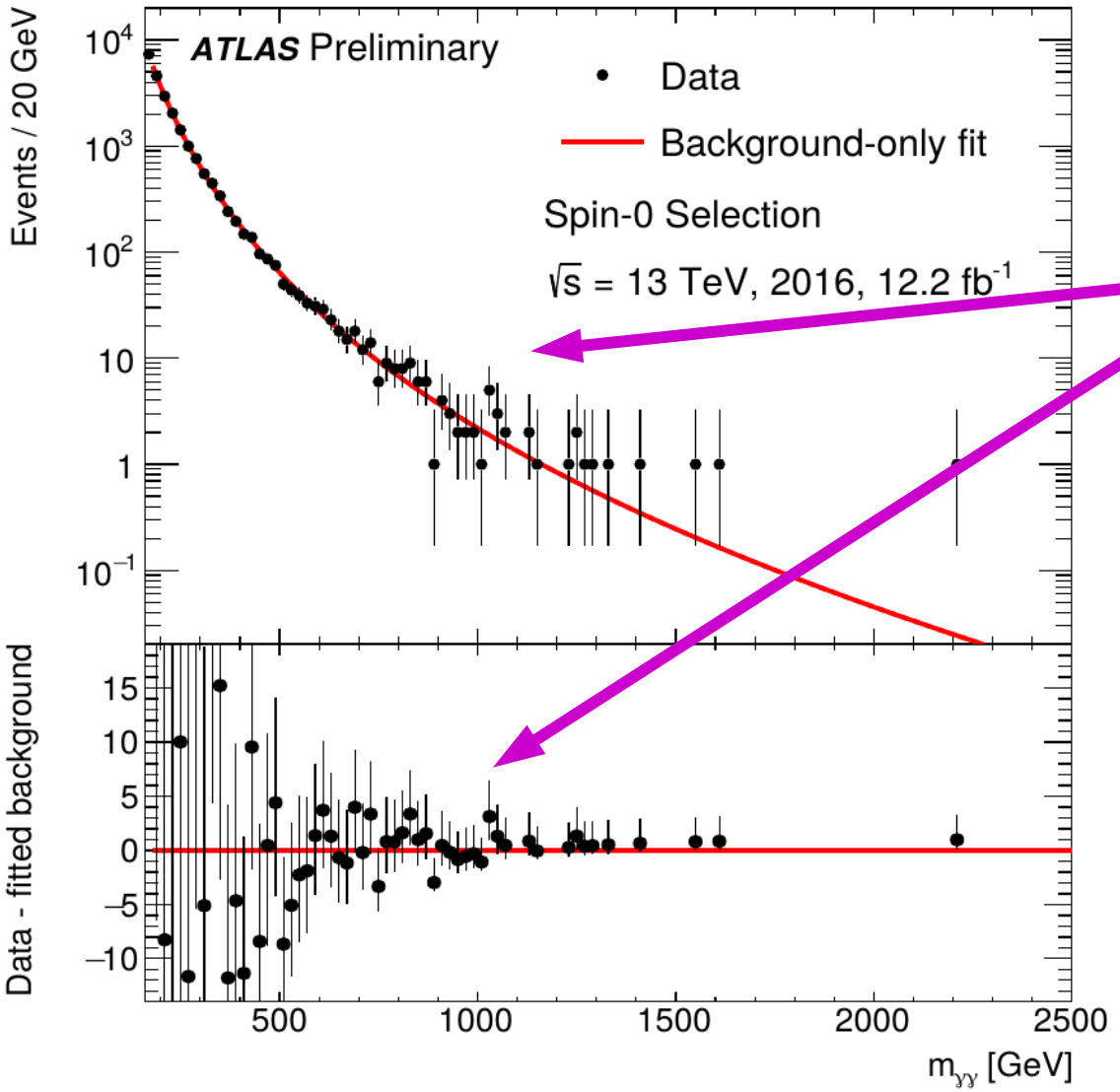
3.9σ ? 2.1σ ?



JHEP 09 (2016) 1

Introduction

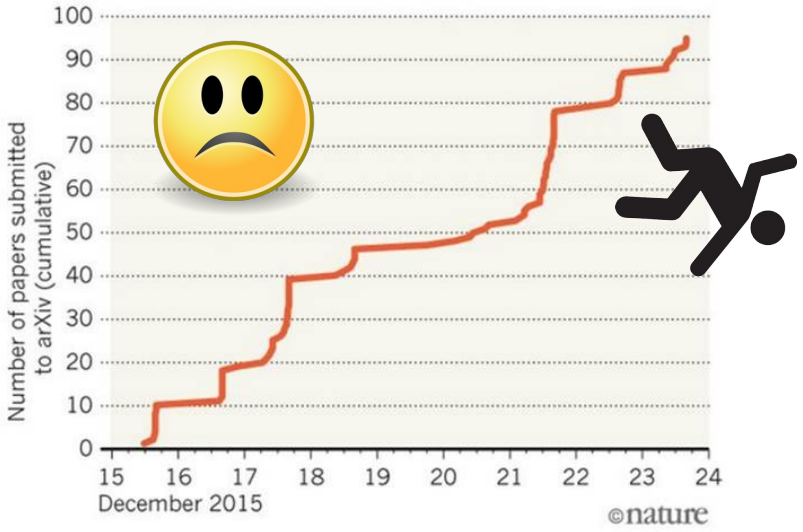
Sometimes difficult to distinguish a bona fide discovery from a **background fluctuation**…



A few months later…

~~**New Physics ?**~~

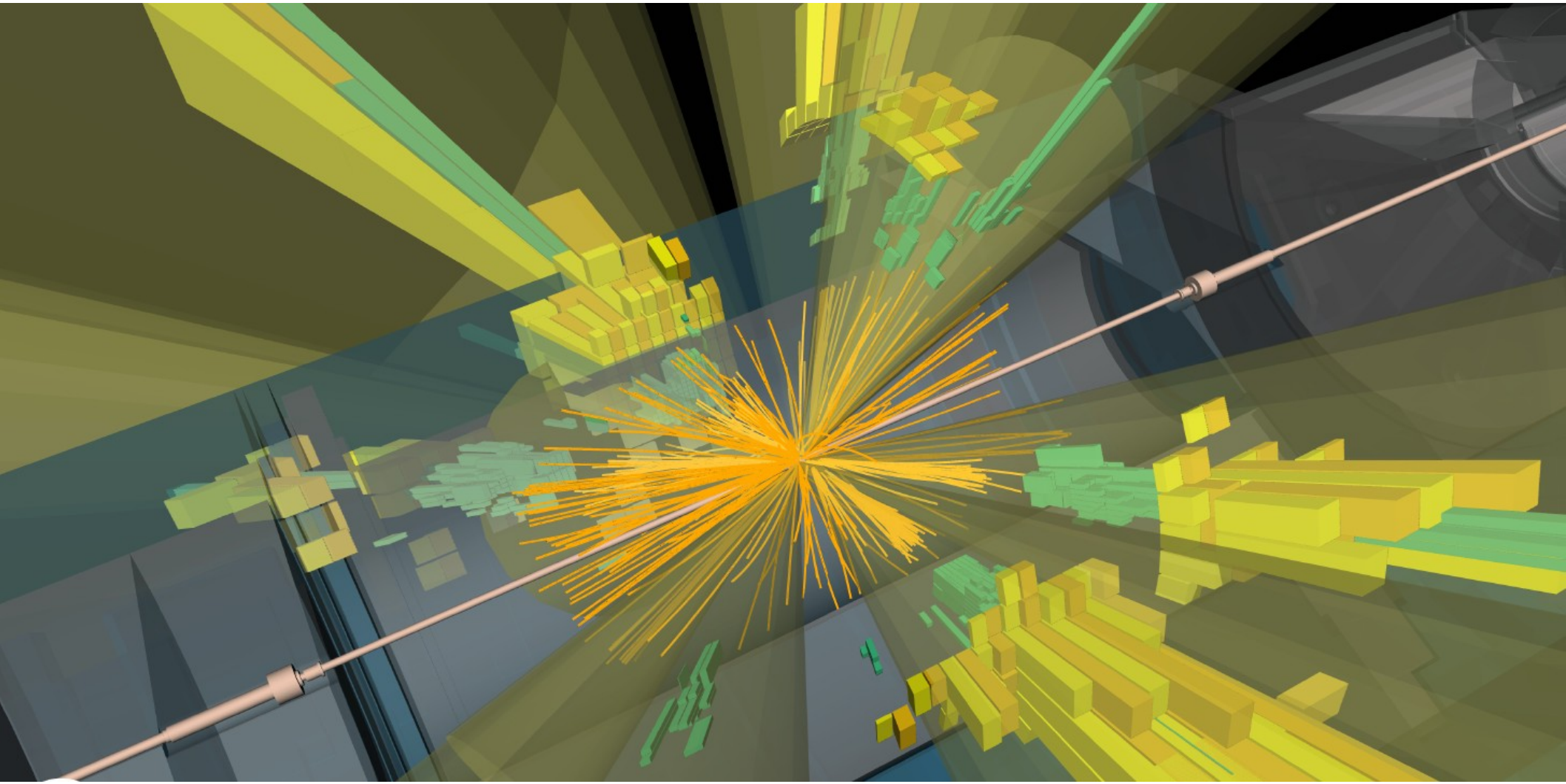
~~**3.9σ ? 2.1σ ?**~~



JHEP 09 (2016) 1

Randomness in High-Energy Physics

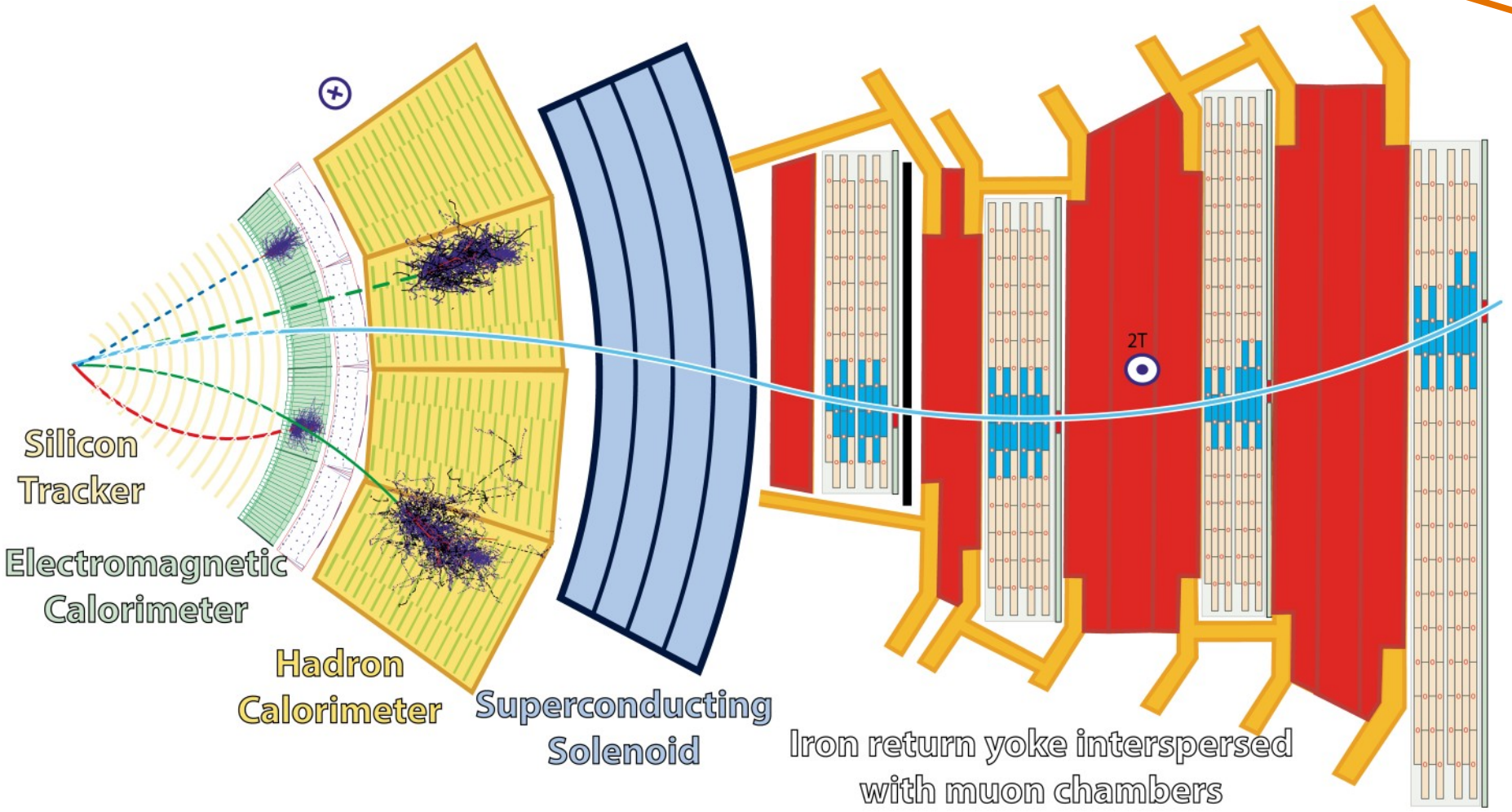
Experimental data is produced by incredibly complex processes



Randomness in High-Energy Physics

More details in Anna Sfyrla's lectures!

Experimental data is produced by incredibly complex processes



Randomness in High-Energy Physics

More details in Anna Sfyrla's lectures!

Experimental data is produced by incredibly complex processes

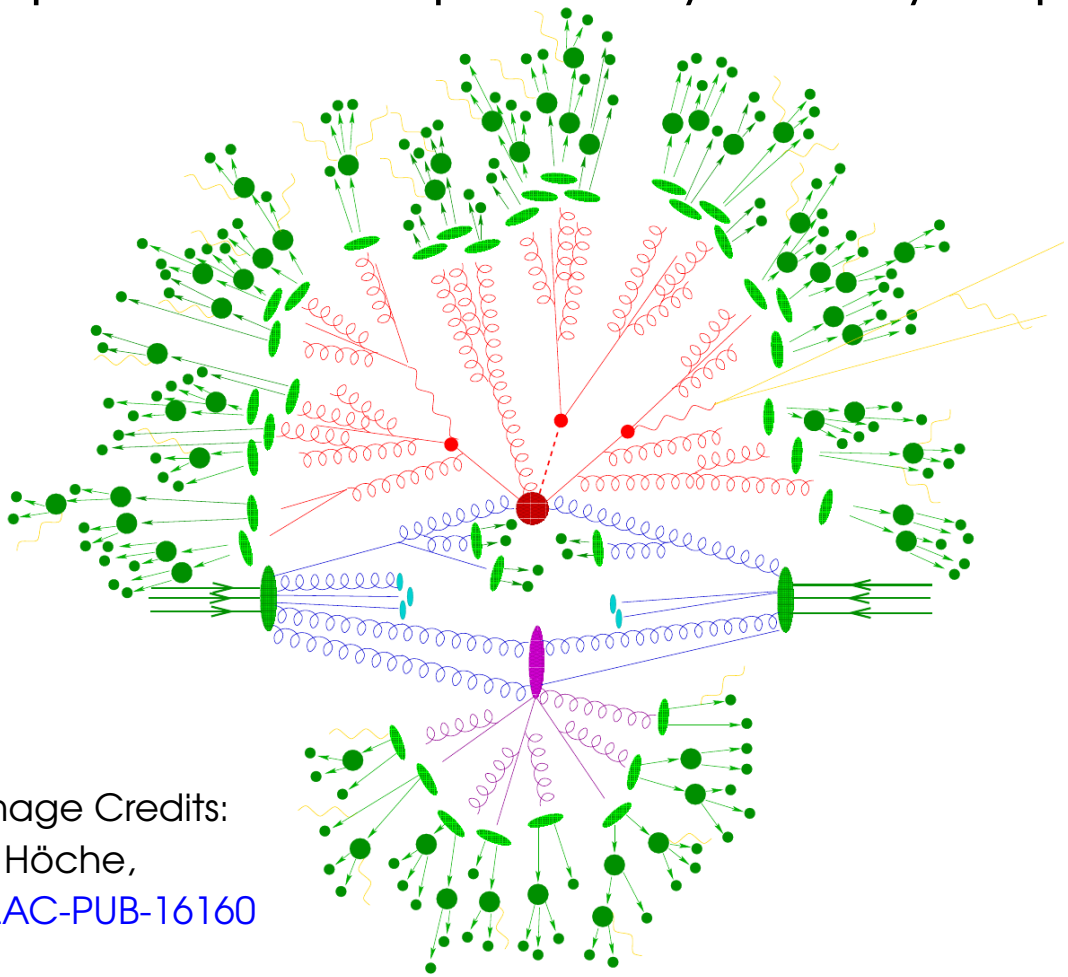


Image Credits:
S. Höche,
SLAC-PUB-16160

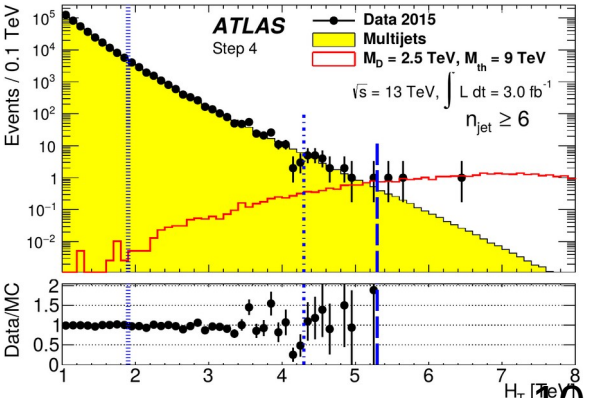
Hard scattering

PDFs, Parton shower, Pileup

Decays

Detector response

Reconstruction



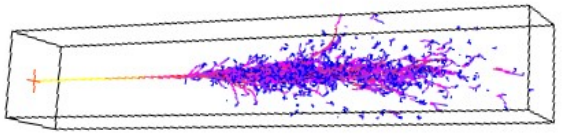
Randomness involved in all stages

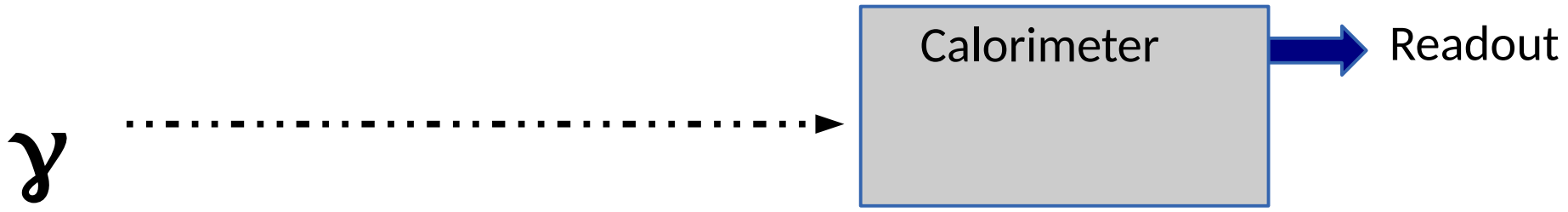
→ **Classical** randomness: detector reponse

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

Measurement Errors: Energy measurement

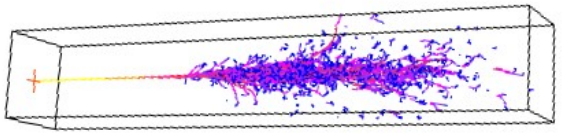
Example: measuring the energy of a photon in a calorimeter

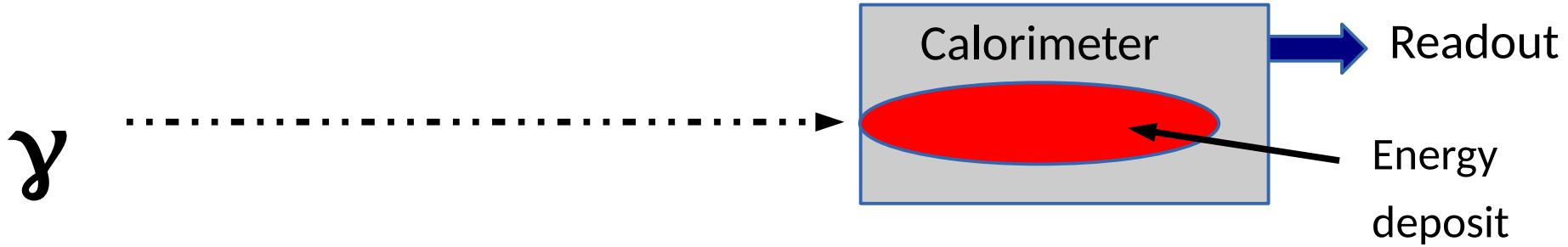


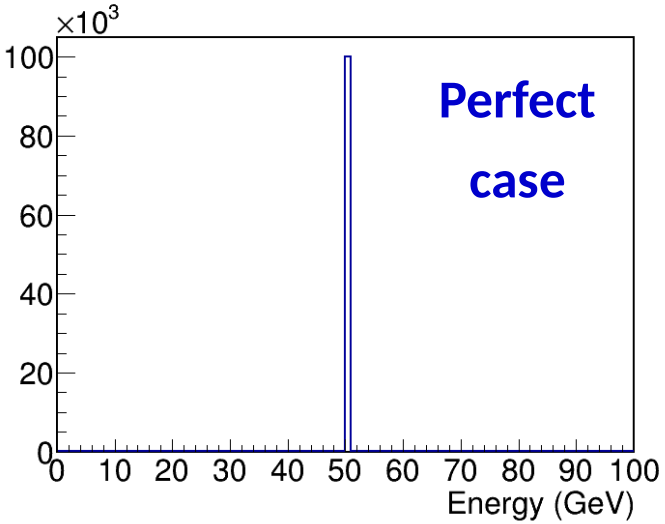


Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

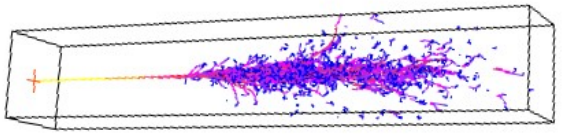


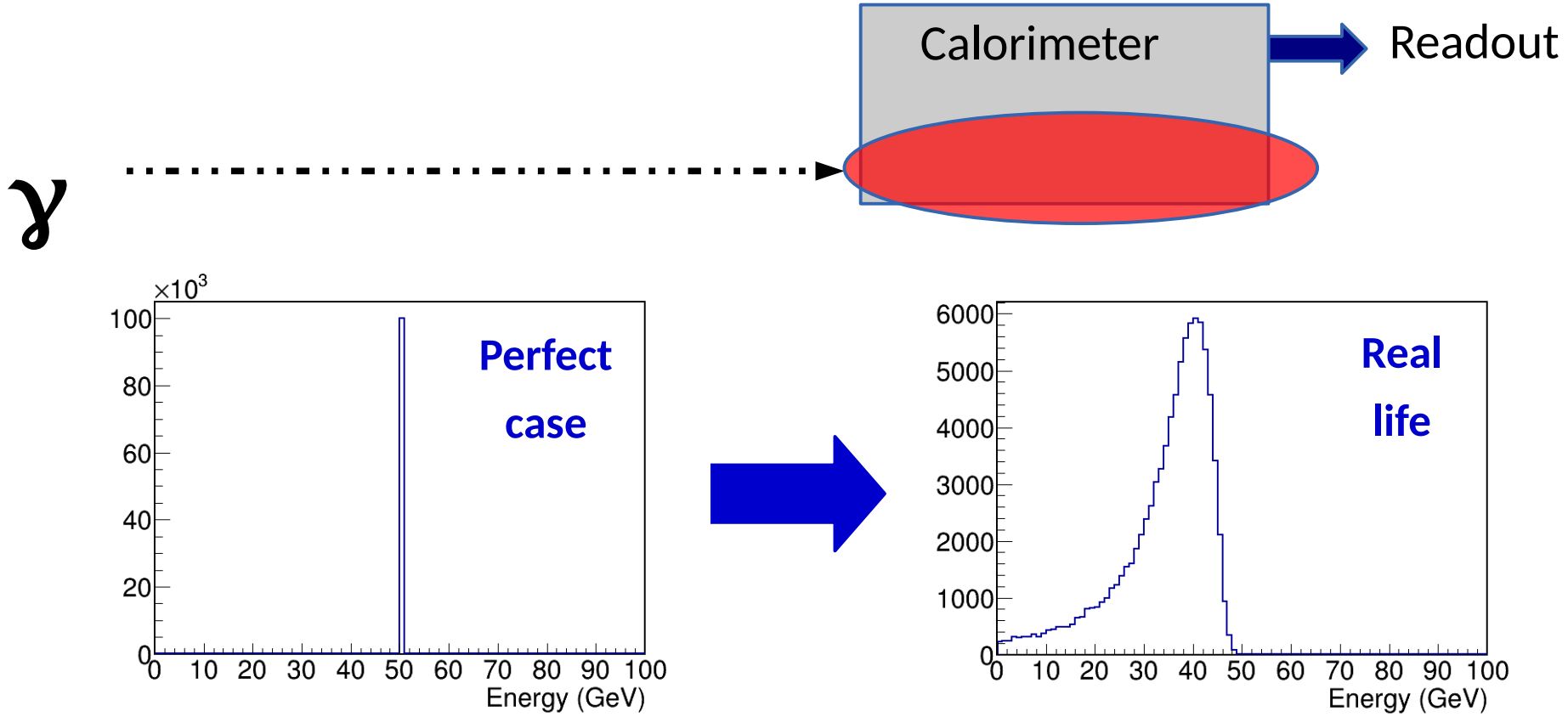




Measurement Errors: Energy measurement

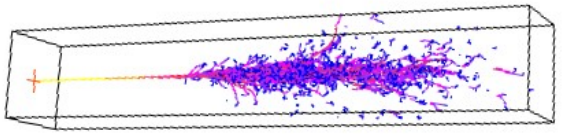
Example: measuring the energy of a photon in a calorimeter





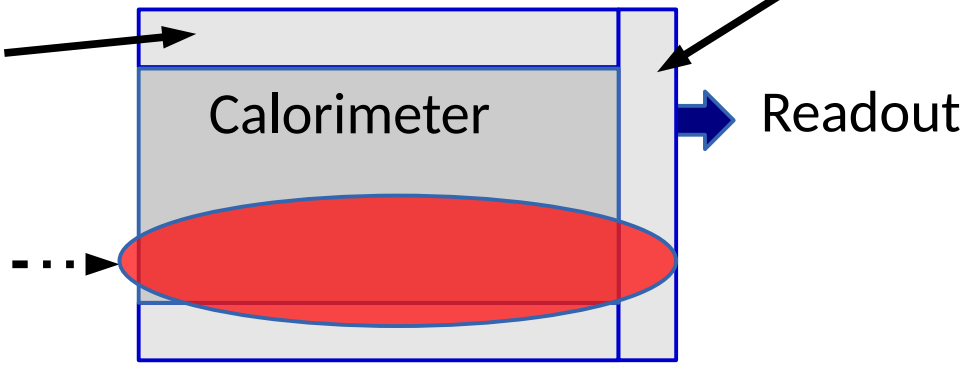
Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter

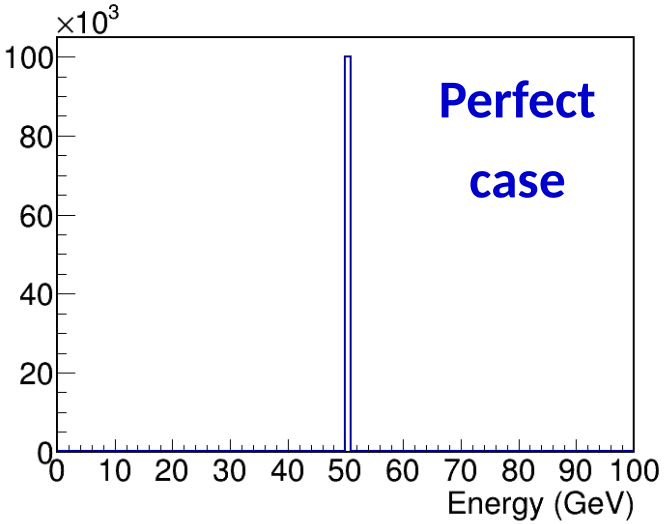


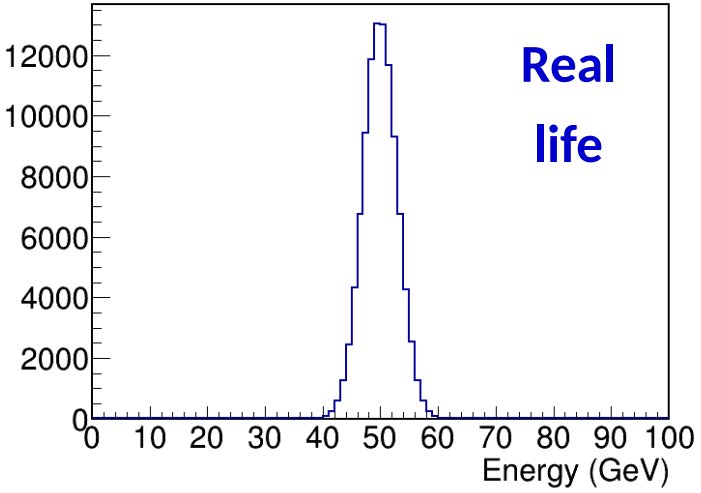
Measure leakage behind calorimeter

Measure leakage into neighboring cells



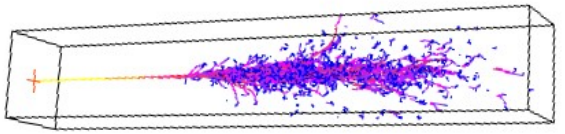
γ

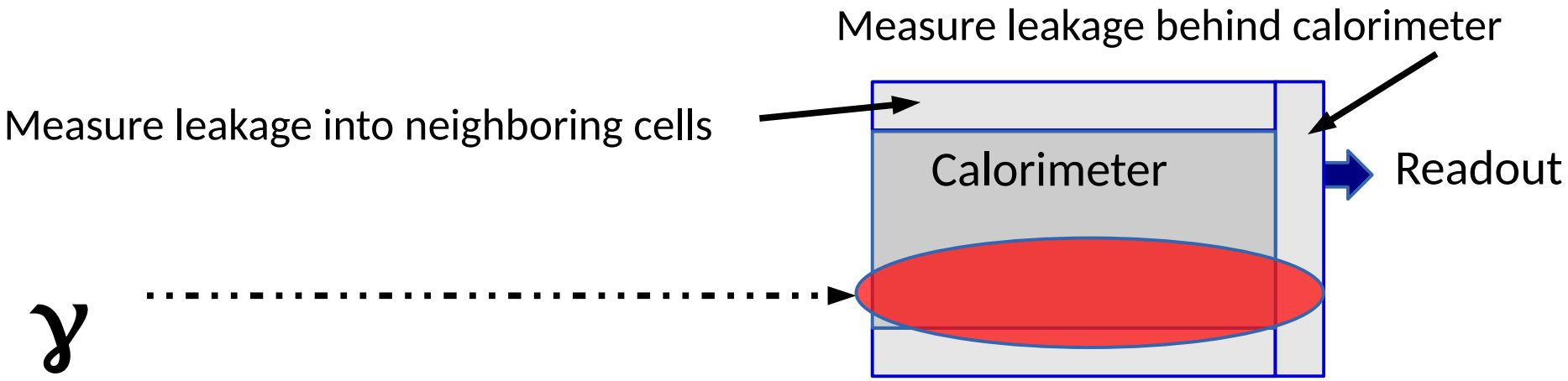


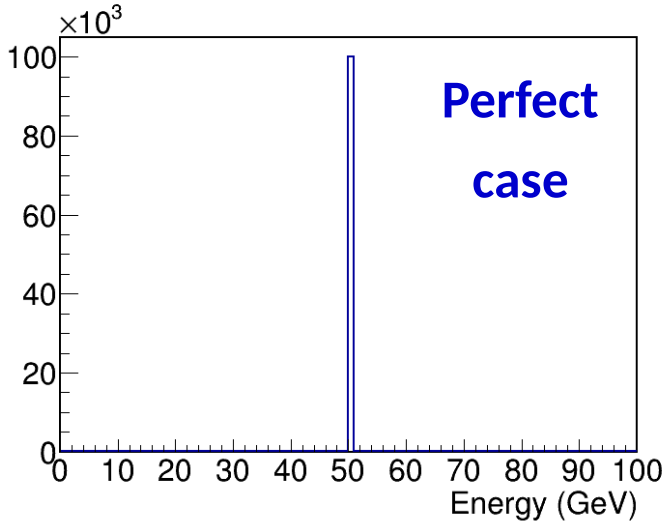


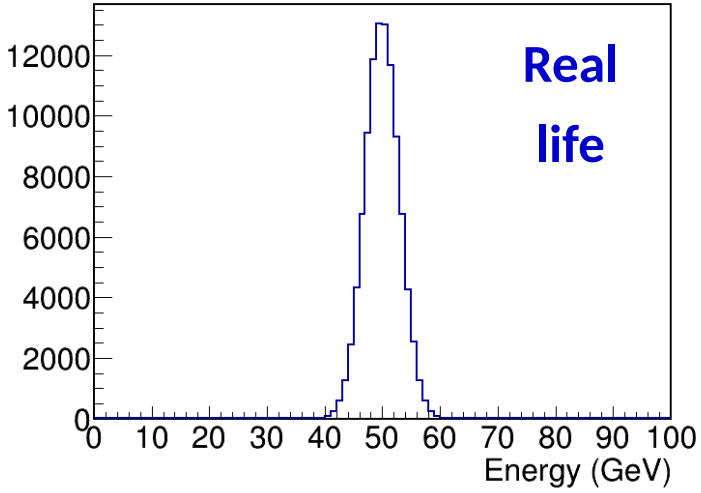
Measurement Errors: Energy measurement

Example: measuring the energy of a photon in a calorimeter







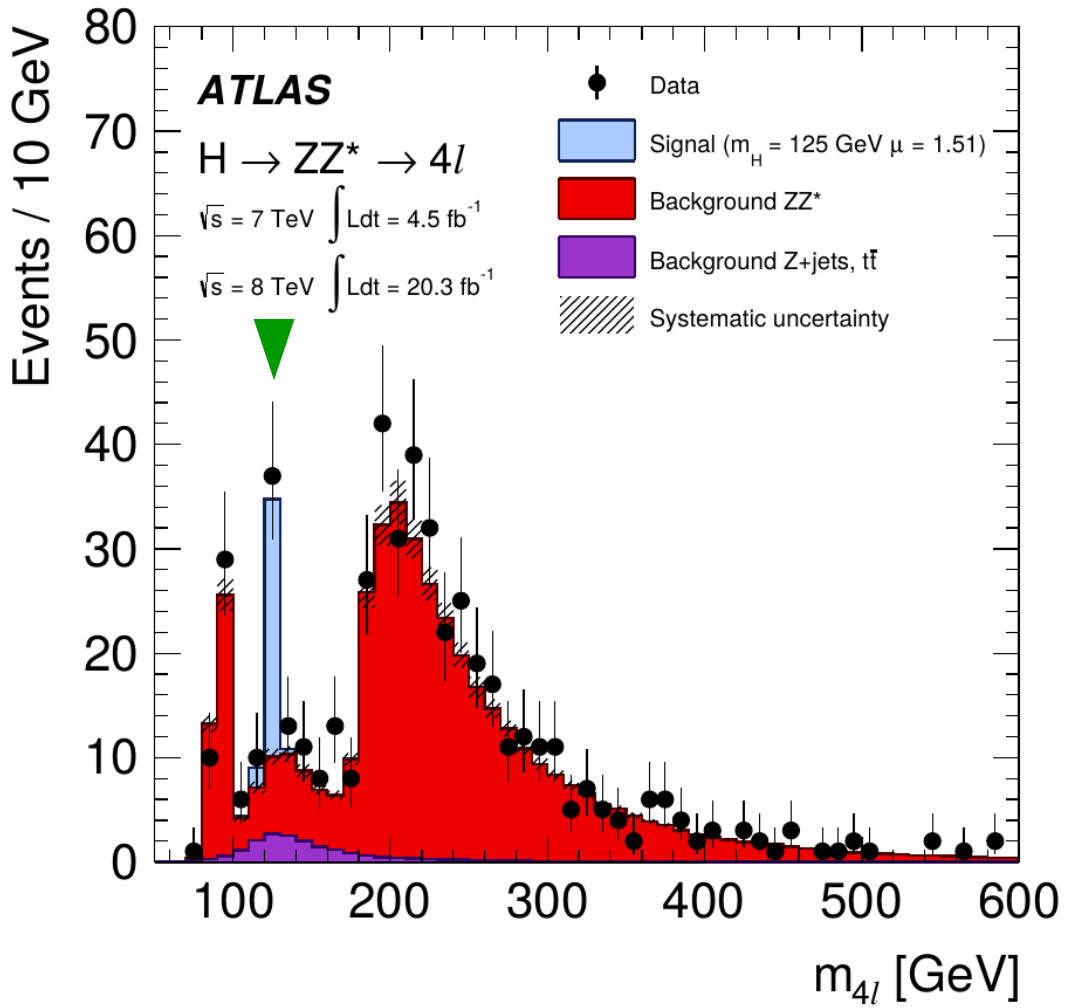


Cannot predict the measured value for a given event

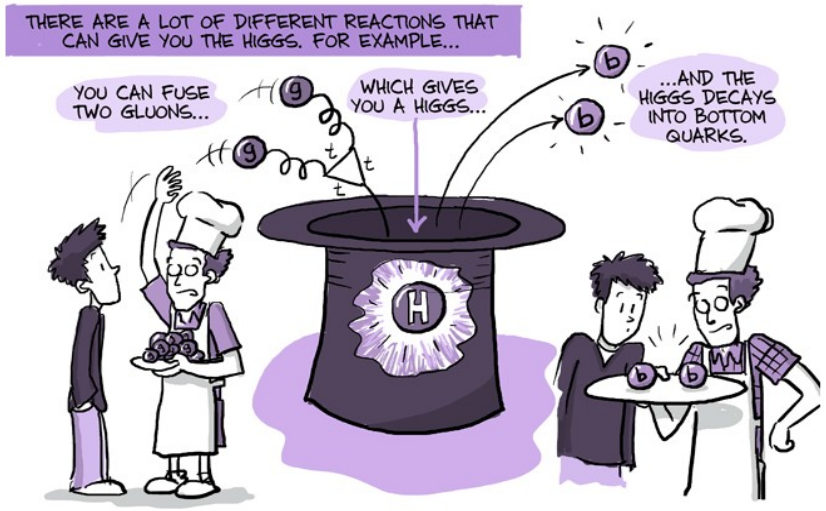
⇒ **Random process** ⇒ **Need a probabilistic description**

Quantum Randomness: H→ZZ*→4l

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

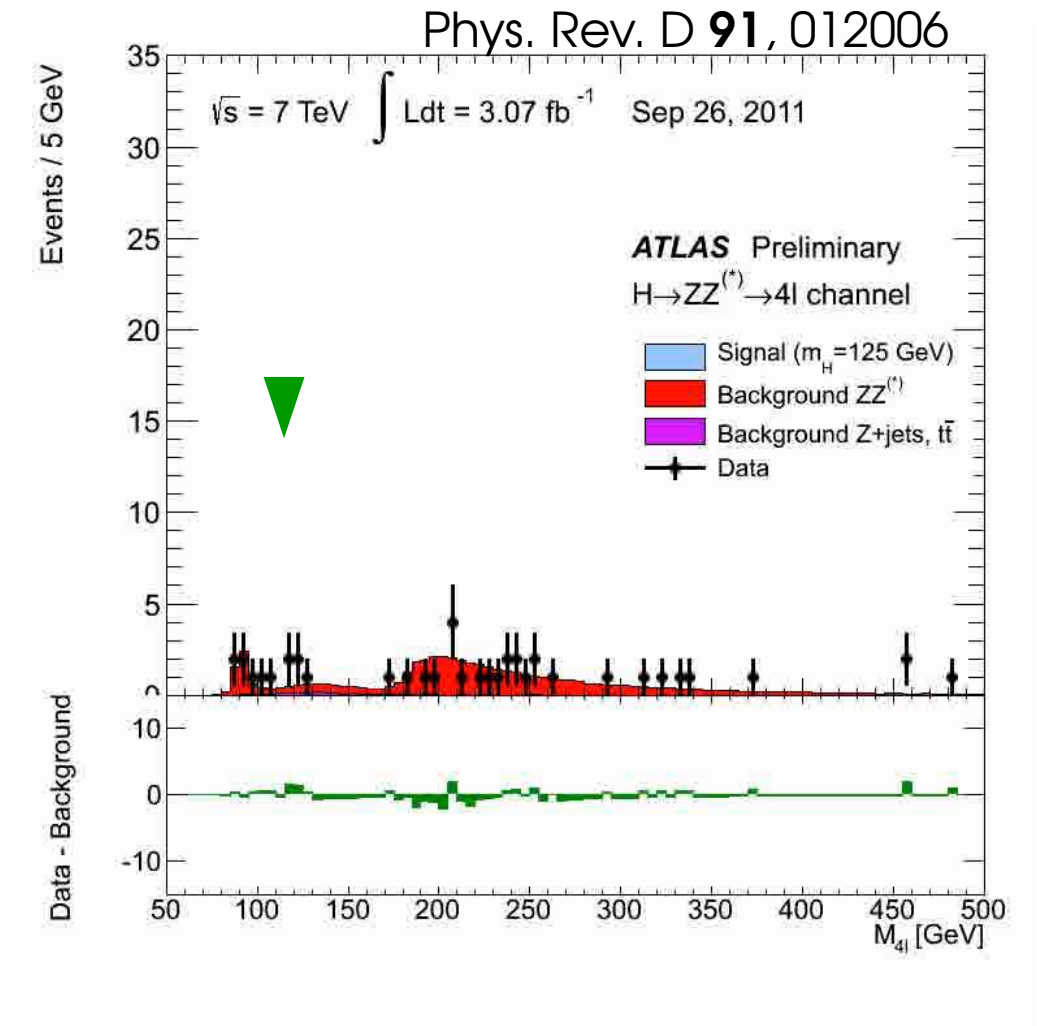


Rare process: Expect 1 signal event every **~6 days**

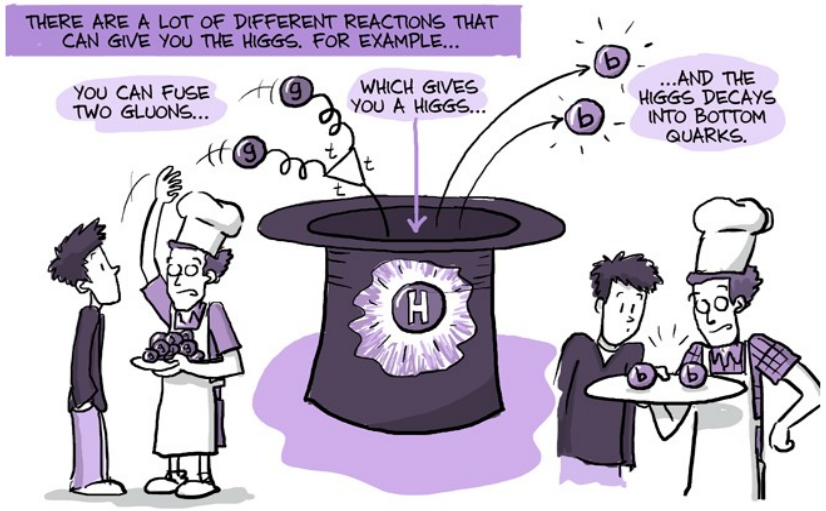


http://www.phdcomics.com/comics/archive.php?comicid=1489

Quantum Randomness: H→ZZ*→4l



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

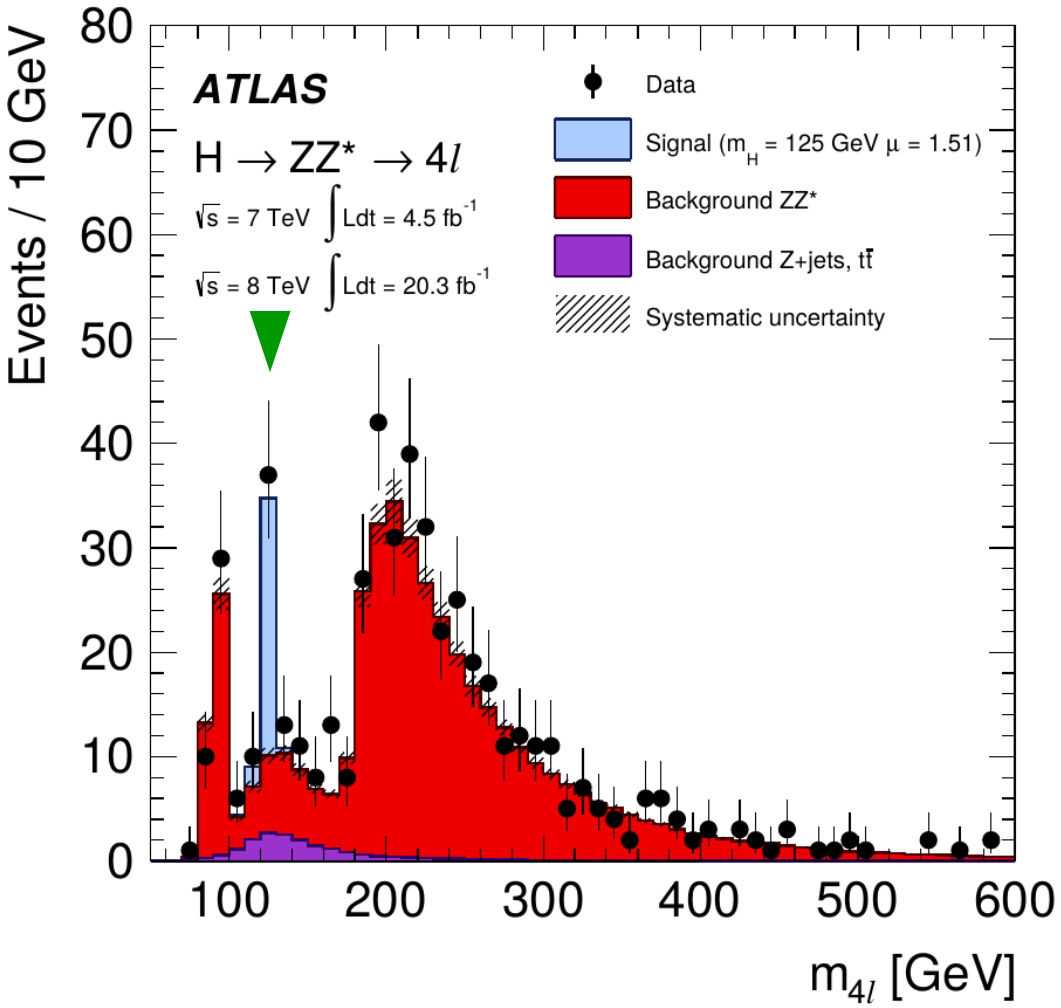


View online

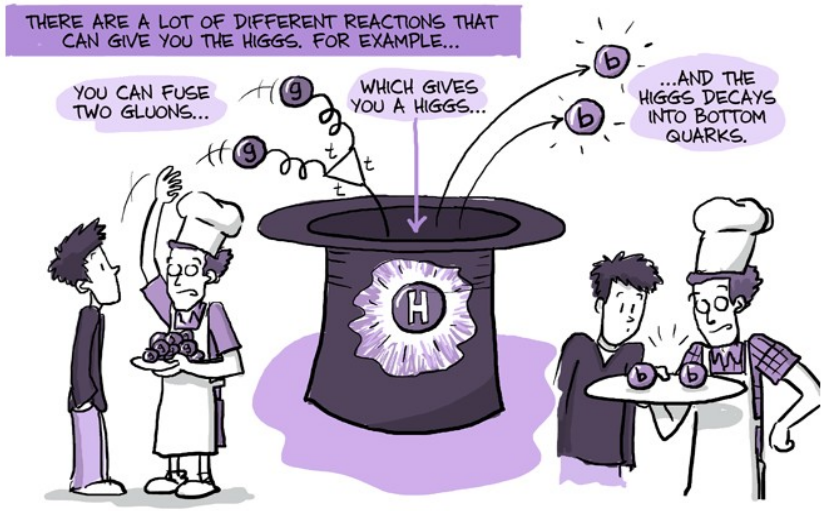
http://www.phdcomics.com/comics/archive.php?comicid=1489

Quantum Randomness: H→ZZ*→4l

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



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



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

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

Statistical Modeling

Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

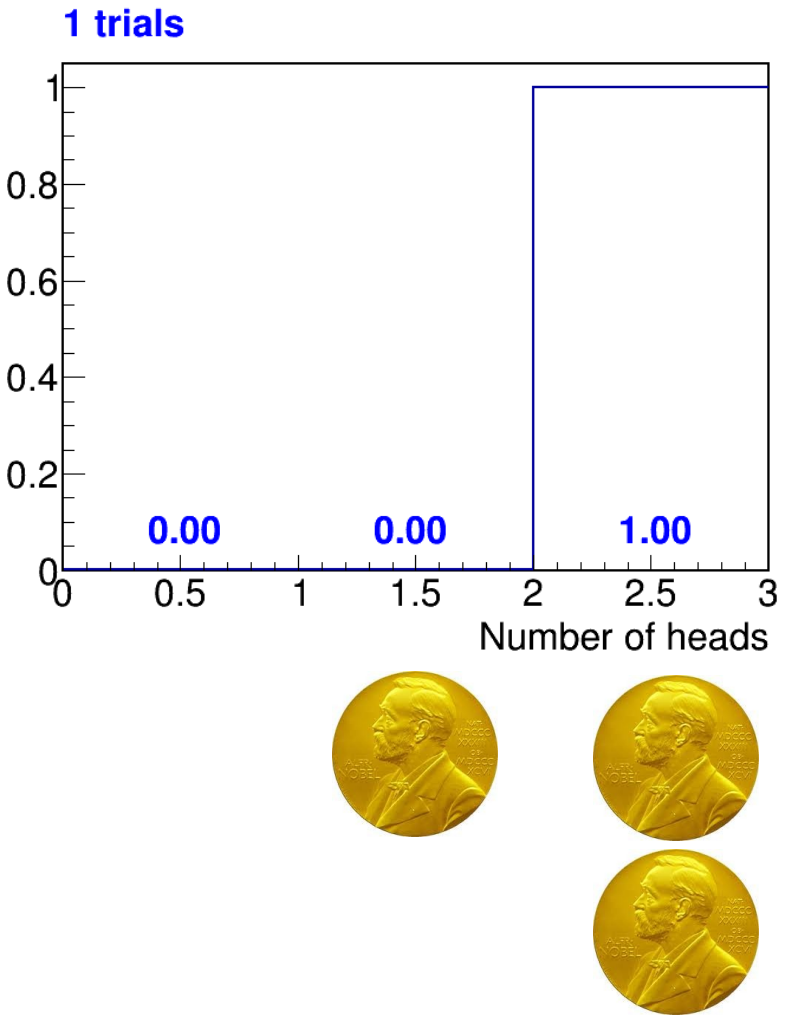
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

Properties

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



Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

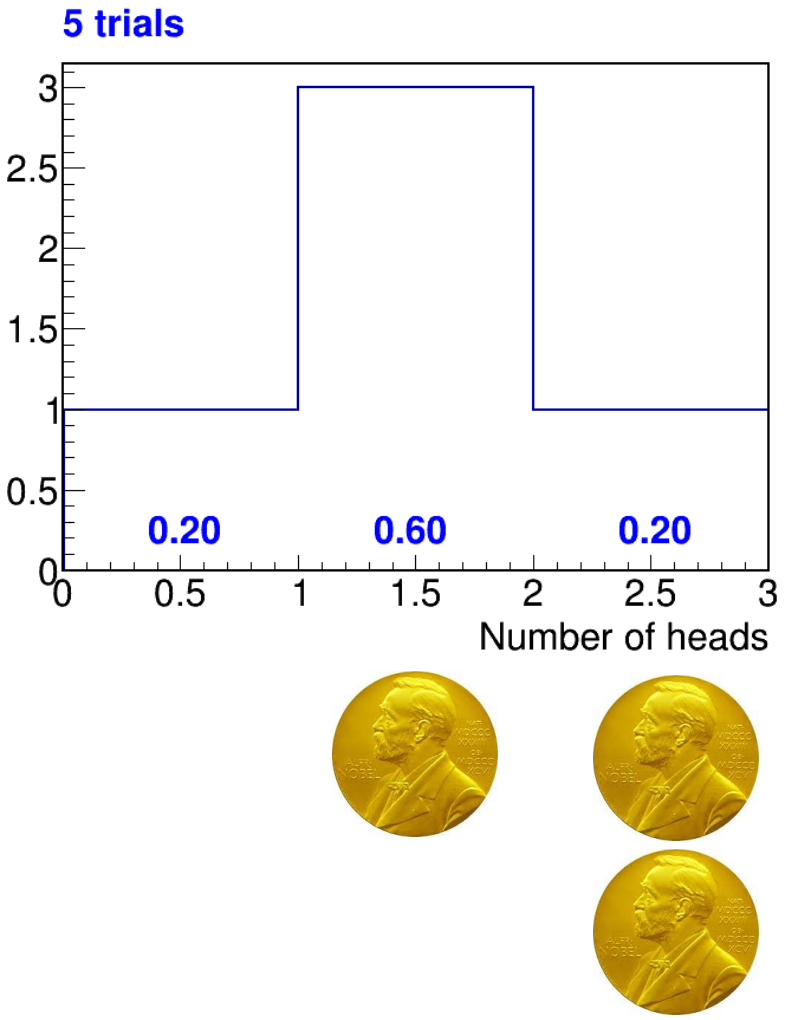
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

Properties

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



Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

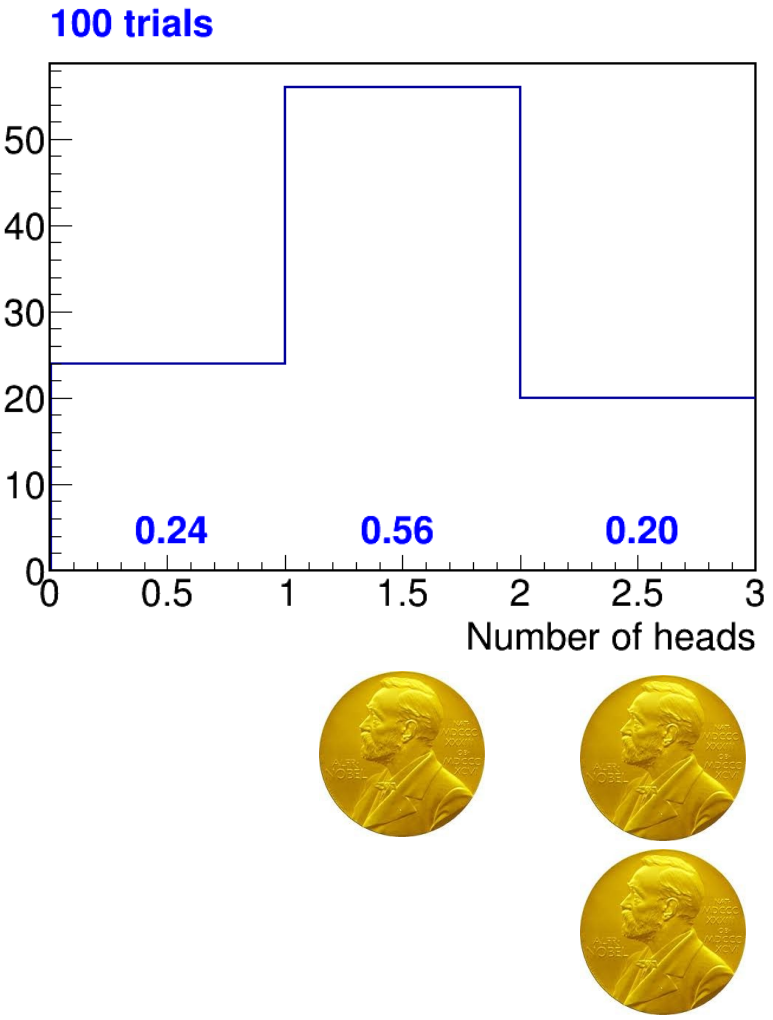
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

Properties

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



Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

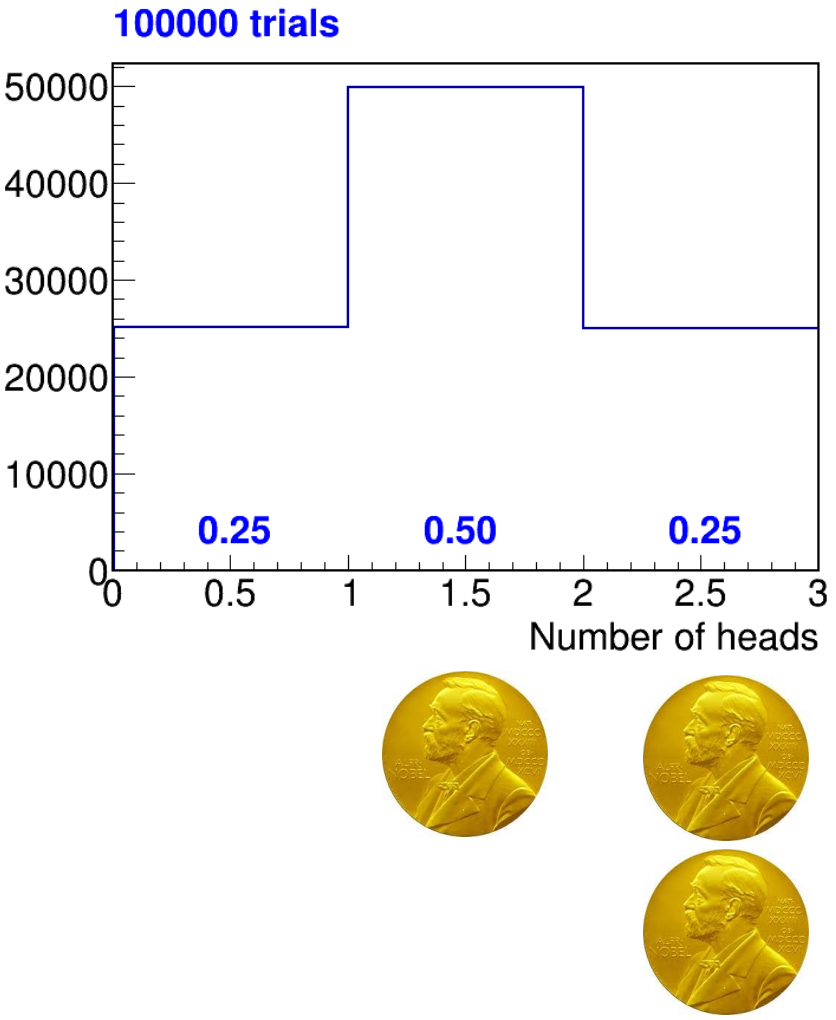
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

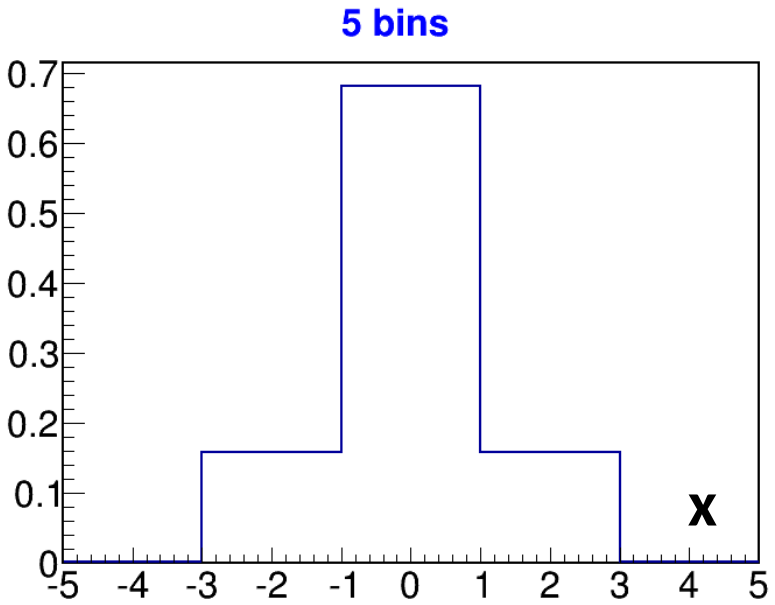
Properties

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



Continuous Variables: PDFs

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



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

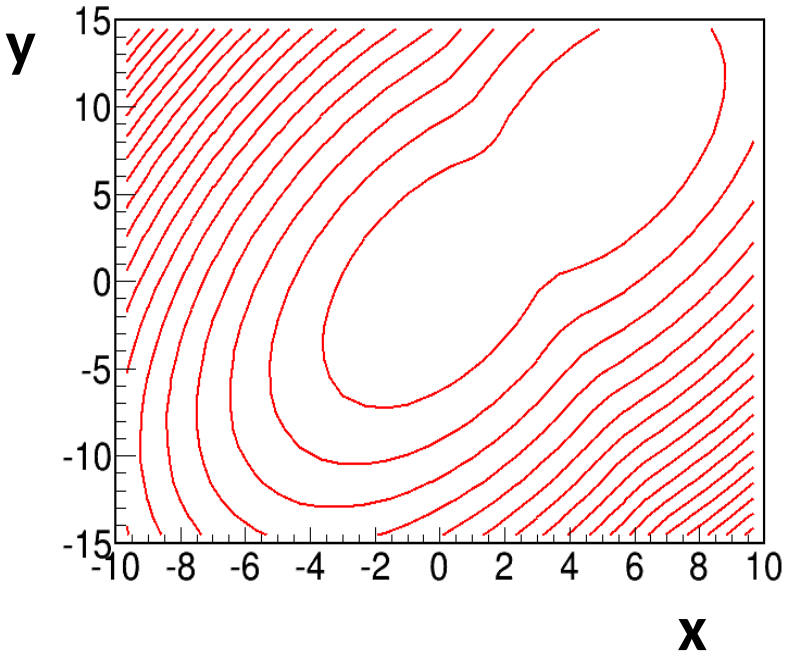
High PDF value

⇒ High chance to get a measurement here

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

Generalizes to **multiple variables** :

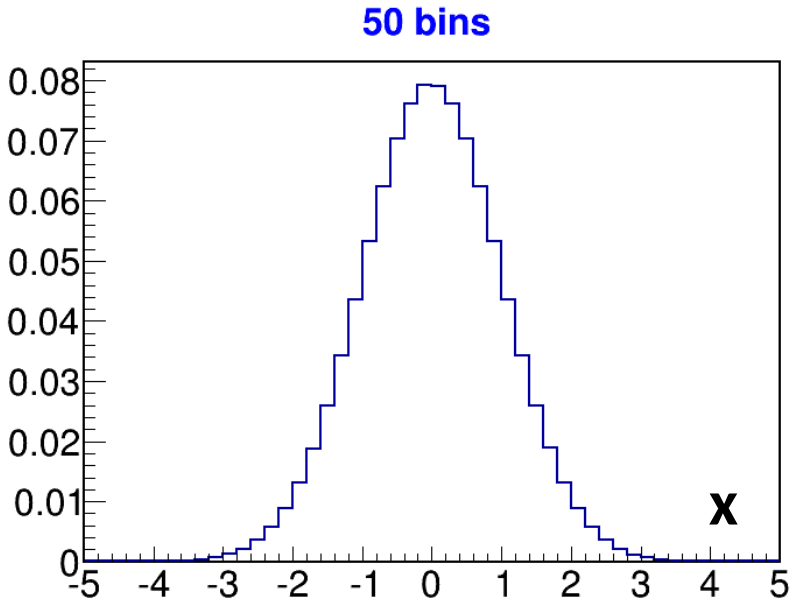
$P(x,y) > 0, \int P(x,y)\, dx\, dy = 1$



Contours: P(x,y)

Continuous Variables: PDFs

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



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

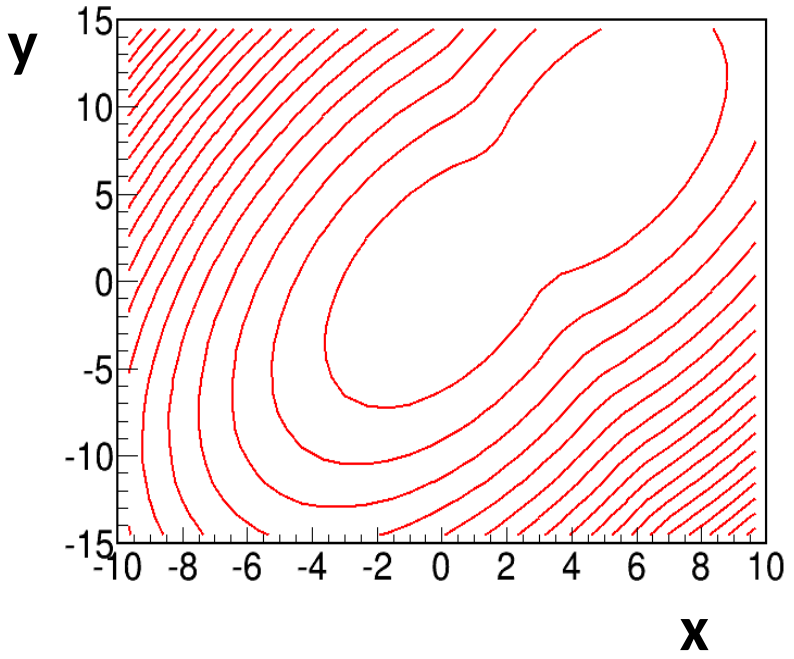
High PDF value

⇒ High chance to get a measurement here

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

Generalizes to **multiple variables** :

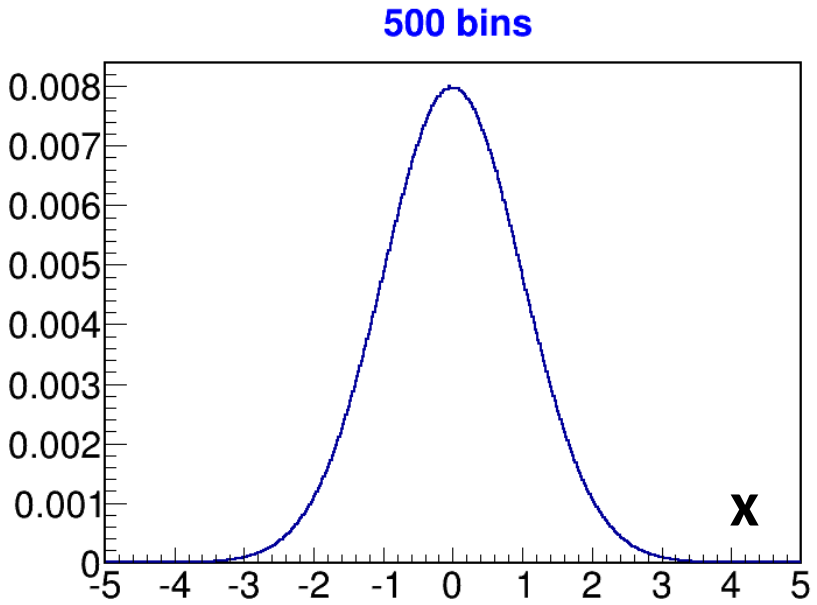
$P(x,y) > 0, \int P(x,y)\, dx\, dy = 1$



Contours: P(x,y)

Continuous Variables: PDFs

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



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

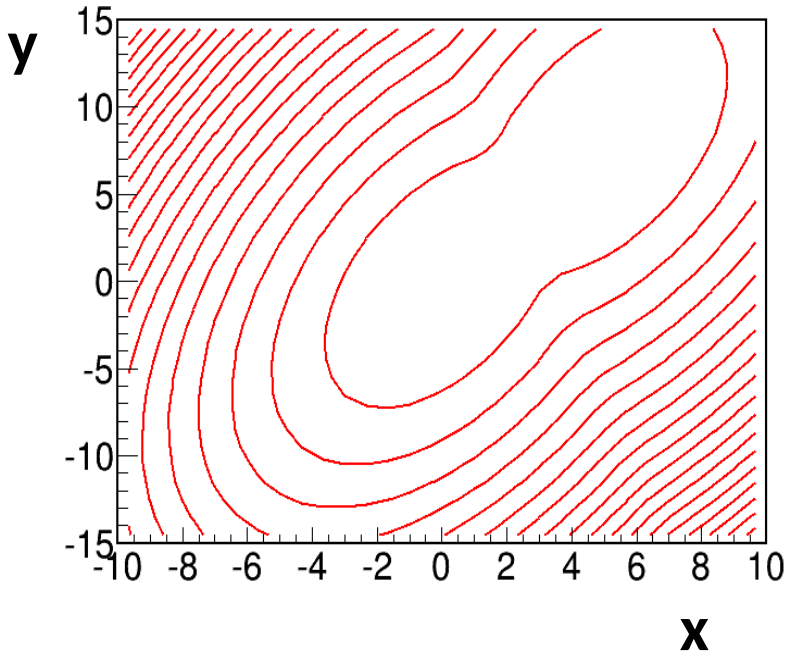
High PDF value

⇒ High chance to get a measurement here

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

Generalizes to **multiple variables** :

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



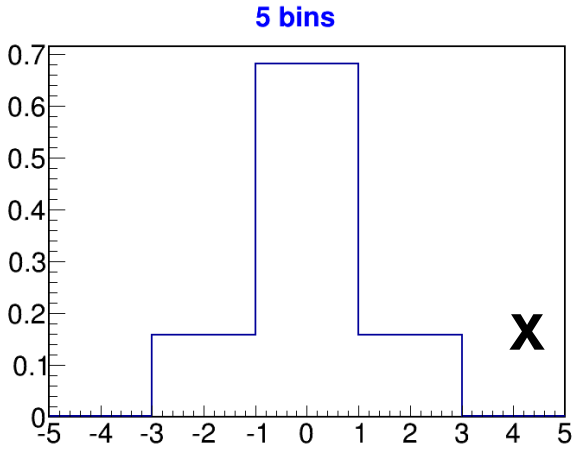
Contours: P(x,y)

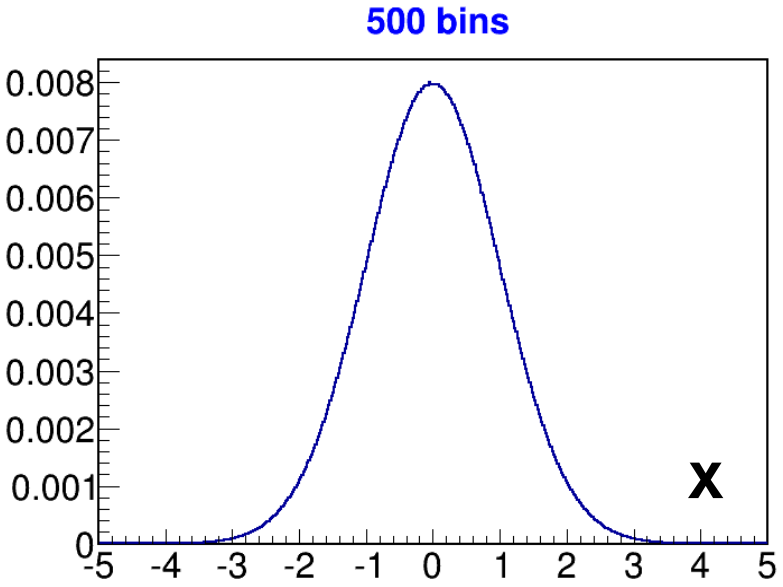
Random Variables

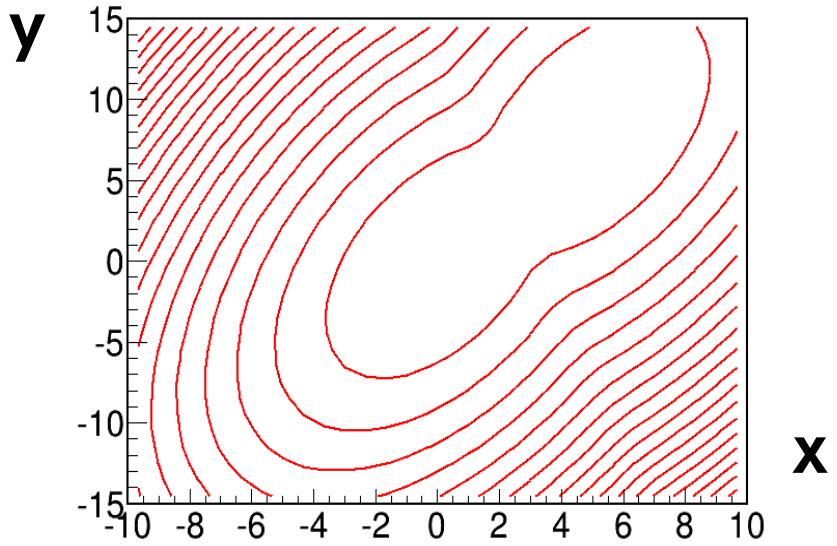
X, Y… are **Random Variables** (continuous or discrete), a.ka. **observables** :
→ X can take any value x, with probability **P(X=x).**

→ P(X=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).







PDF Properties: Mean

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

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

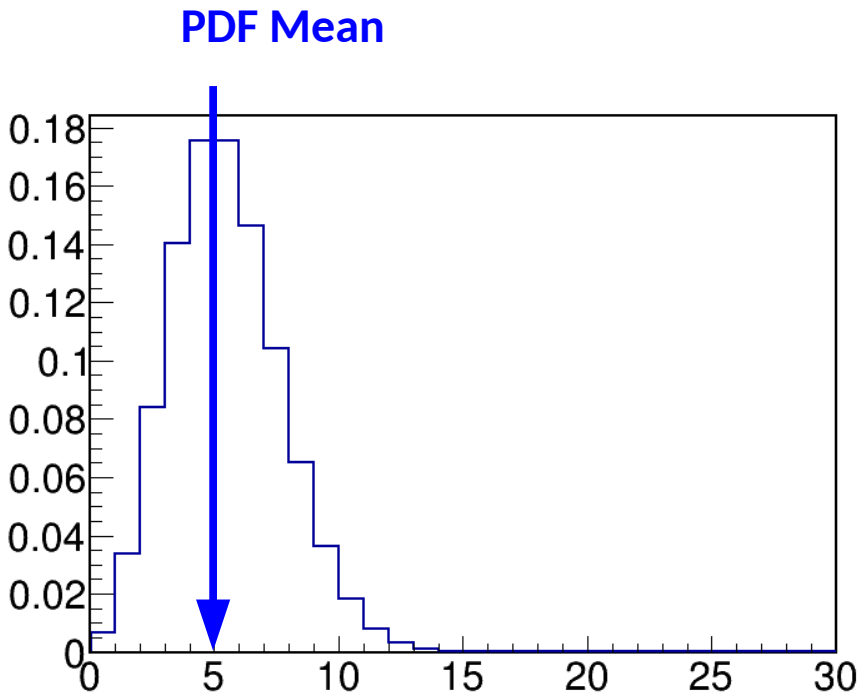
$$\langle X \rangle = \int x \, P(x) \, dx$$

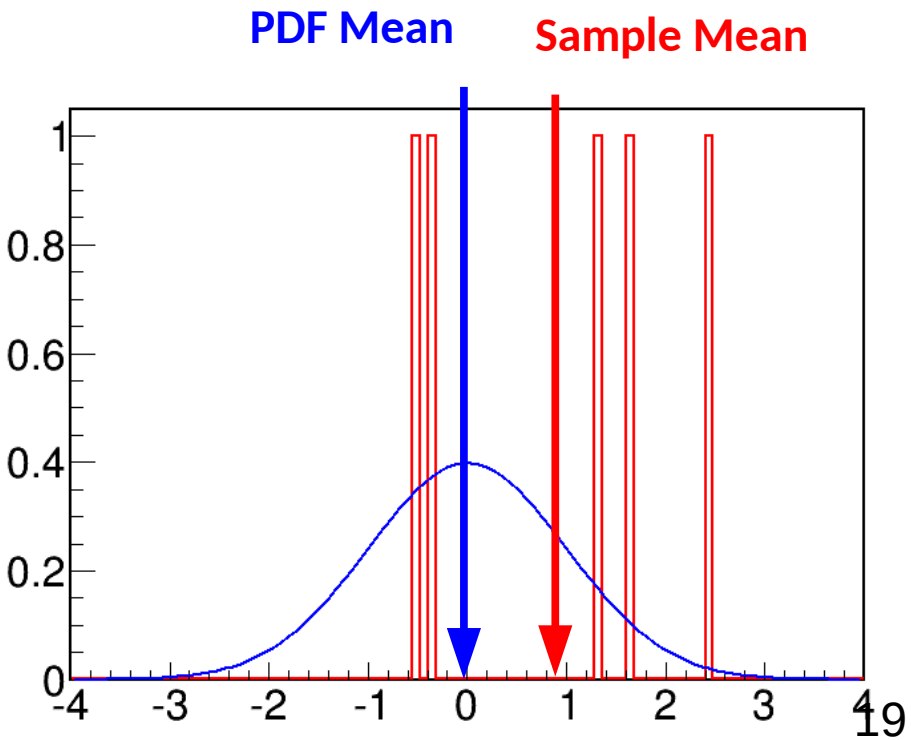
→ Property of the **PDF**

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:

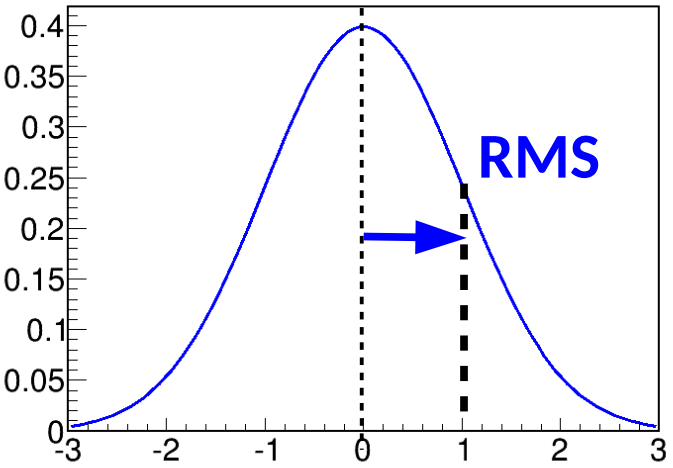
$$\mathbf{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$$

→ Average square of deviation from mean

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

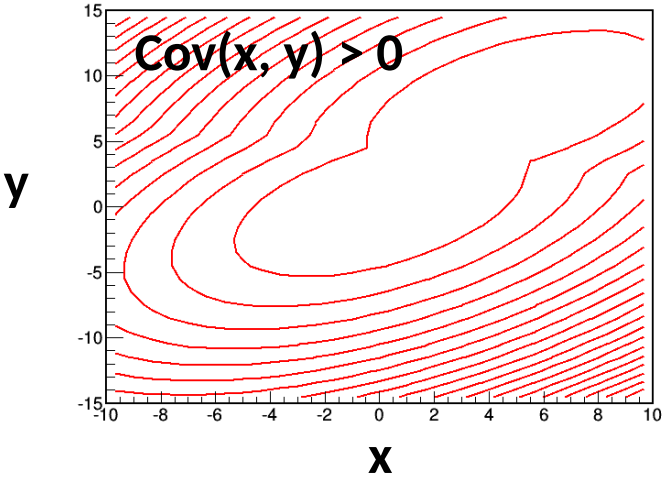
Can be approximated by **sample variance**:

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



Covariance of X and Y:

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



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

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

PDF Properties: (Co)variance

Variance of X:

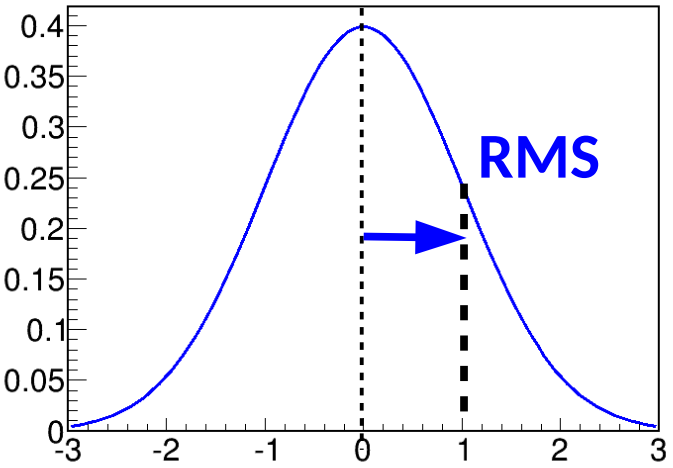
$$\mathrm{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$$

→ Average square of deviation from mean

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

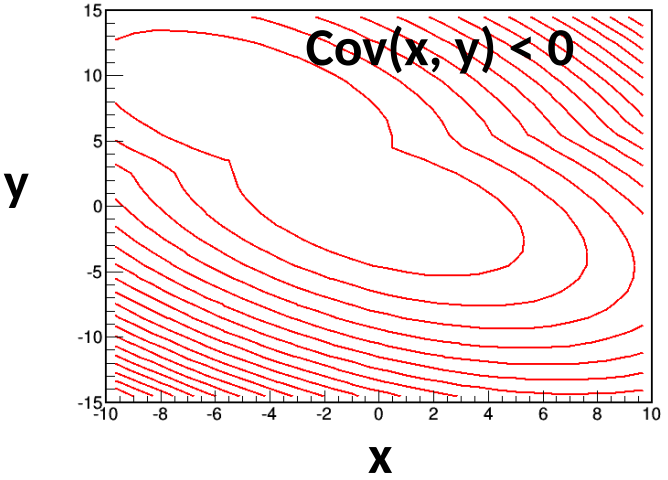
Can be approximated by **sample variance**:

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



Covariance of X and Y:

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



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

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

PDF Properties: (Co)variance

Variance of X:

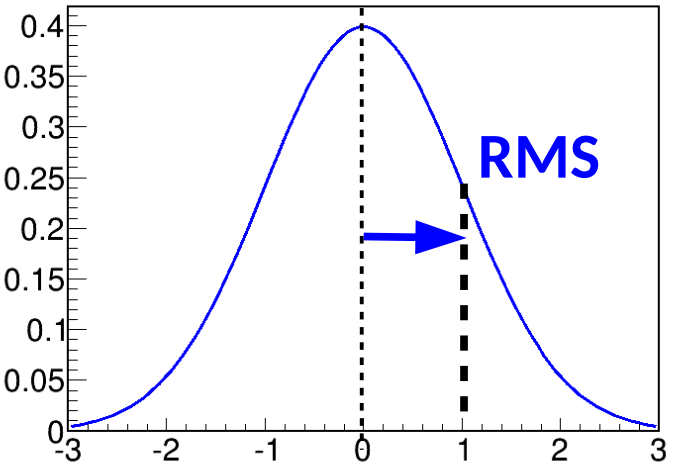
$$\mathbf{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$$

→ Average square of deviation from mean

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

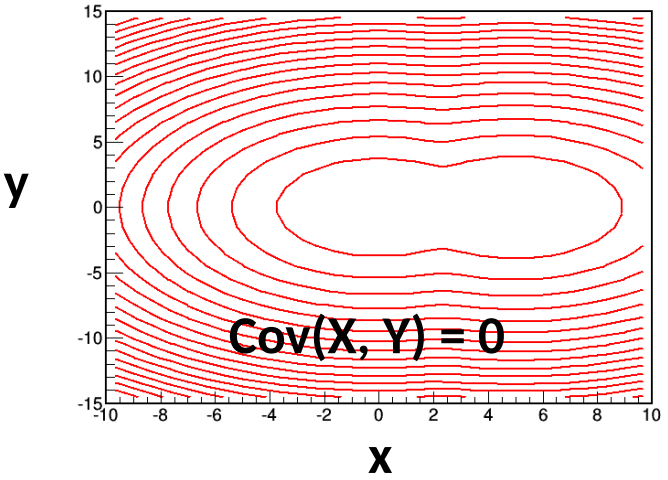
Can be approximated by **sample variance**:

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



Covariance of X and Y:

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

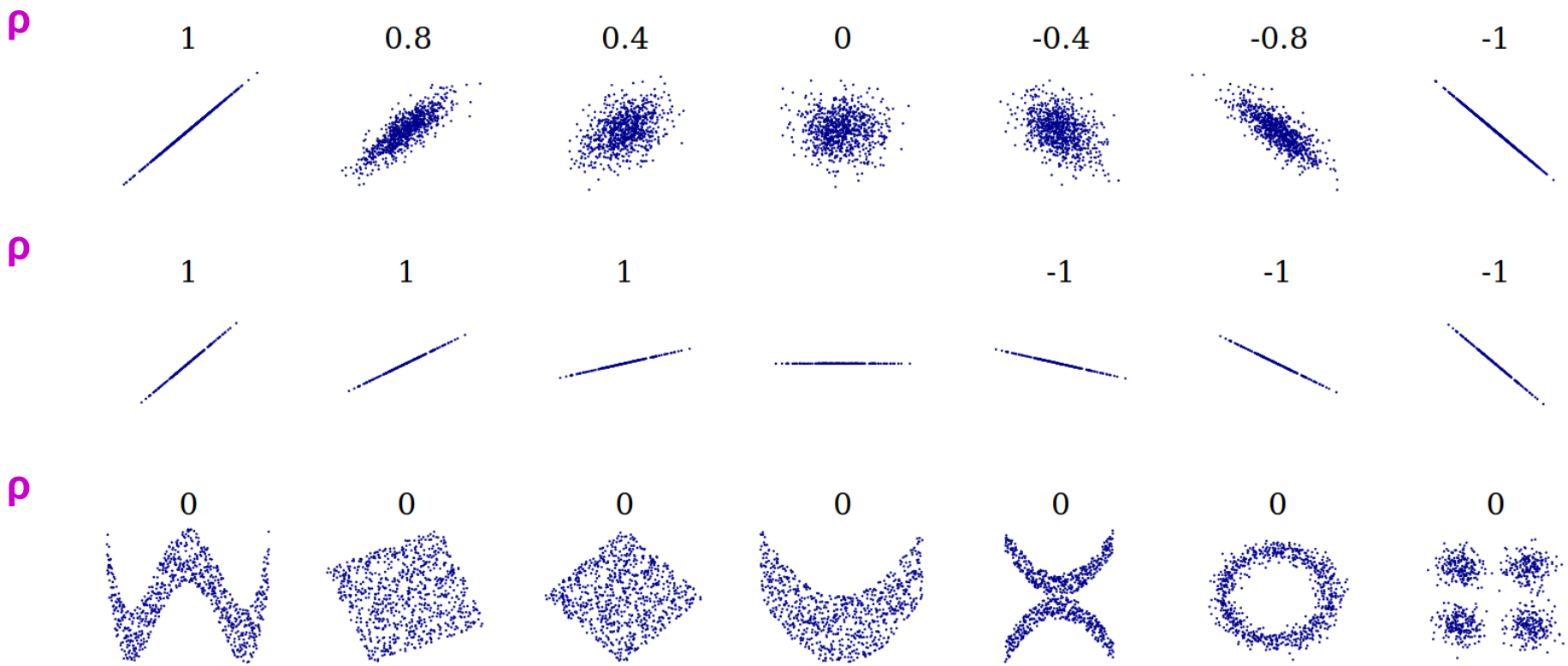


→ Large if variations of X and Y are "synchronized"

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

“Linear” vs. “non-linear” correlations

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



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

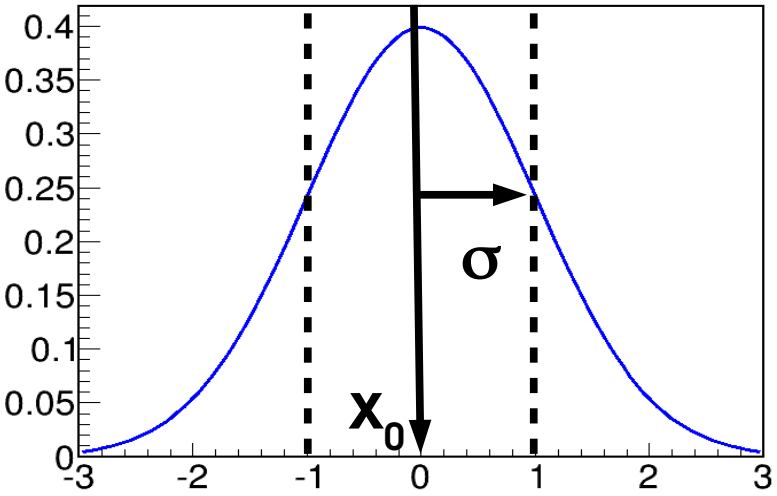
Source: Wikipedia

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

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**:

→ **Mean** : X_0

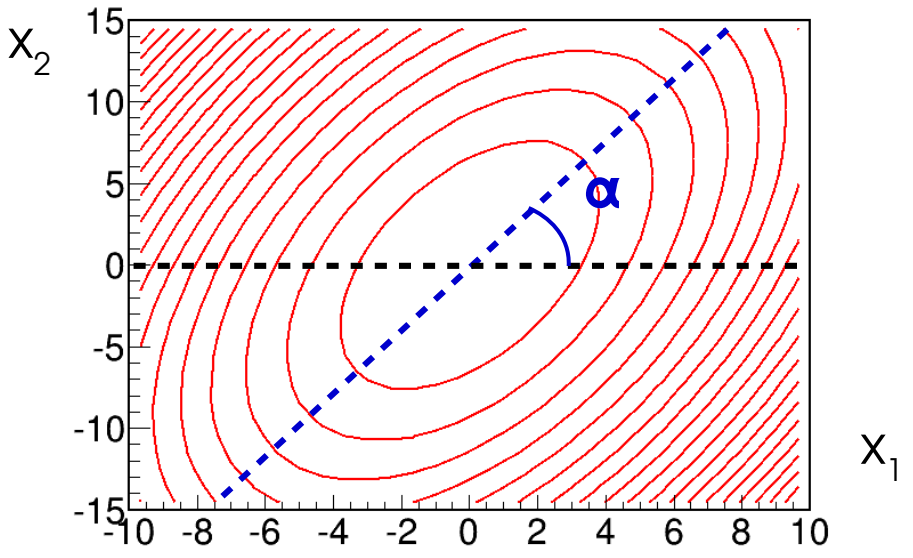
→ **Covariance matrix** :

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

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

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

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



Gaussian Quantiles

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

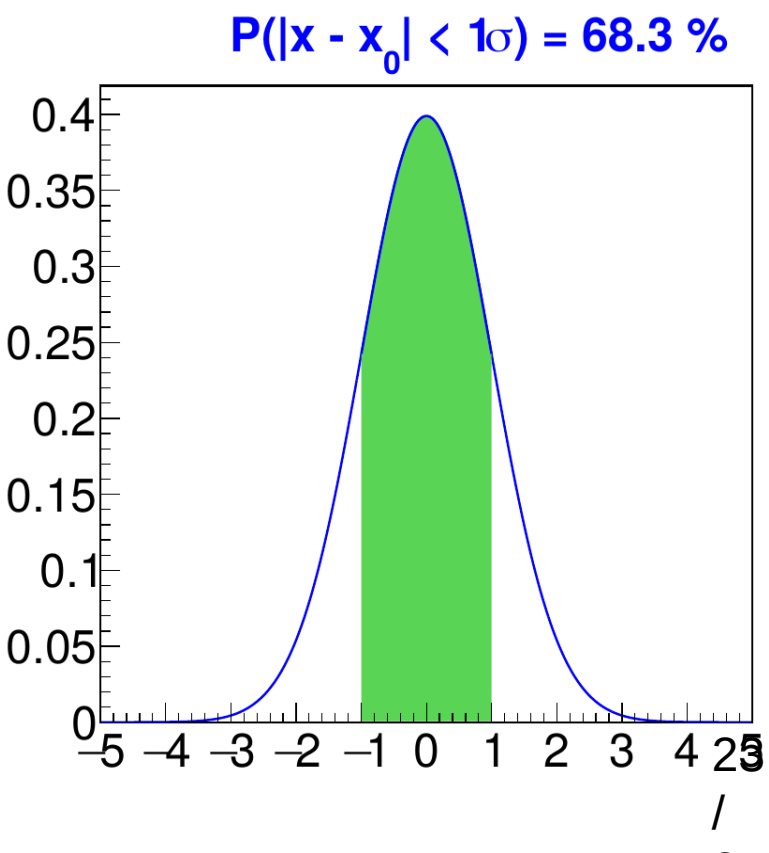
$G(x;x_0,\sigma)$ depends only on $z \sim G(z; 0,1)$

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)** :

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



Gaussian Quantiles

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

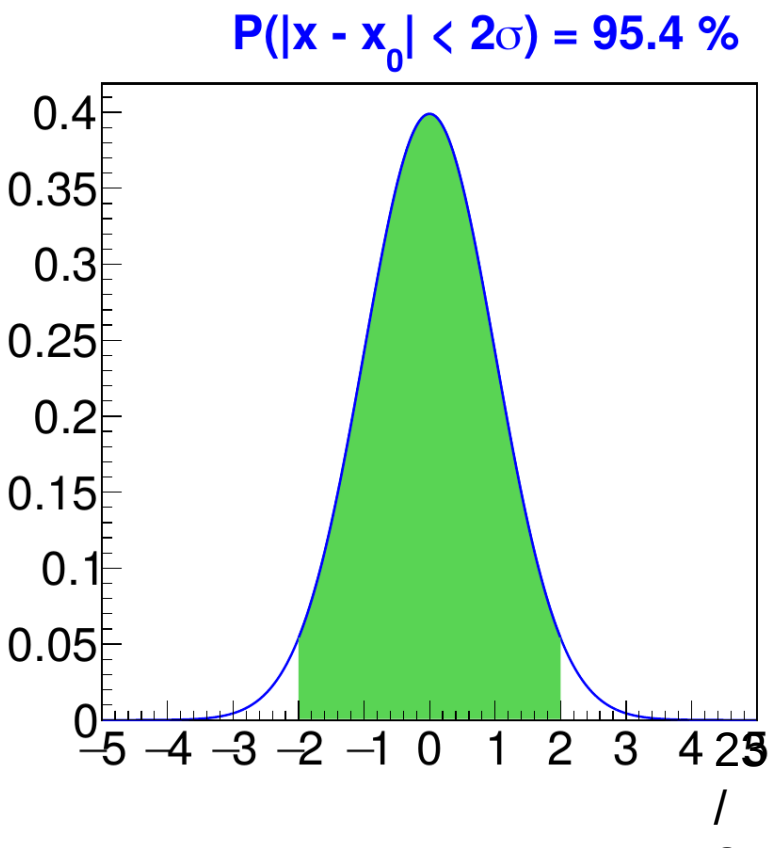
$G(x;x_0,\sigma)$ depends only on $z \sim G(z; 0,1)$

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)** :

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



Gaussian Quantiles

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

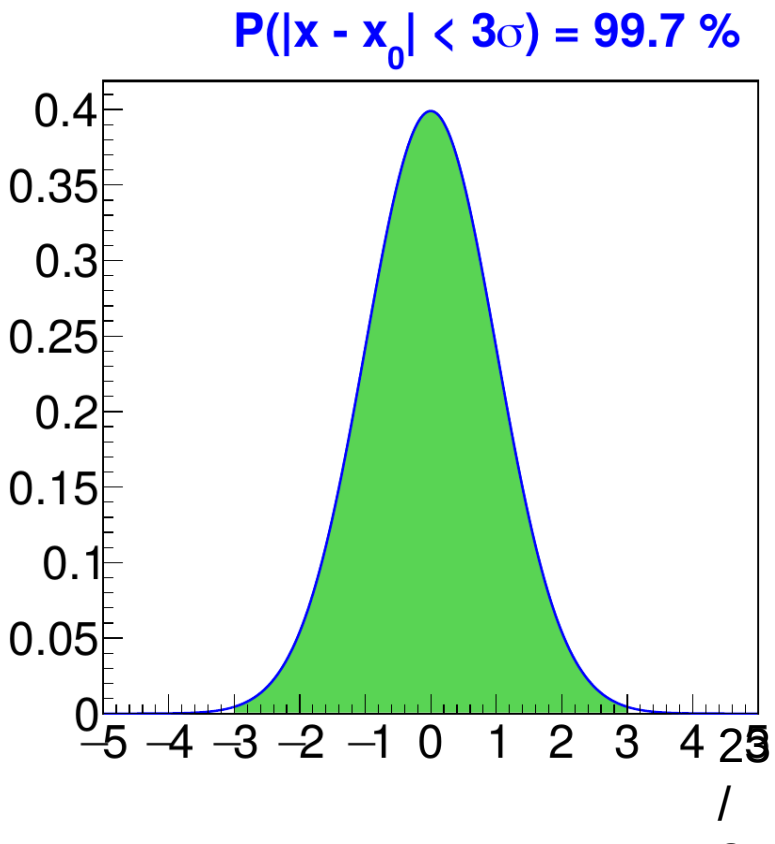
$G(x;x_0,\sigma)$ depends only on $z \sim G(z; 0,1)$

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)** :

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



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

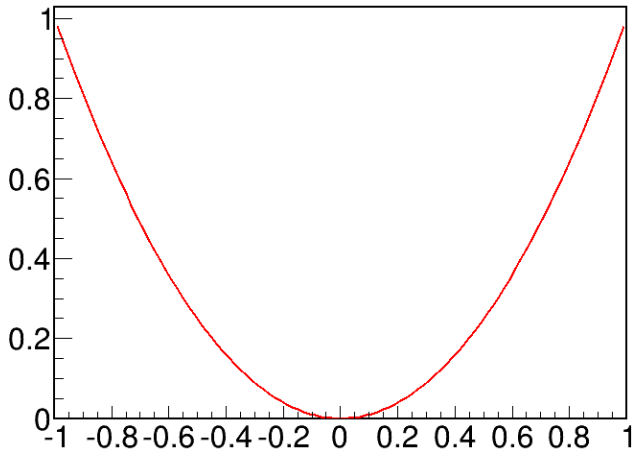
Another version:

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

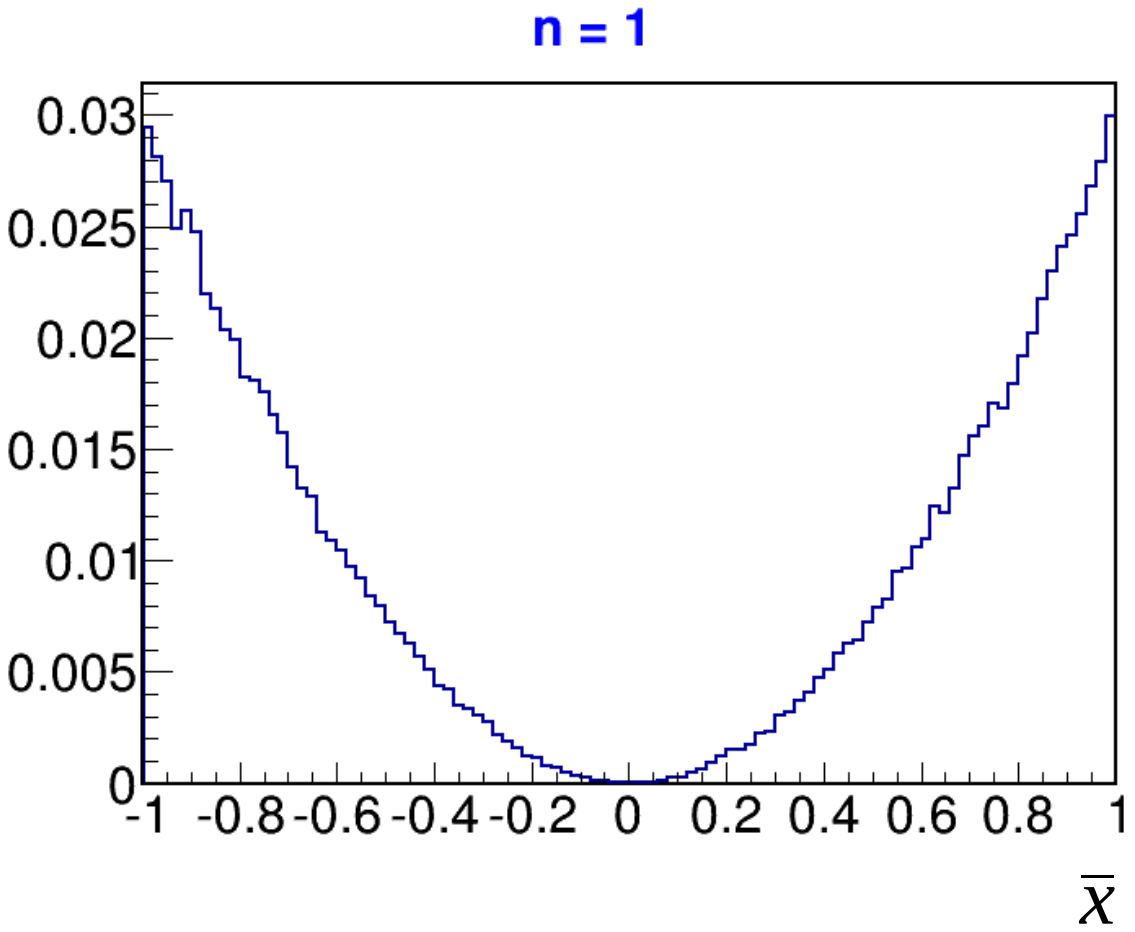
Mean scales like n, but RMS only like √n

Central Limit Theorem in action

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



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

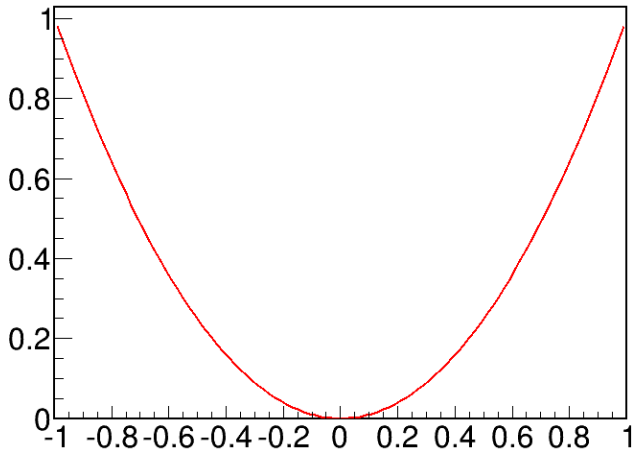


Distribution becomes Gaussian, although very non-Gaussian originally

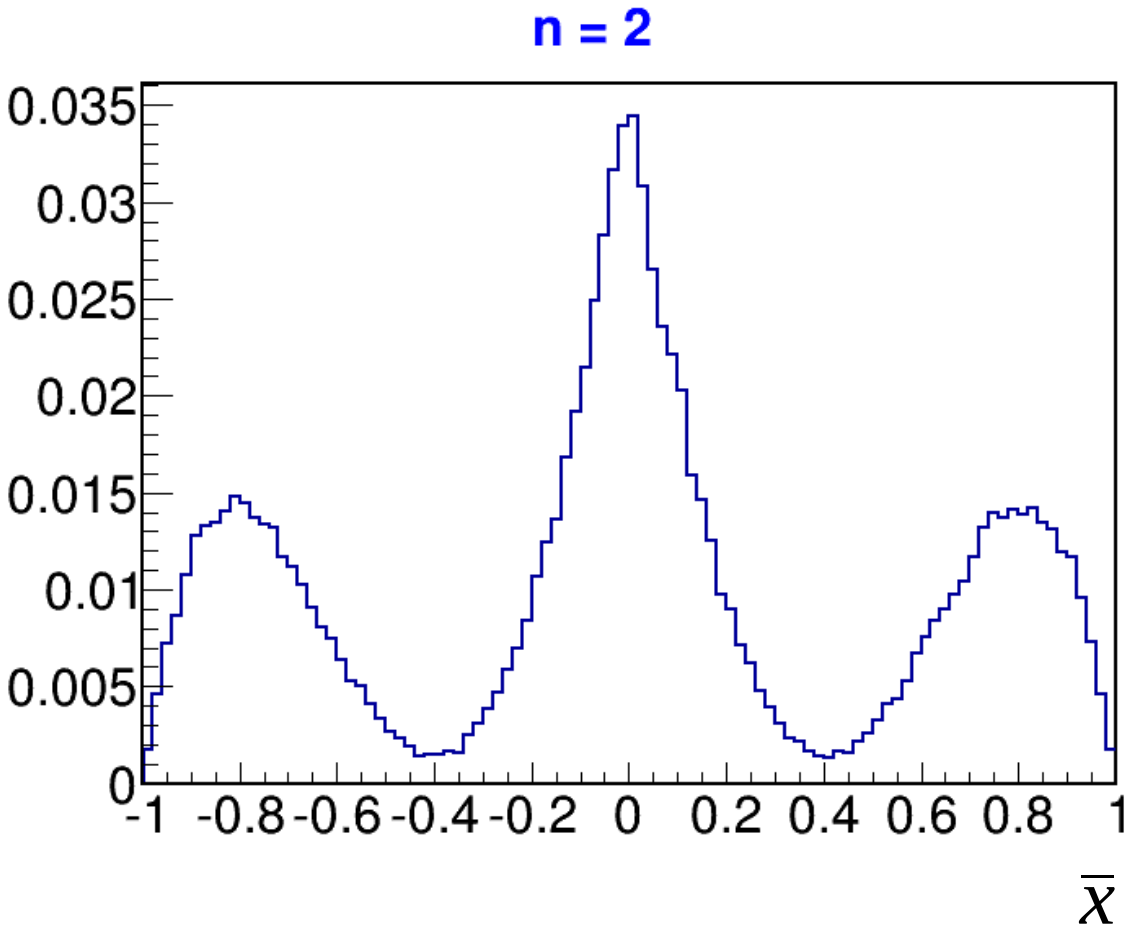
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

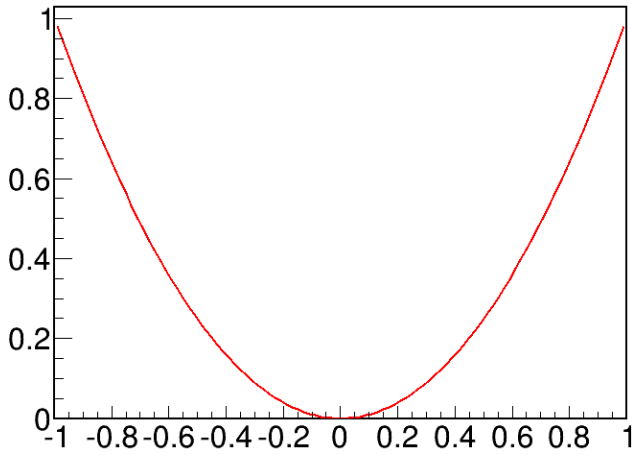


Distribution becomes Gaussian, although very non-Gaussian originally

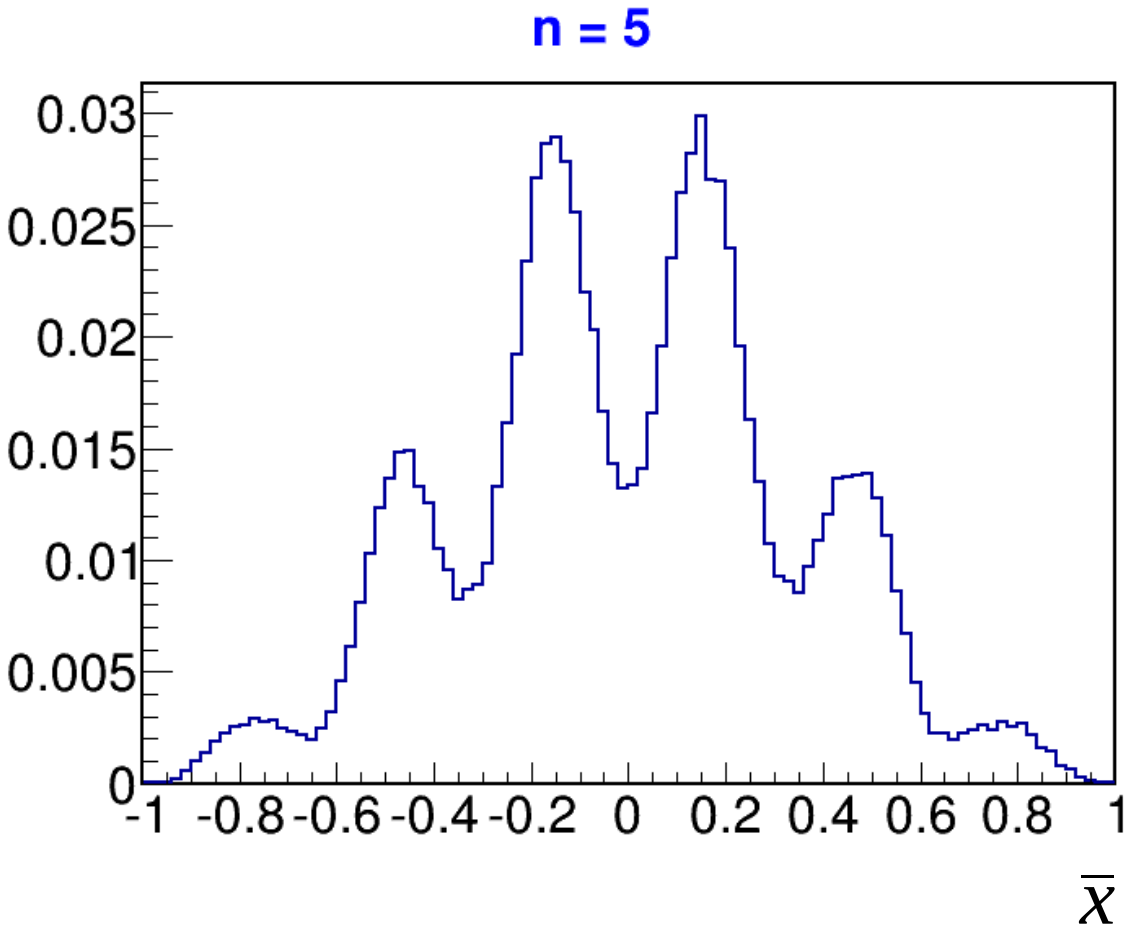
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

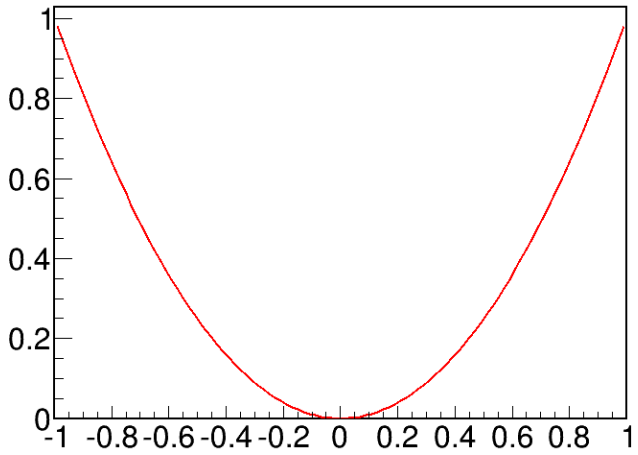


Distribution becomes Gaussian, although very non-Gaussian originally

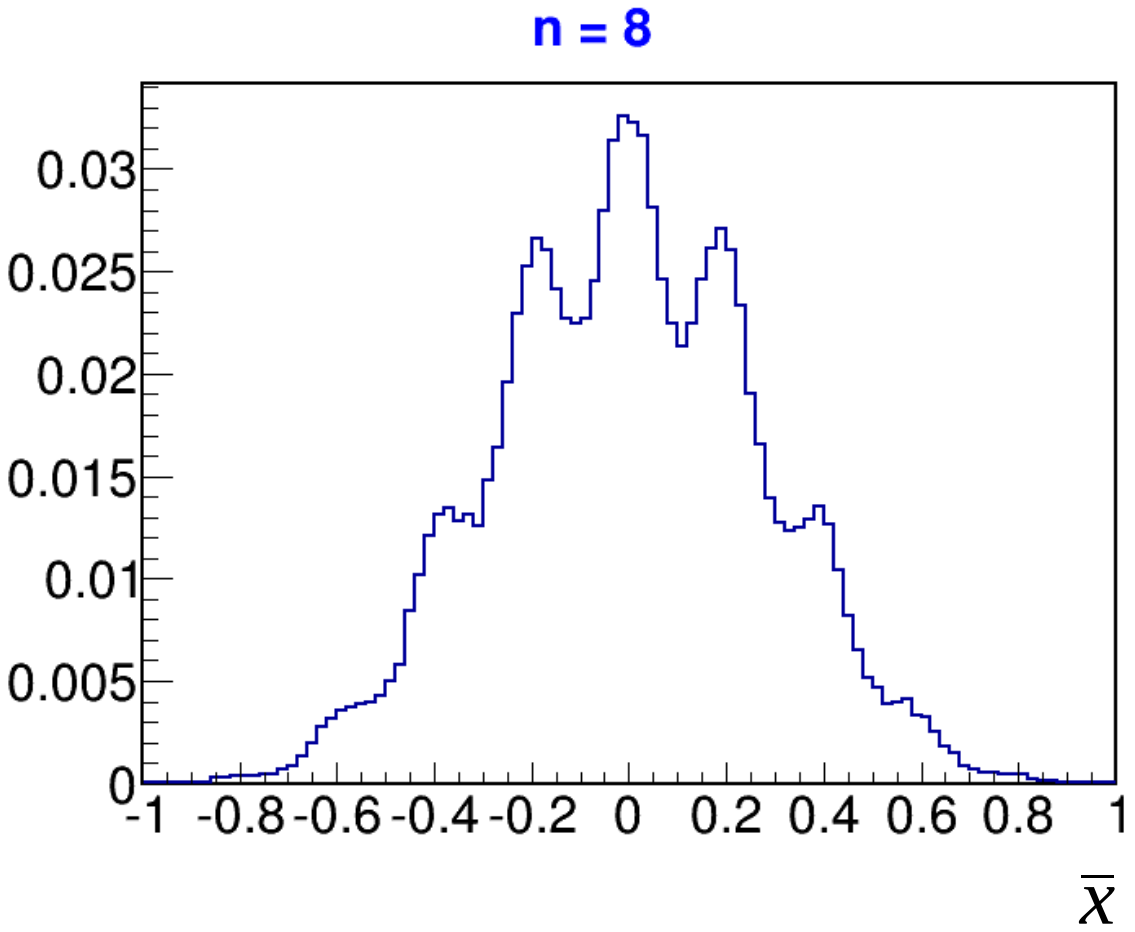
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

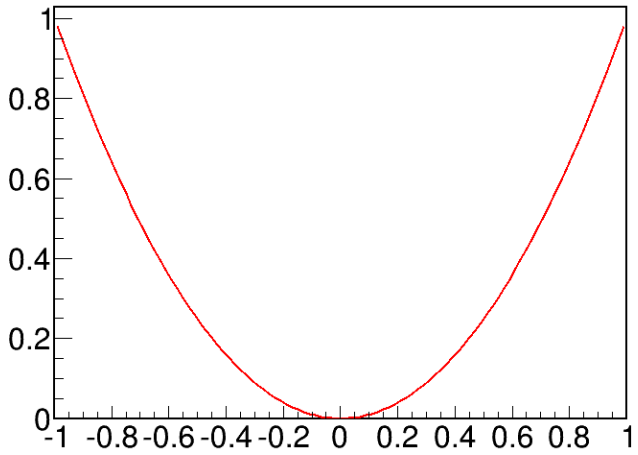


Distribution becomes Gaussian, although very non-Gaussian originally

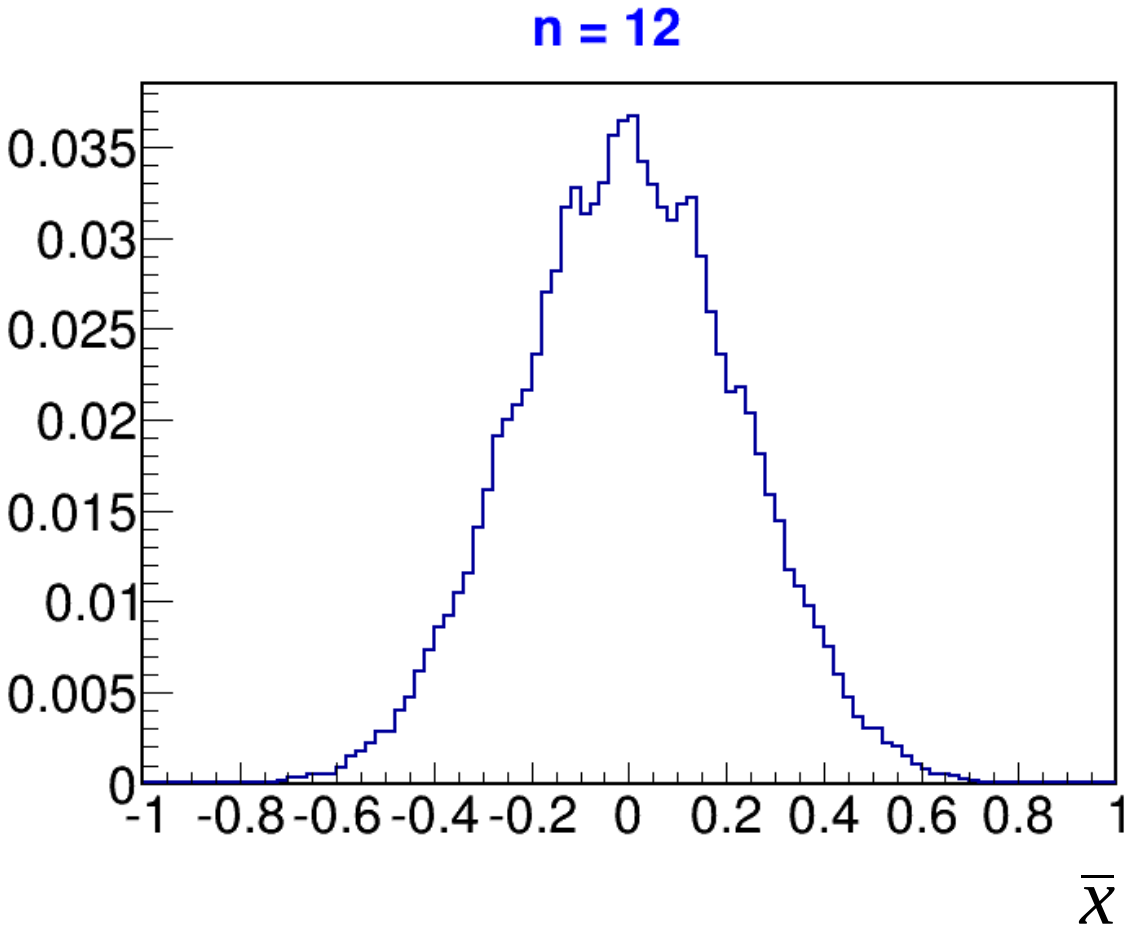
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

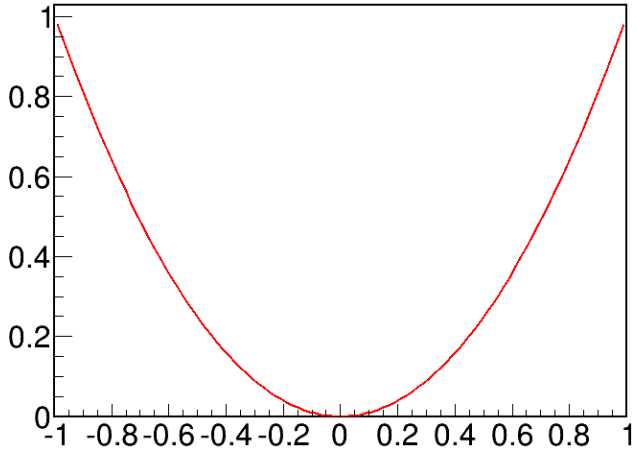


Distribution becomes Gaussian, although very non-Gaussian originally

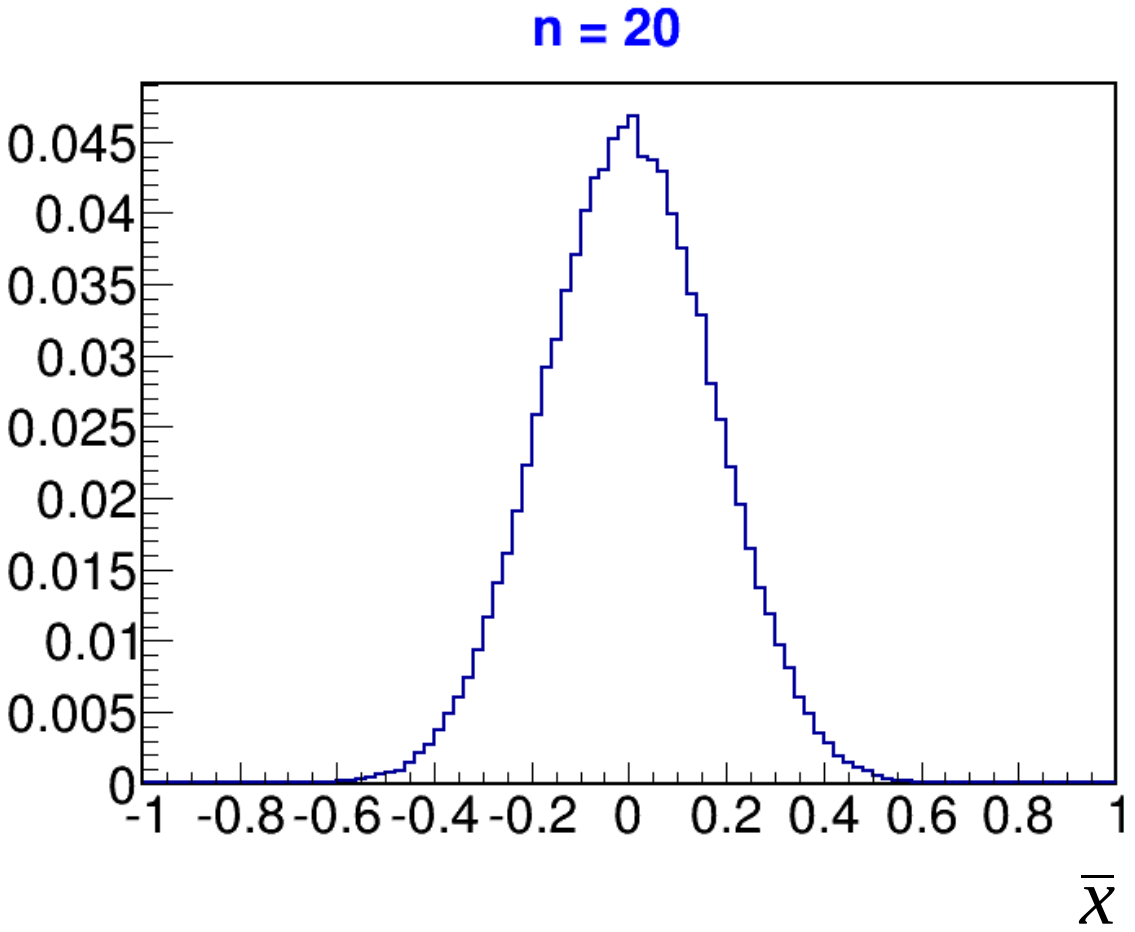
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

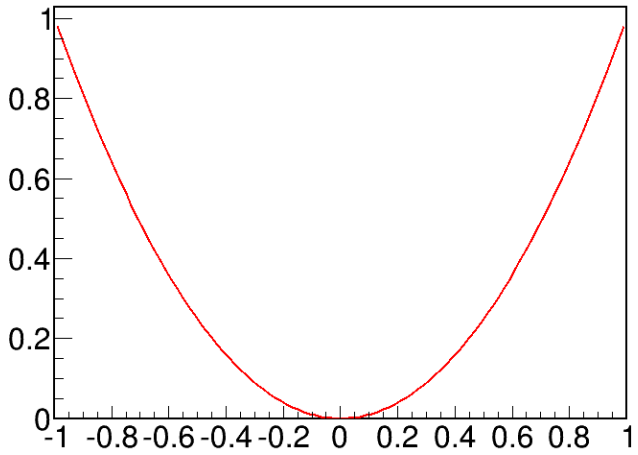


Distribution becomes Gaussian, although very non-Gaussian originally

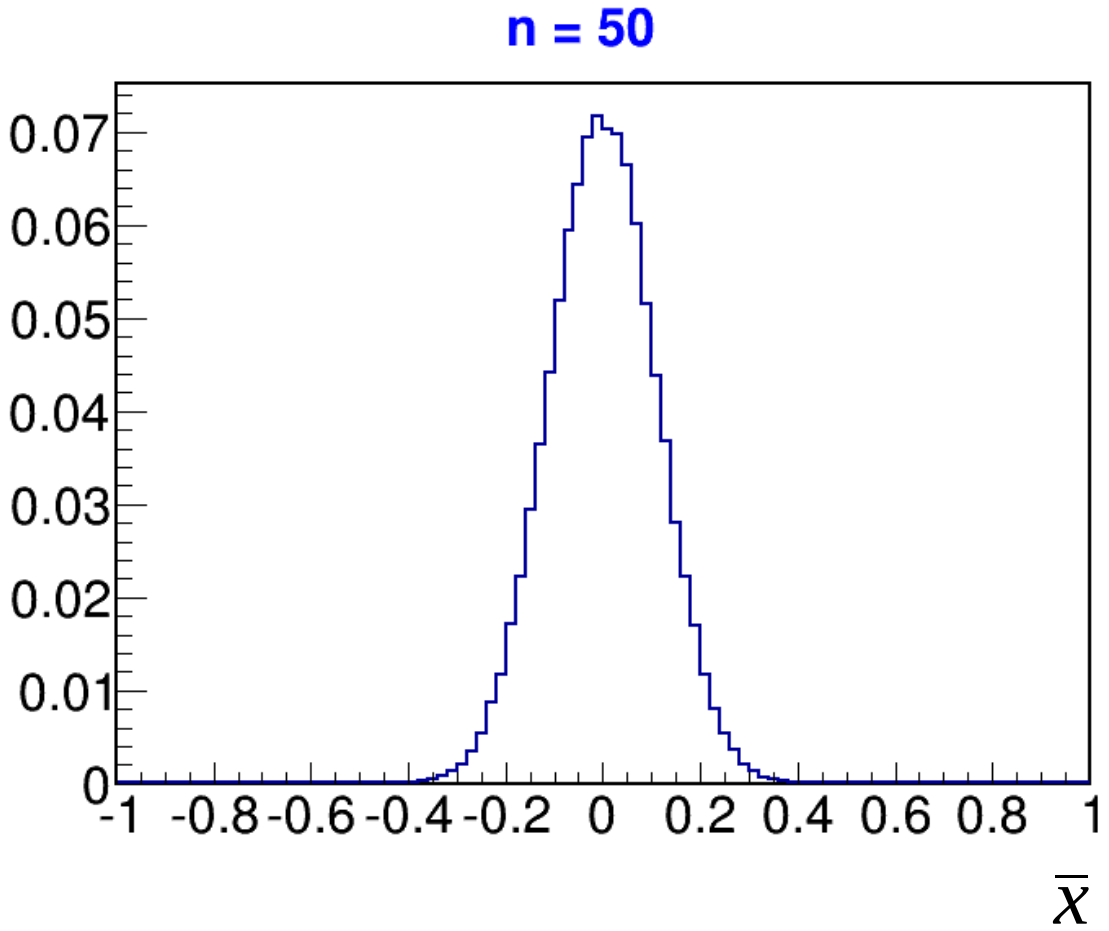
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

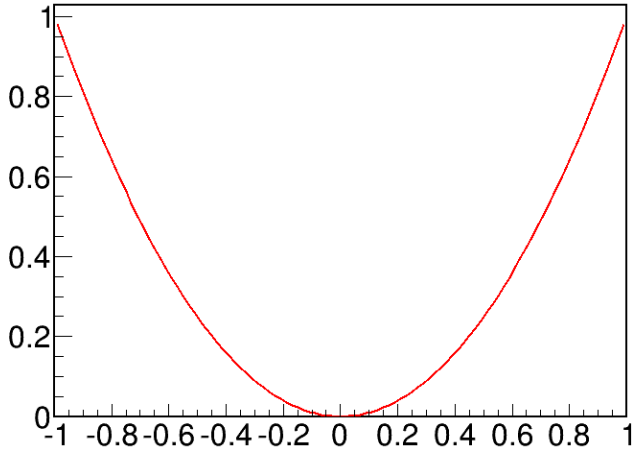


Distribution becomes Gaussian, although very non-Gaussian originally

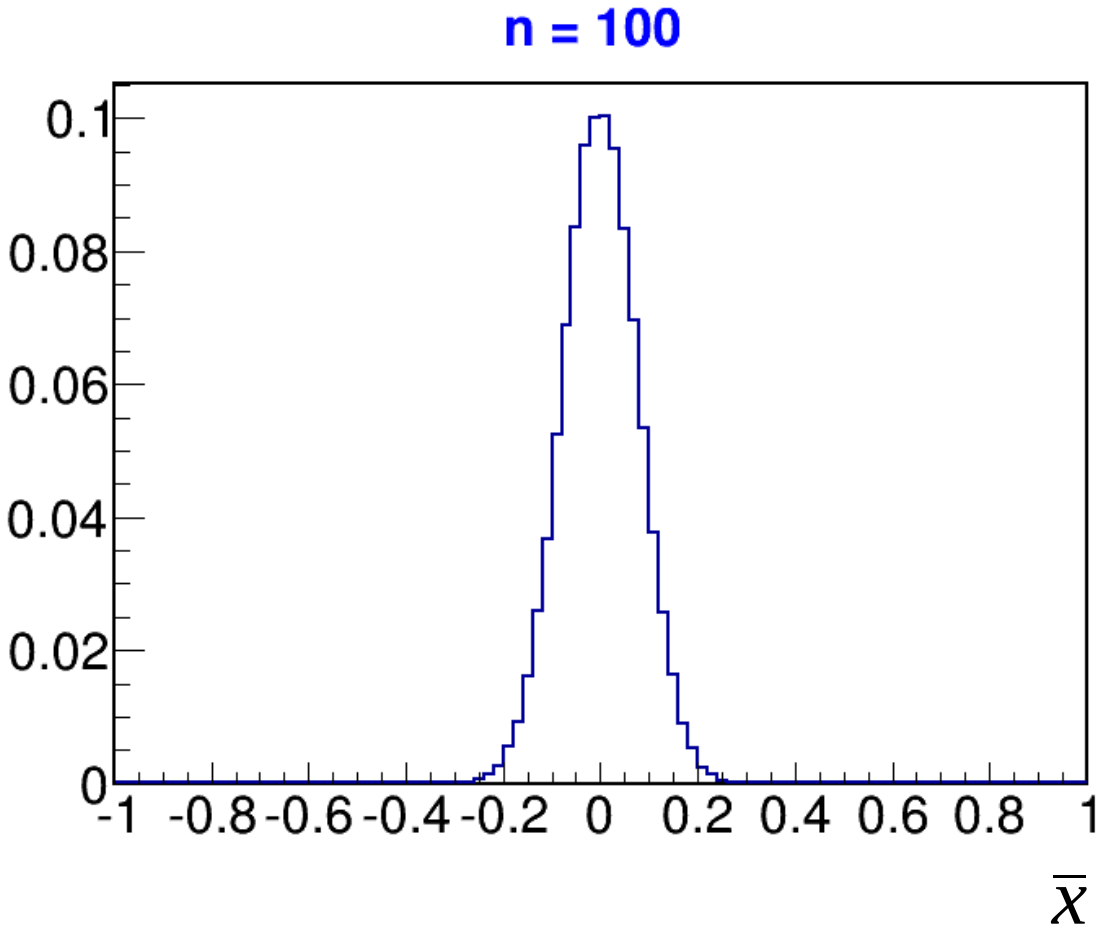
Distribution becomes narrower as expected (as 1/√n)

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)

Chi-squared

Multiple Independent Gaussian variables x_i: Define

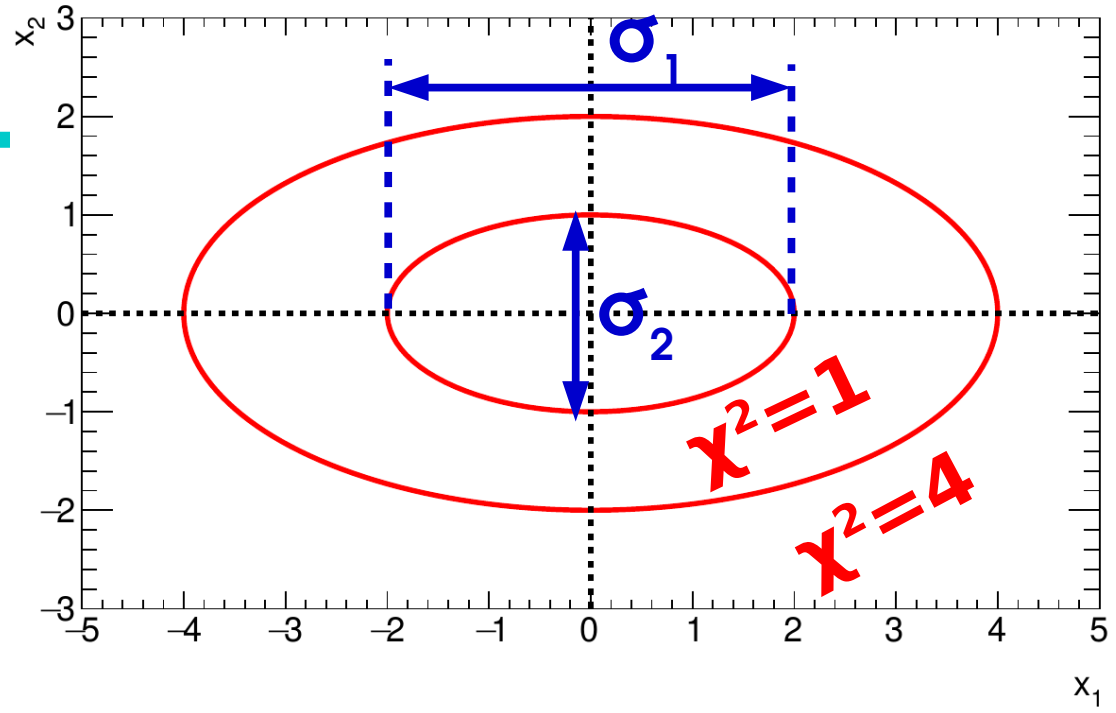
$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - x_i^0}{\sigma_i} \right)^2$$

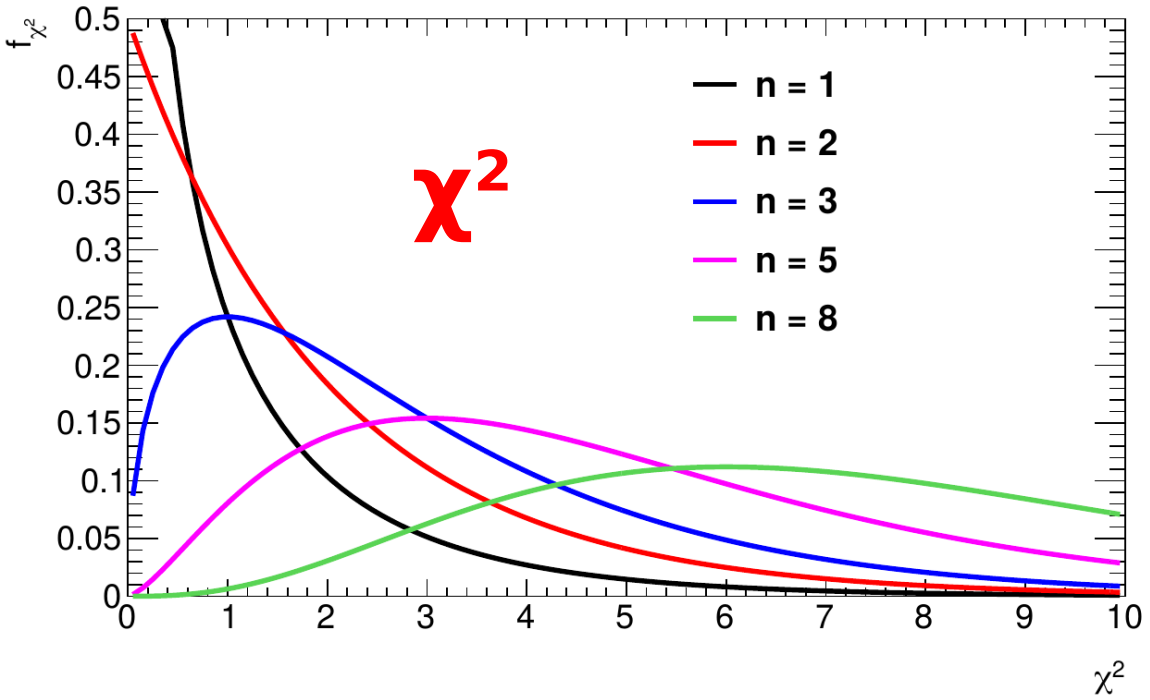
Measures global distance from reference point (x_1^0 .... x_n^0)

Distribution depends on n :

Rule of thumb:

χ^2/n **should be** $\lesssim 1$





Chi-squared

Multiple Independent Gaussian variables x_i: Define

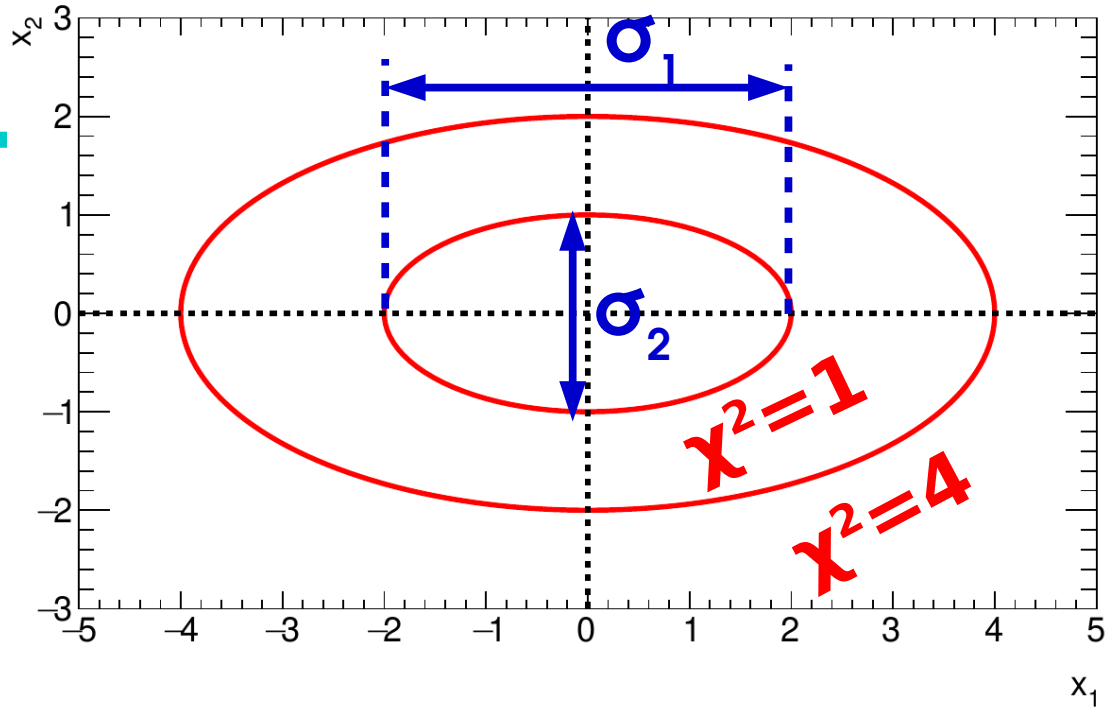
$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - x_i^0}{\sigma_i} \right)^2$$

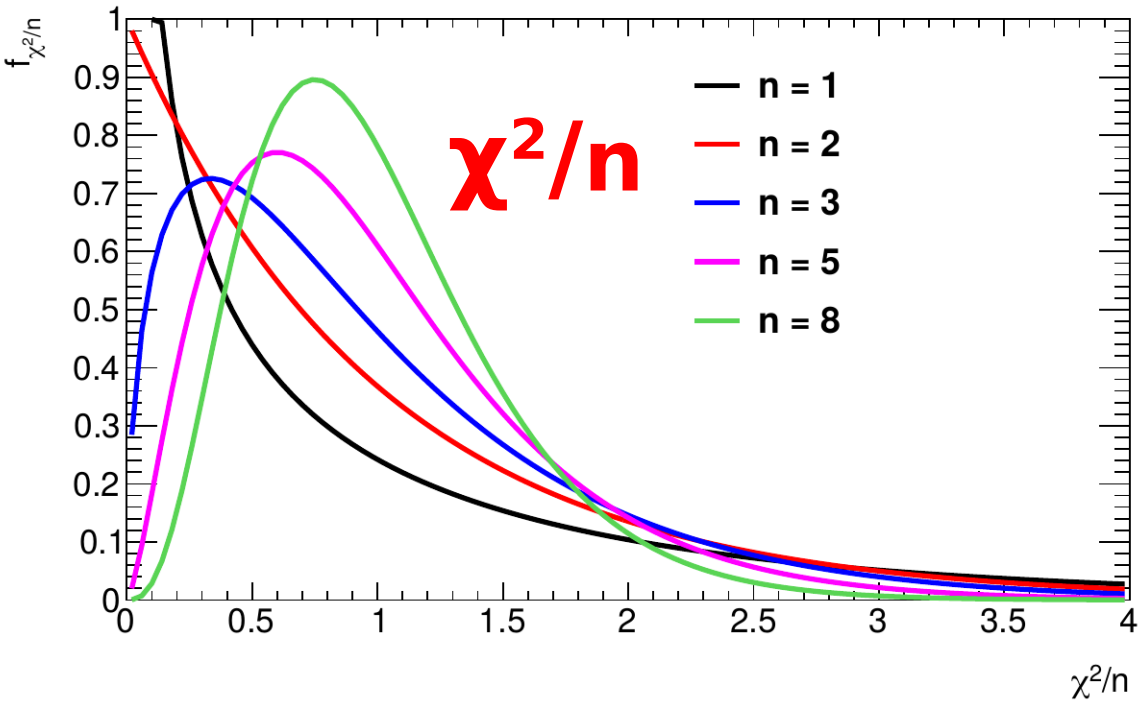
Measures global distance from reference point (x_1^0 .... x_n^0)

Distribution depends on n :

Rule of thumb:

χ^2/n **should be** $\lesssim 1$

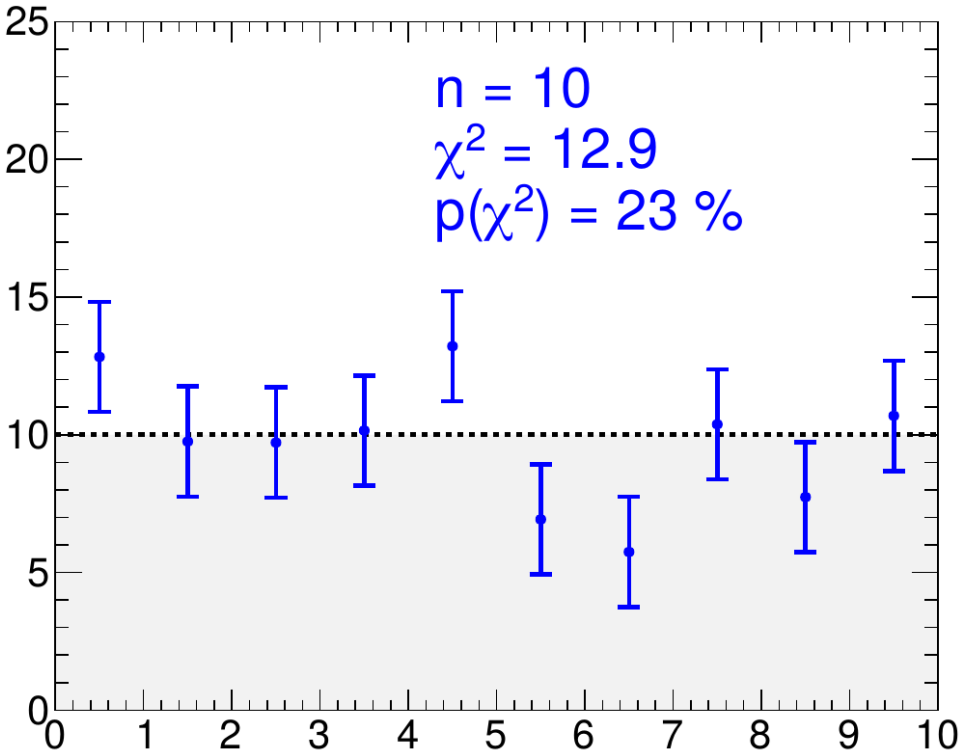




Histogram Chi-squared

Histogram 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)



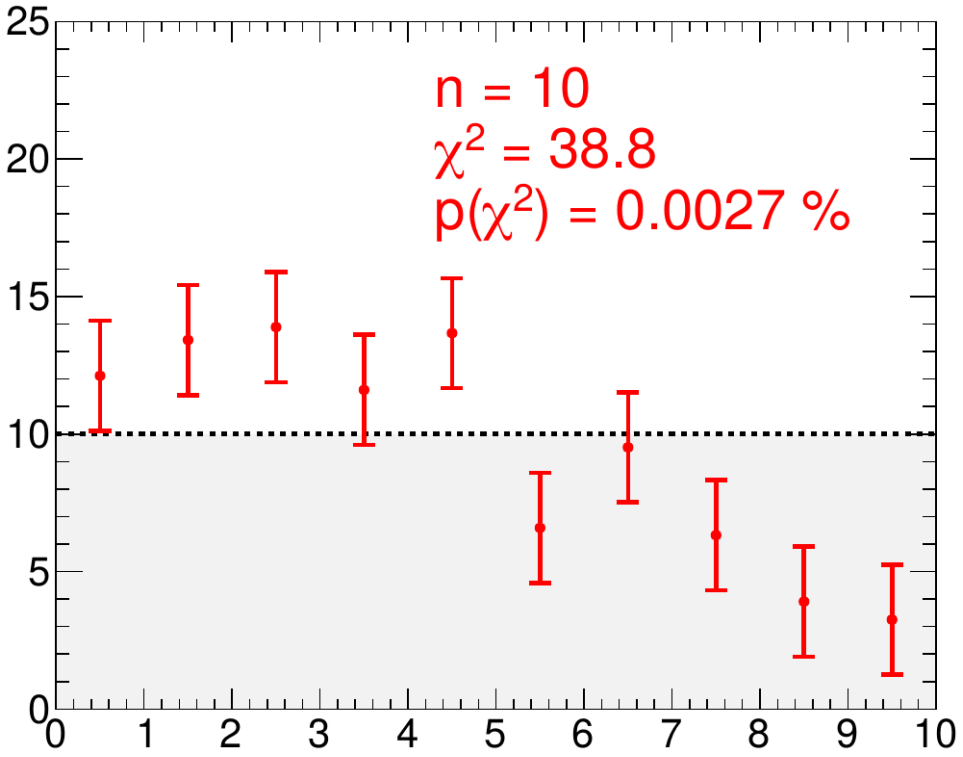
BLUE histogram vs. flat reference

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

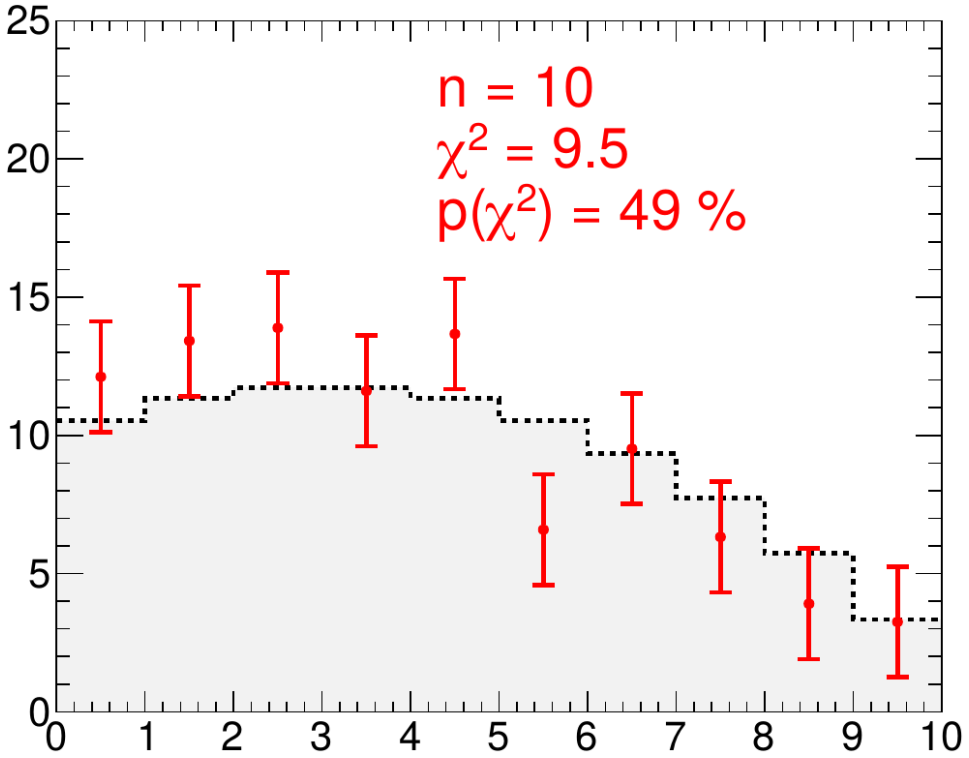
$\chi^2 = 38.8$, $p(\chi^2=38.8, n=10) =$ **0.003%**

✘

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

χ2 = 9.5, p(χ2=9.5, n=10) = 49%

Error Bars

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

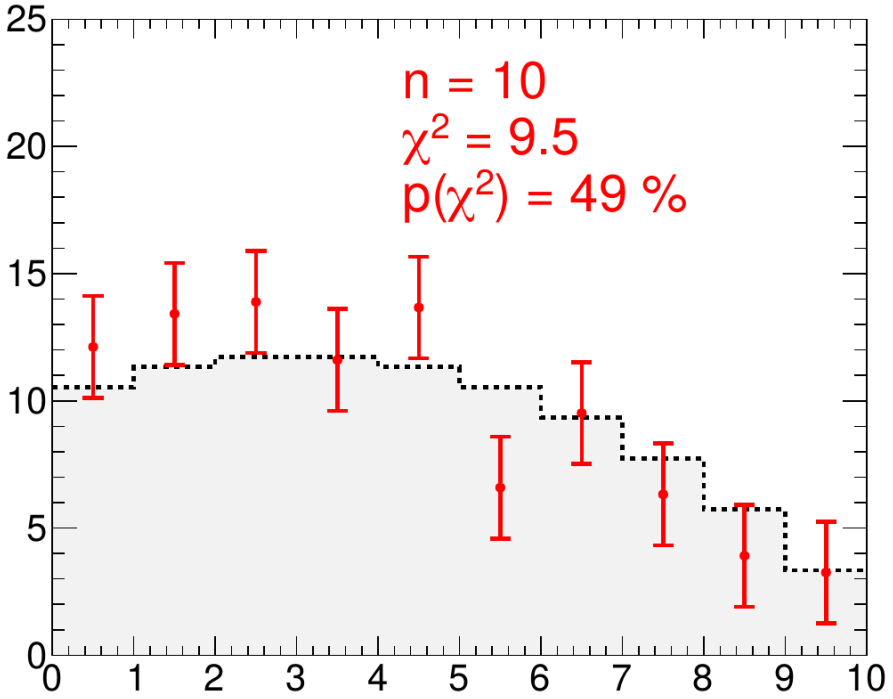
→ ***Bin central value*** *~ mean of the bin PDF*

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

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

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

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



Error Bars

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

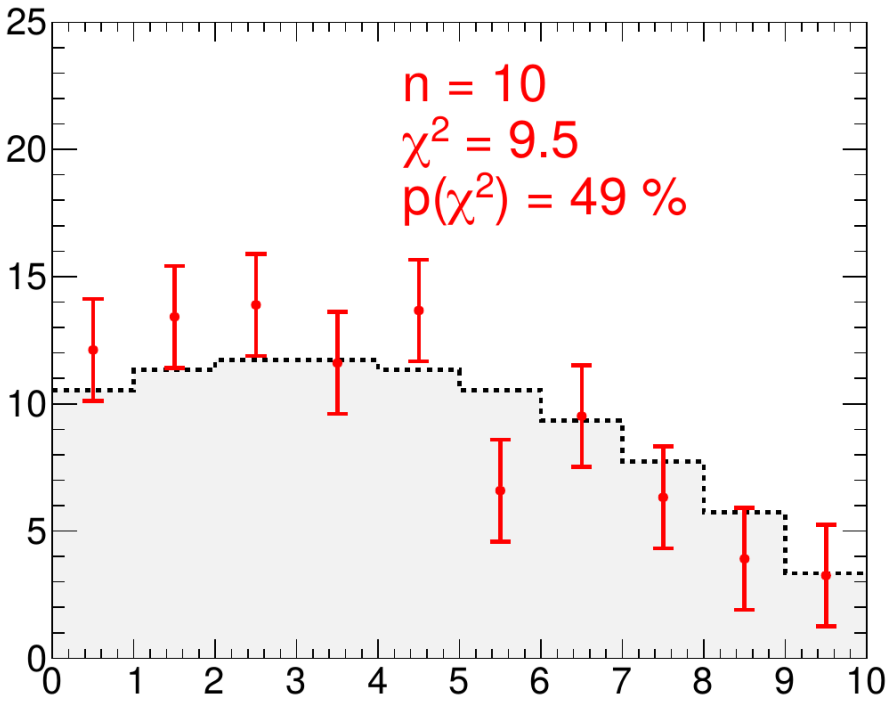
→ ***Bin central value*** *~ mean of the bin PDF*

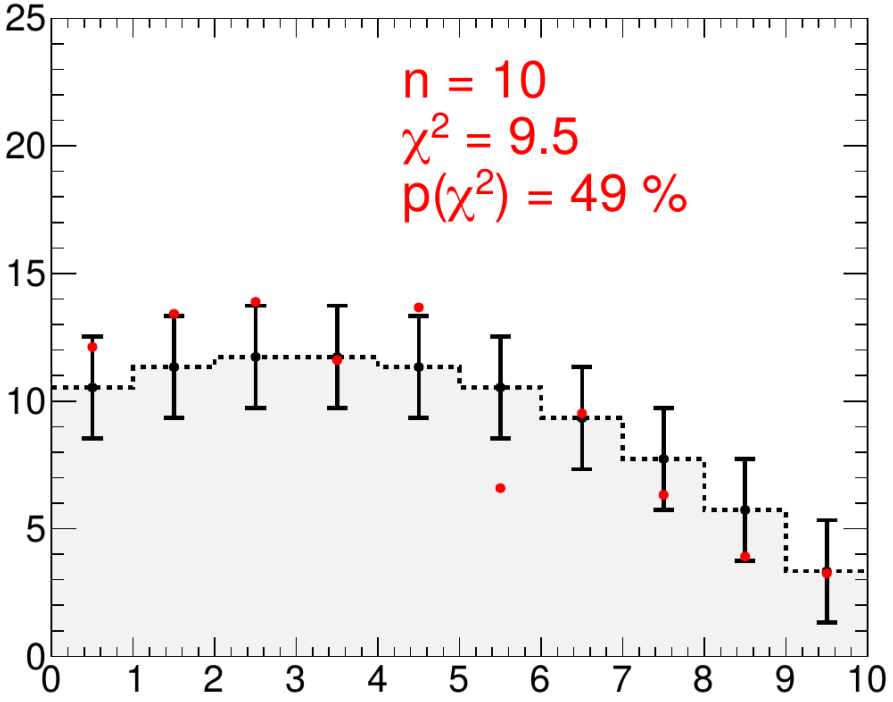
→ ***Bin uncertainty*** *~ RMS of the bin PDF*

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

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

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



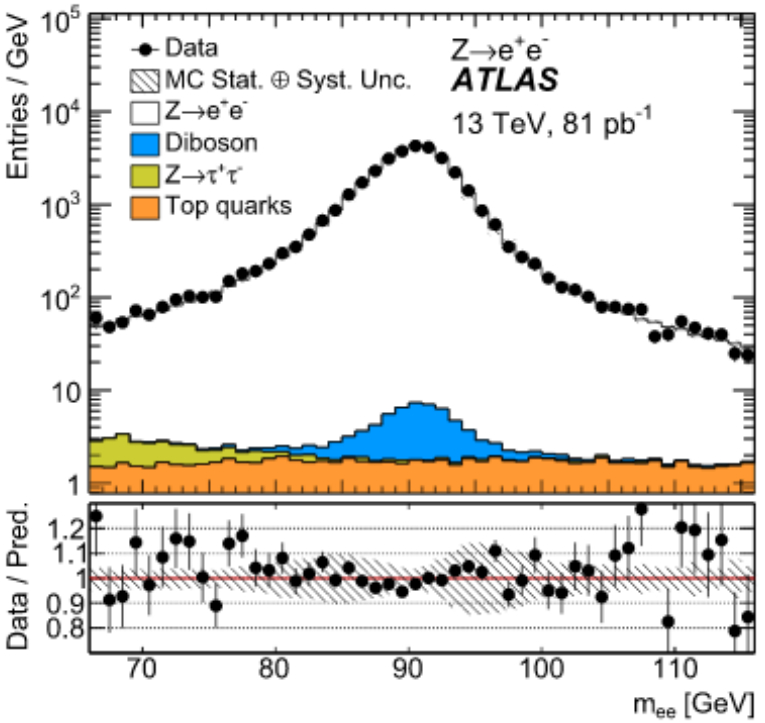


Statistical Modeling

Example 1: Z counting

Phys. Lett. B 759 (2016) 601

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



35000 ± 187

175 ± 8

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

(81 ± 2) pb^{-1}

0.552 ± 0.006

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

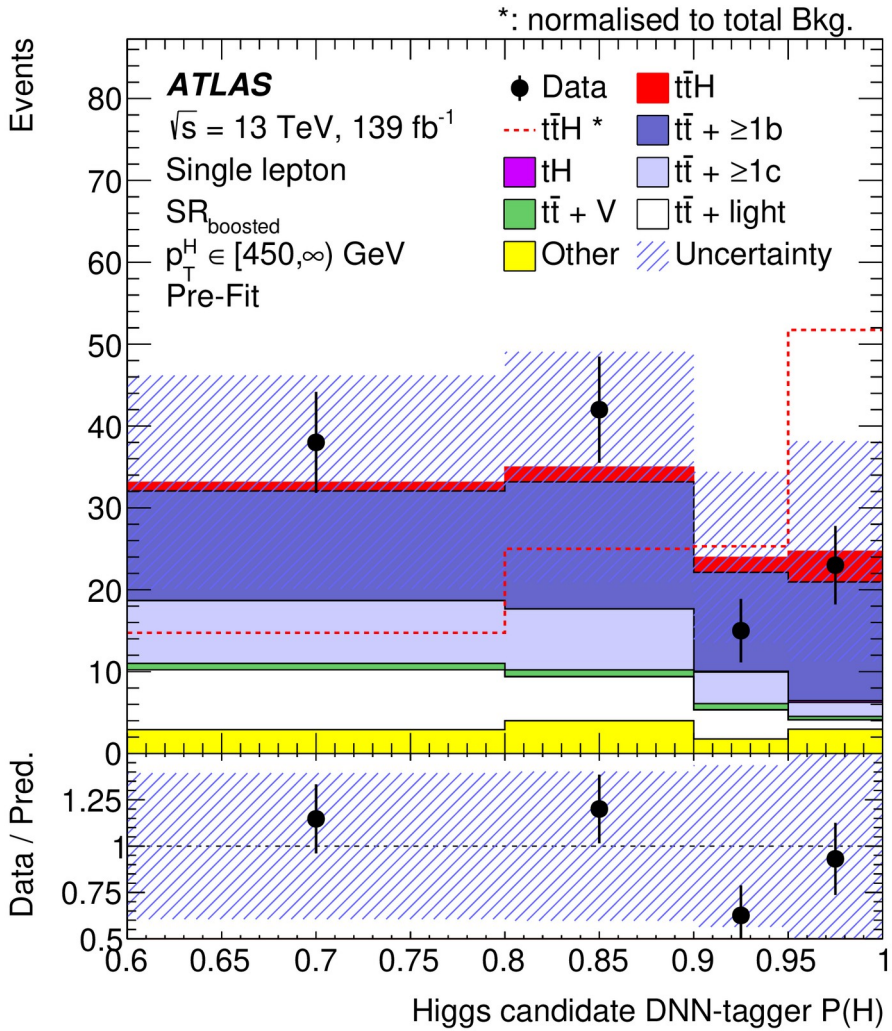
Fluctuations in the data counts

Other uncertainties (assumptions, parameter values)

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

Example 2: ttH→bb

arXiv:2111.06712



Event counting in different regions:

Multiple-bin counting

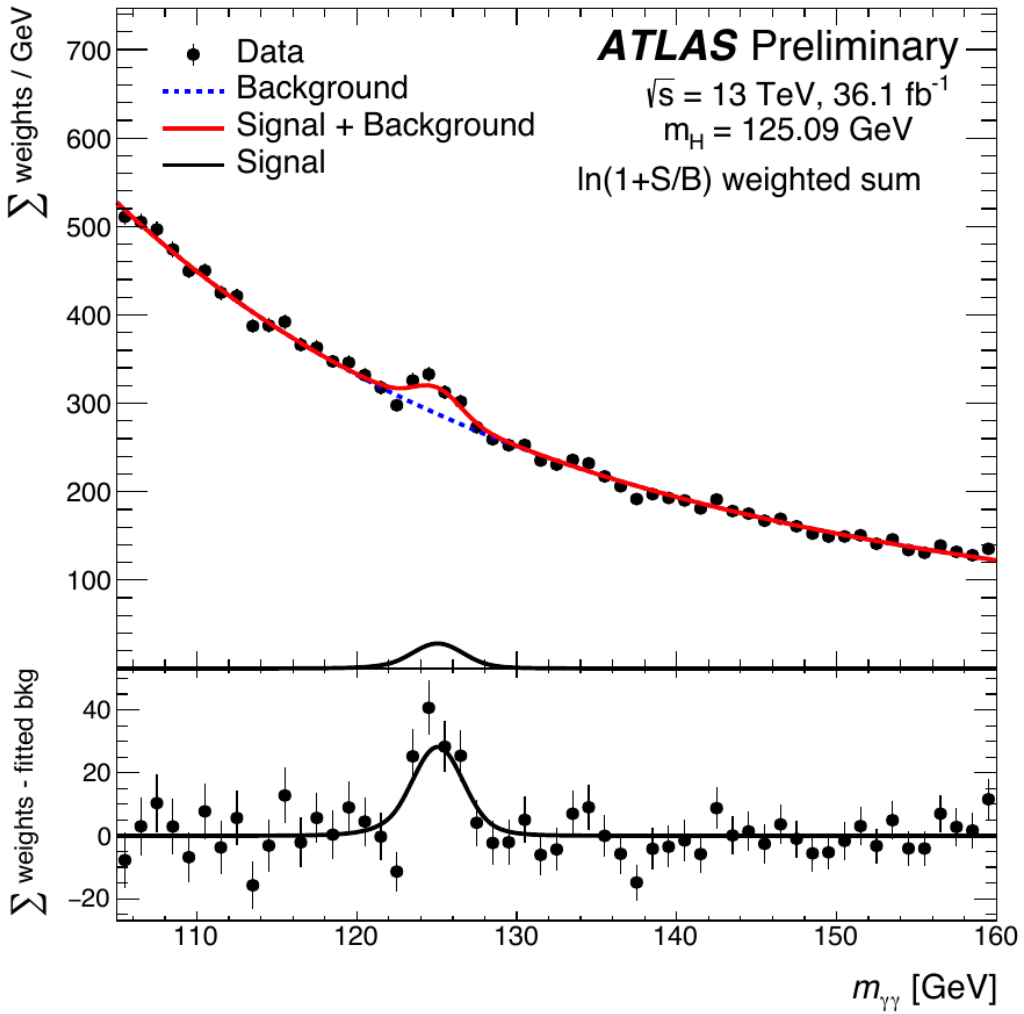
Lots of information available

→ Potentially higher sensitivity

→ How to make optimal use of it ?

Example 3: unbinned modeling

ATLAS-CONF-2017-045



All modeling done using continuous distributions:

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

How to count

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

→ In principle, binomial process

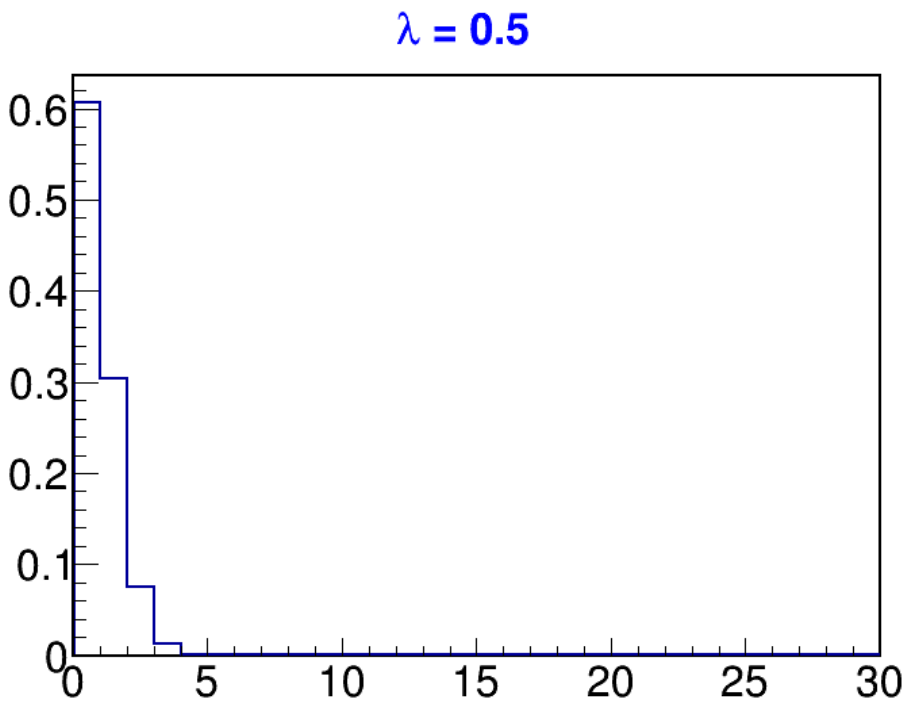
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √Mean

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

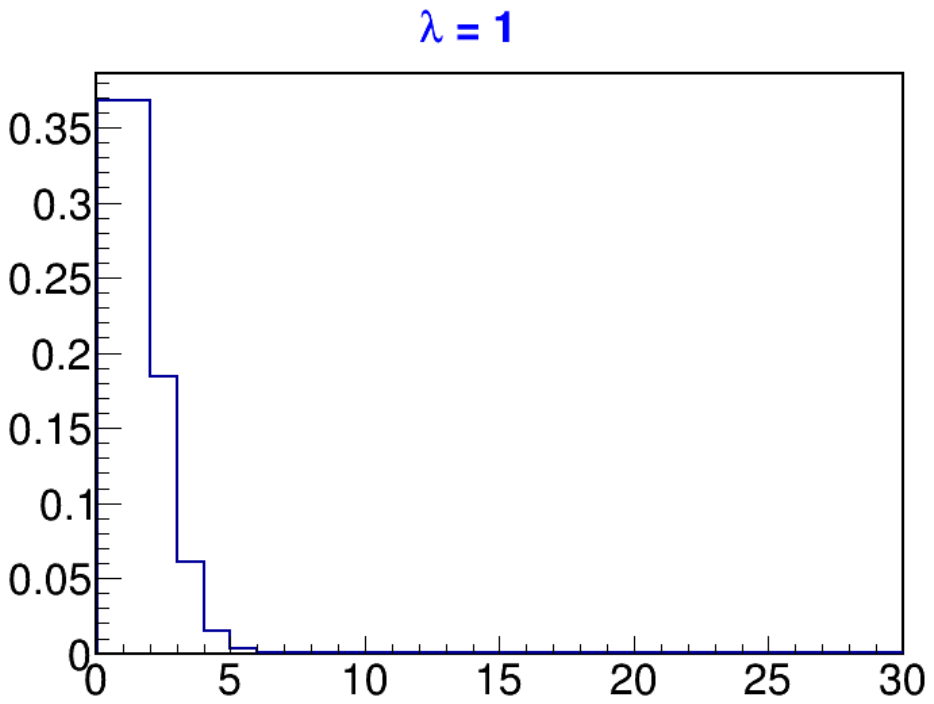
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √Mean

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

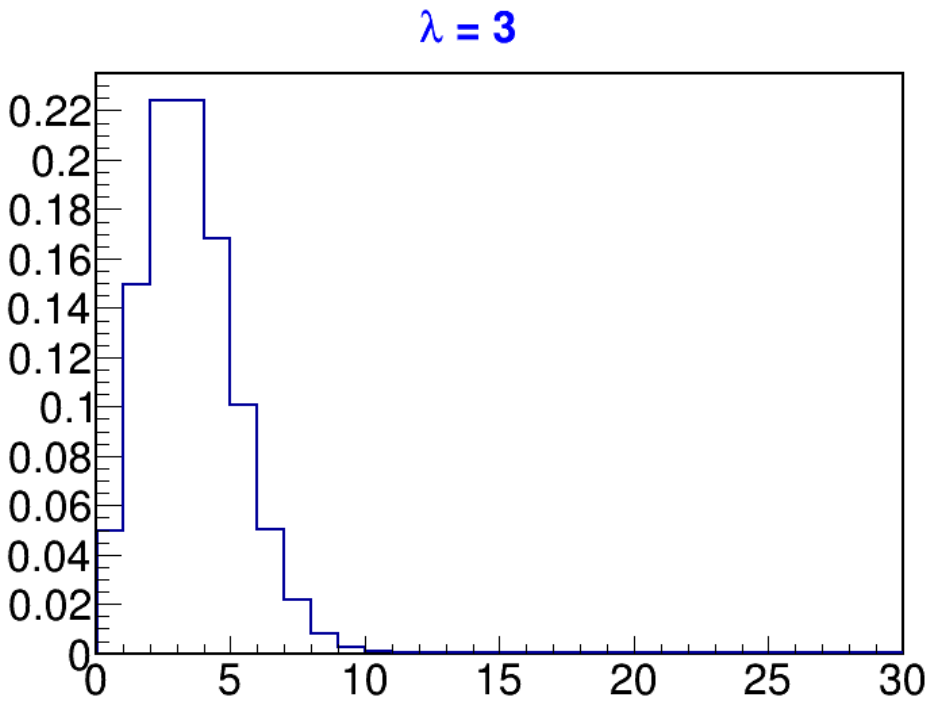
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √Mean

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

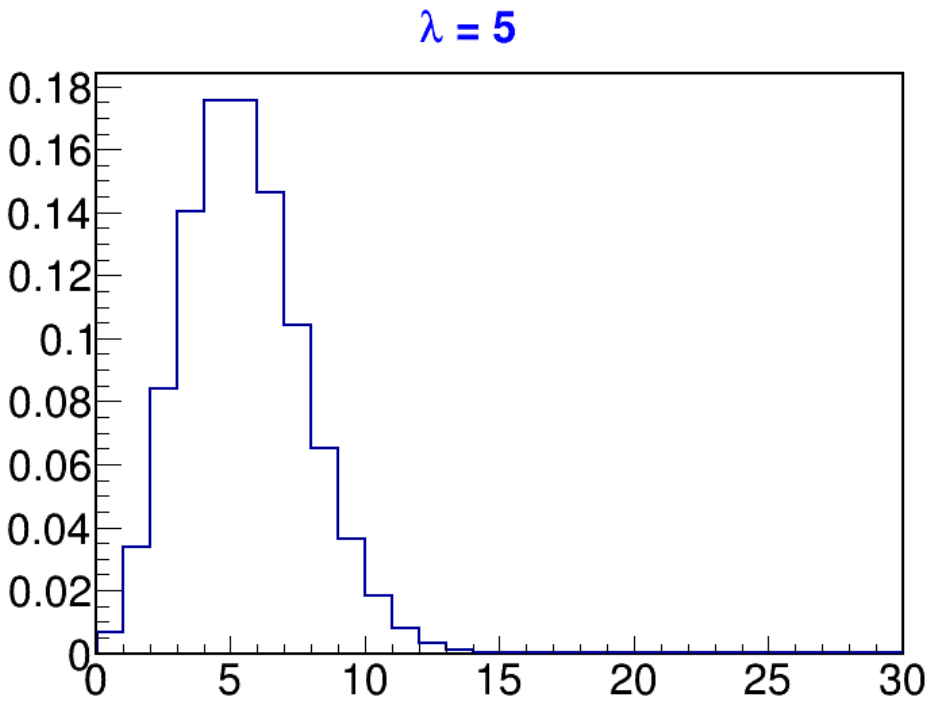
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √Mean

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

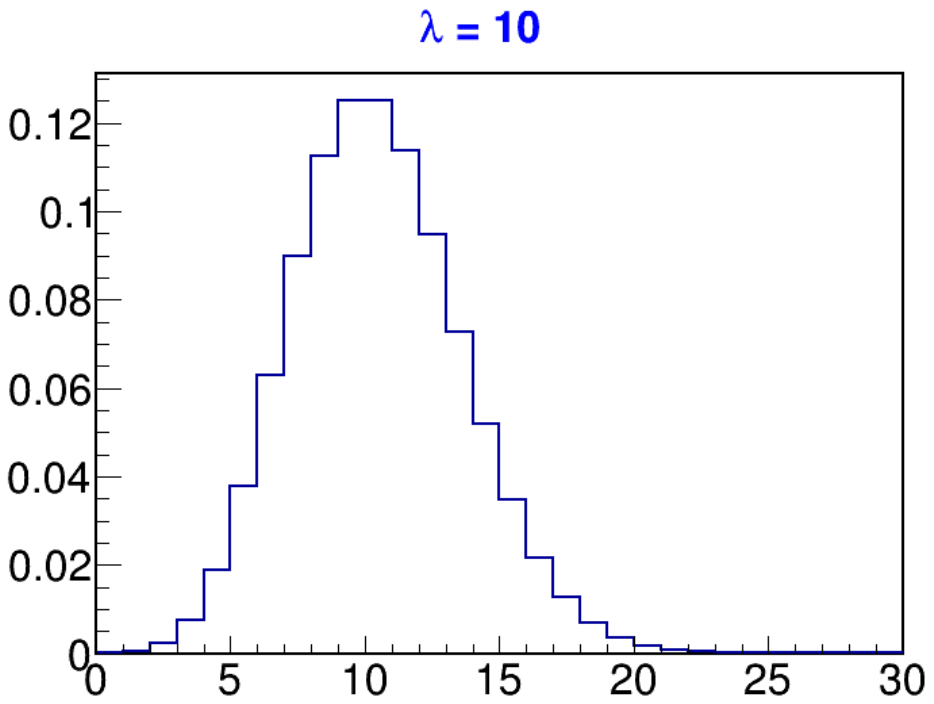
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

σ = √λ

For a counting measurement, RMS = √Mean

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

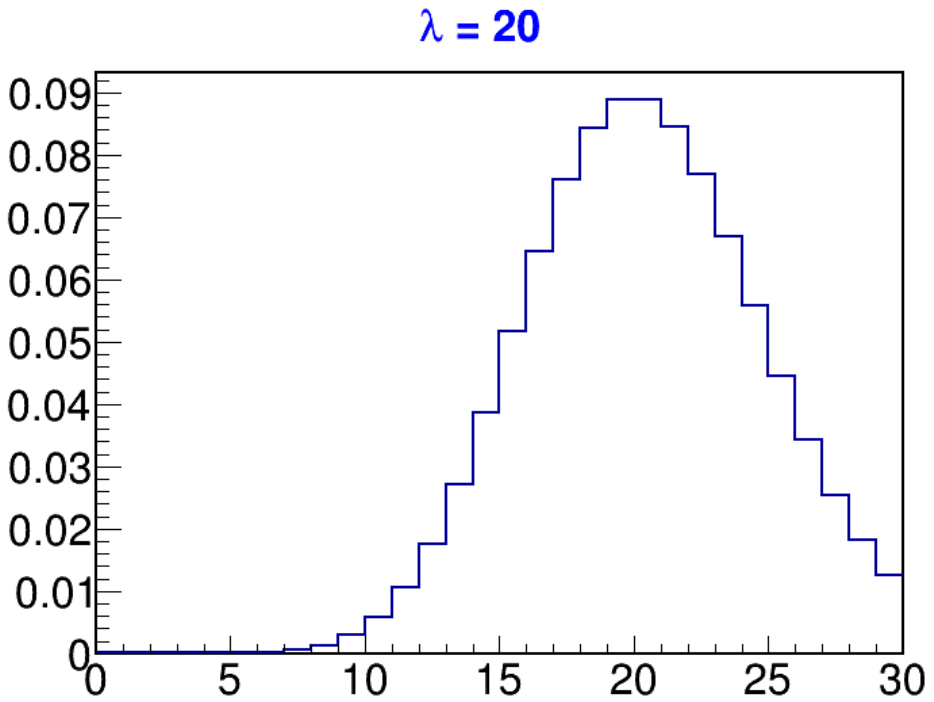
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

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

Poisson distribution

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

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



Mean = λ

Variance = λ

σ = √λ

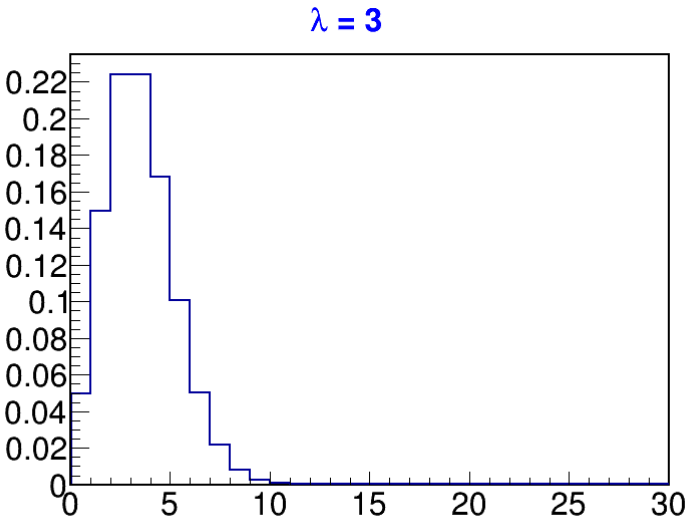
For a counting measurement, RMS = √Mean

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.**

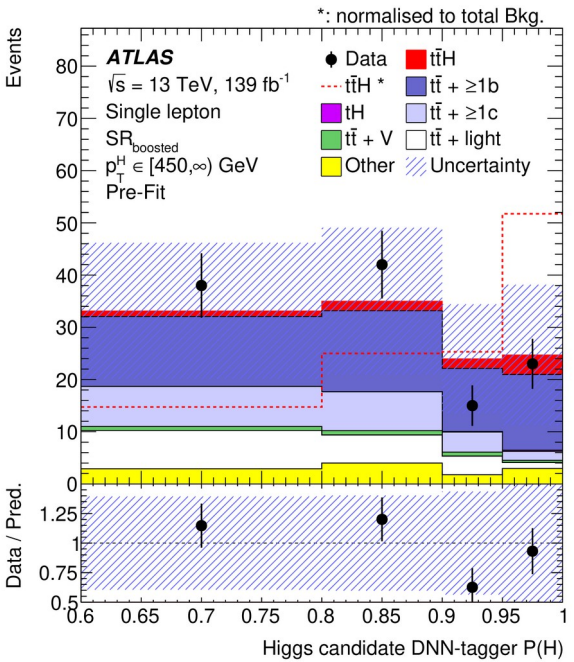
Multiple counting bins

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

(N : number of bins)

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

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



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

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

However not always possible to generate sufficiently large MC samples

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

Model Parameters

Model typically includes:

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

  → **S**, $\mathbf{m_W}$, …

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

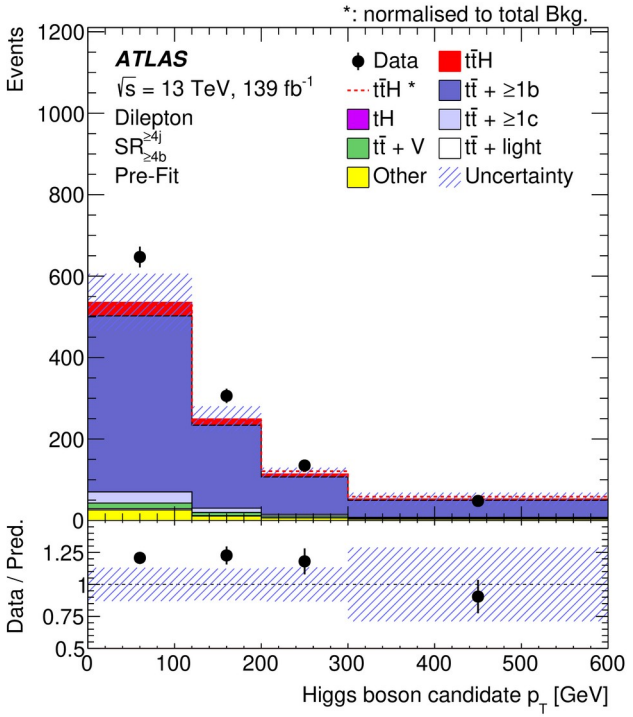
  → Background levels (**B**)

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

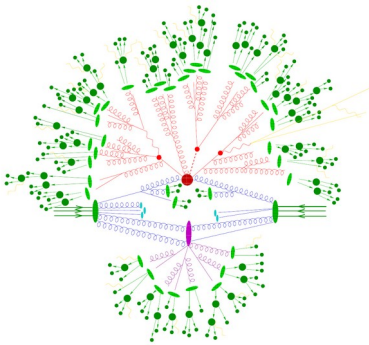
NPs must be either:

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

→ **Constrained by the data**



Takeaways



Random data must be described using a statistical model:

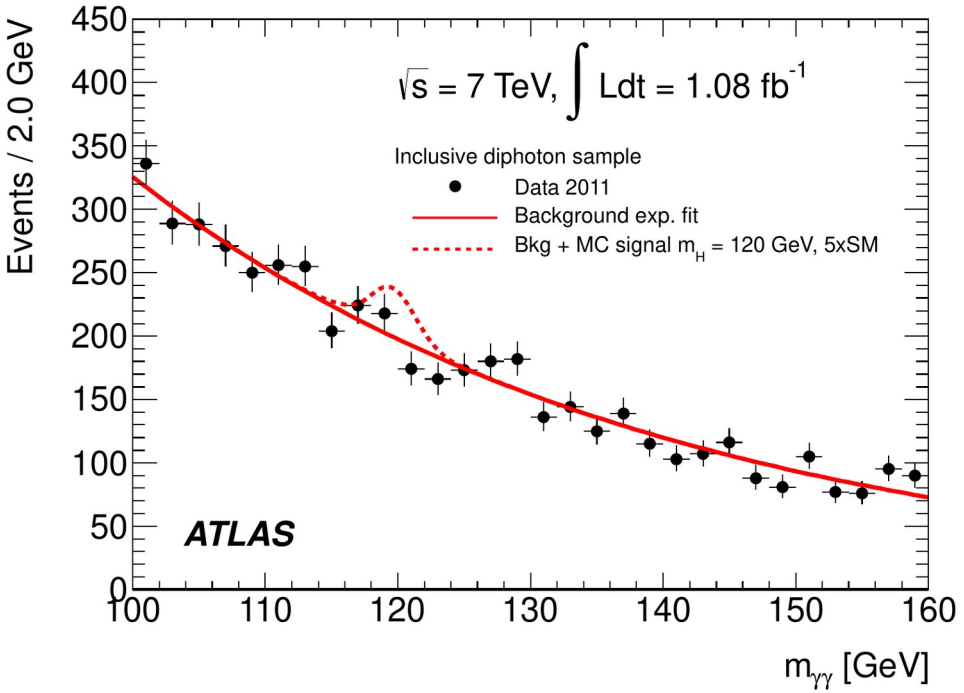
Description	Observable	Likelihood
Counting	n	**Poisson** $P(n;S,B)=e^{-(S+B)}\frac{(S+B)^n}{n!}$
Binned shape analysis	n_i, i = 1 .. N_{bins}	**Poisson product** $P(n_i;S,B)=\prod_{i=1}^{n_{bins}} e^{-(S f_i^{sig}+B f_i^{bkg})}\frac{(S f_i^{sig}+B f_i^{bkg})^{n_i}}{n_i!}$
Unbinned shape analysis	m_i,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

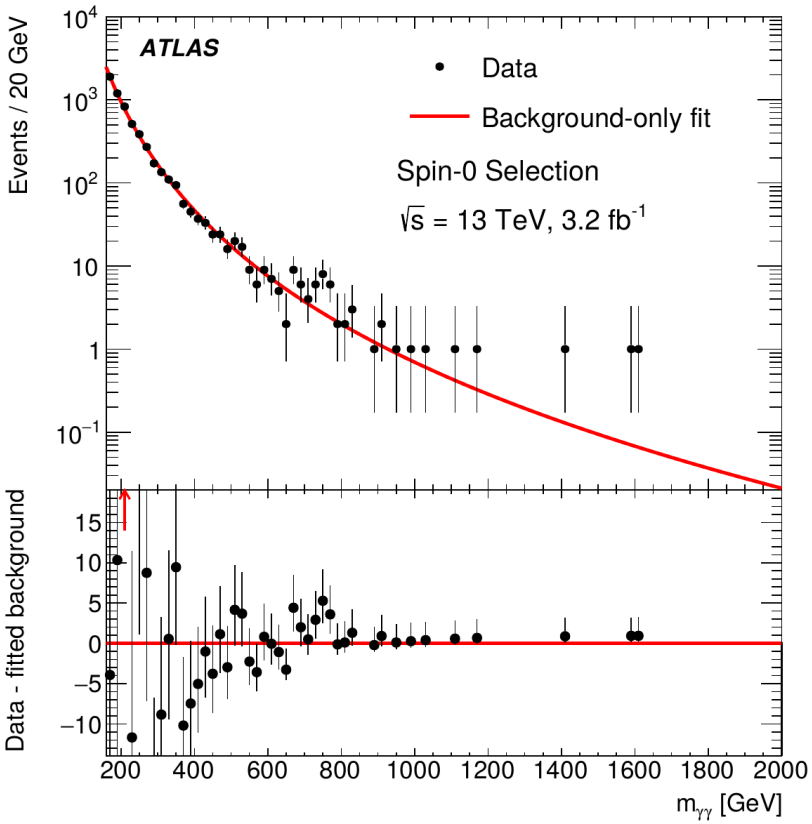
Model can include multiple **categories**, each with a separate description

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

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

Hypothesis Testing and discovery

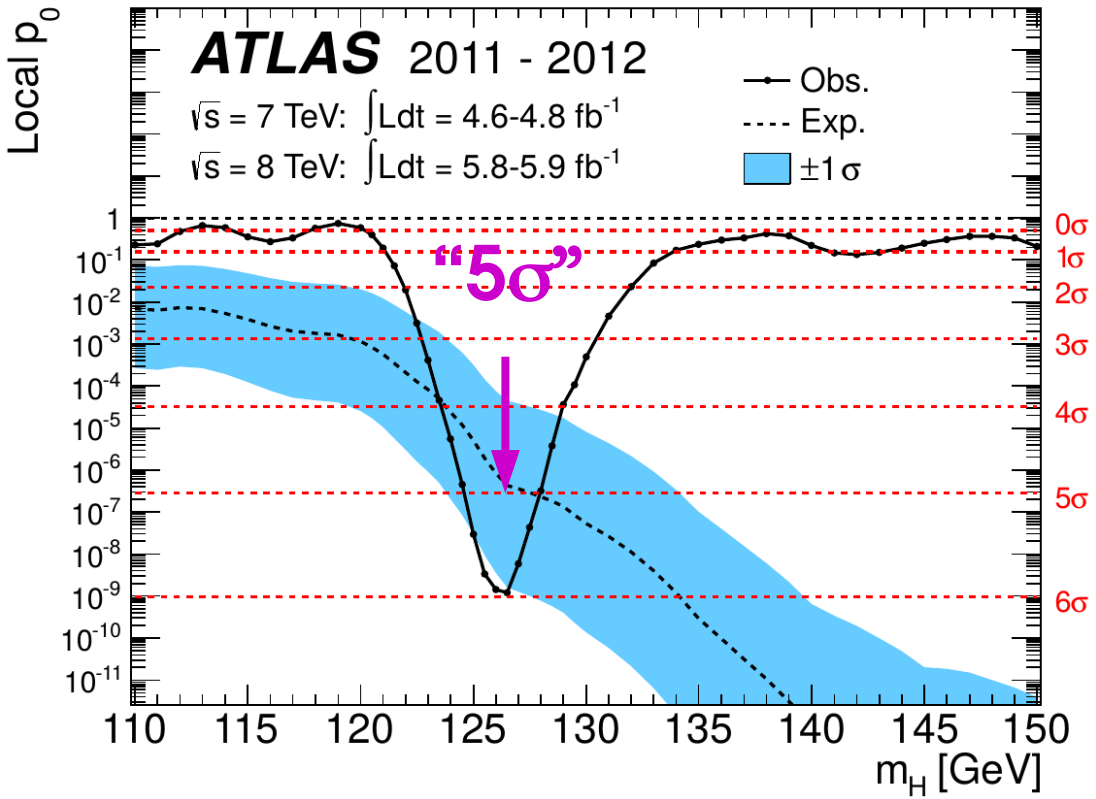




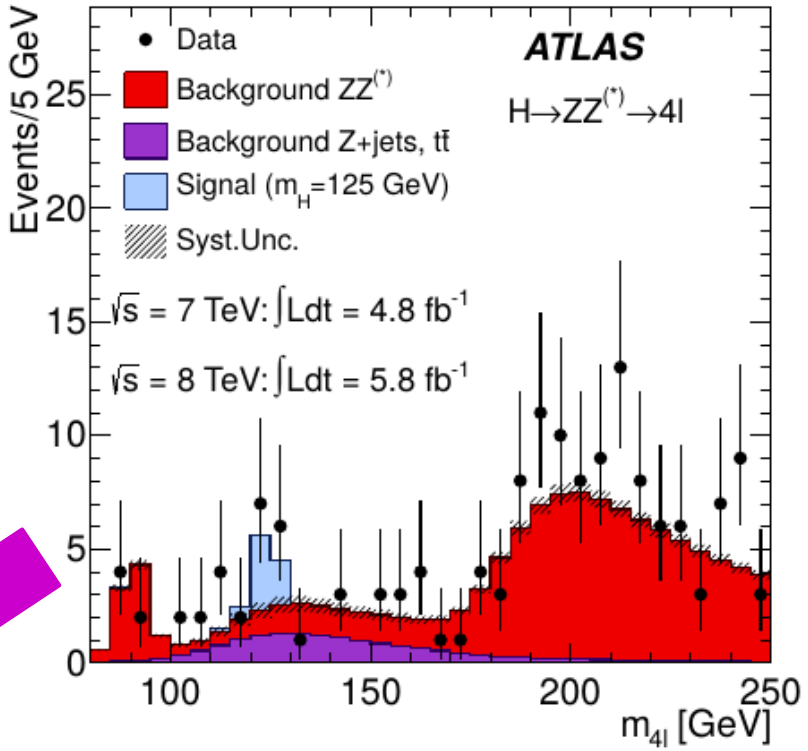
Discovery Testing

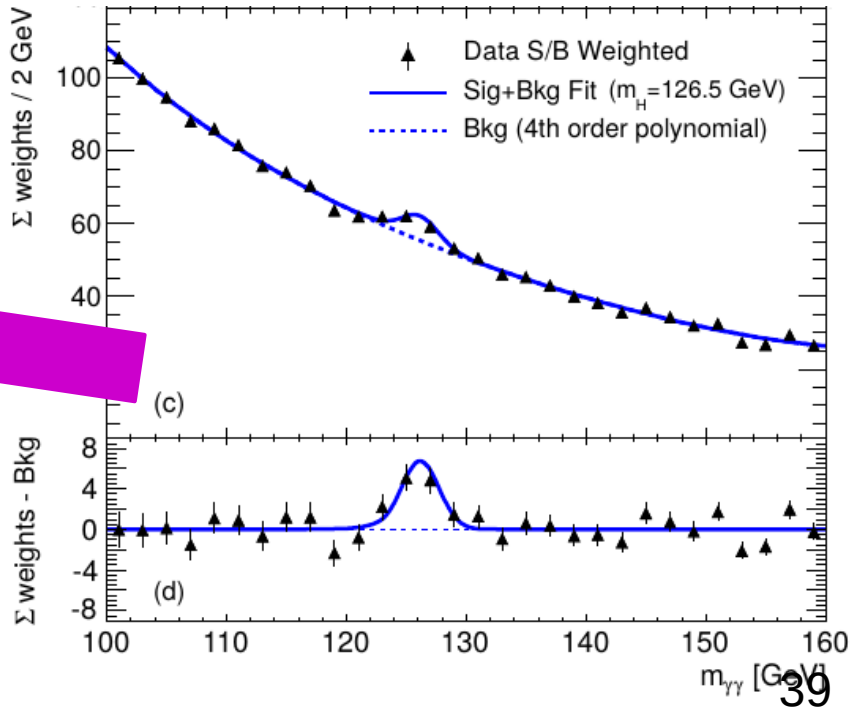
We see an unexpected feature in our data, is it a signal for new physics or a fluctuation ?

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

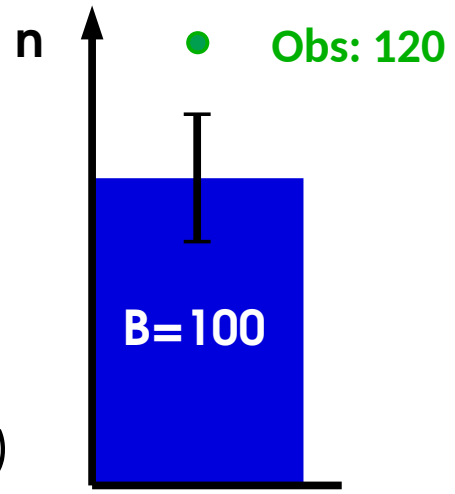


Phys. Lett. B 716 (2012) 1-29





Discovery Testing



Say we have a Gaussian measurement with a background **B=100**, and we measure **n=120**

Did we just discover something ? ***Maybe :-)*** *(but not very likely)*

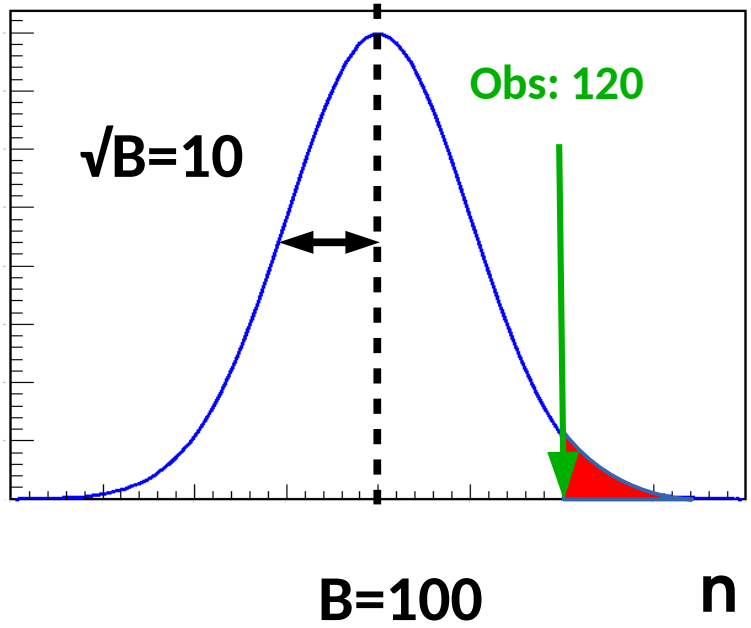
The measured signal is **S = 20**. $S = n_{obs} - B$

Uncertainty on B is √B = 10

⇒ Significance Z = 2

⇒ we are **~2σ** away from S=0.

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



Gaussian quantiles :

Z = 2 happens p_0 **~ 2.3%** of the time if S=0

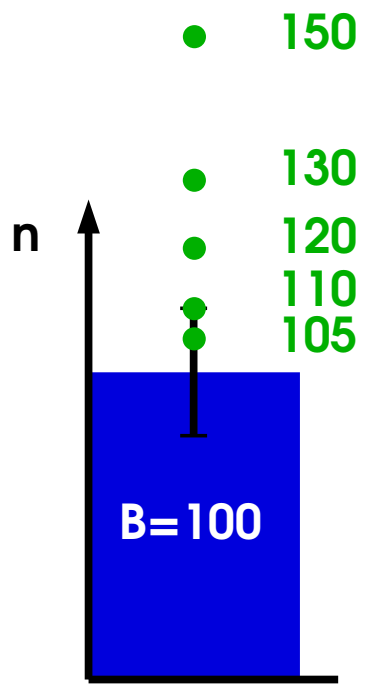
P-value:

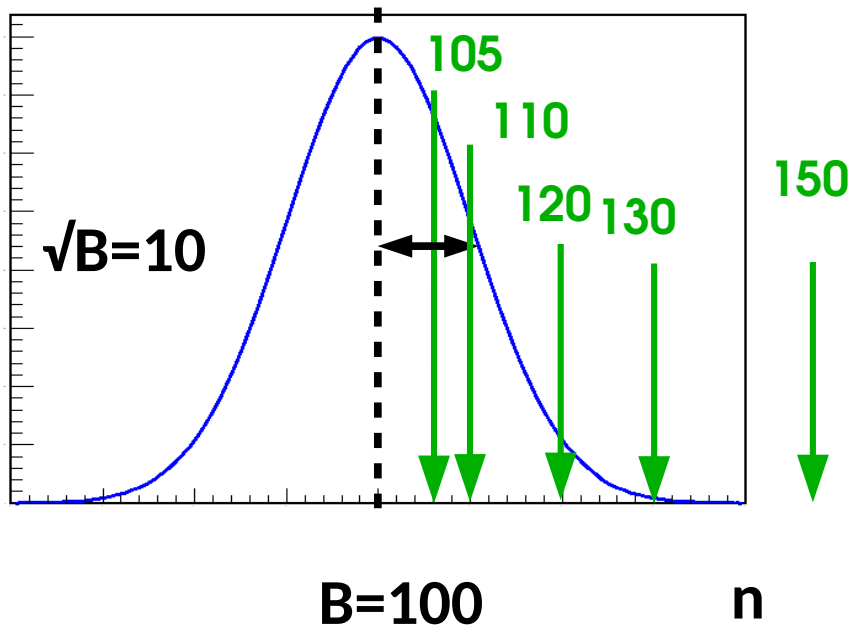
$$p_0 = 1 - \Phi(Z)$$

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

⇒ Rare, but not exceptional

Discovery Testing





n_{obs}	S	Z	p_0
105	5	0.5σ	31%
110	10	1σ	16%
120	20	2σ	2.3%
130	**30**	**3σ**	**0.1%**

Straightforward in this Gaussian case

Need to be able to do the same in more complex cases:

- **Determine S**
- **Compute Z and p_0**

Evidence

Discovery

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

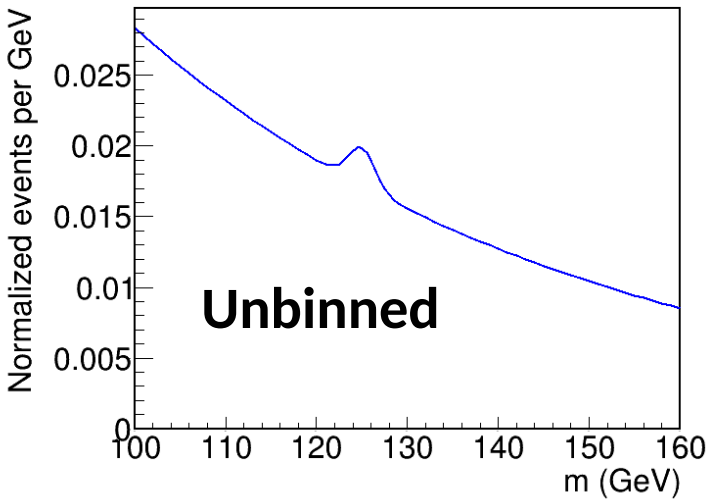
Can draw random events according to PDF : **generate** ***pseudo-data***

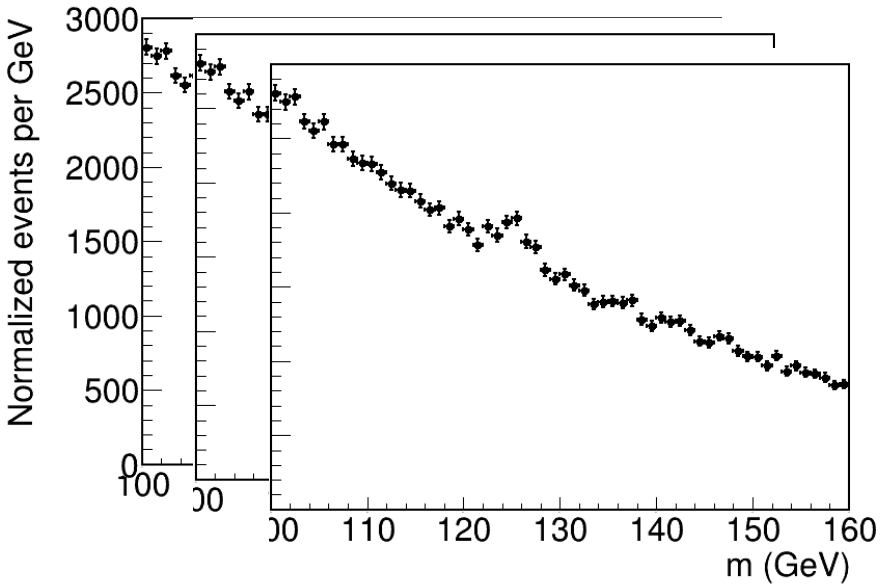
$$P(\lambda = 5)$$

2, 5, 3, 7, 4, 9, ….

Each entry = separate “experiment”

Generate





What a PDF is also for: Likelihood

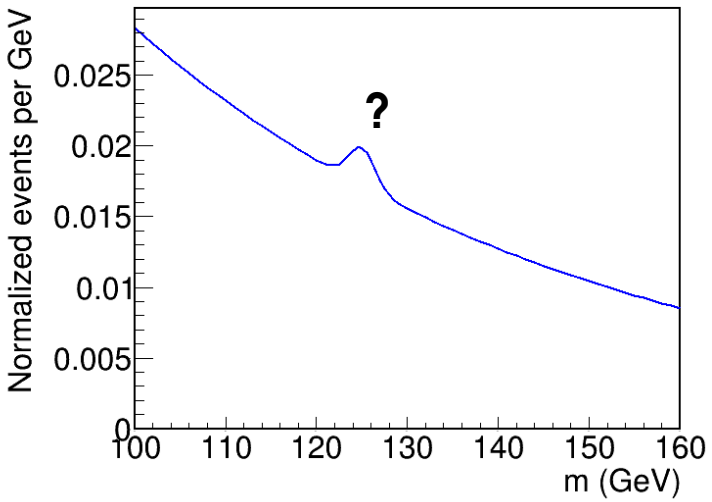
Model describes the distribution of the observable: **P(data; parameters)**

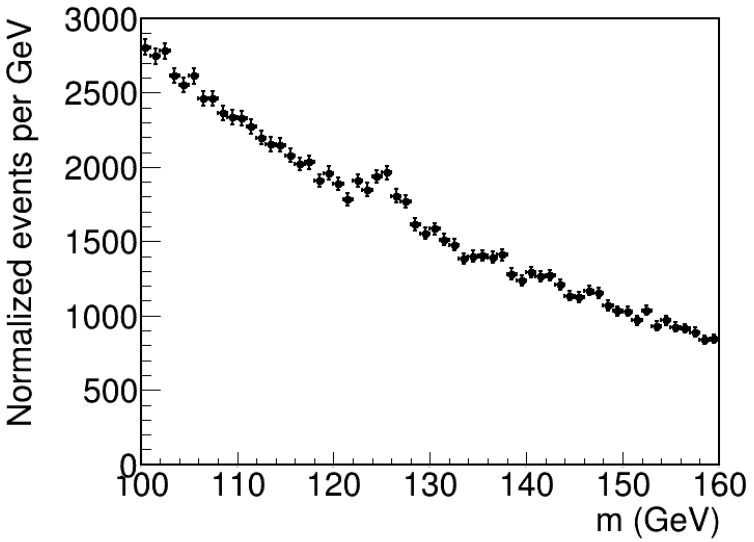
⇒ Possible outcomes of the experiment, for given parameter values

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

$P(\lambda = ?)$ ⟵ **2**

Estimate ⟵





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

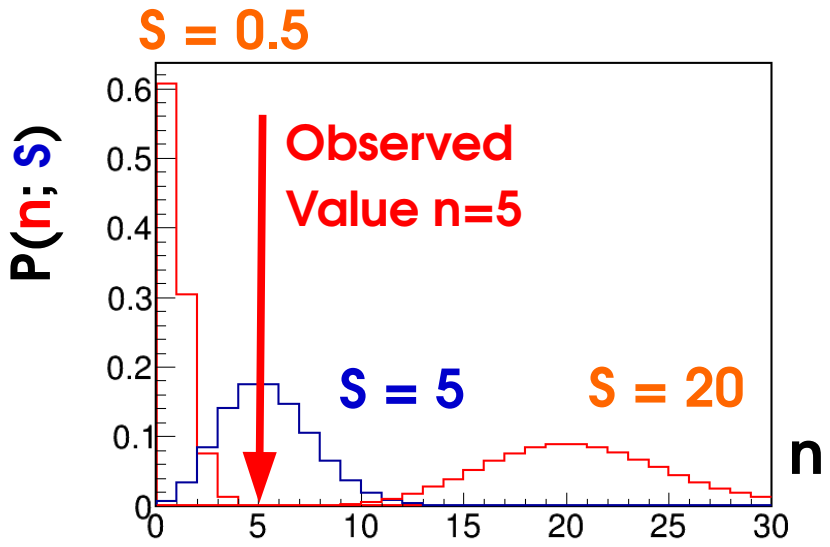
→ **same as the PDF,** but seen as function of the parameters

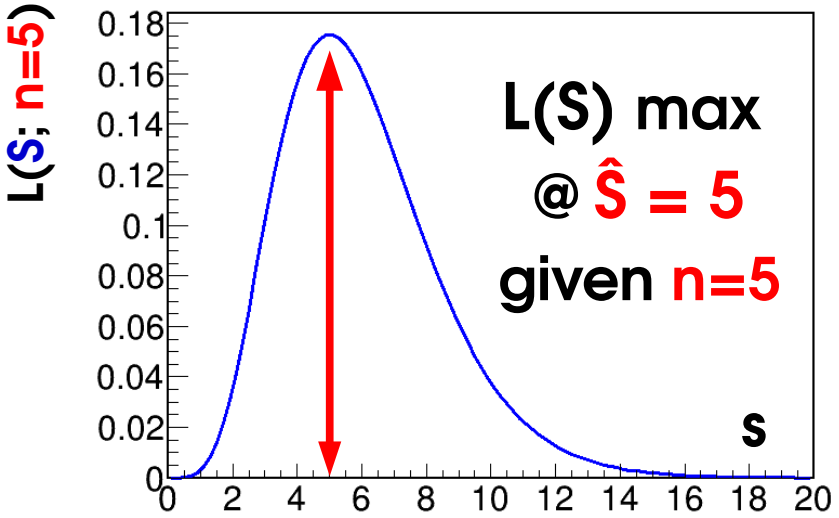
Maximum Likelihood Estimation

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

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

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



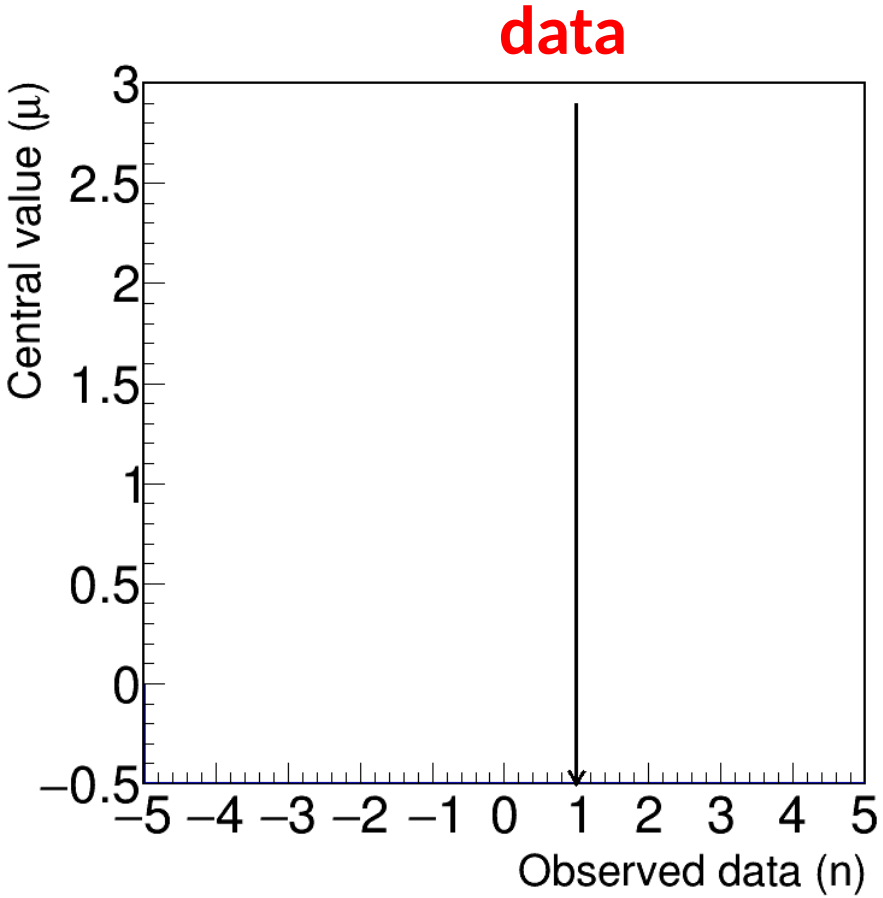


MLE: the value of μ for which **this data** was ***most likely to occur***

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

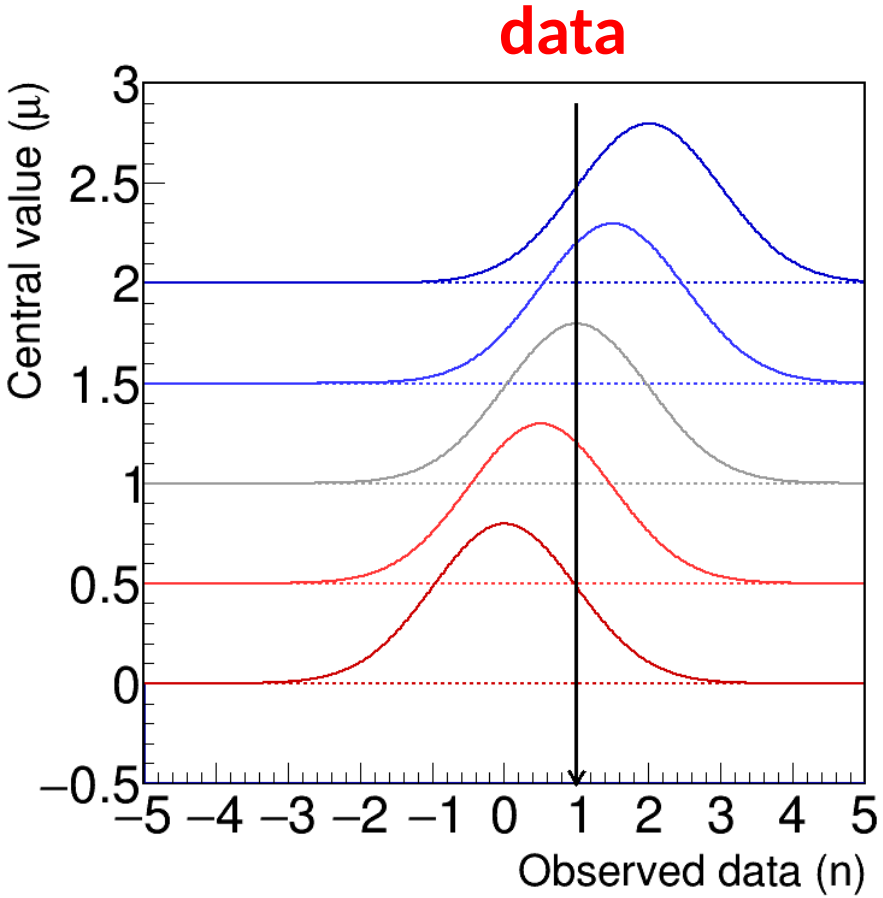
No guarantee it is the true value (data may be "unlikely") but sensible estimate

Gaussian case



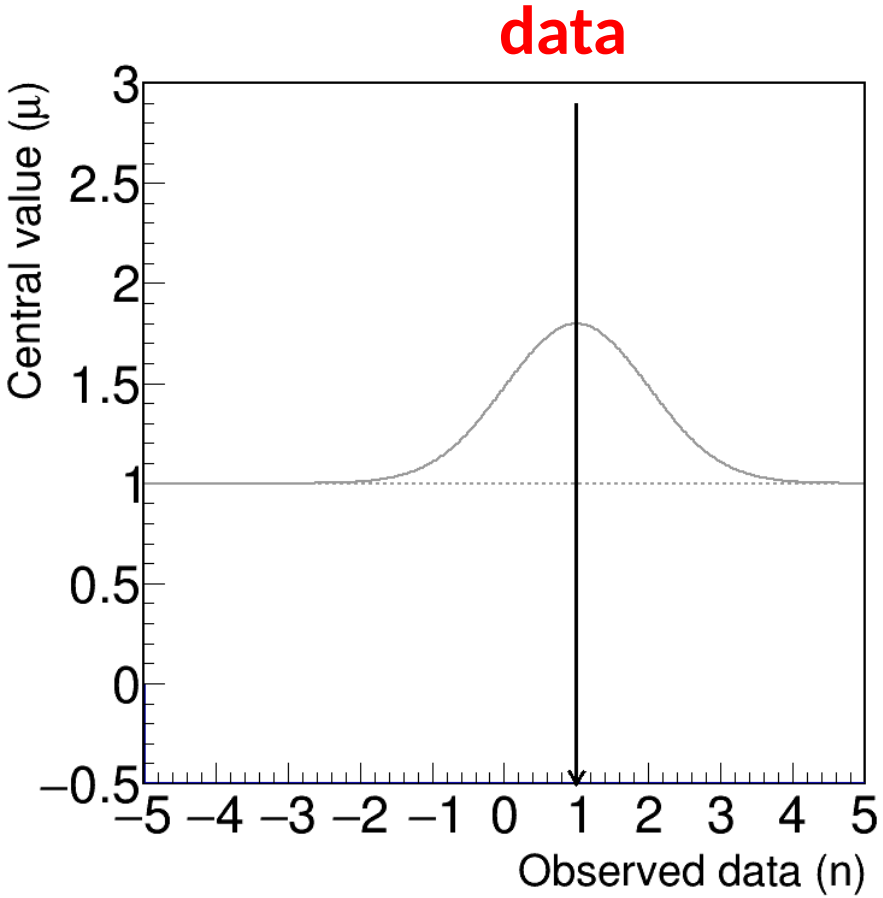
Best-fit of Gaussian PDF mean to observed data

Gaussian case



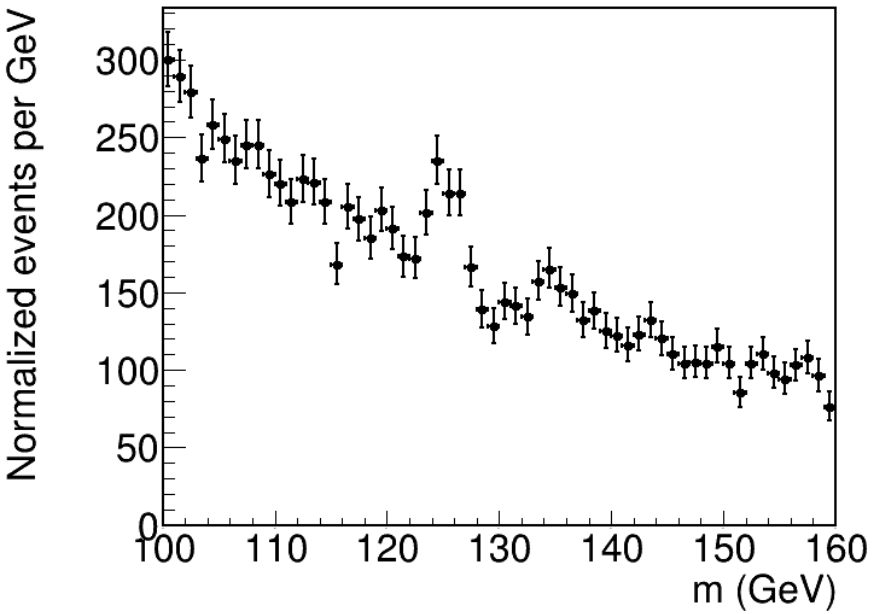
Best-fit of Gaussian PDF mean to observed data

Gaussian case



Best-fit of Gaussian PDF mean to observed data

Multiple Gaussian bins



-2 log Likelihood:

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

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

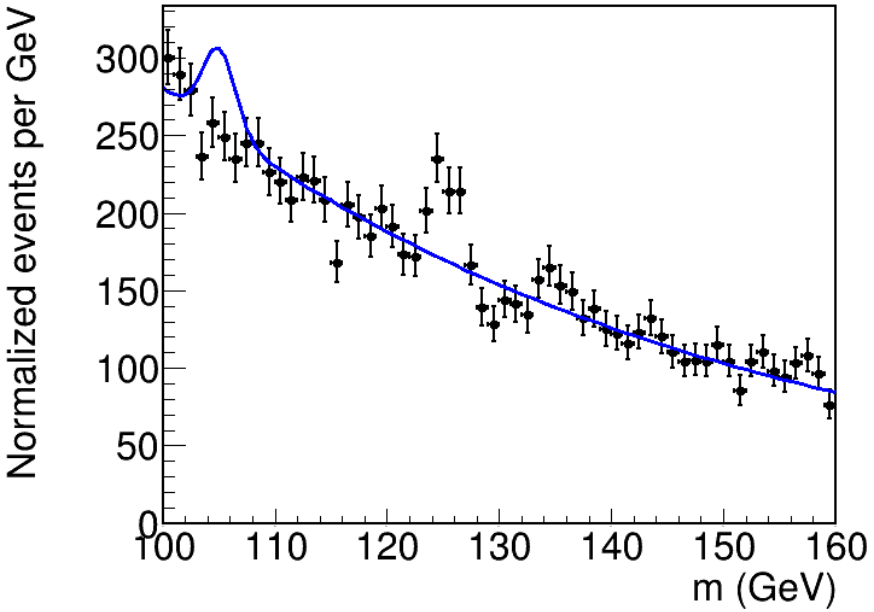
However typically need to perform non-linear minimization in other cases.

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:

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

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

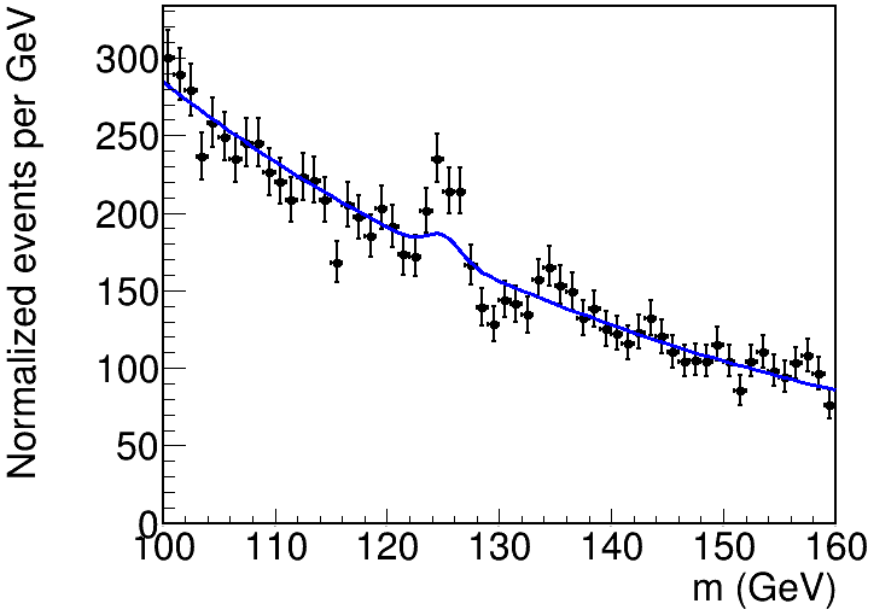
However typically need to perform non-linear minimization in other cases.

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:

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

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

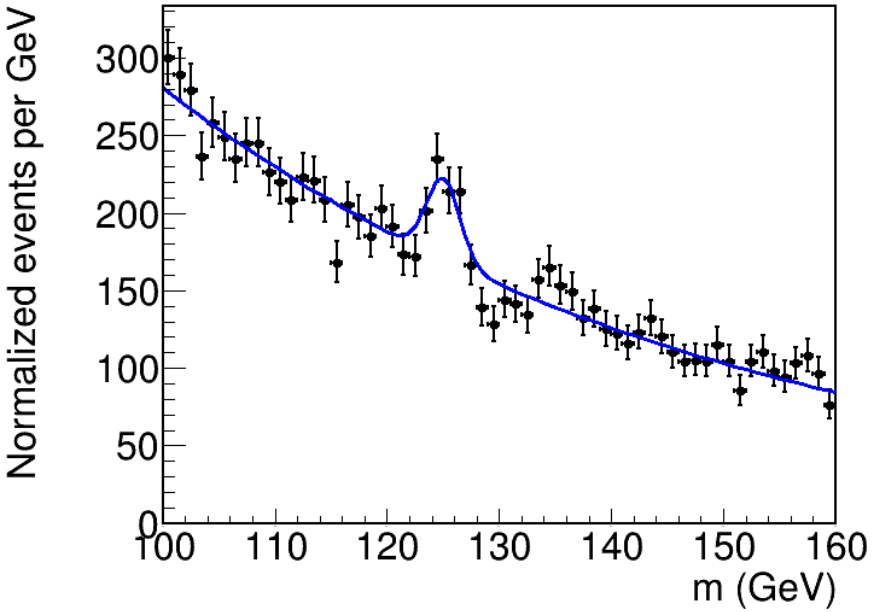
However typically need to perform non-linear minimization in other cases.

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:

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

Maximum likelihood ⇔ Minimum χ^2

⇔ Least-squares minimization

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

HEP practice:

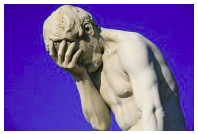
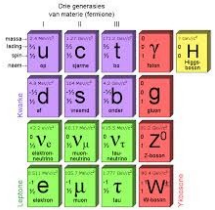
- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/… backends

  → Many algorithms – gradient-based, etc.

Hypothesis Testing

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

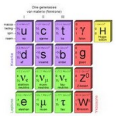
→ **Goal** : decide if H_0 is favored or disfavored using a test based on the data

Possible outcomes:	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	**Discovery!**  	**Missed discovery** 
H_0 is true (Nothing new)	**False discovery** 	**No new physics, None found** 

"… the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis." – R. A. Fisher

Hypothesis Testing

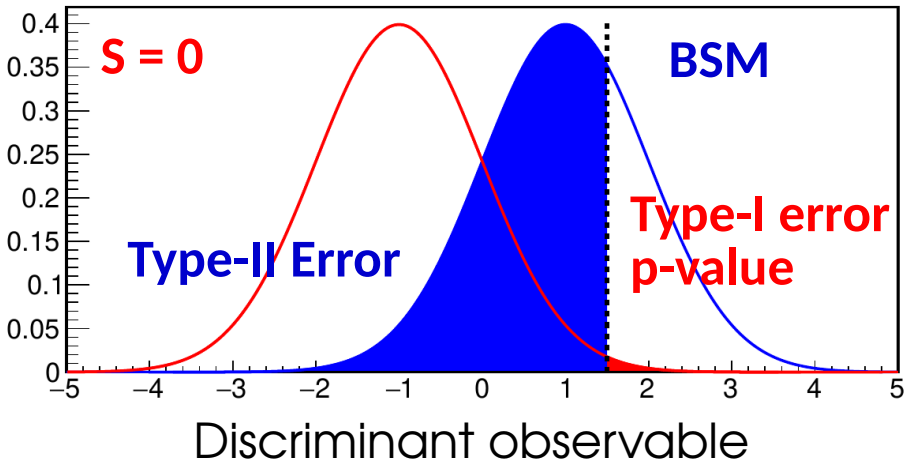
Hypothesis: assumption on model parameters, say value of S (e.g. **H_0 : S=0**)

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

p-value, significance

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

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



ROC Curves

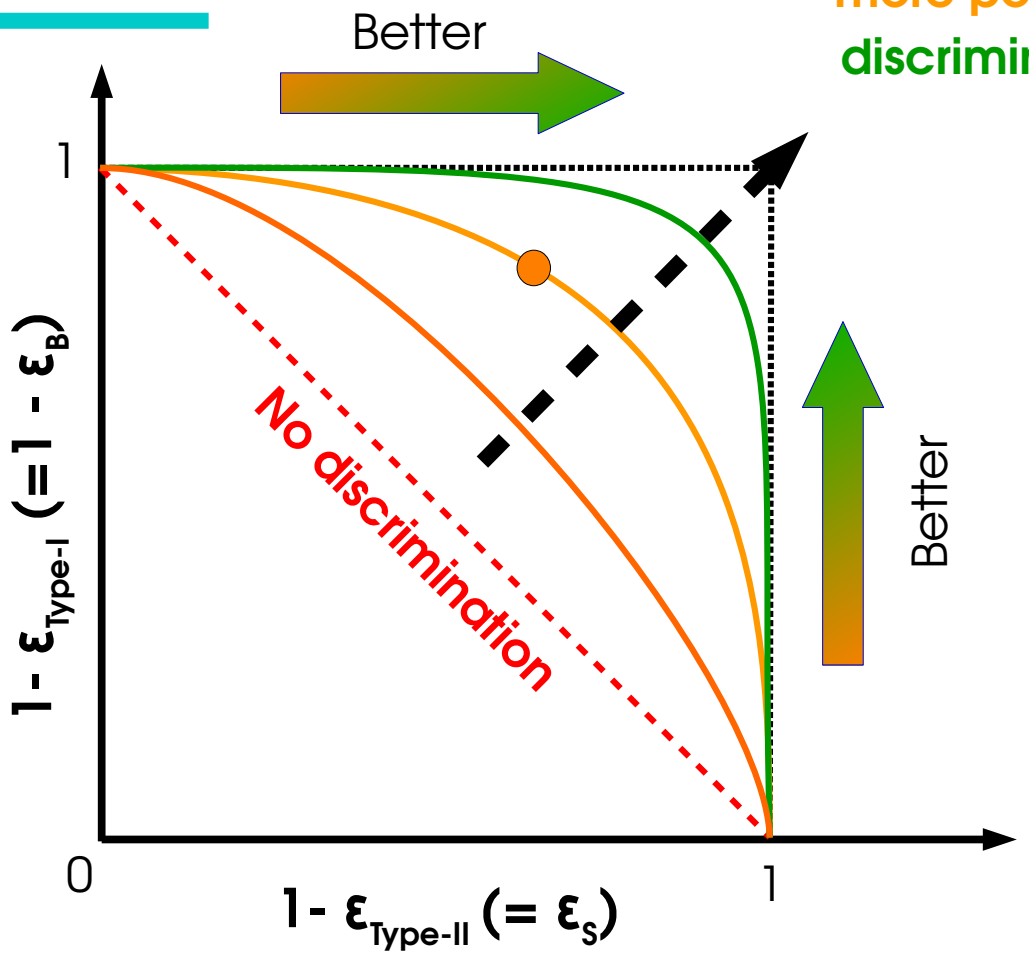
Increasingly more powerful discriminators

"Receiver operating characteristic" (ROC) Curve:

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

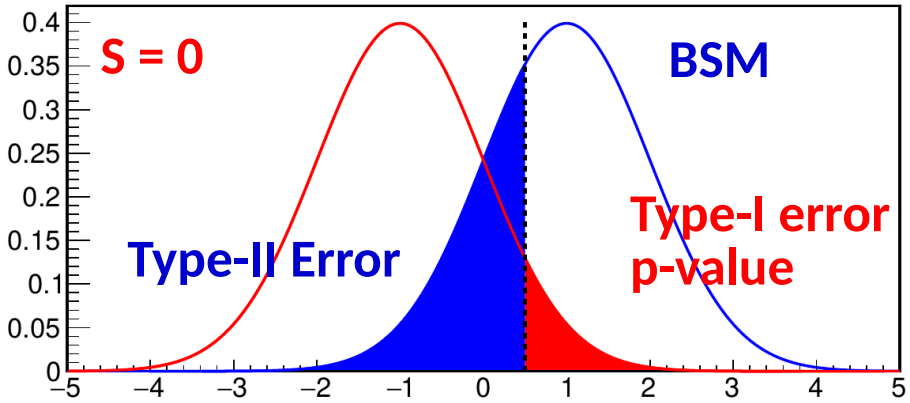
→ All curves monotonically decrease from (0,1) to (1,0)

→ Better discriminators more bent towards (1,1)



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

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



ROC Curves

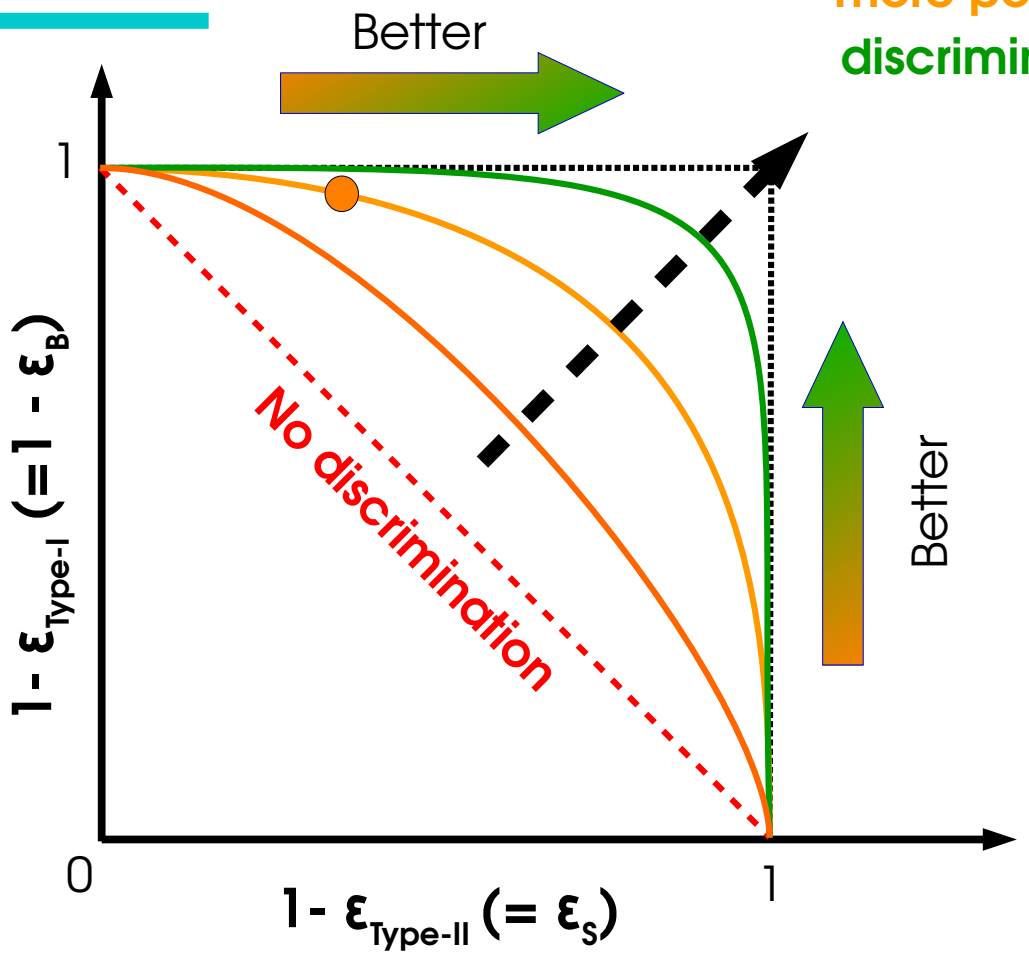
Increasingly more powerful discriminators

"Receiver operating characteristic" (ROC) Curve:

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

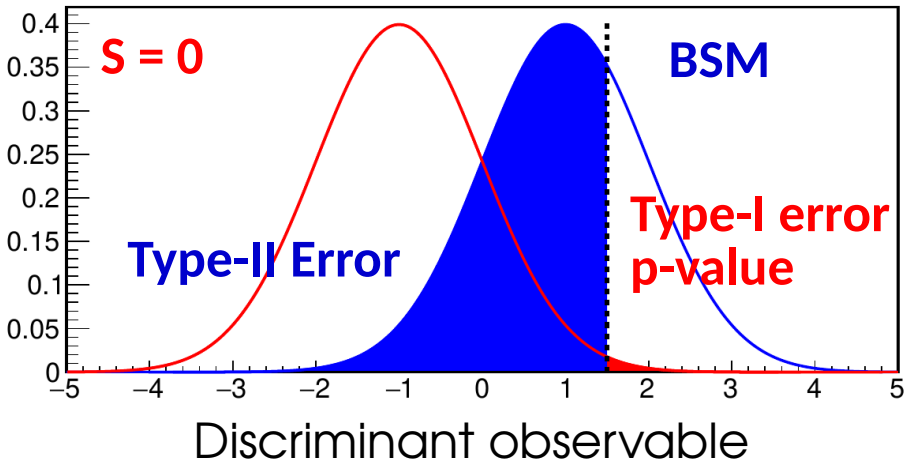
→ All curves monotonically decrease from (0,1) to (1,0)

→ Better discriminators more bent towards (1,1)



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

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



ROC Curves

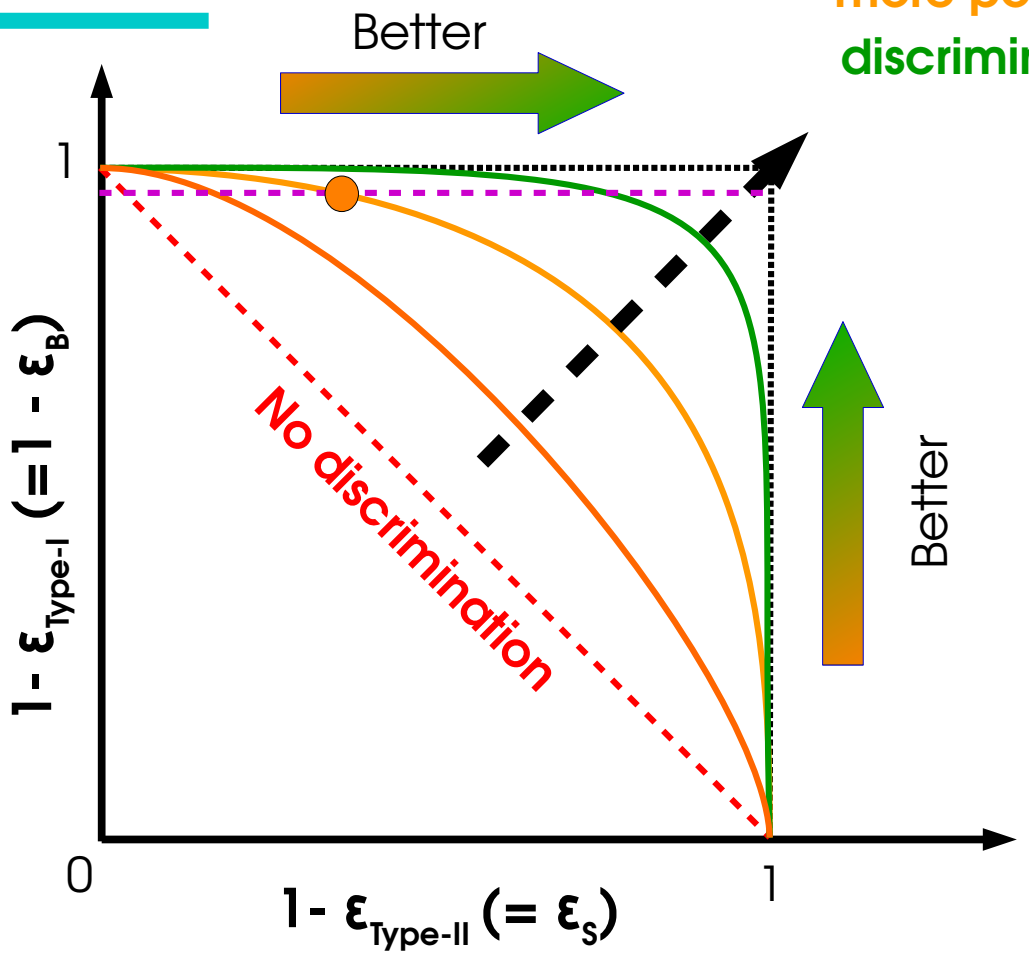
Increasingly more powerful discriminators

"Receiver operating characteristic" (ROC) Curve:

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

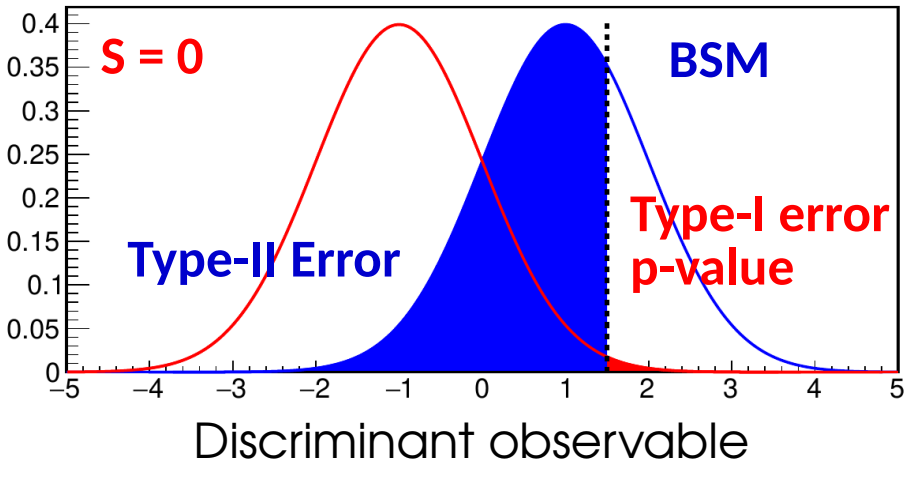
→ All curves monotonically decrease from (0,1) to (1,0)

→ Better discriminators more bent towards (1,1)

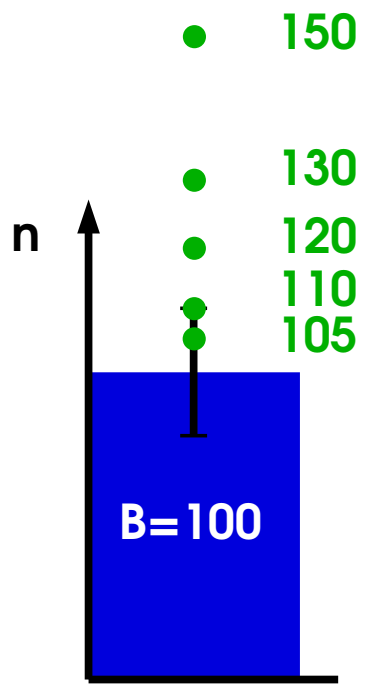


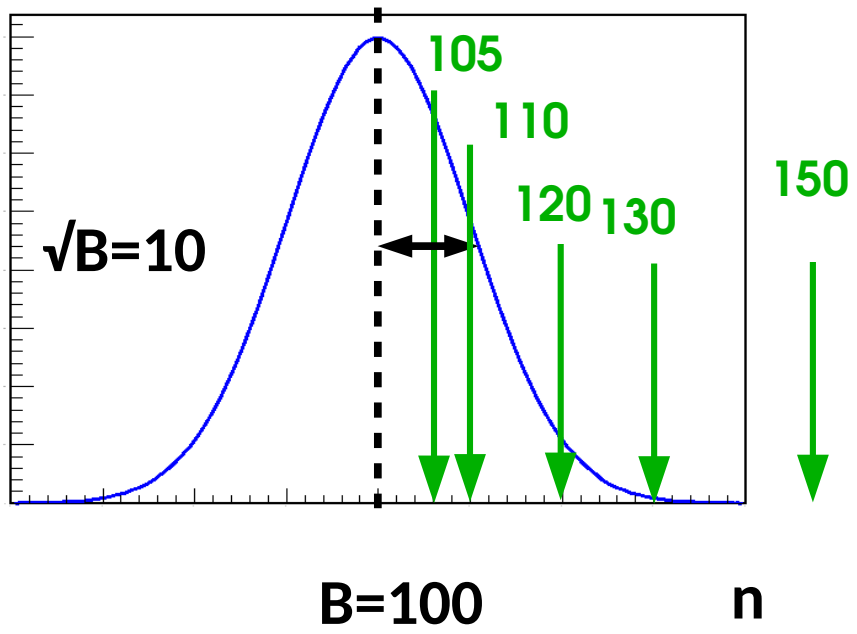
→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.

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



Discovery Testing





n_{obs}	S	Z	p_0
105	5	0.5σ	31%
110	10	1σ	16%
120	20	2σ	2.3%
130	**30**	**3σ**	**0.1%**

Straightforward in this Gaussian case

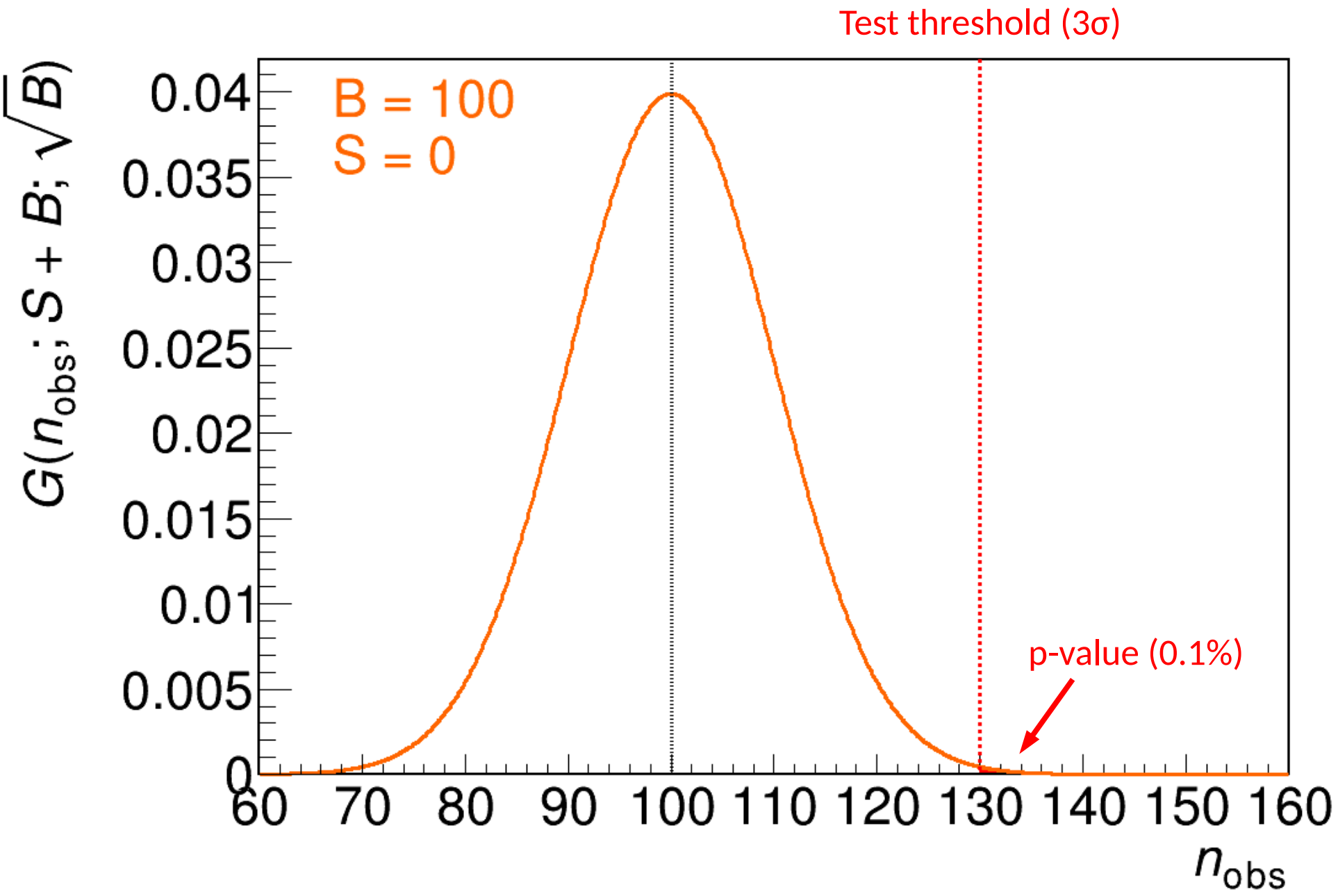
Need to be able to do the same in more complex cases:

- **Determine S**
- **Compute Z and p_0**

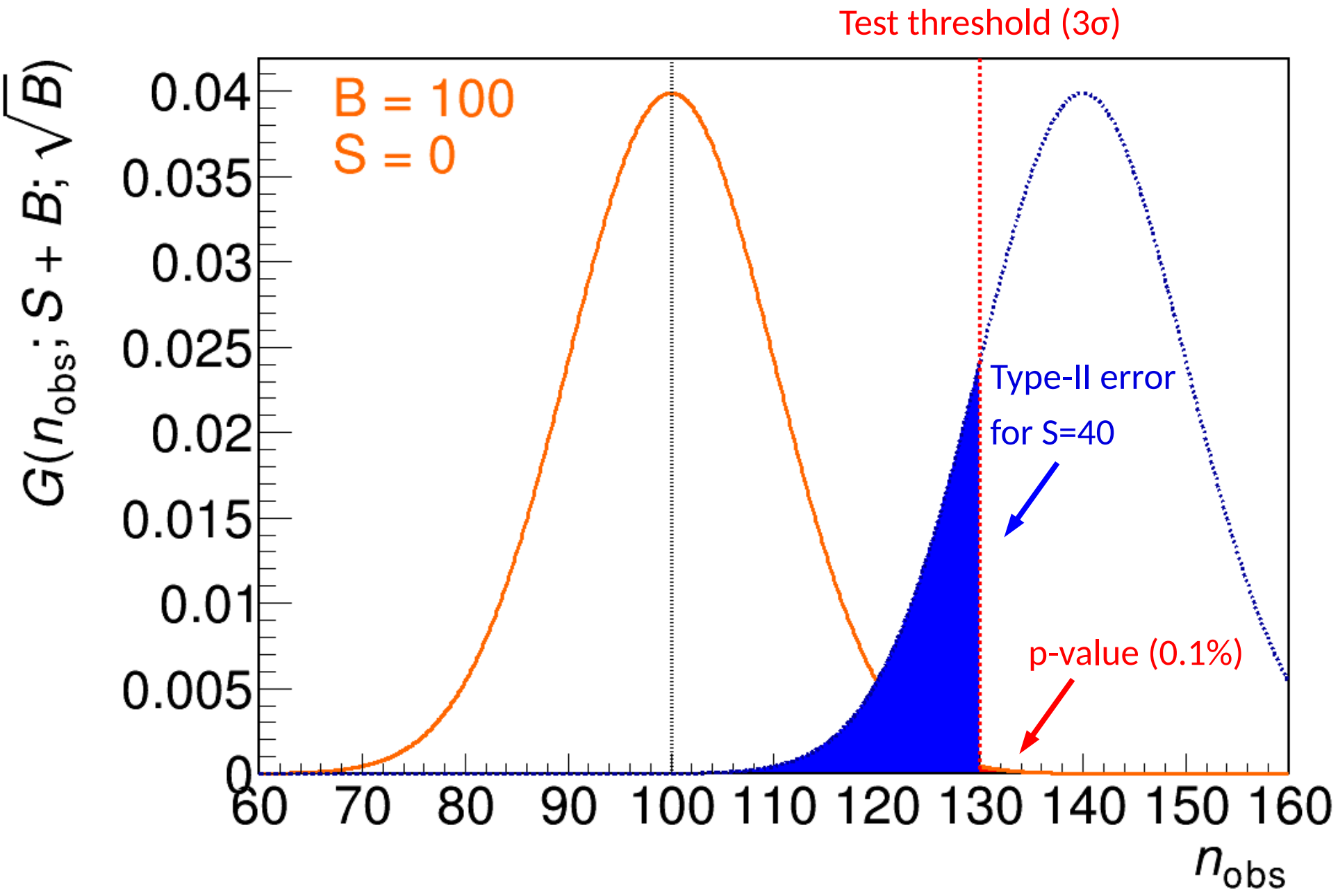
Evidence

Discovery

Testing for Evidence in Gaussian counting



Testing for Evidence in Gaussian counting



Hypothesis Testing with Likelihoods

Neyman-Pearson Lemma

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

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

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

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

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

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

→ **Minimizes Type-II uncertainties** for given level of Type-I uncertainties

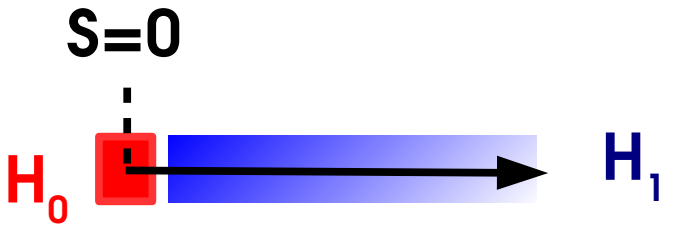
→ **In the following**: all tests based on LR, will focus on p-values (Type-I errors), trusting that Type-II errors are anyway as small as they can be…

Discovery: Test Statistic

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

Discovery :

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



→ For H_1, any S > 0 is possible, which to use ? **The one preferred by the data, $\hat{S}$.**

⇒ Use Likelihood ratio: $\dfrac{L(S=0)}{L(\hat{S})}$

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

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

Discovery p-value

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

⇒ observed $\hat{S}$ is far from 0

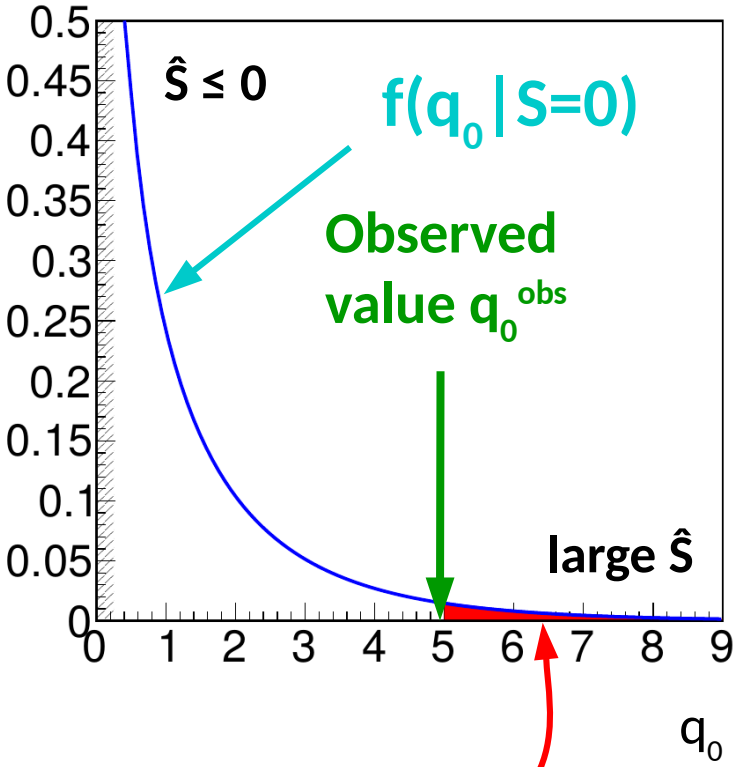
⇒ $H_0(S=0)$ ***disfavored*** compared to $H_1(S\neq 0)$.

⇒ Large $\hat{S}$!

Compute *p-value* in the tail of the distribution

to exclude H_0 (… and **claim a discovery!**)

data prefer S = 0 ⟷ data prefer S > 0



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

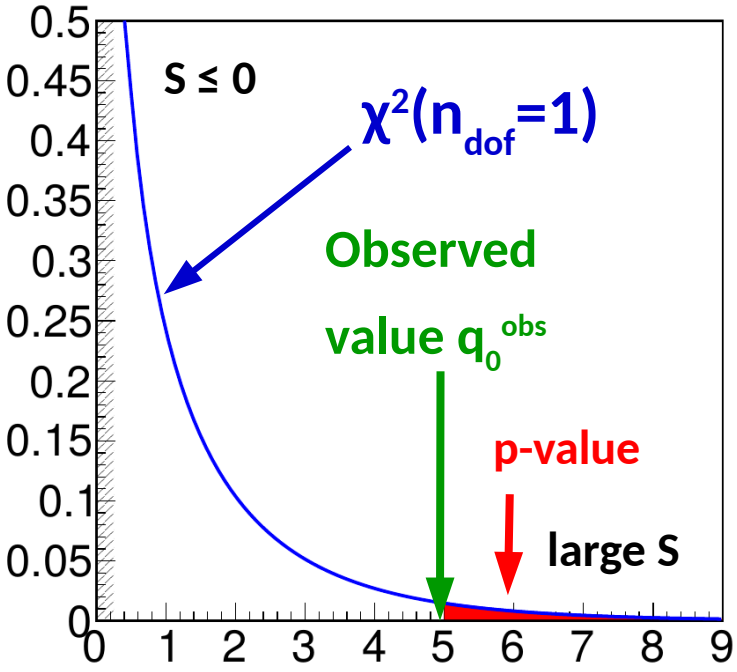
Need to know $f(q_0 \mid S=0)$, the distribution of the test statistic…

Asymptotic distribution of q_0

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

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

Wilk's Theorem: **q_0 distributed as χ^2 (n_{par}) for S = 0**



⇒ n_{par} = 1 : **$\sqrt{q_0}$ is distributed as a Gaussian**

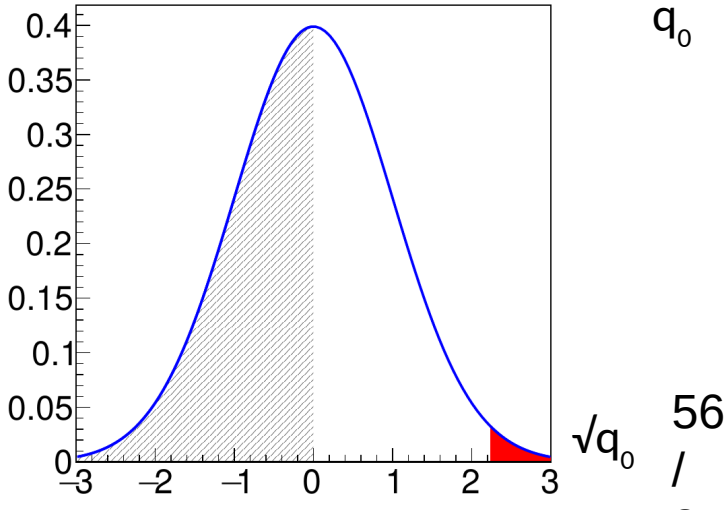
⇒ Can compute p-values from Gaussian quantiles

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

⇒ Even more simply, the significance is:

$$Z = \sqrt{q_0}$$

Typically works well already for for event counts of O(5) and above ⇒ Widely applicable



(*) 1-line "proof" : asymptotically L and S are Gaussian, so

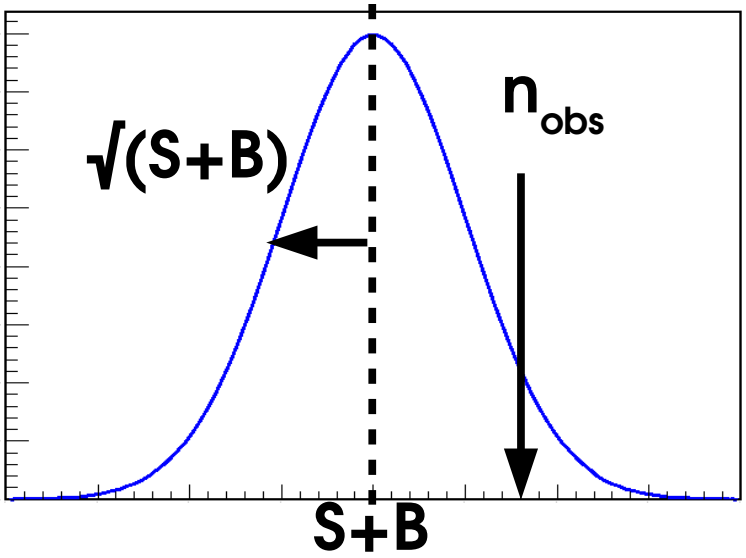
$$L(S) = \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}}{\sigma}\right)^2\right] \Rightarrow q_0 = \left(\frac{\hat{S}}{\sigma}\right)^2 \Rightarrow \sqrt{q_0} = \frac{\hat{S}}{\sigma} \sim G(0,1) \Rightarrow q_0 \sim \chi^2(n_{dof}=1)$$

Homework 1: Gaussian Counting

Count number of events n in data

→ Assume n large enough so process is Gaussian

→ Assume B is known, and we measure S



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

→ Find the best-fit value (MLE) $\hat{S}$ for the signal

(can use $\lambda = -2 \log L$ instead of L for simplicity)

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

→ Find the expression for the significance

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

Homework 2: Poisson Counting

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

$$L(S;n) = e^{-(S+B)}(S+B)^n$$

(Can remove the n! constant since we're only dealing with L ratios)

→ As before, compute $\hat{S}$, and q_0

→ Compute $Z = \sqrt{q_0}$, assuming asymptotic behavior

Solution:

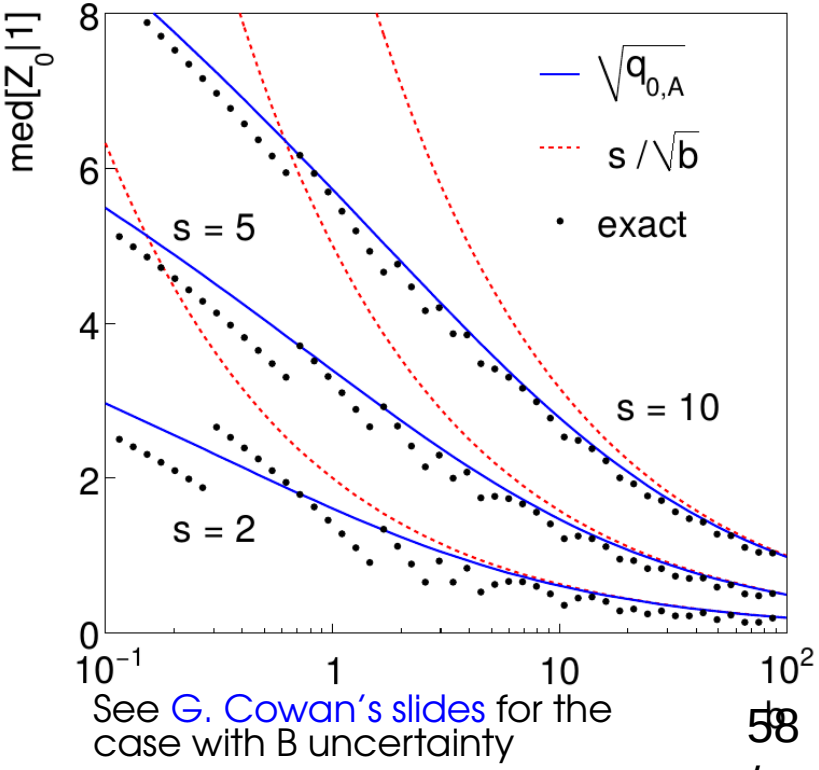
$$Z = \sqrt{2\left[(\hat{S}+B)\log\left(1+\frac{\hat{S}}{B}\right) - \hat{S}\right]}$$

Exact result can be obtained using

pseudo-experiments → close to $\sqrt{q_0}$ result

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

Eur.Phys.J.C71:1554,2011



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

Discovery Thresholds

Evidence : $3\sigma \Leftrightarrow p_0 = 0.3\% \Leftrightarrow$ 1 chance in 300

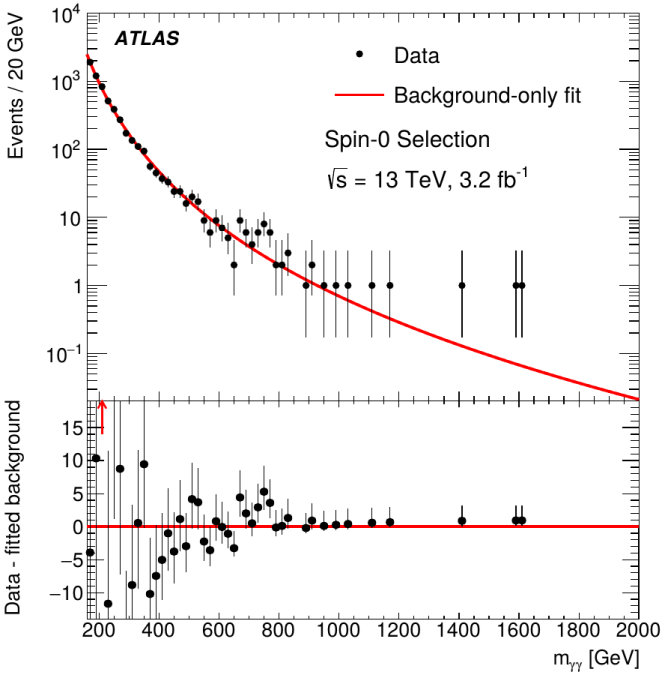
Discovery: $5\sigma \Leftrightarrow p_0 = 3\ 10^{-7} \Leftrightarrow$ 1 chance in 3.5M

Why so high thresholds ? (from Louis Lyons):

- **Look-elsewhere effect**: searches typically cover multiple independent regions ⇒ Higher chance to have a fluctuation "somewhere"

  N_{trials} **~ 1000 : local 5σ ⇔ O(10^{-4})** more reasonable

- **Mismodeled systematics**: factor 2 error in syst-dominated analysis ⇒ factor 2 error on Z…



- **History**: 3σ and 4σ excesses do occur regularly, for the reasons above

Extraordinary claims require extraordinary evidence!

Extra Slides

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,)

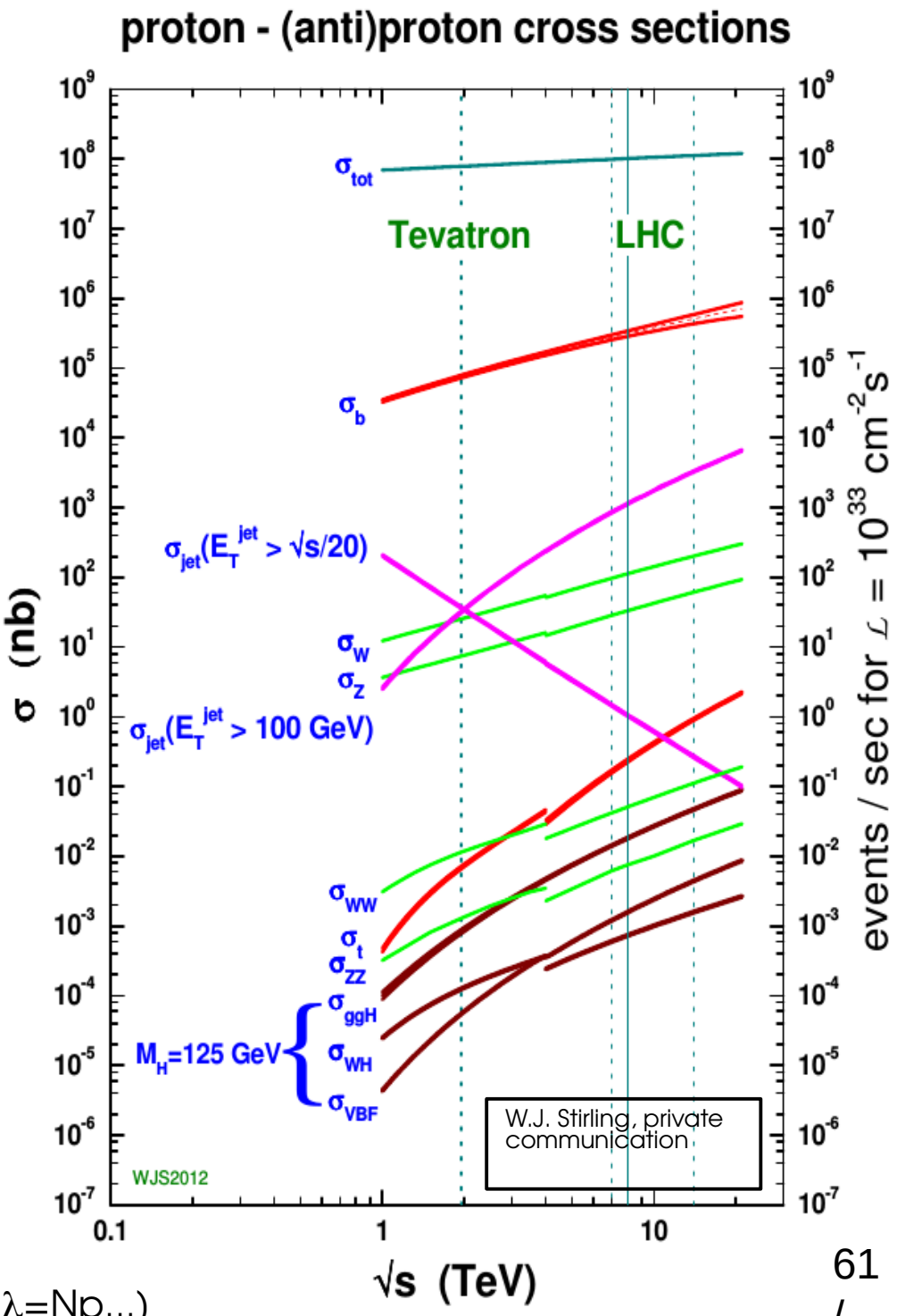
- **Trigger rate ~ 1 kHz**

  (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.



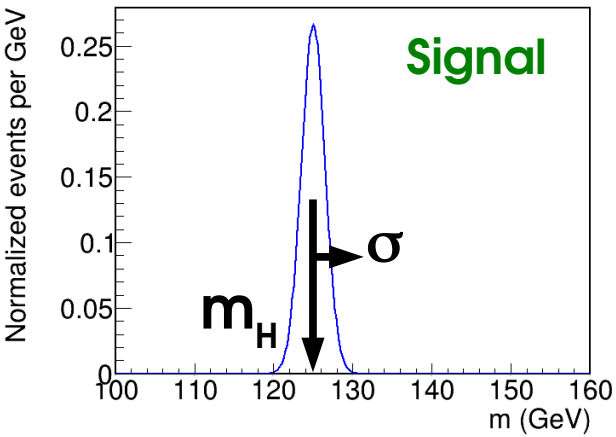
(Large N = design requirement, to get not-too-small λ=Np…)

Unbinned Shape Analysis

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

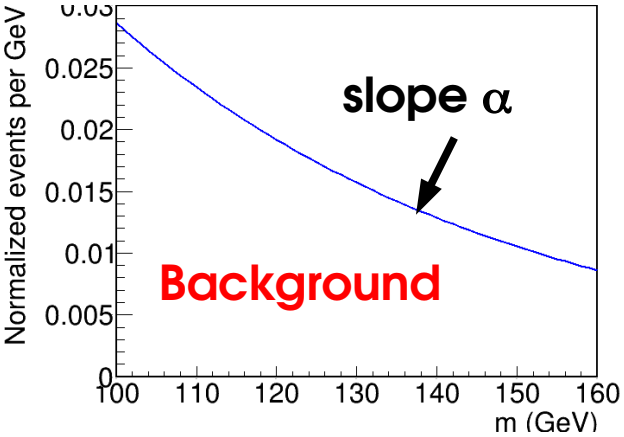
→ Describe shape of the **distribution of m**

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



H→γγ-inspired example:

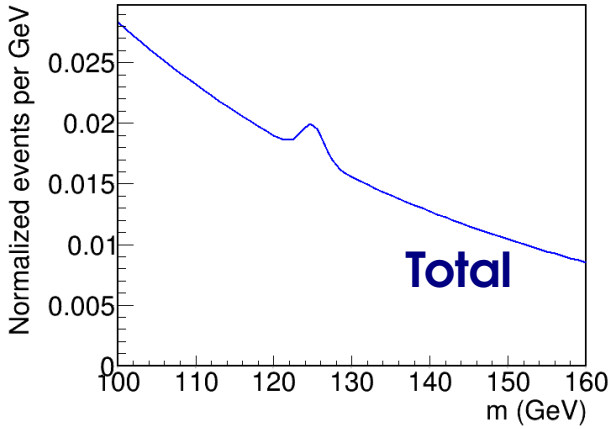
- Gaussian signal $P_{\text{signal}}(m) = G(m; m_H, \sigma)$
- Exponential bkg $P_{\text{bkg}}(m) = \alpha\, e^{-\alpha m}$



Expected yields : **S**, **B**

⇒ **Total PDF for a single event**:

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



⇒ **Total PDF for a dataset**

Probability to observe the value m_i

Probability to observe n events

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

Poisson Example

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

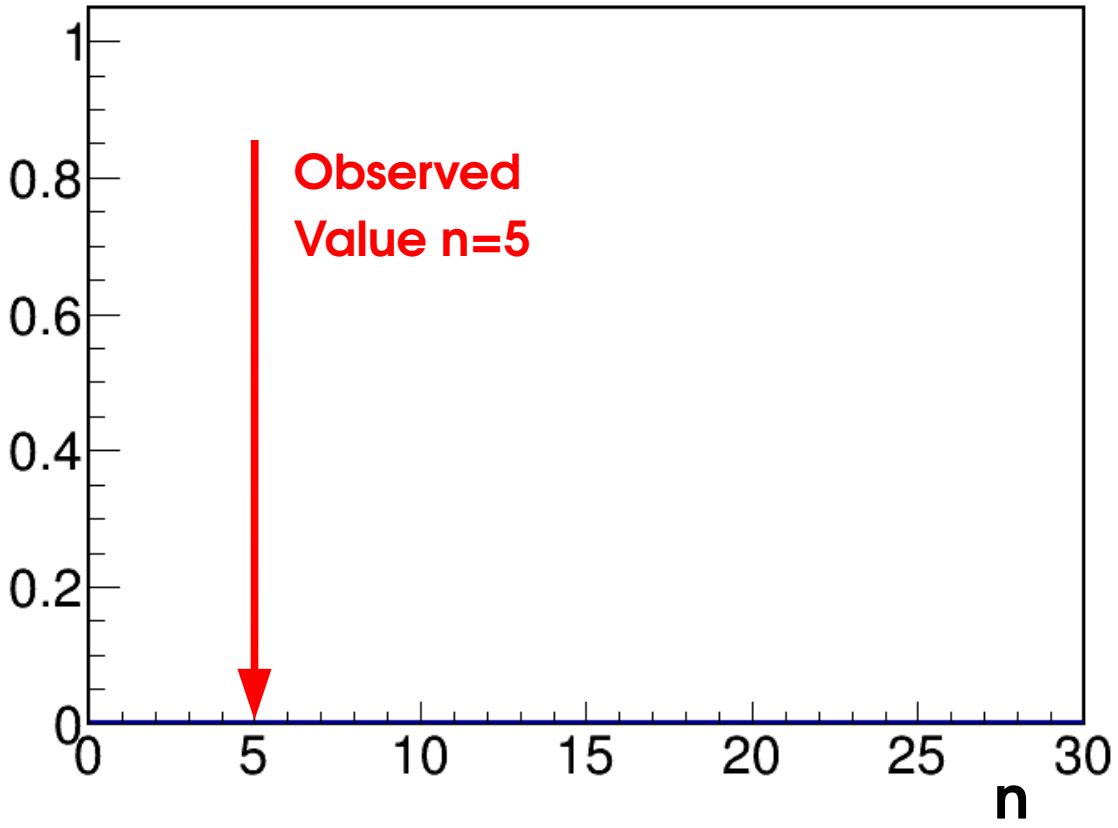
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

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



Poisson Example

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

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

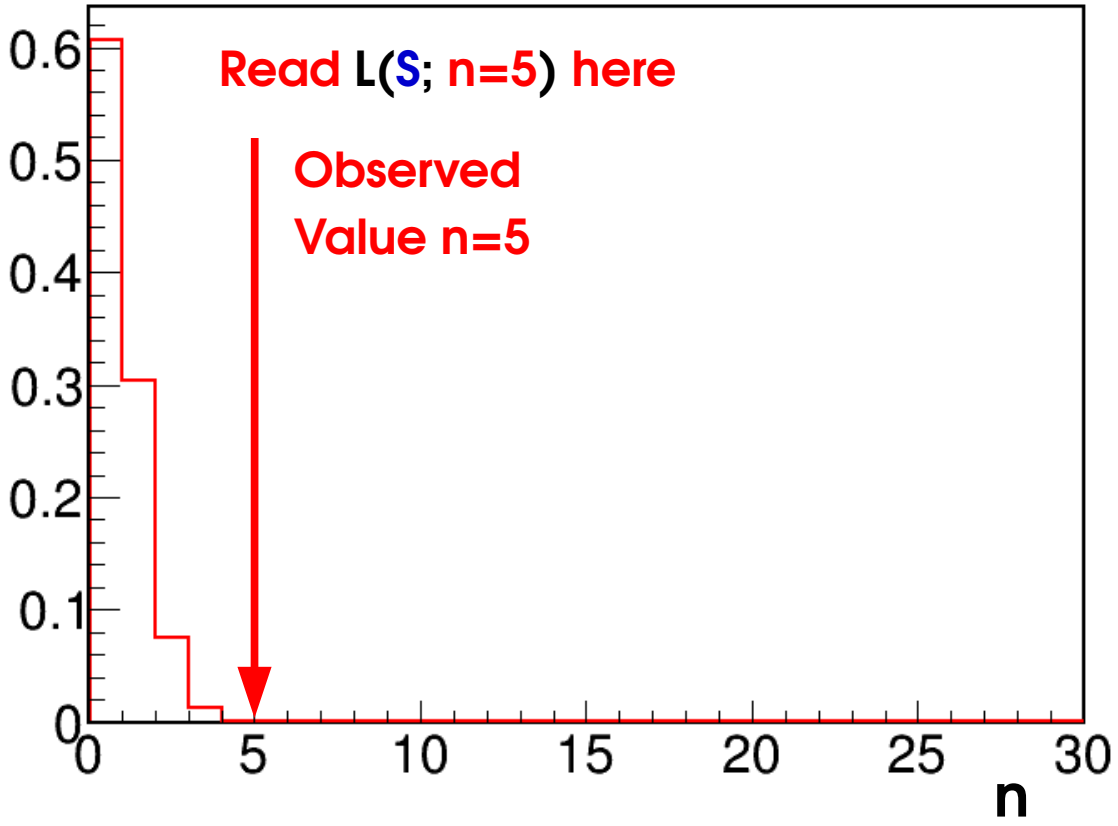
Say we **observe** n=5, want to infer information on the parameter S

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

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

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

P(S = 0.5)
Low
likelihood



Poisson Example

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

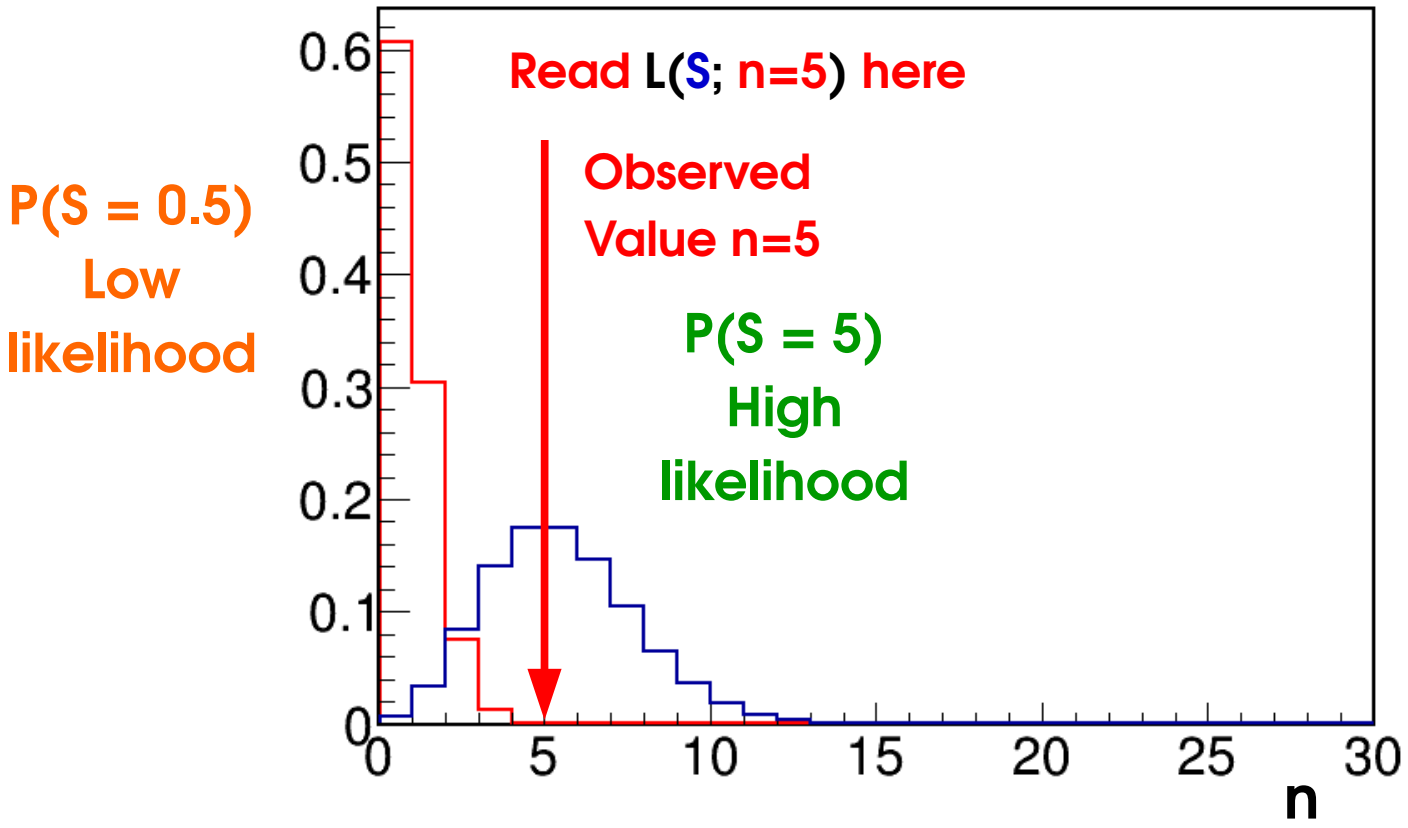
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

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



Poisson Example

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

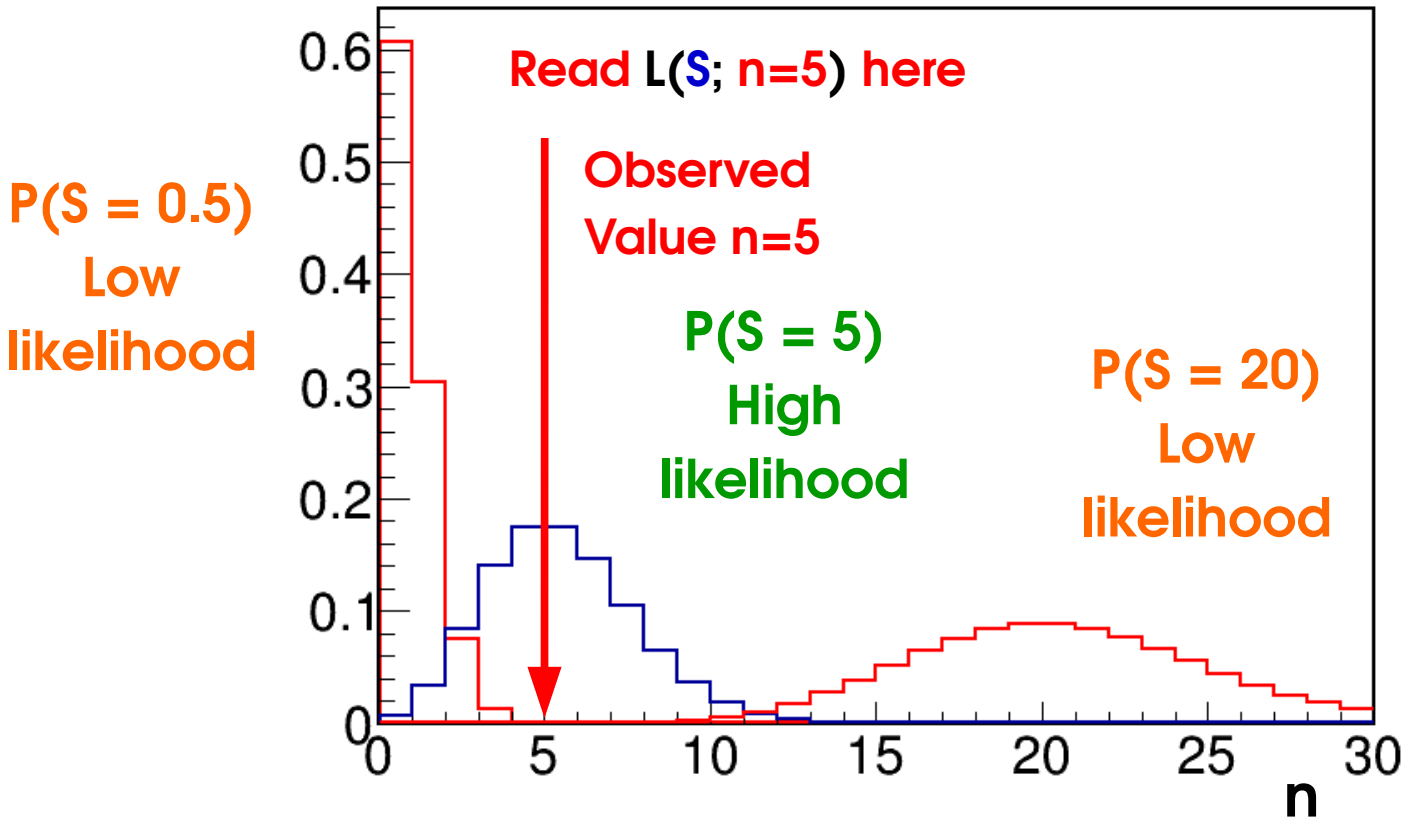
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

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

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

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



Poisson Example

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

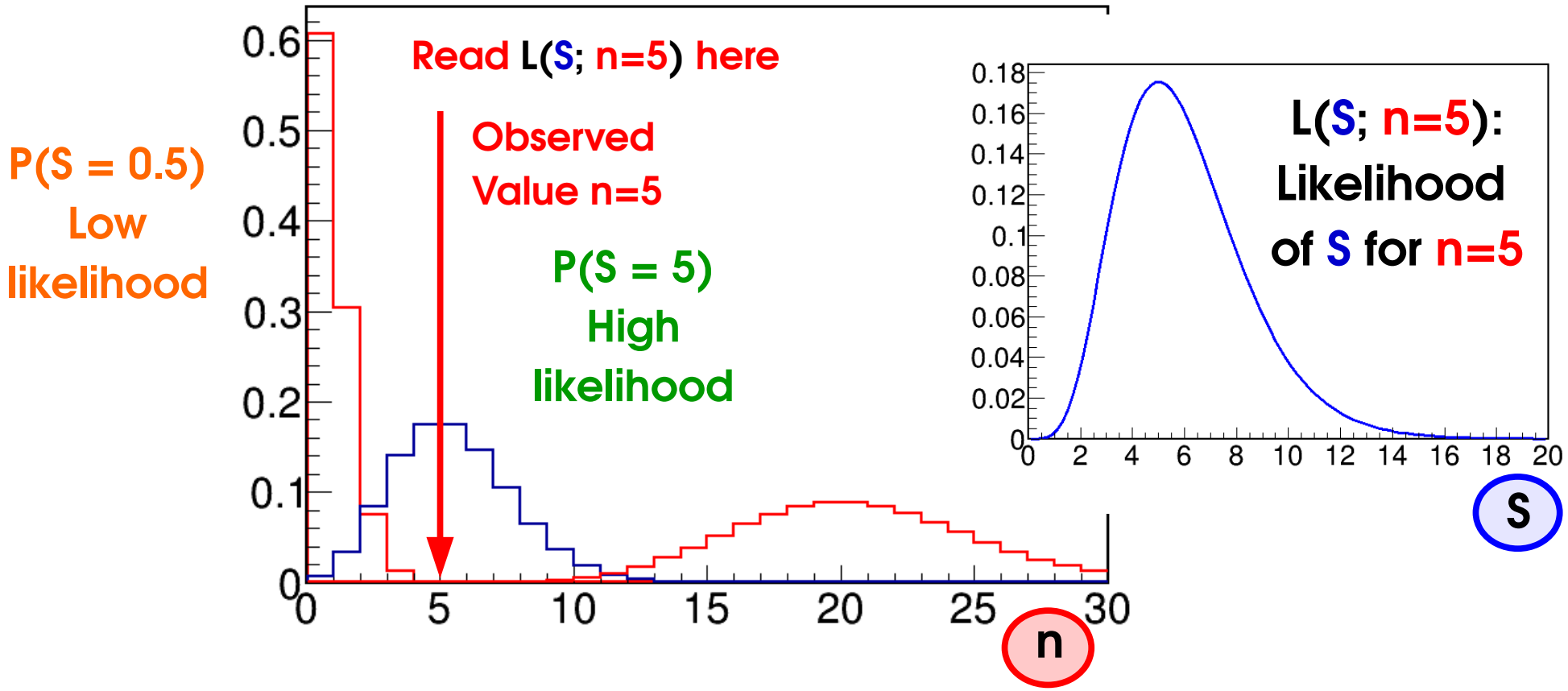
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

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

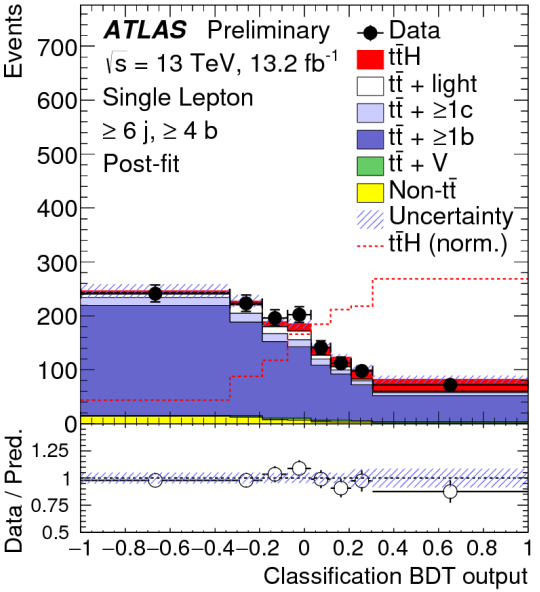
→ Try different values of S for a fixed data value n=5

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

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



MLEs in Shape Analyses



Binned shape analysis:

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

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

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

Needs a computer...

In the Gaussian limit

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

χ^2 formula!

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

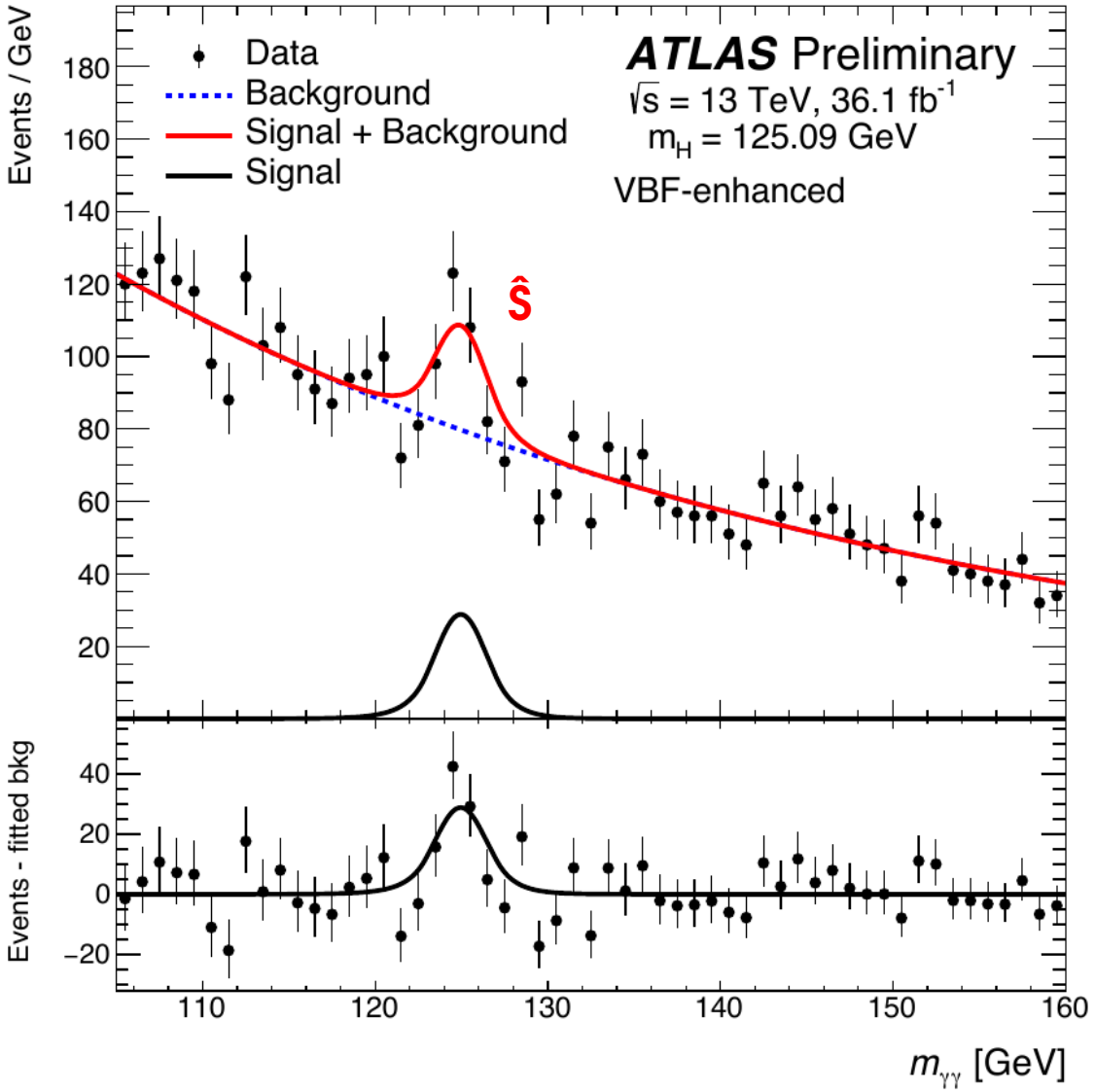
→ **Poisson MLE** (min λ_{Pois}) : ***Best fit value*** in a *likelihood* fit (in `ROOT`, fit option "`L`")

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

In both cases, MLE ⇔ *Best Fit*

H→γγ

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



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

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

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

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

ATLAS-CONF-2017-045

MLE Properties

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

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

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

for large enough datasets

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

  → MLE captures all the available information in the data

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

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

  → Usually more efficient numerically

  → For Gaussian L, λ is parabolic:

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

Extra: Fisher Information

Fisher Information:

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

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

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

→ smaller σ_{Gauss} ⇒ more information.

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

For any estimator μ̃.

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

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

Gaussian case:

- For a Gaussian estimator μ̃

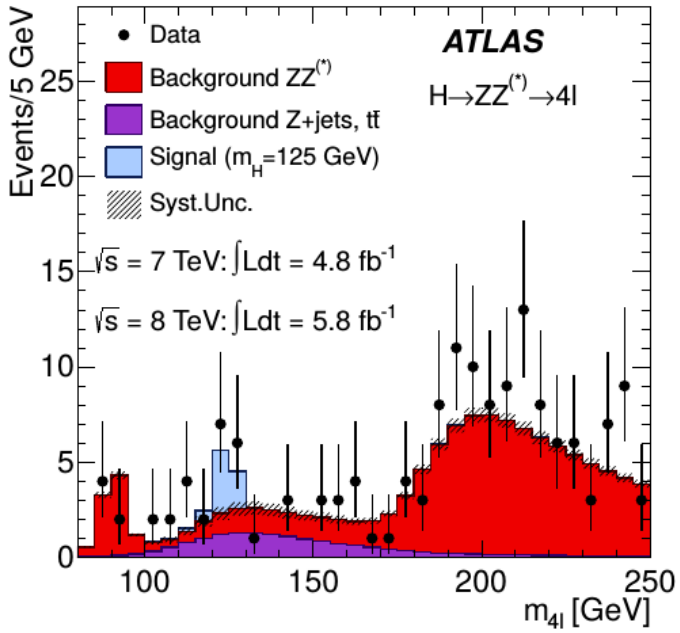
$$P(\tilde{\mu}) \propto \exp\left(-\frac{(\tilde{\mu} - \mu^*)^2}{2\sigma^2_{\tilde{\mu}}} \right)$$

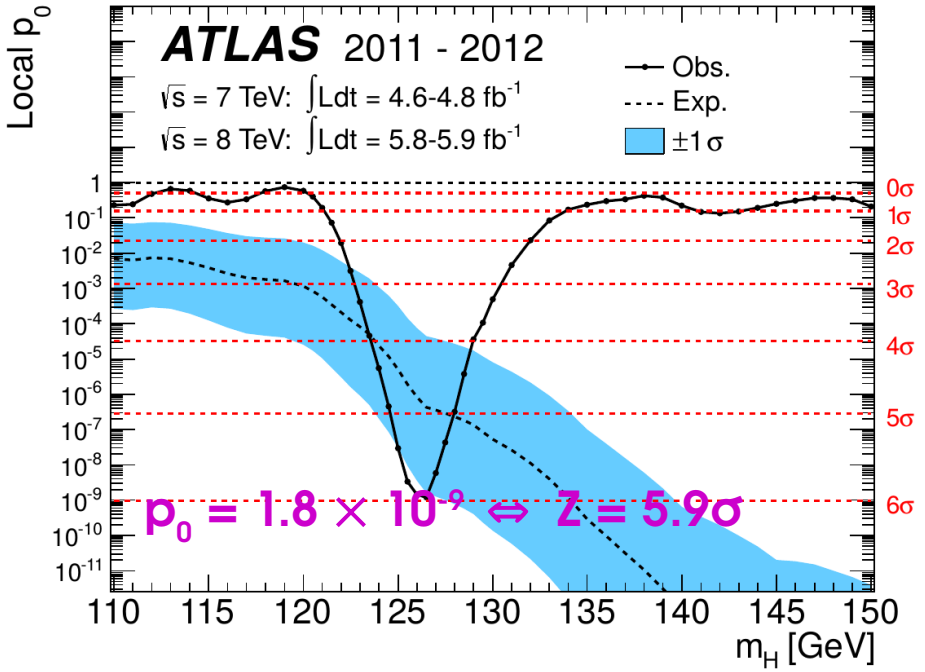
- **MLE**: $\text{Var}(\hat{\mu}) = \sigma_{\hat{\mu}}^2$

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

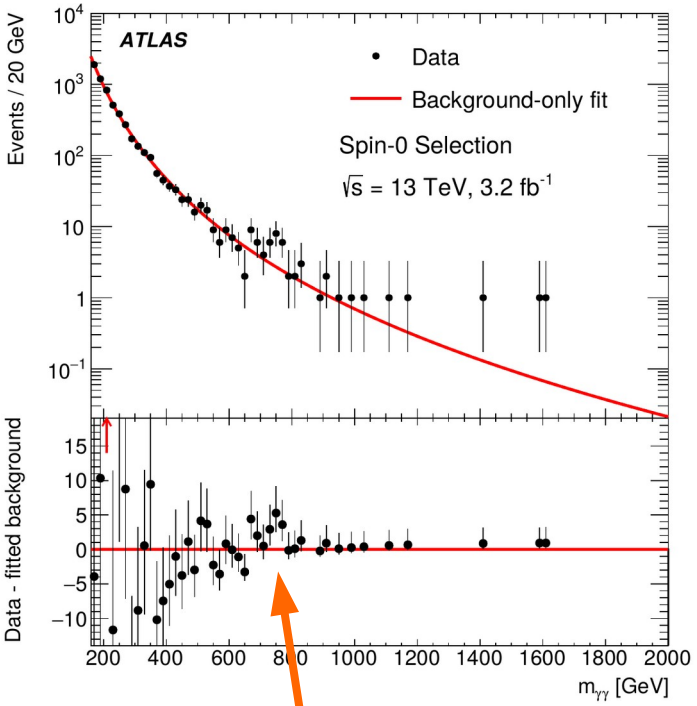
Some Examples

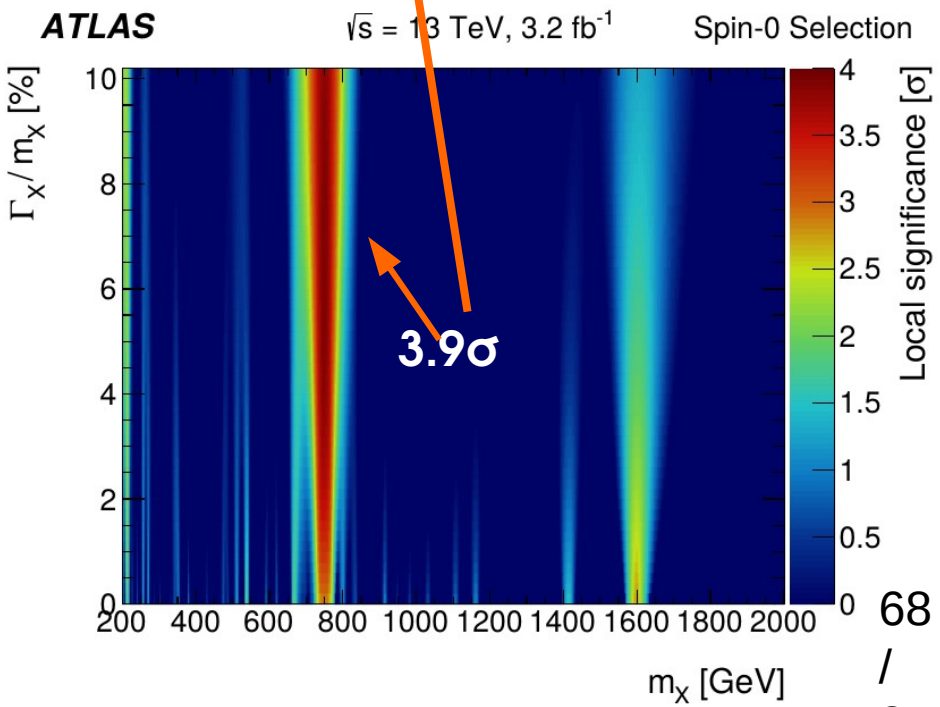
Higgs Discovery: Phys. Lett. B 716 (2012) 1-29





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





Upper Limit Pathologies

Upper limit: $S_{up} \sim \hat{S} + 1.64\,\sigma_S$.

Problem: for negative $\hat{S}$, get **very** good observed limit.

→ For $\hat{S}$ sufficiently negative, even $S_{up} < 0$!

How can this be ?

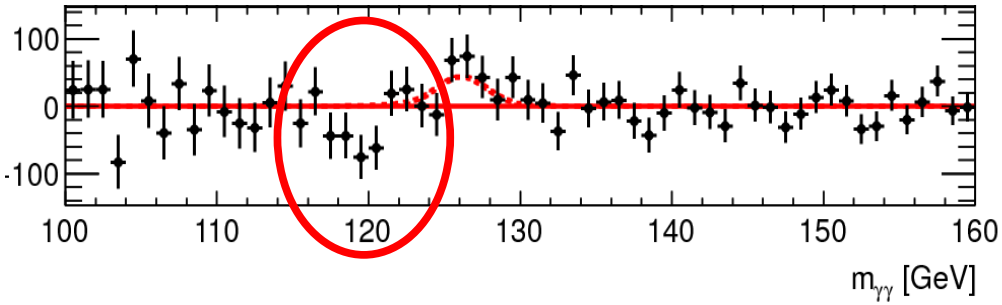
→ **Background modeling issue ?...** Or:

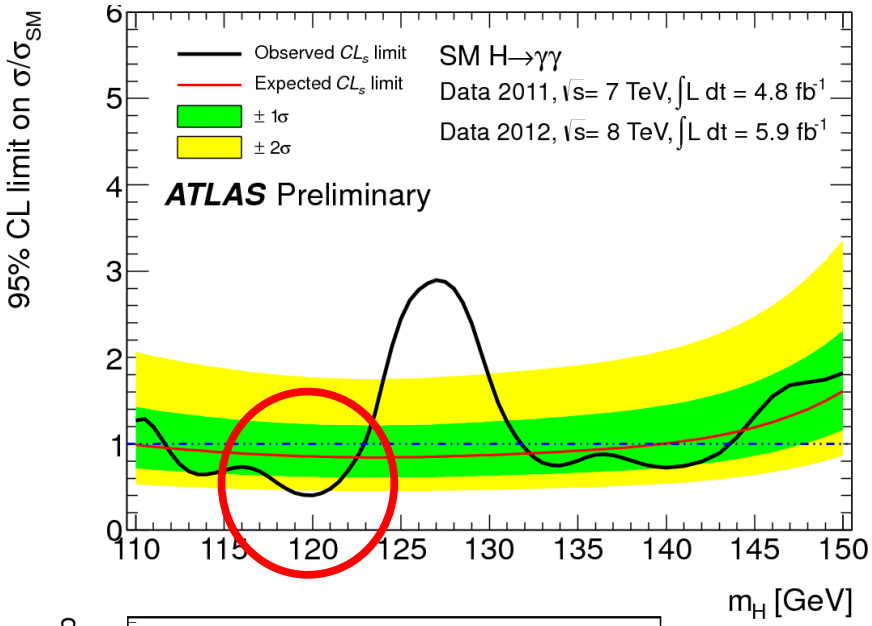
→ This is a **95%** limit ⇒ **5% of the time, the limit wrongly excludes the true value,** *e.g.* S*=0.

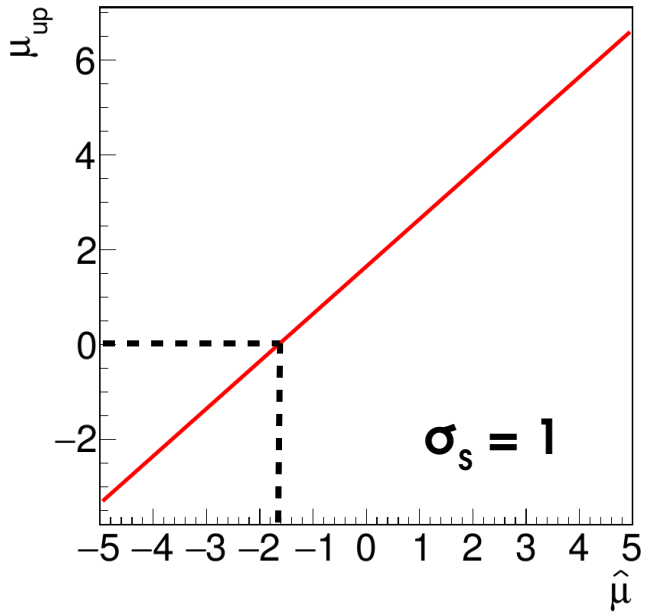
Options

→ **live with it**: sometimes report limit < 0

→ **Special procedure to avoid these cases,** since if we assume S must be >0, we know a priori this is just a fluctuation.







CL_s

A. Read, J.Phys. G28 (2002) 2693-2704

Usual solution in HEP : **CL_s**.

→ Compute modified p-value

$$p_{CL_s} = \frac{p_{S_0}}{(1 - p_B)}$$

The usual p-value under H(S=S_0) (=5%)

The p-value computed under H(S=0)

⇒ **Rescale** exclusion at S_0 by exclusion at S=0.

→ Somewhat ad-hoc, but good properties...

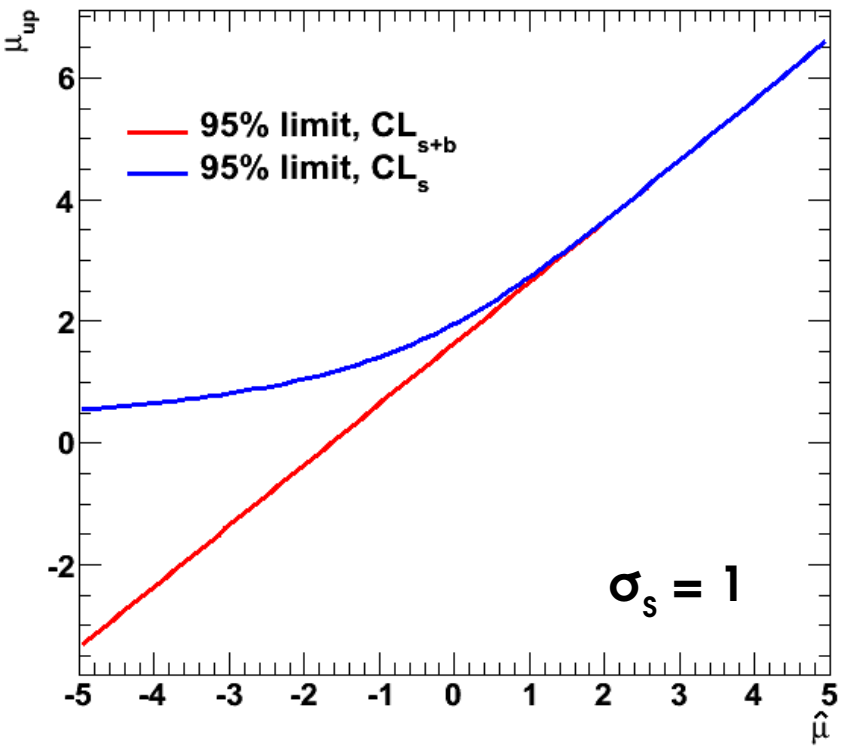
$\hat{S}$ compatible with 0 : $p_B \sim O(1)$

$p_{CLs} \sim p_{S0}$ **~ 5%, no change**.

Far-negative $\hat{S}$: $1 - p_B \ll 1$

$p_{CLs} \sim p_{S0}/(1-p_B) \gg 5\%$

→ lower exclusion ⇒ higher limit, usually >0 as desired

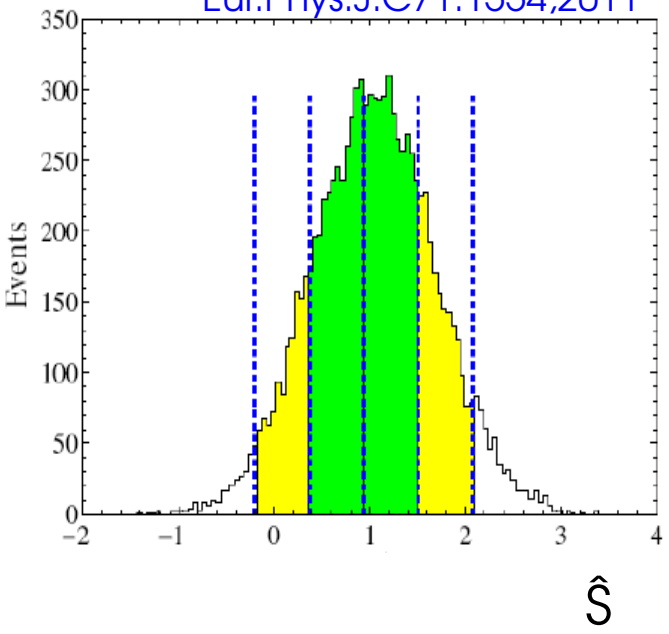


Drawback: ***overcoverage***

→ limit is claimed to be 95% CL, but actually >95% CL for small $1-p_B$.

CL_s : Gaussian Bands

Eur.Phys.J.C71:1554,2011



Usual Gaussian counting example with known B:

95% CL_s upper limit on S:

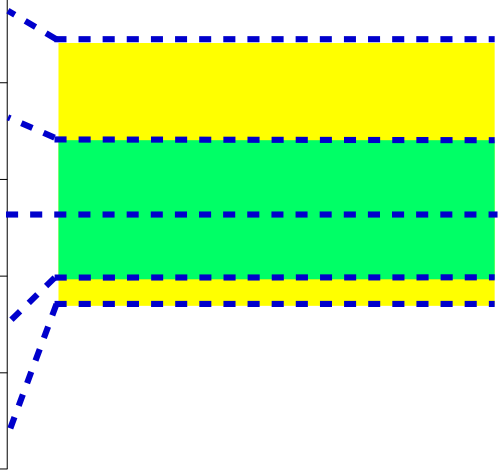
$$S_{up} = \hat{S} + \left[\Phi^{-1}\left(1 - 0.05\ \Phi\left(\hat{S}/\sigma_S\right)\right)\right]\sigma_S \qquad \text{with } \sigma_S = \sqrt{B}$$

Compute expected bands for S=0:

→ **Asimov dataset ⇔ $\hat{S} = 0$** : $S^0_{up,exp} = 1.96\ \sigma_S$

→ **± nσ bands**: $S^{\pm n}_{up,exp} = \left(\pm n + \left[1 - \Phi^{-1}\left(0.05\ \Phi(\mp n)\right)\right]\right)\sigma_S$

n	$S_{exp}^{\pm n} / \sqrt{B}$
+2	3.66
+1	2.72
0	1.96
-1	1.41
-2	1.05



CLs :

- Positive bands somewhat reduced,
- Negative ones more so

Band width from $\sigma^2_{S,A} = \dfrac{S^2}{q_S(\text{Asimov})}$ depends on S, for non-Gaussian cases,different values for each band...

Comparison with LEP/TeVatron definitions

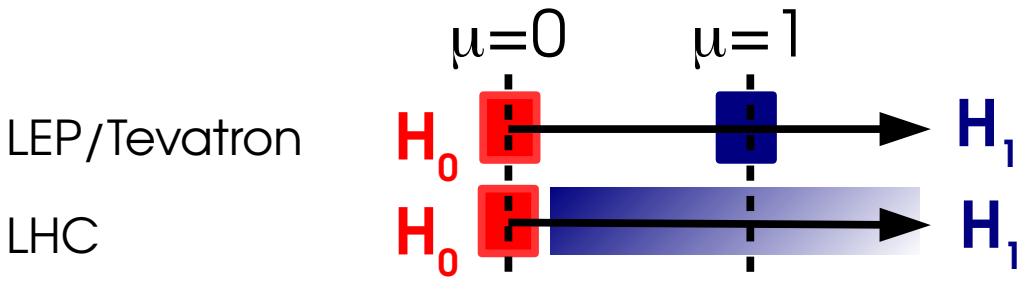
Likelihood ratios are not a new idea:

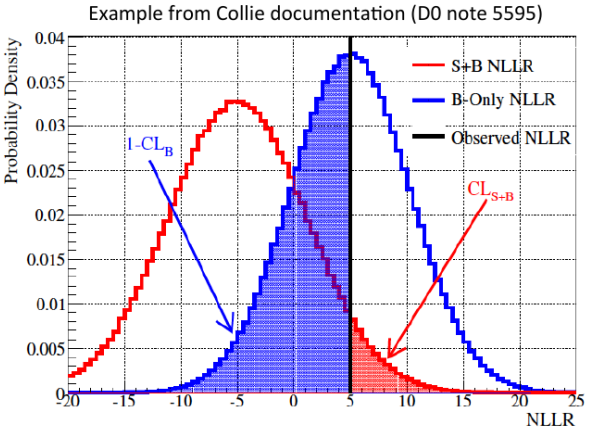
- **LEP**: Simple LR with NPs from MC
  - Compare μ=0 and μ=1
- **Tevatron**: PLR with profiled NPs

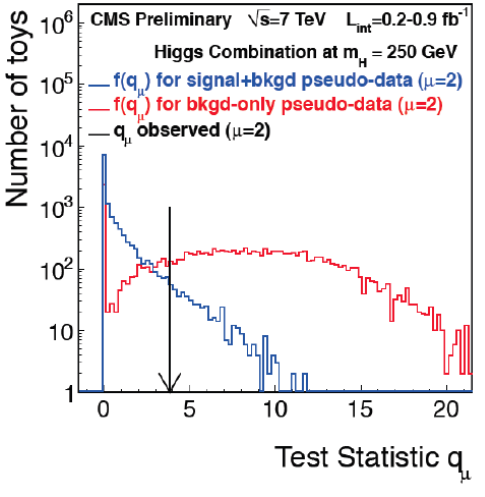
$$q_{LEP} = -2\log\frac{L(\mu=0,\tilde{\theta})}{L(\mu=1,\tilde{\theta})}$$

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

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







→ Asymptotically:

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

→ Still use TeVatron-style for discrete cases

Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

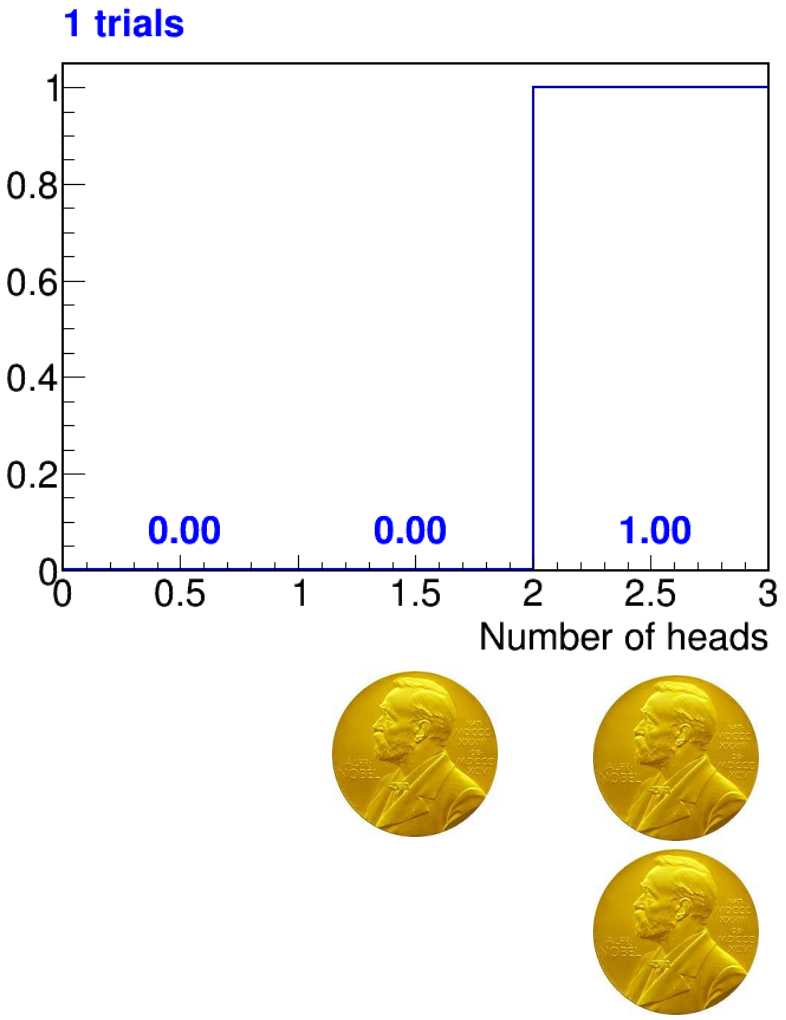
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

Properties

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



Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

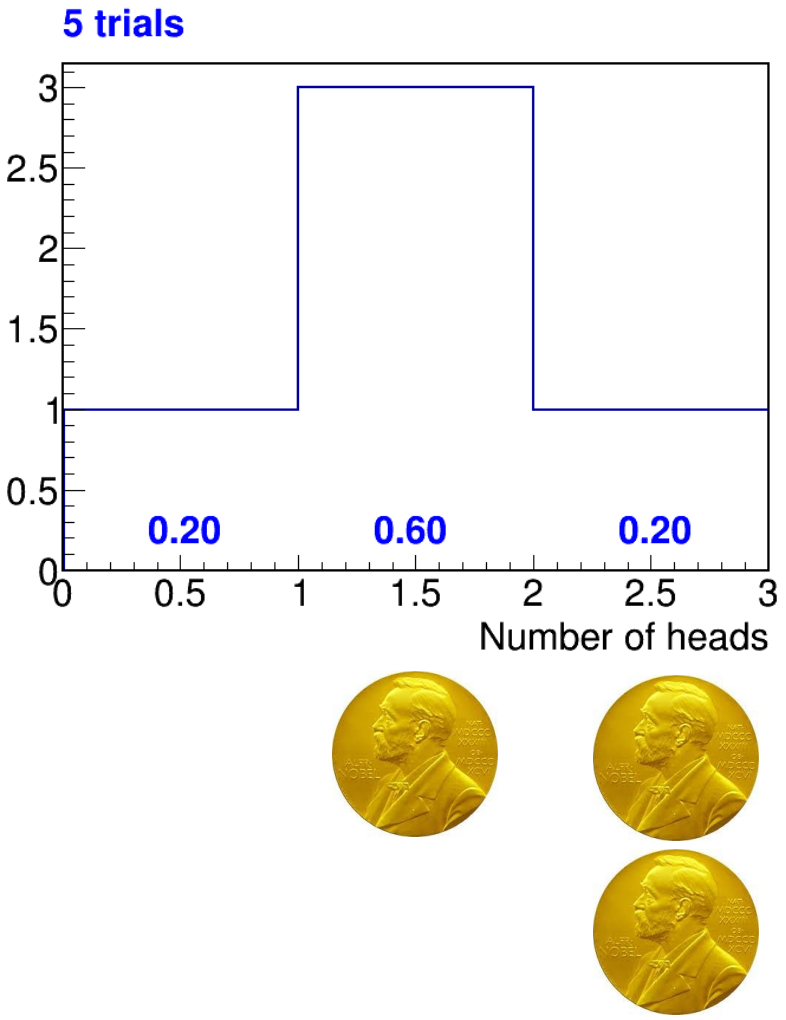
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

Properties

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



Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

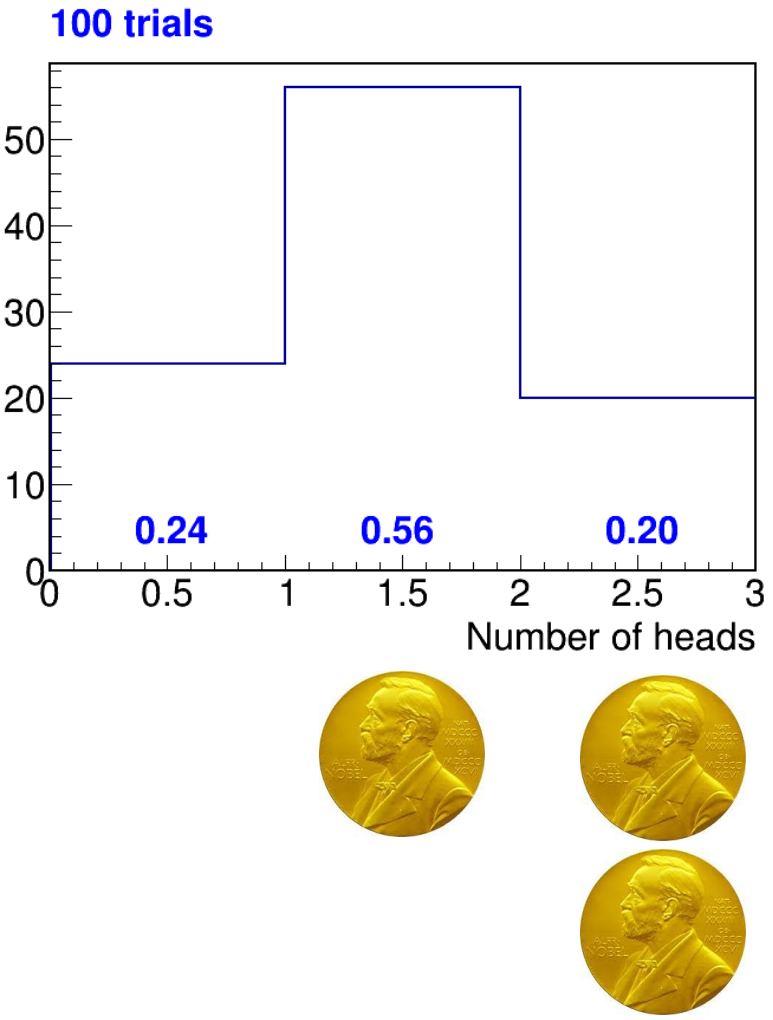
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

Properties

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



Probability Distributions

Probabilistic treatment of possible outcomes

⇒ ***Probability Distribution***

Example: two-coin toss

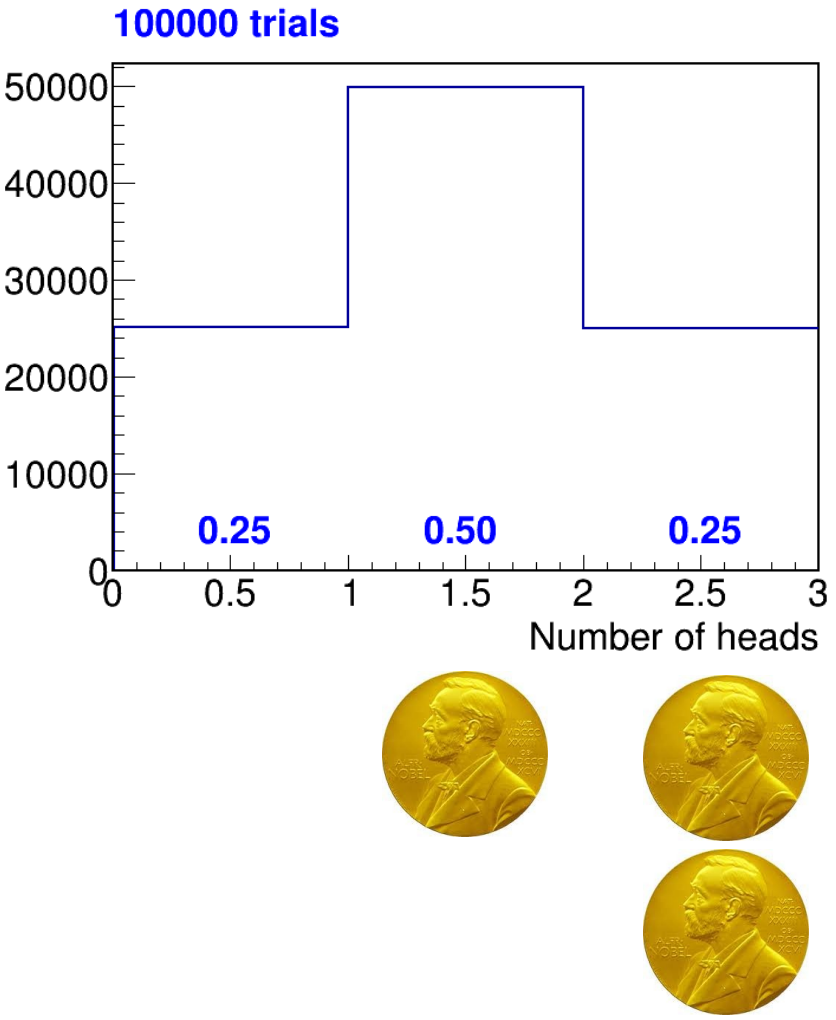
→ Fractions of events in each bin i converge to a limit p_i

Probability distribution :

{ P_i } for i = 0, 1, 2

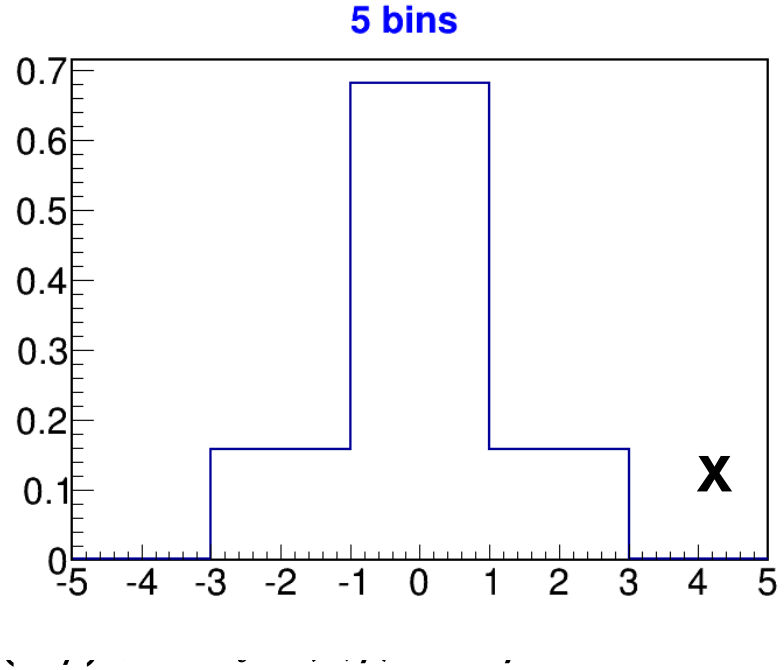
Properties

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



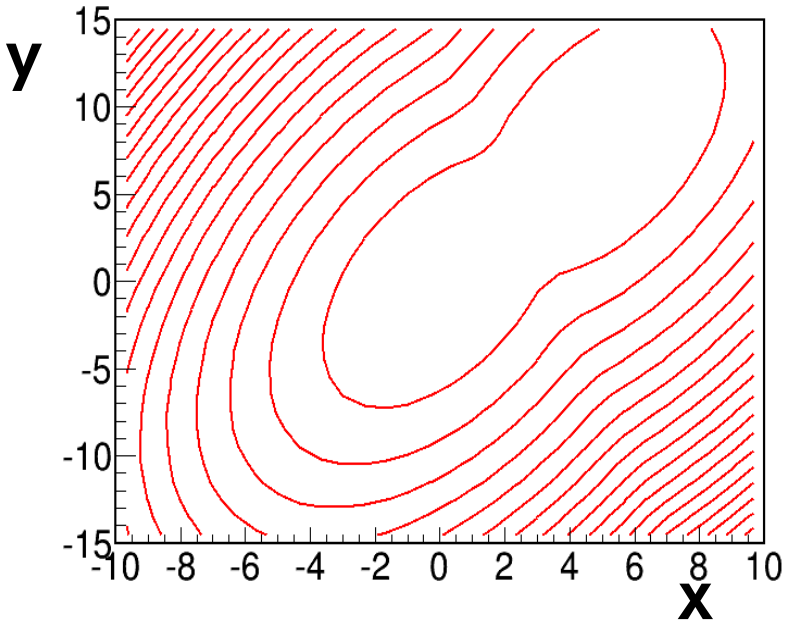
Continuous Variables: PDFs

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



Bin size → 0 :
Probability distribution function P(x)
→ High values ⇔ high chance to get a measurement here

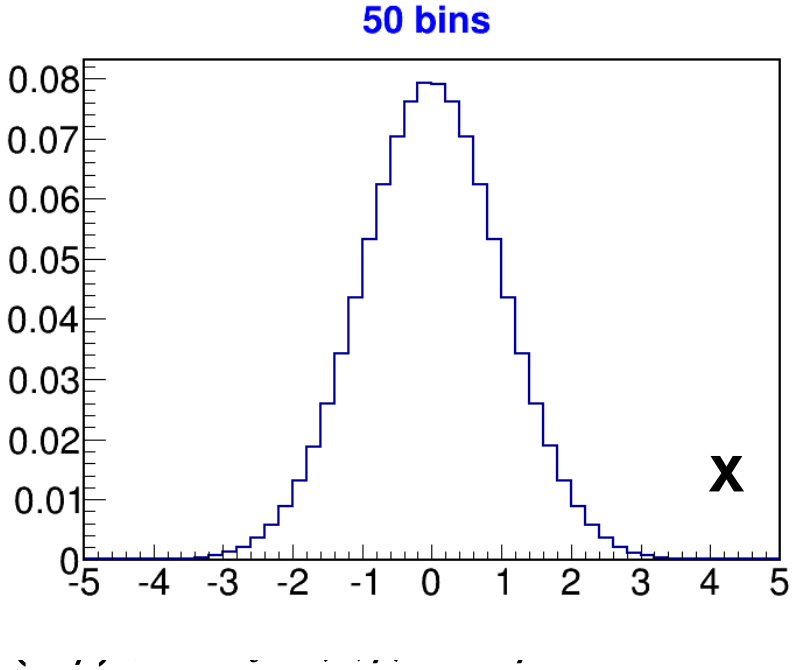
$P(x) > 0, \int P(x)\, dx = 1$



Contours: P(x,y)

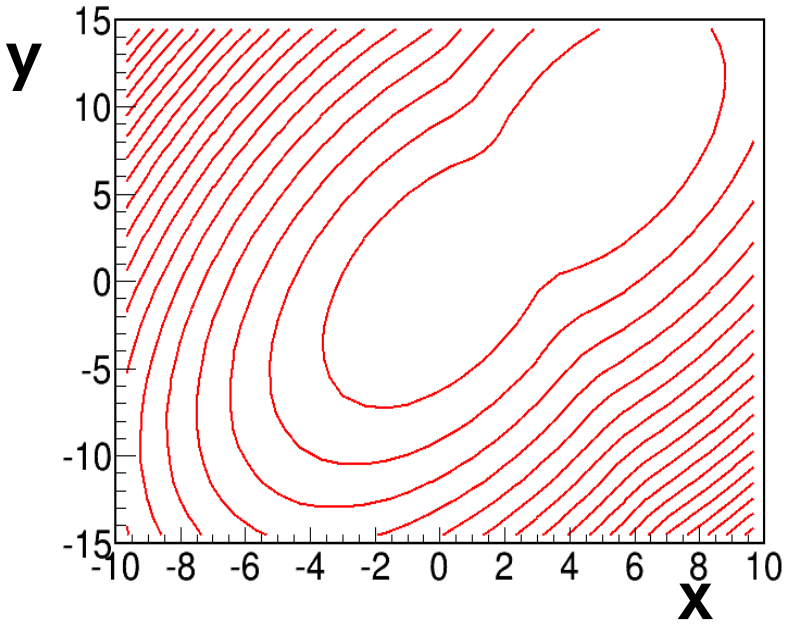
Continuous Variables: PDFs

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



Bin size → 0 :
Probability distribution function P(x)
→ High values ⇔ high chance to get a measurement here

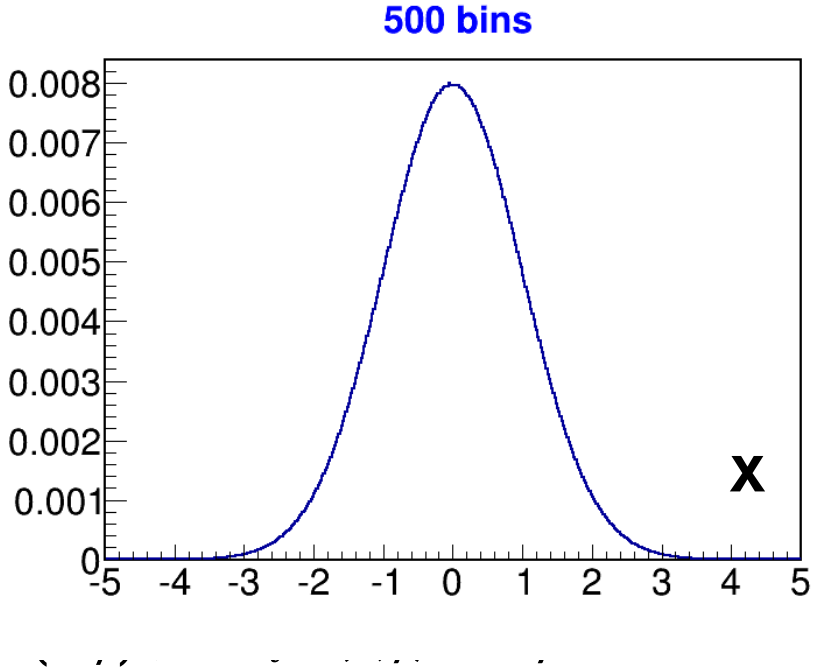
$P(x) > 0, \int P(x)\, dx = 1$



Contours: P(x,y)

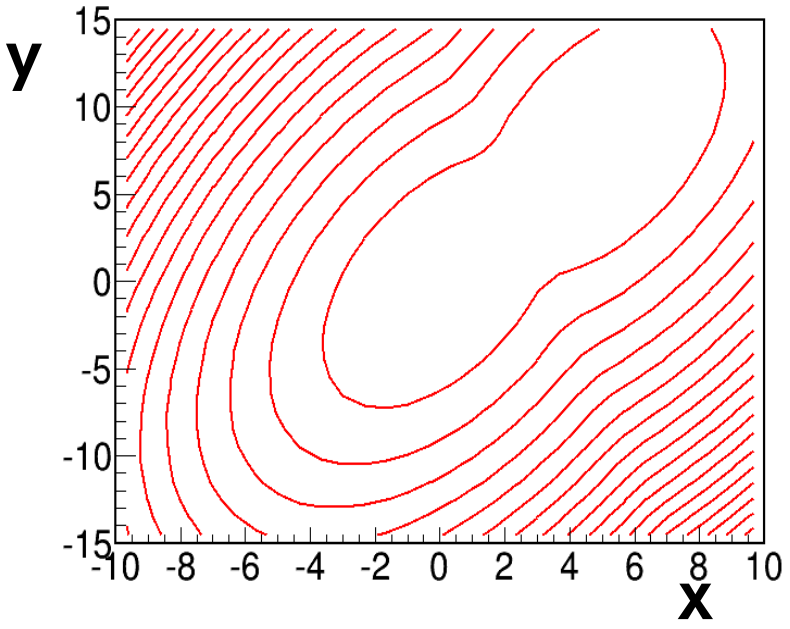
Continuous Variables: PDFs

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



Bin size → 0 :
Probability distribution function P(x)
→ High values ⇔ high chance to get a measurement here

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



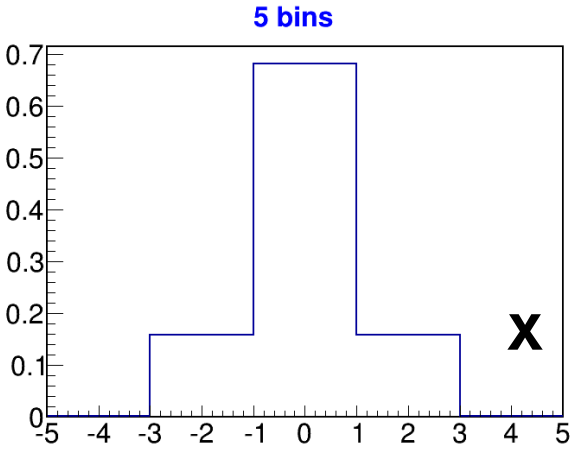
Contours: P(x,y)

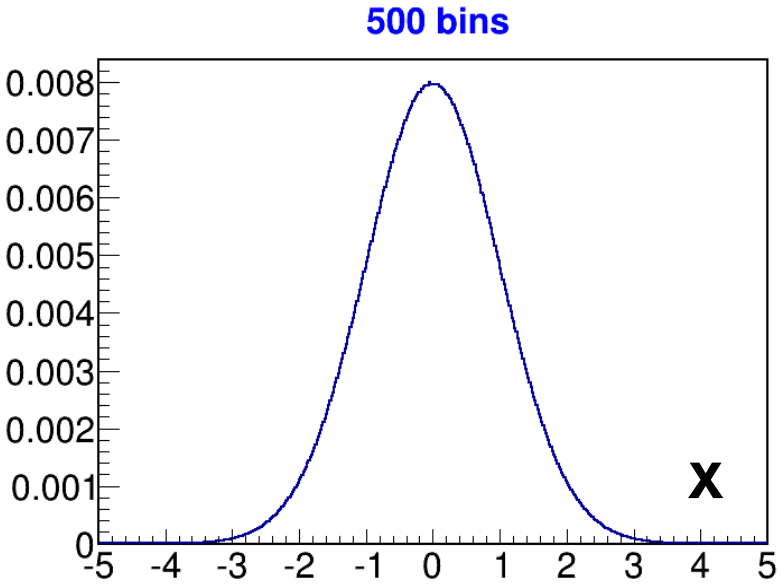
Random Variables

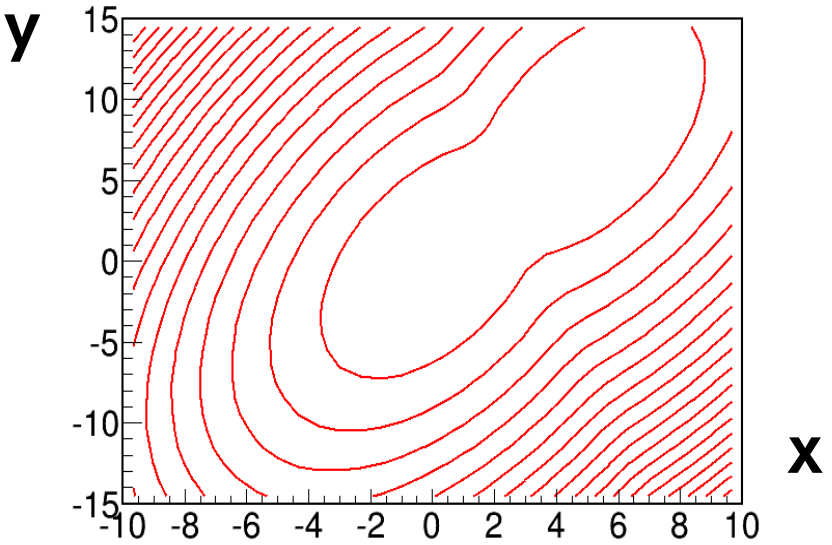
X, Y… are **Random Variables** (continuous or discrete), a.ka. **observables** :
→ X can take any value x, with probability **P(X=x).**

→ 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**

$$\langle X \rangle = \sum x_i \, P_i$$ or

→ Property of the **PDF**

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

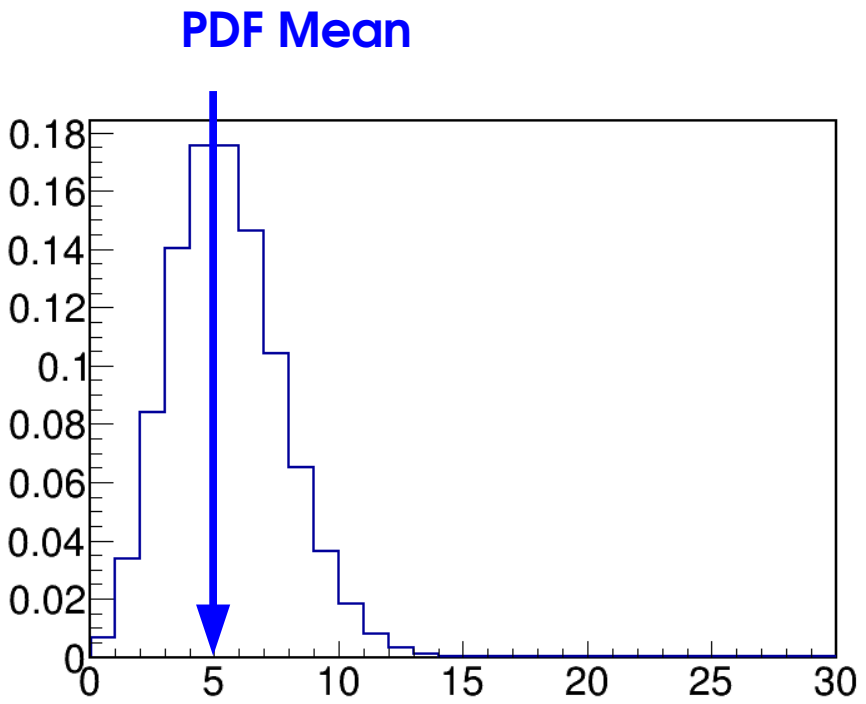
For measurements $x_1 \ldots x_n$,

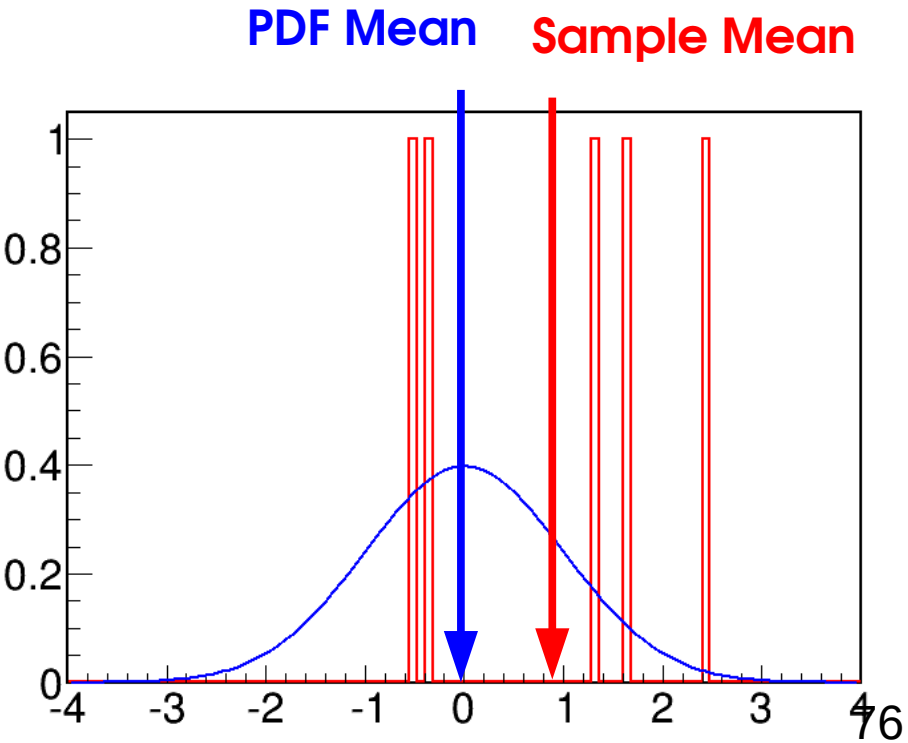
then can compute the **Sample mean**:

→ Property of the **sample**

→ approximates the PDF mean.

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





PDF Properties: (Co)variance

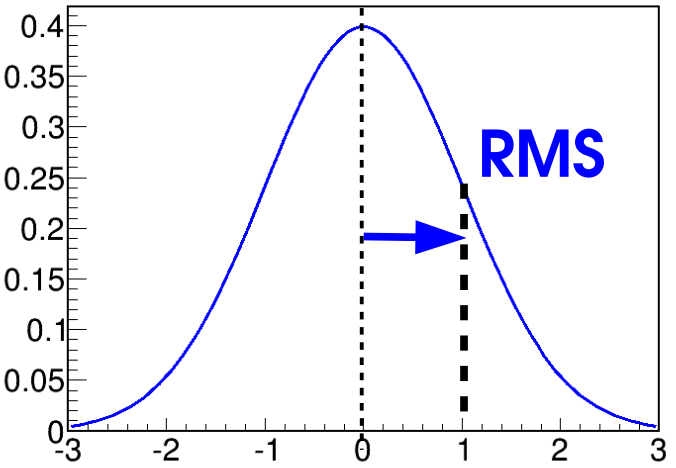
Variance of X:

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

→ Average square of deviation from mean

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

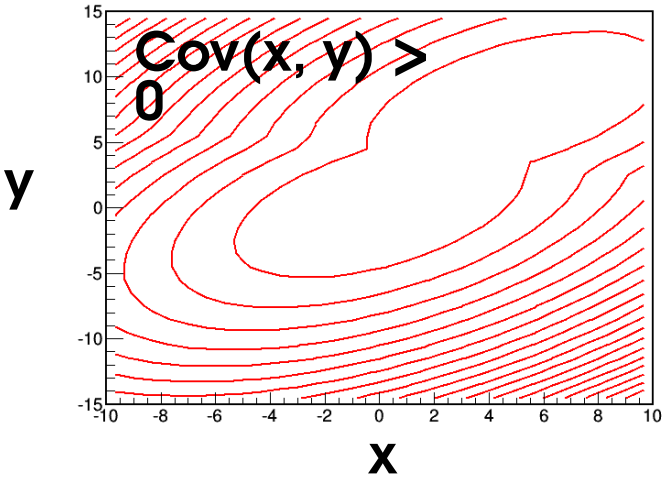
Can be approximated by **sample variance**:



Covariance of X and Y:

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

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



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

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

PDF Properties: (Co)variance

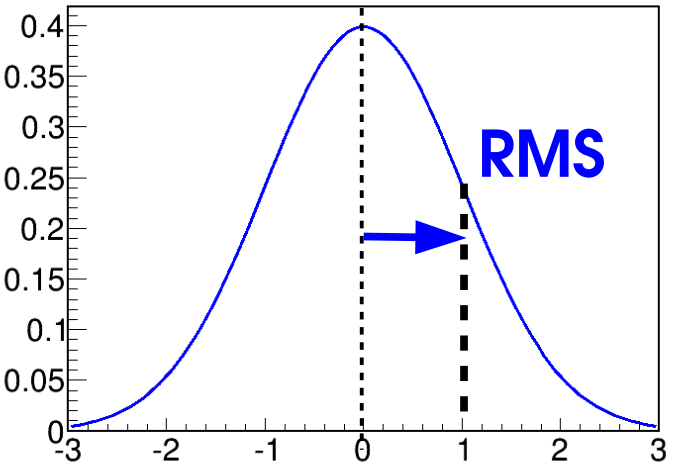
Variance of X:

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

→ Average square of deviation from mean

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

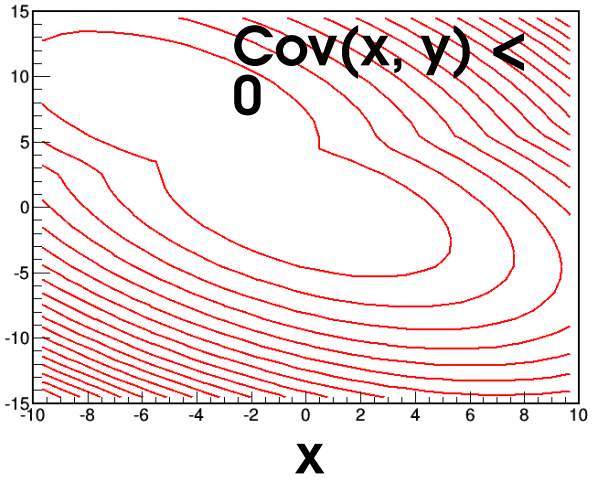
Can be approximated by **sample variance**:



Covariance of X and Y:

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

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



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

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

PDF Properties: (Co)variance

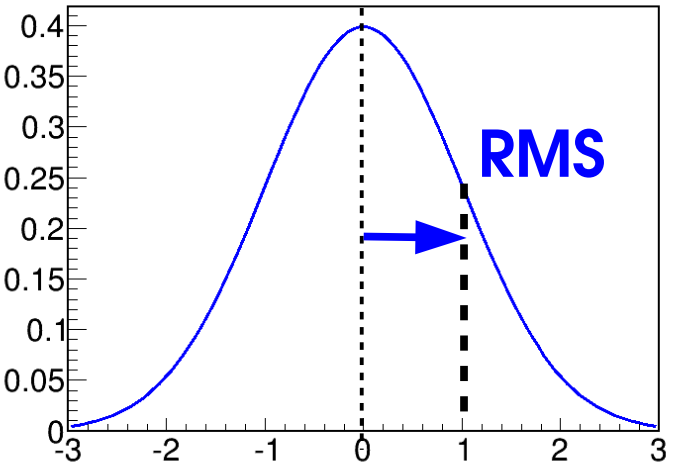
Variance of X:

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

→ Average square of deviation from mean

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

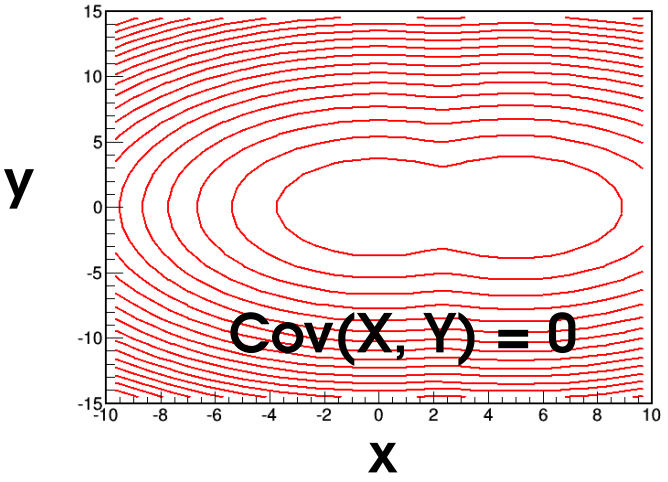
Can be approximated by **sample variance**:



Covariance of X and Y:

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

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

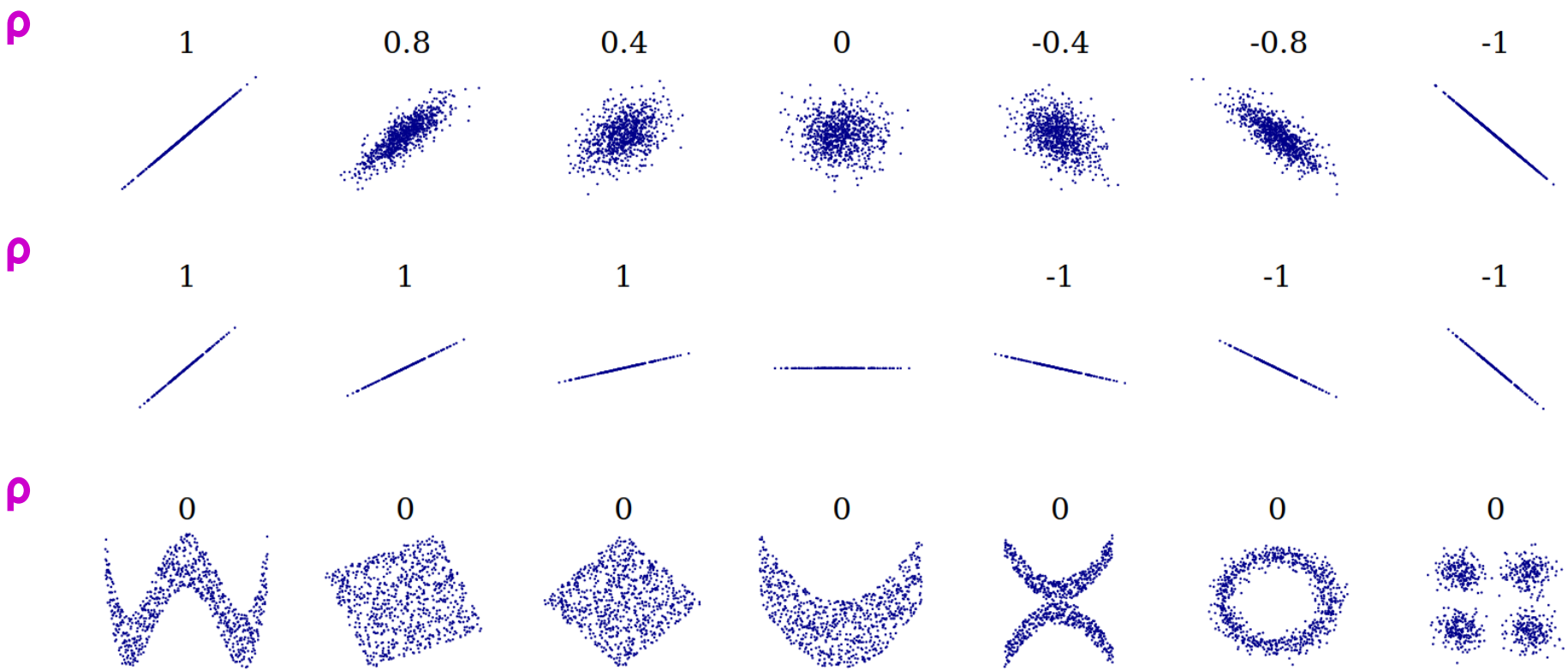


$$\mathbf{Cov}(X, Y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$$

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

“Linear” vs. “non-linear” correlations

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



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

Source: Wikipedia

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

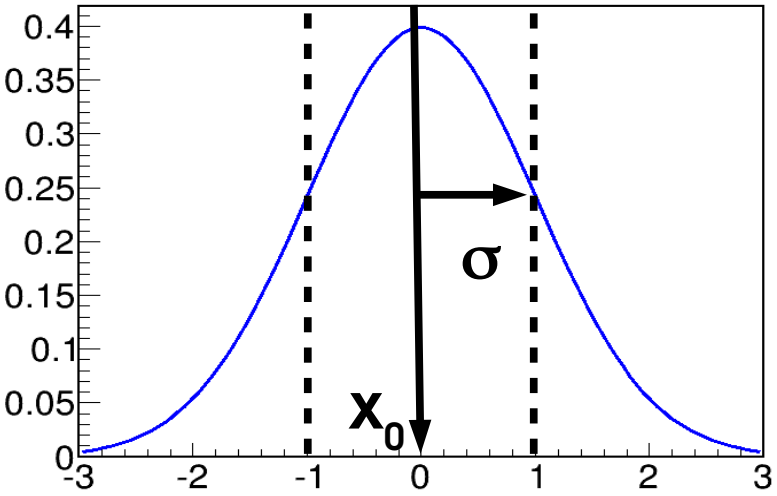
Gaussian PDF

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

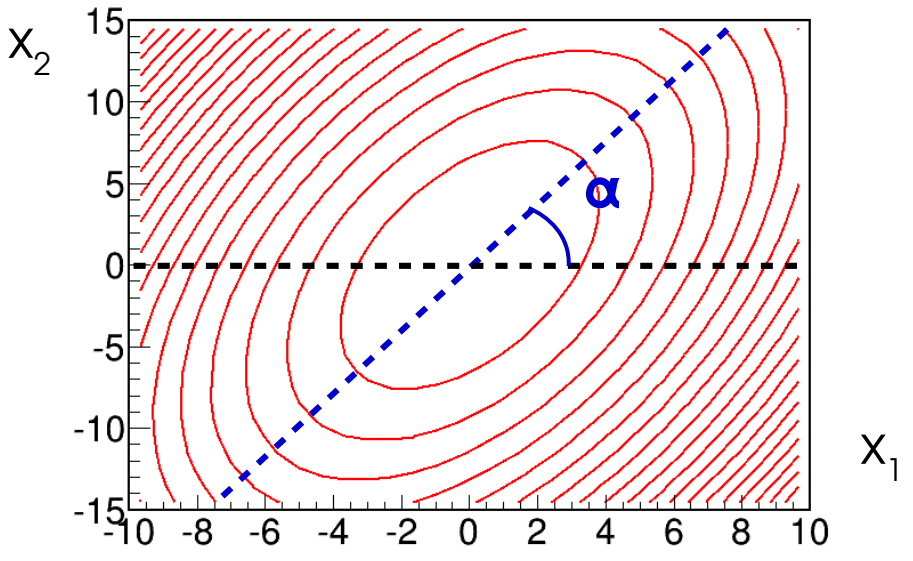
→ Covariance matrix :

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

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

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

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



Gaussian Quantiles

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

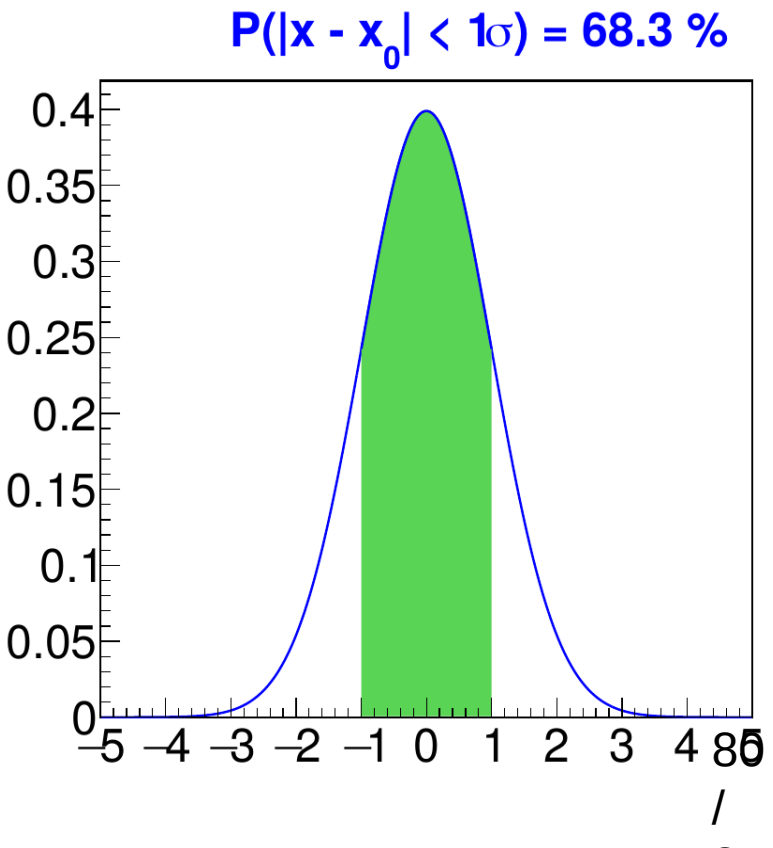
$G(x;x_0,\sigma)$ depends only on $z \sim G(z; 0,1)$

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

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

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

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



Gaussian Quantiles

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

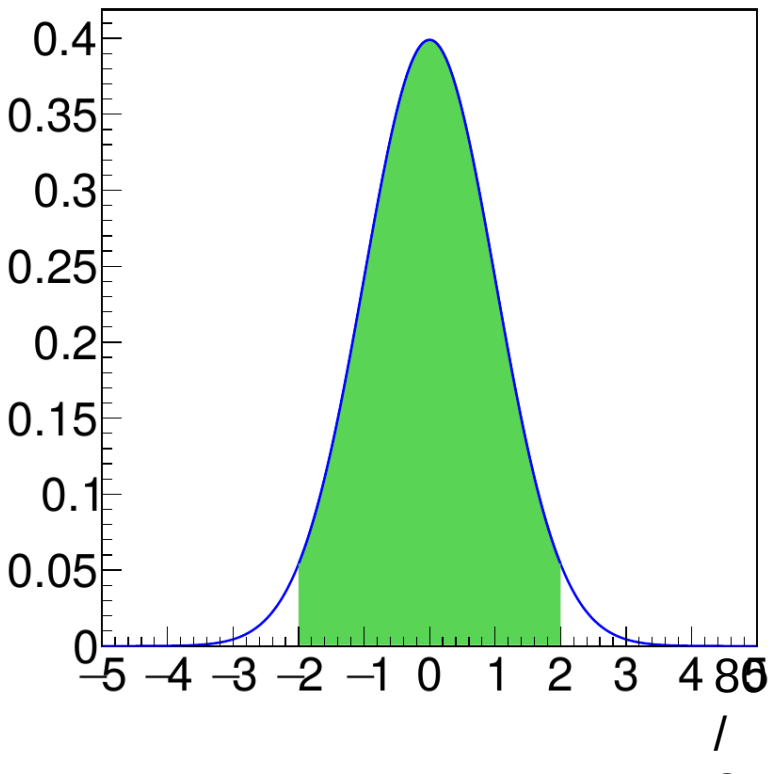
$G(x;x_0,\sigma)$ depends only on $z \sim G(z; 0,1)$

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

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

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

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



Gaussian Quantiles

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

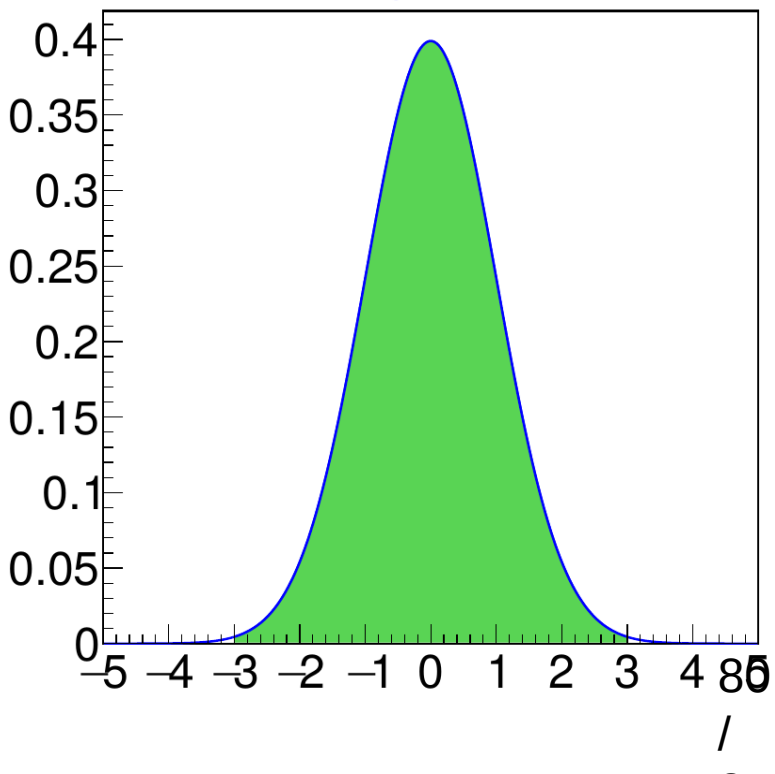
$G(x;x_0,\sigma)$ depends only on $z \sim G(z; 0,1)$

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

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

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

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



Central Limit Theorem

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

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

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

- **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

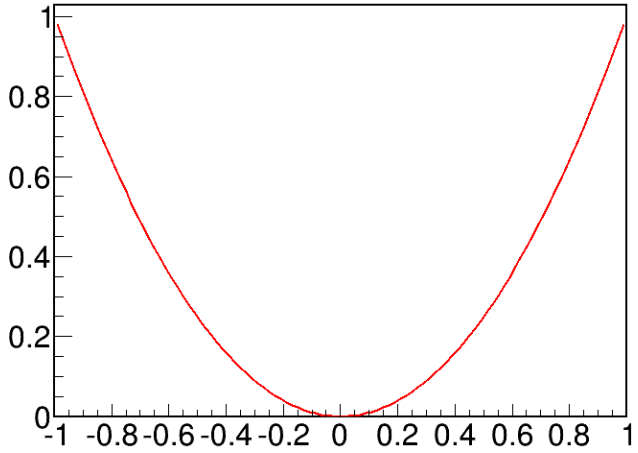
Another version:

Mean scales like n, but RMS only like √n

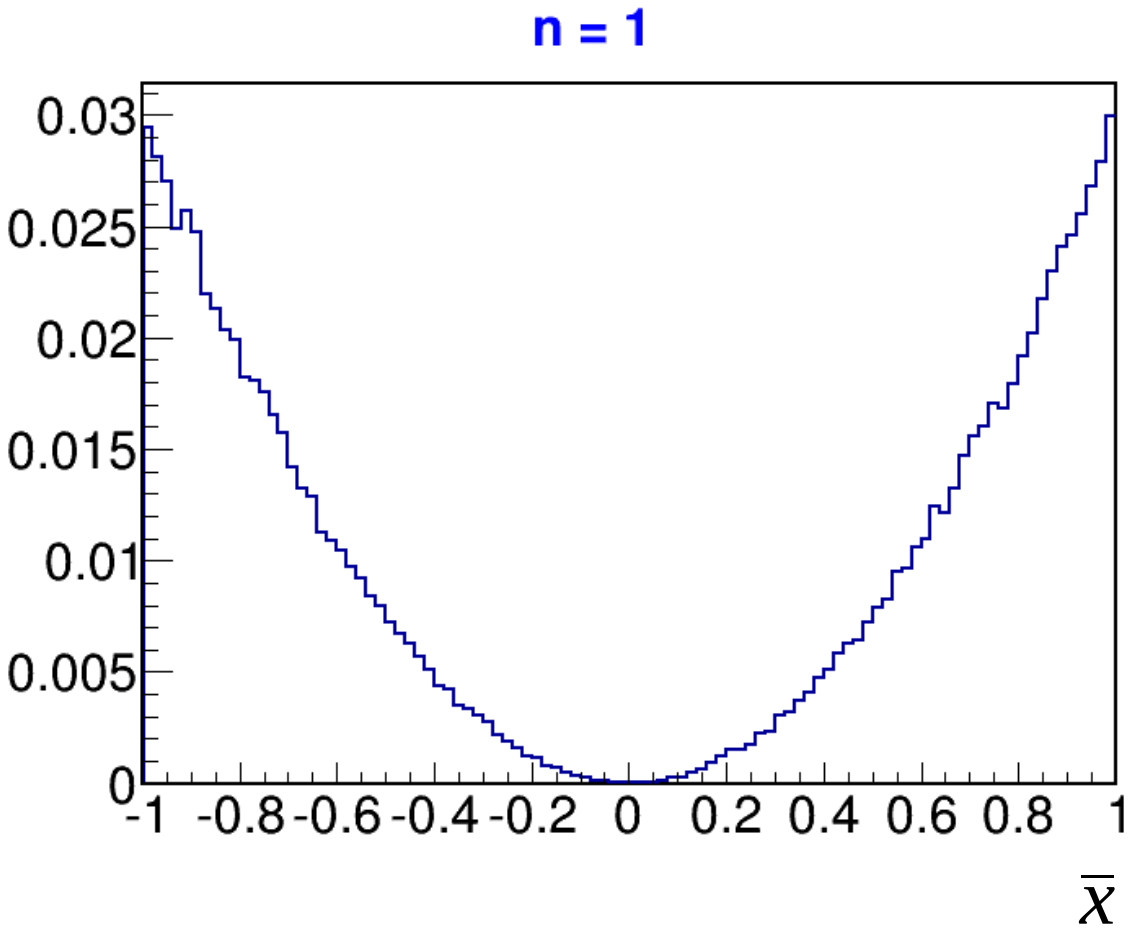
$$\sum_{i=1}^{n} x_i \overset{n\to\infty}{\sim} G\left(n\langle X\rangle, \sqrt{n}\,\sigma_X\right)$$

Central Limit Theorem in action

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



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

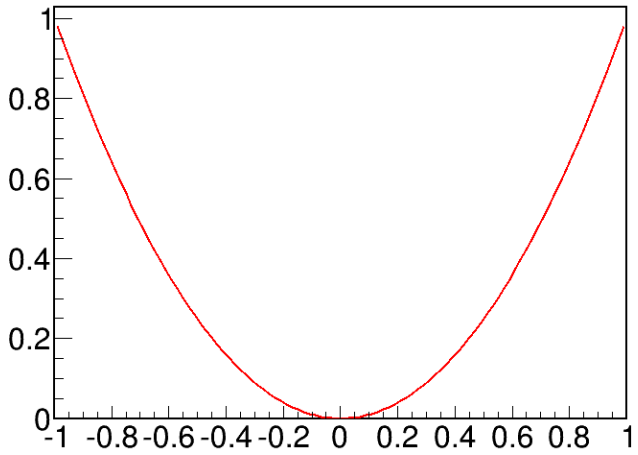


Distribution becomes Gaussian, although very non-Gaussian originally

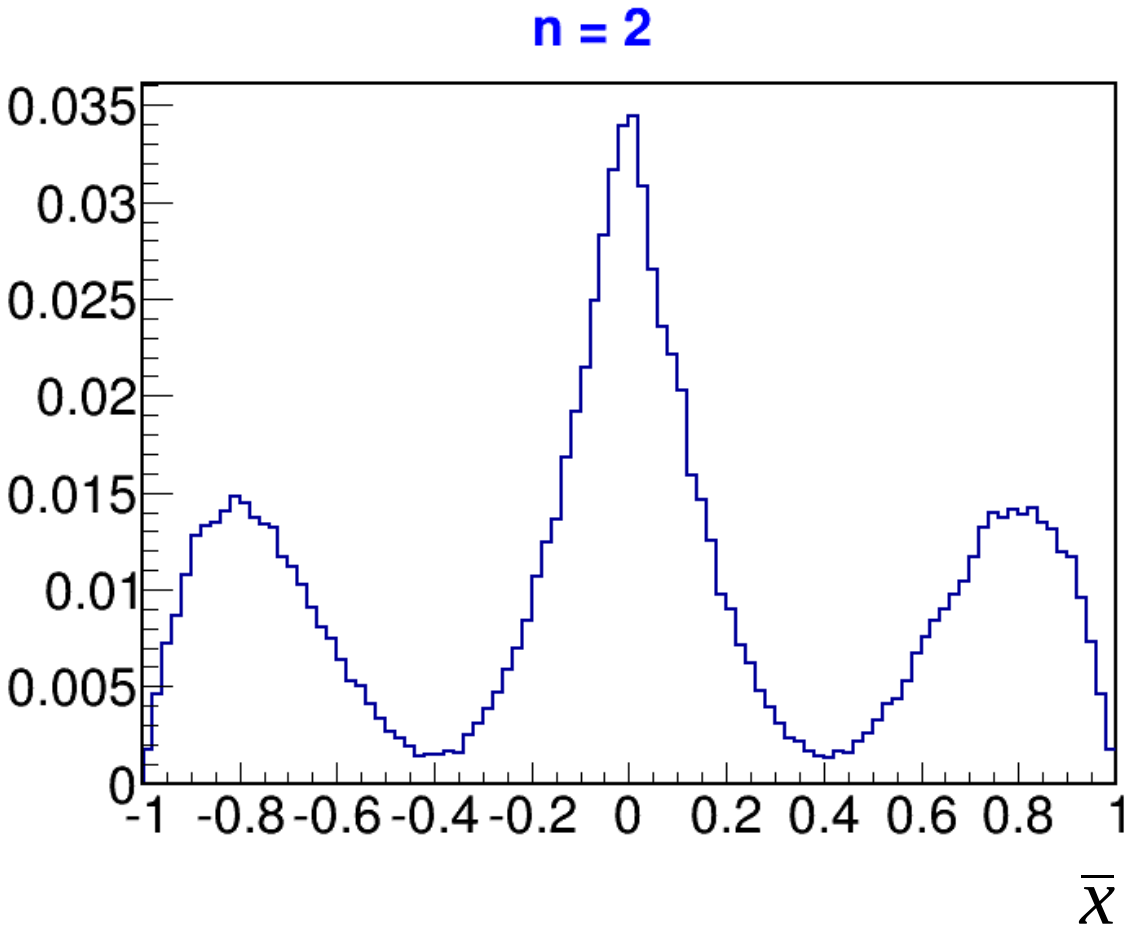
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

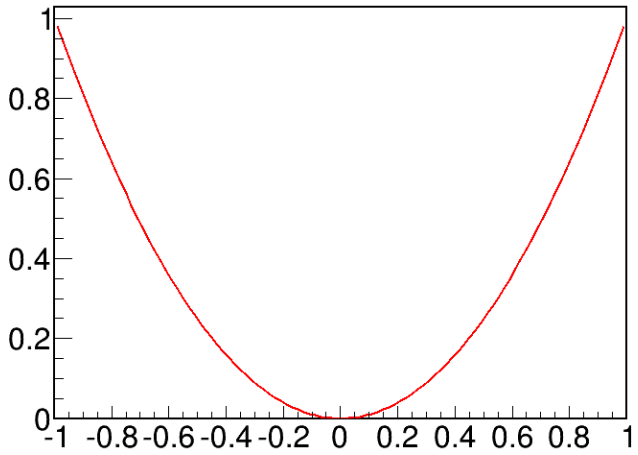


Distribution becomes Gaussian, although very non-Gaussian originally

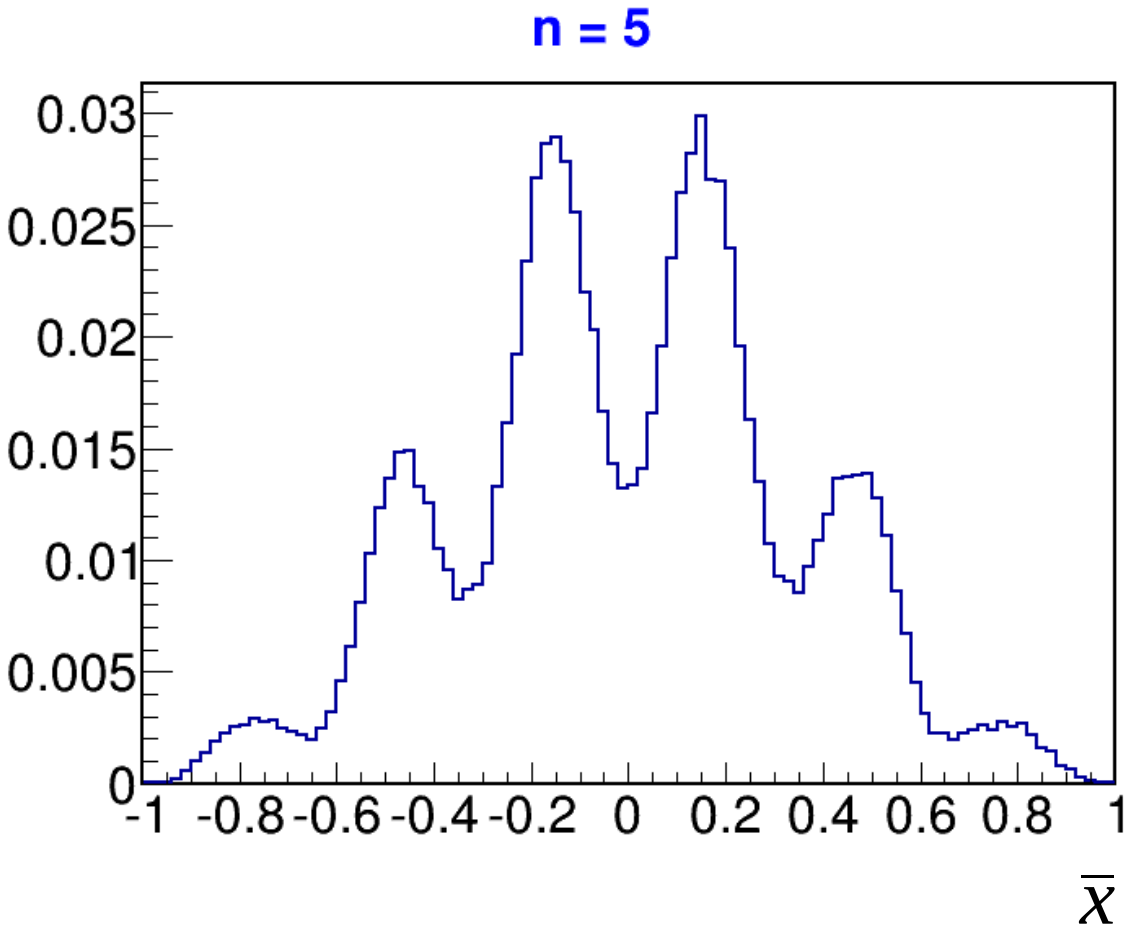
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

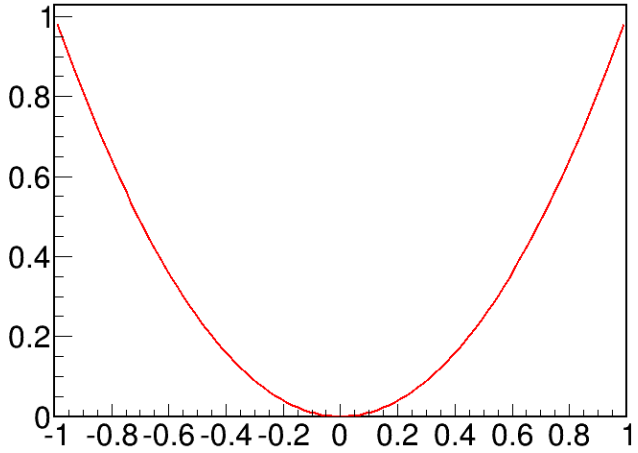


Distribution becomes Gaussian, although very non-Gaussian originally

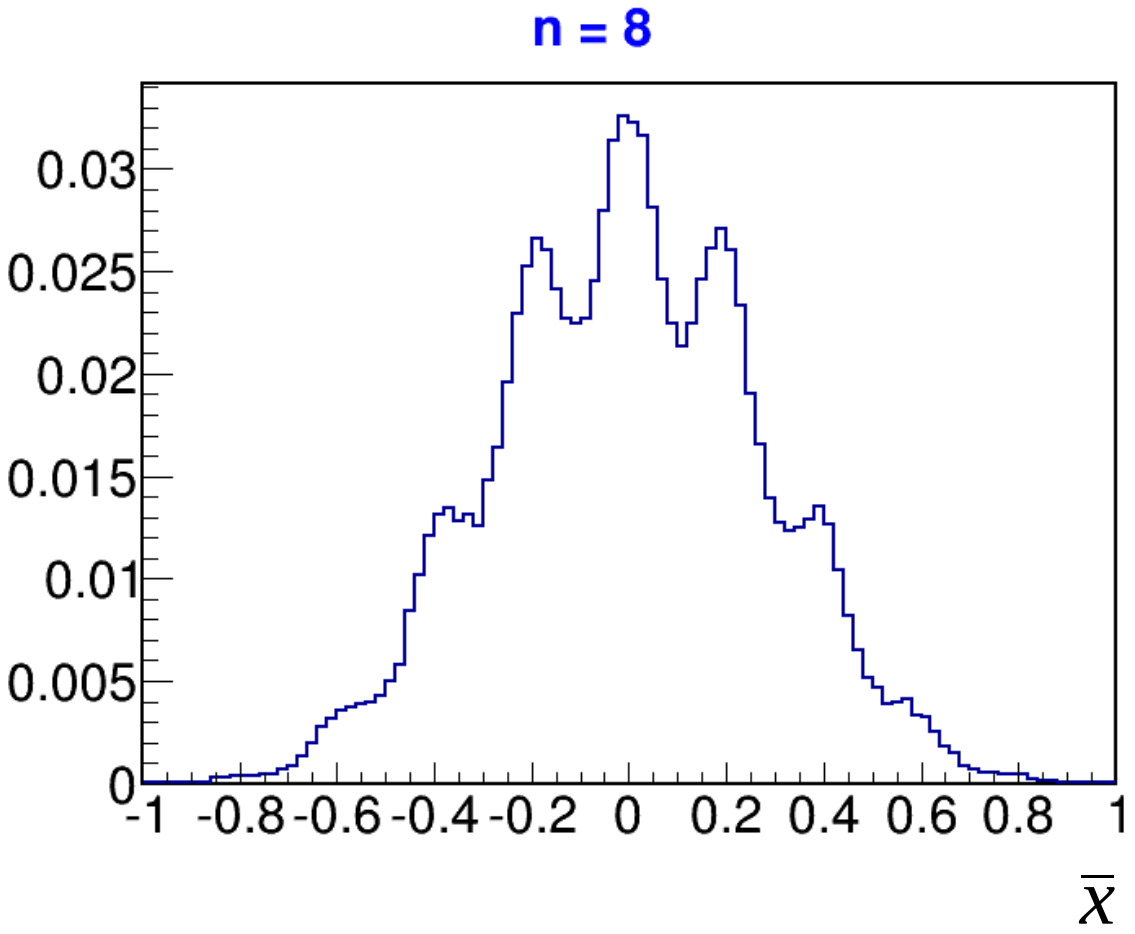
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

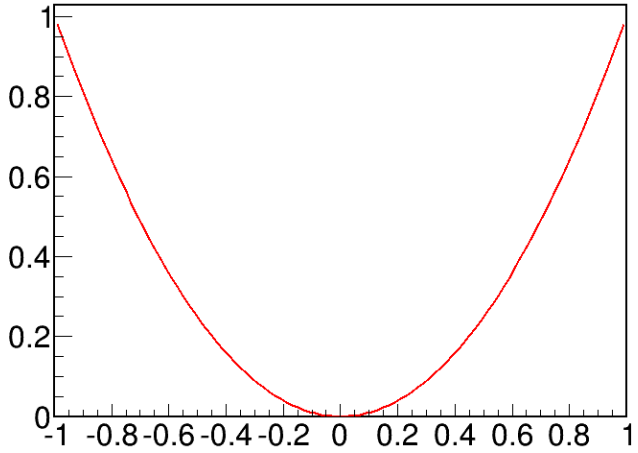


Distribution becomes Gaussian, although very non-Gaussian originally

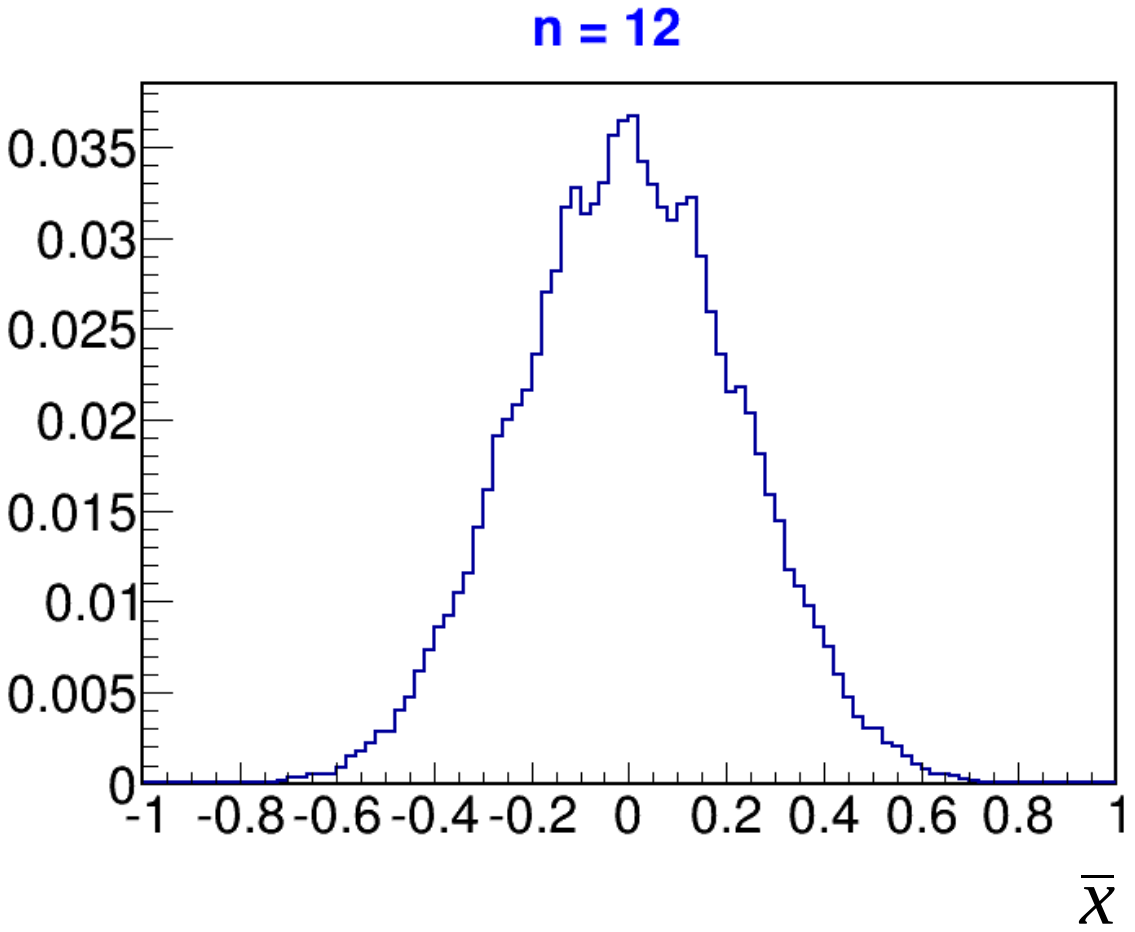
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

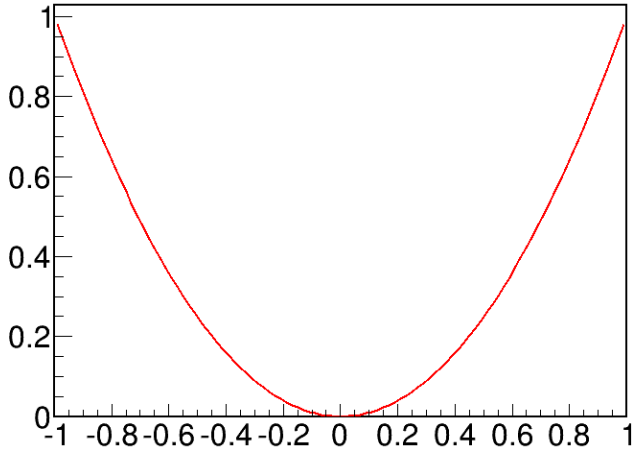


Distribution becomes Gaussian, although very non-Gaussian originally

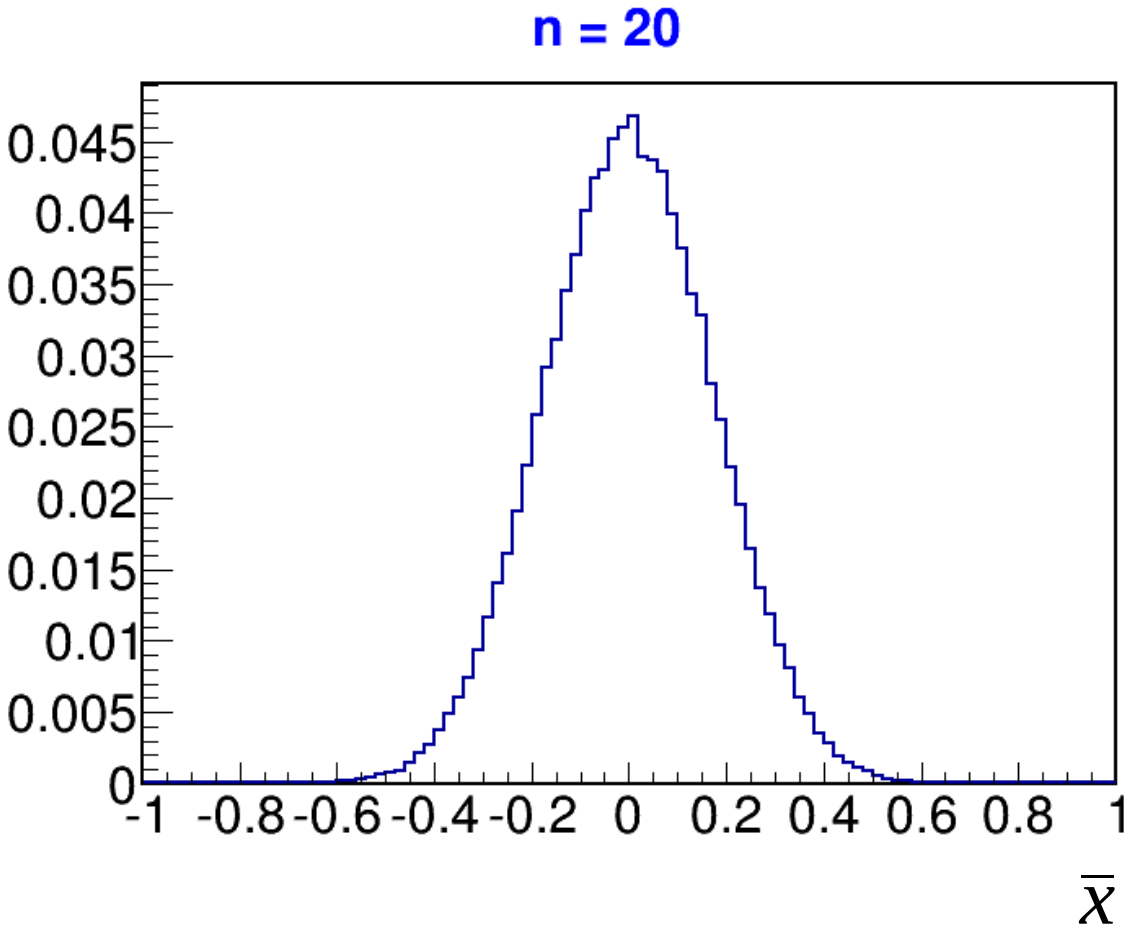
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

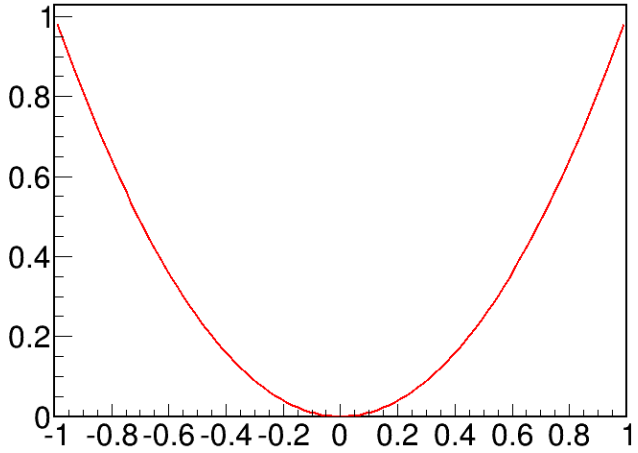


Distribution becomes Gaussian, although very non-Gaussian originally

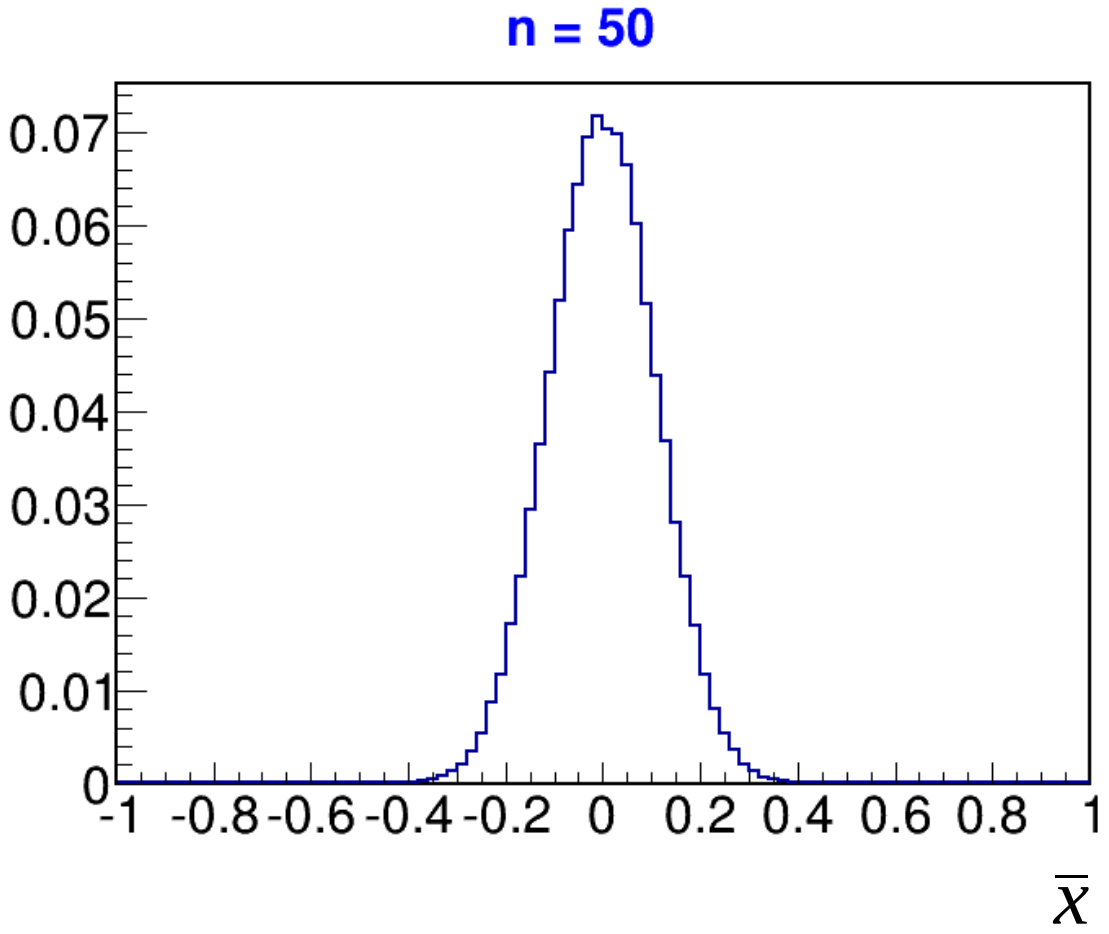
Distribution becomes narrower as expected (as 1/√n)

Central Limit Theorem in action

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



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

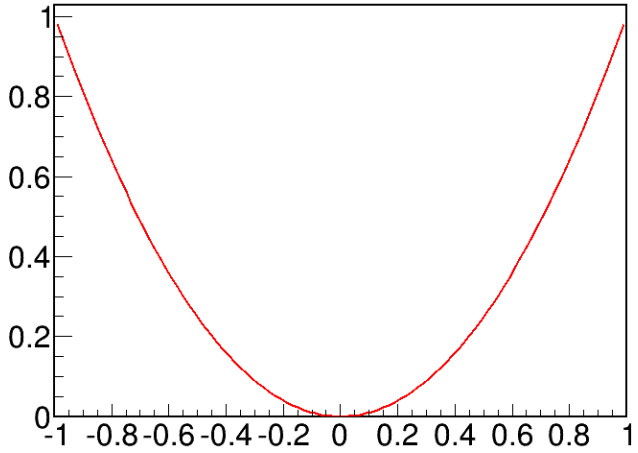


Distribution becomes Gaussian, although very non-Gaussian originally

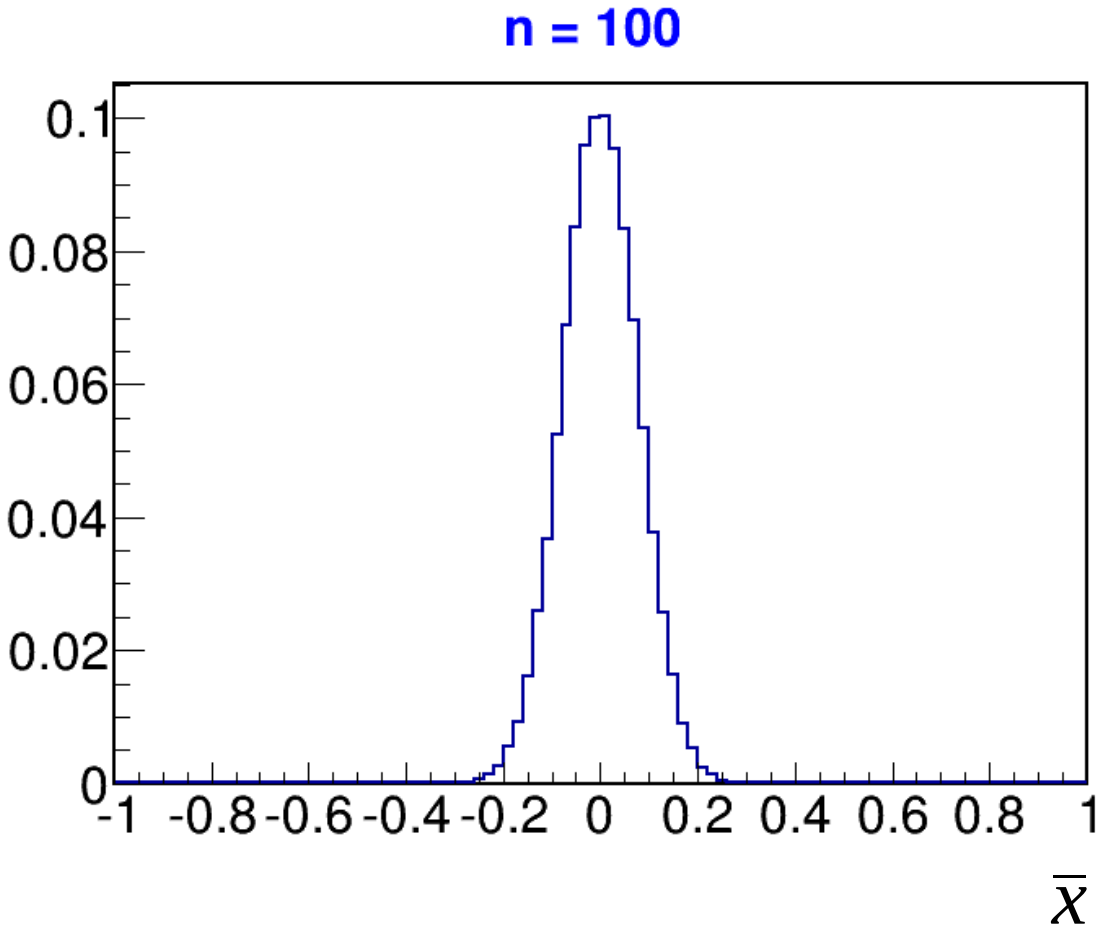
Distribution becomes narrower as expected (as 1/√n)

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)

Chi-squared

Multiple Independent Gaussian variables x_i: Define

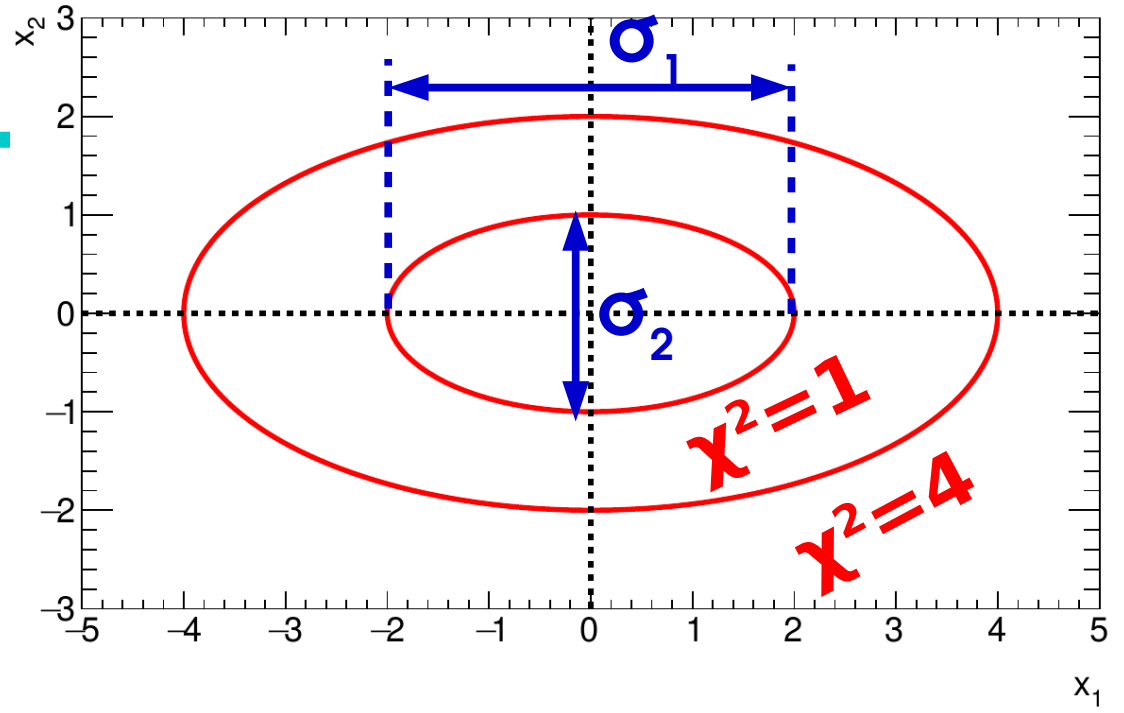
$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - x_i^0}{\sigma_i} \right)^2$$

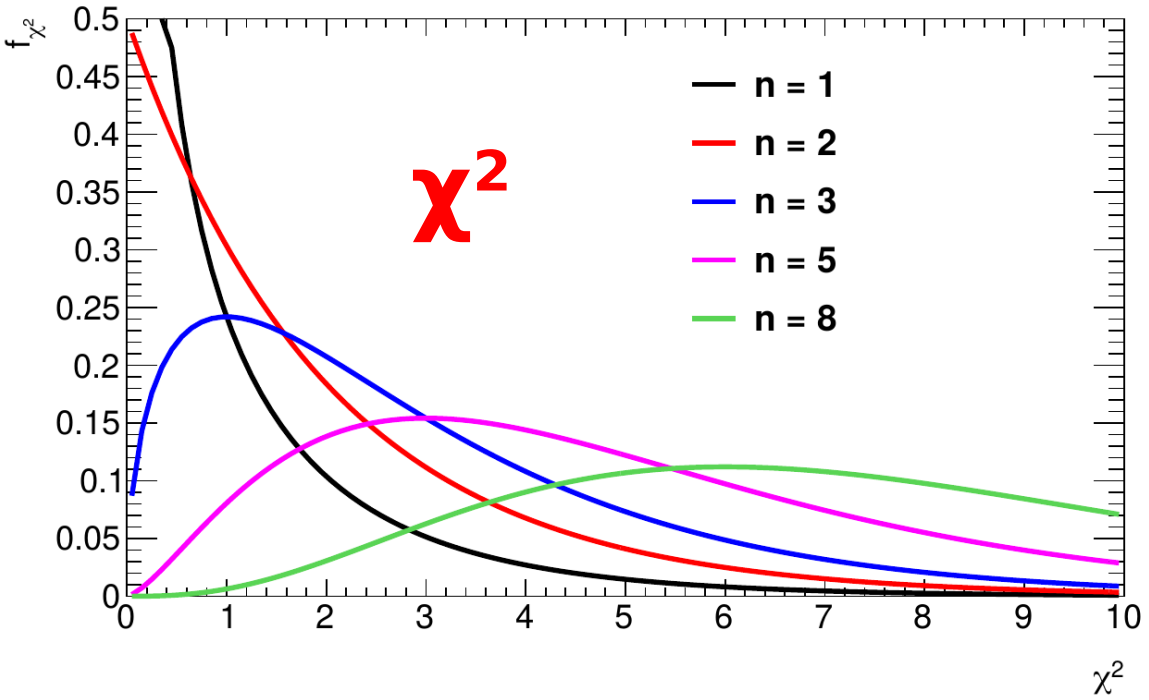
Measures global distance from reference point (x_1^0 .... x_n^0)

Distribution depends on n :

Rule of thumb:

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





Chi-squared

Multiple Independent Gaussian variables x_i: Define

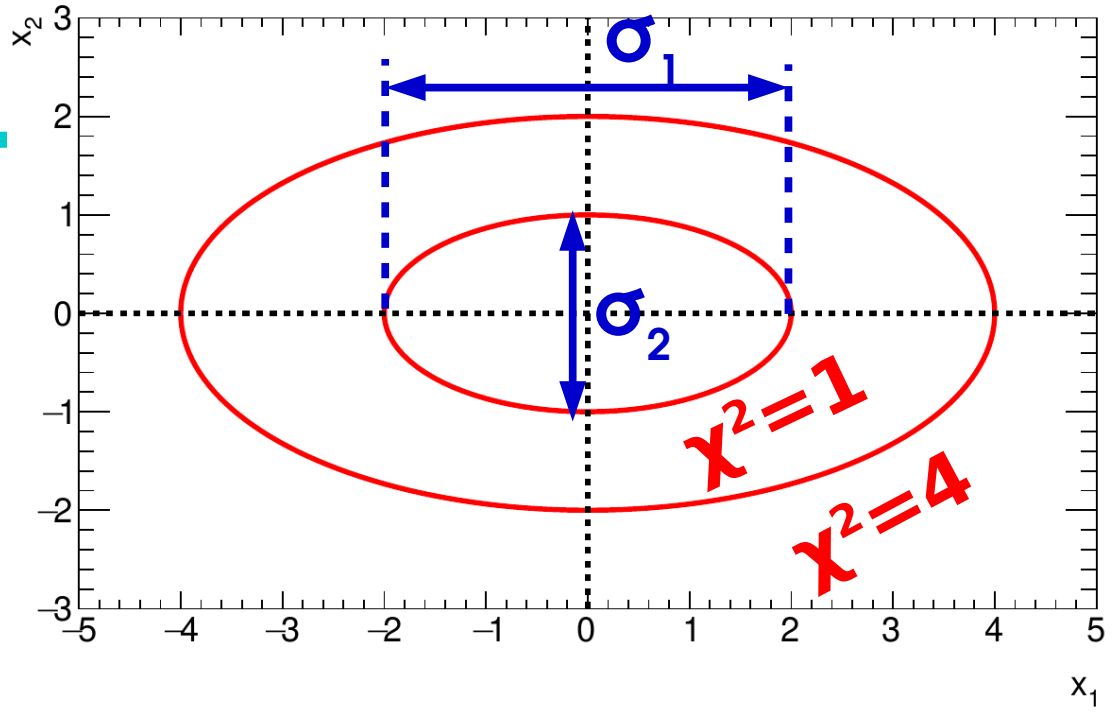
$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - x_i^0}{\sigma_i} \right)^2$$

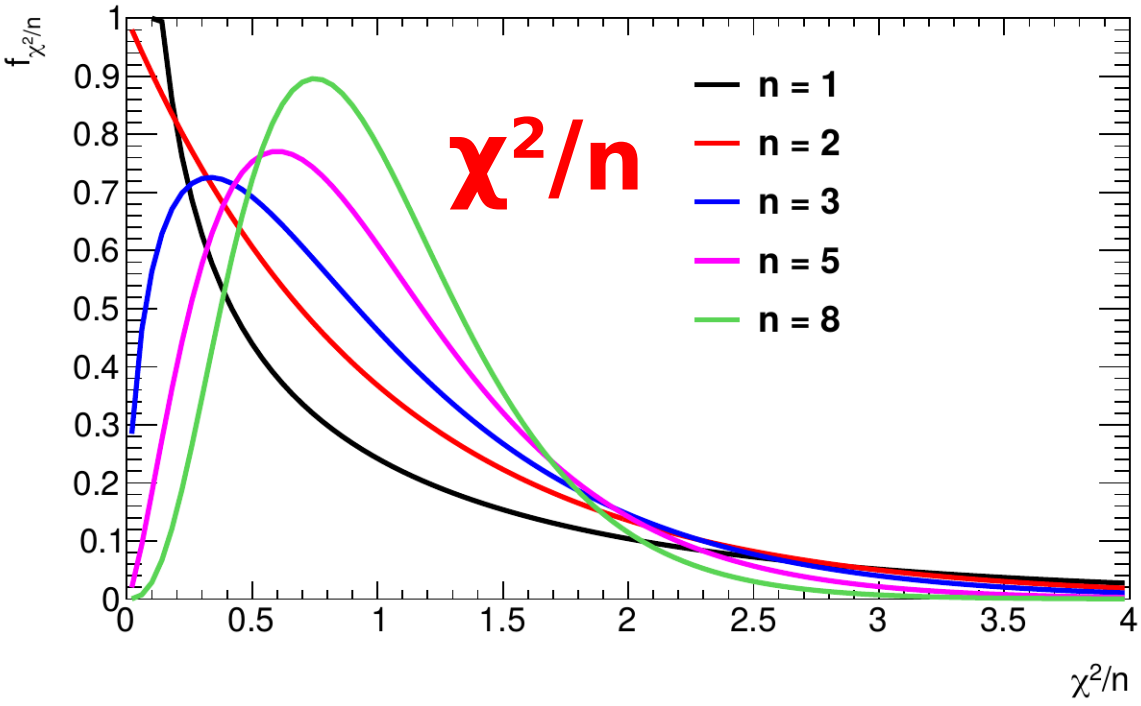
Measures global distance from reference point (x_1^0 .... x_n^0)

Distribution depends on n :

Rule of thumb:

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

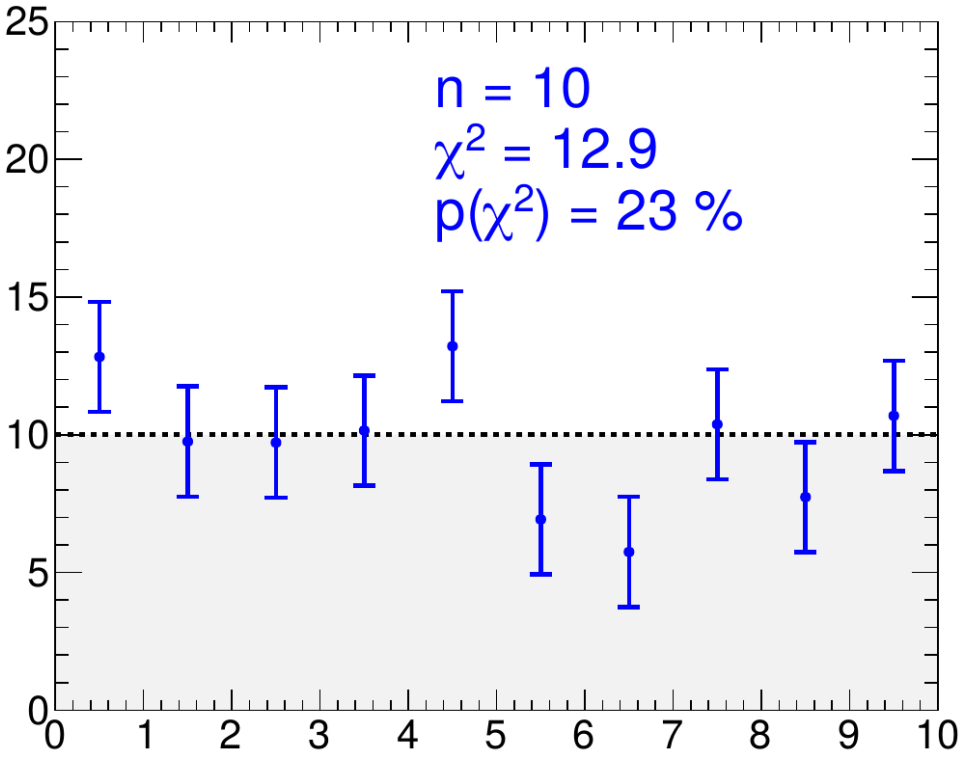




Histogram Chi-squared

Histogram 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)



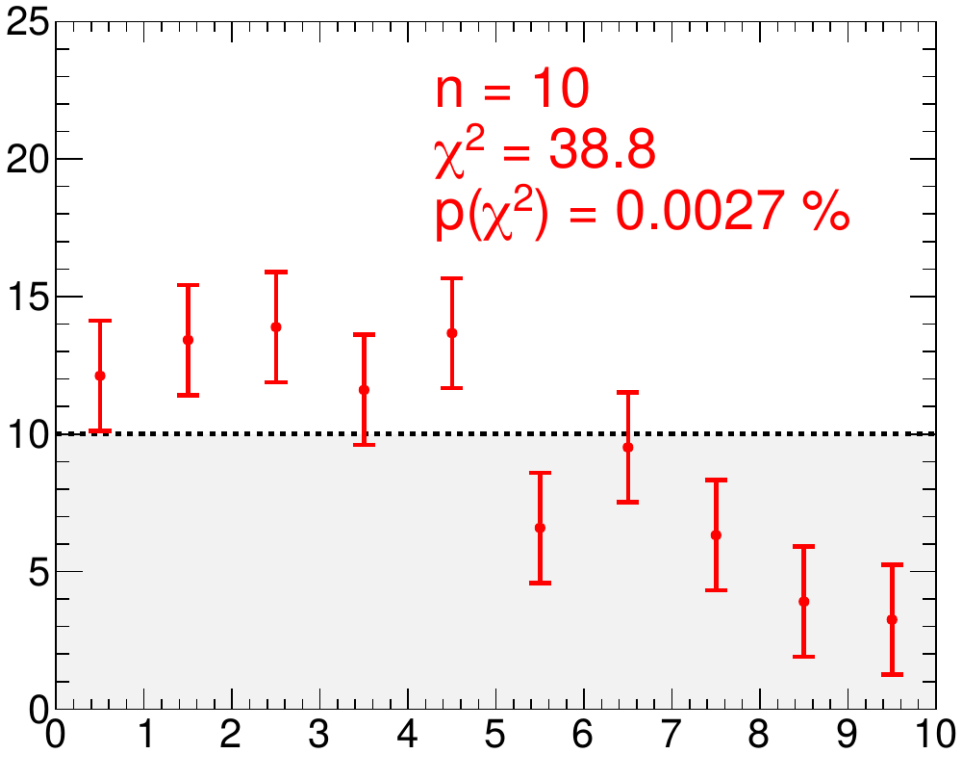
BLUE histogram vs. flat reference

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

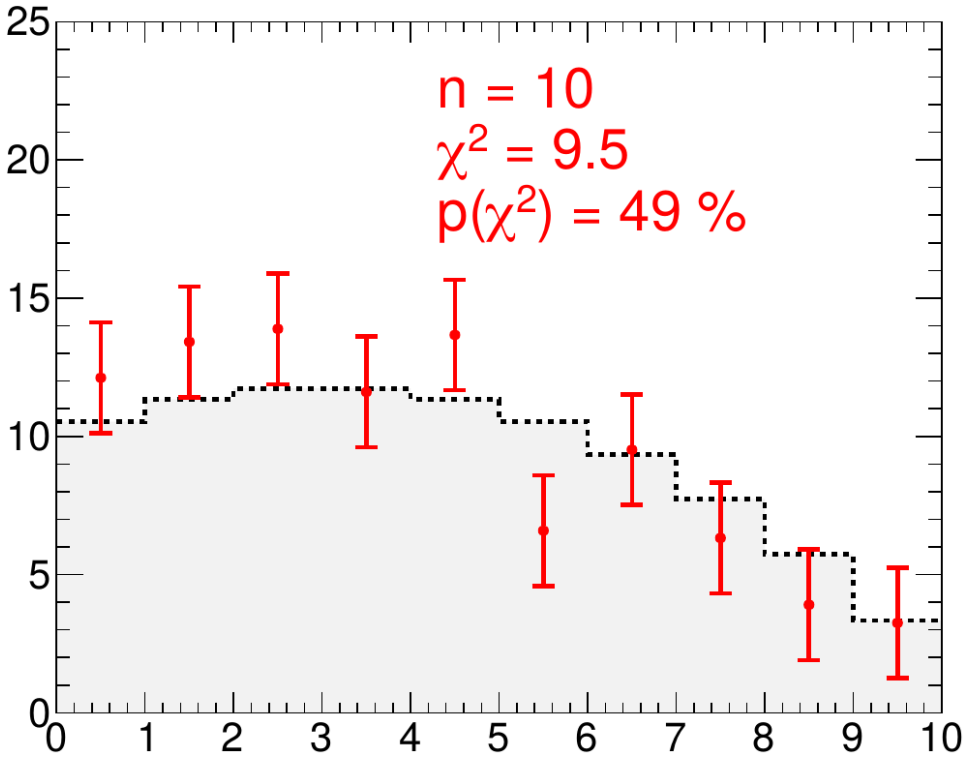
RED histogram vs. flat reference

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

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

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%

Error Bars

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

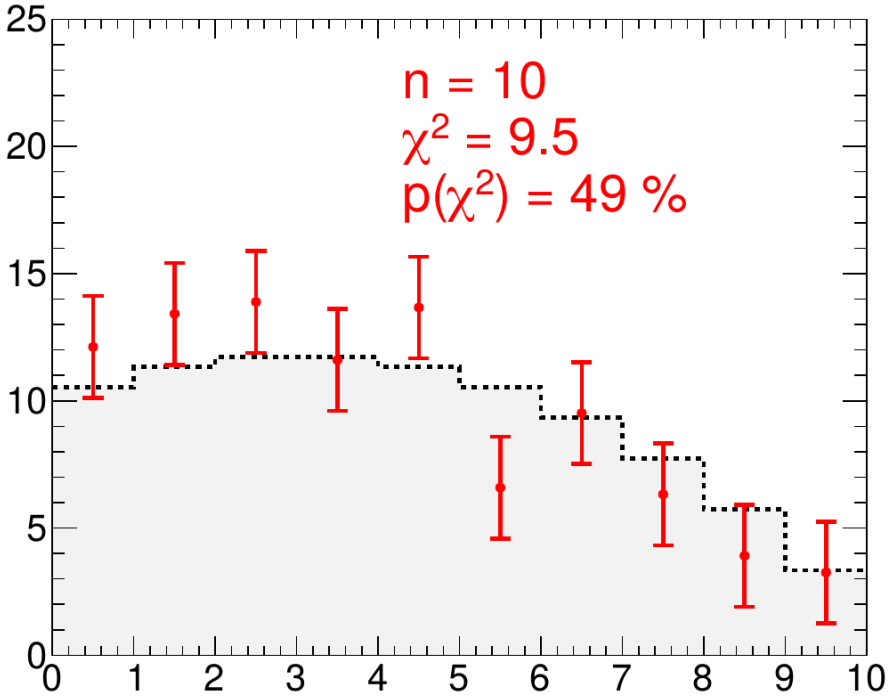
→ ***Bin central value*** ~ *mean of the bin PDF*

→ ***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**.