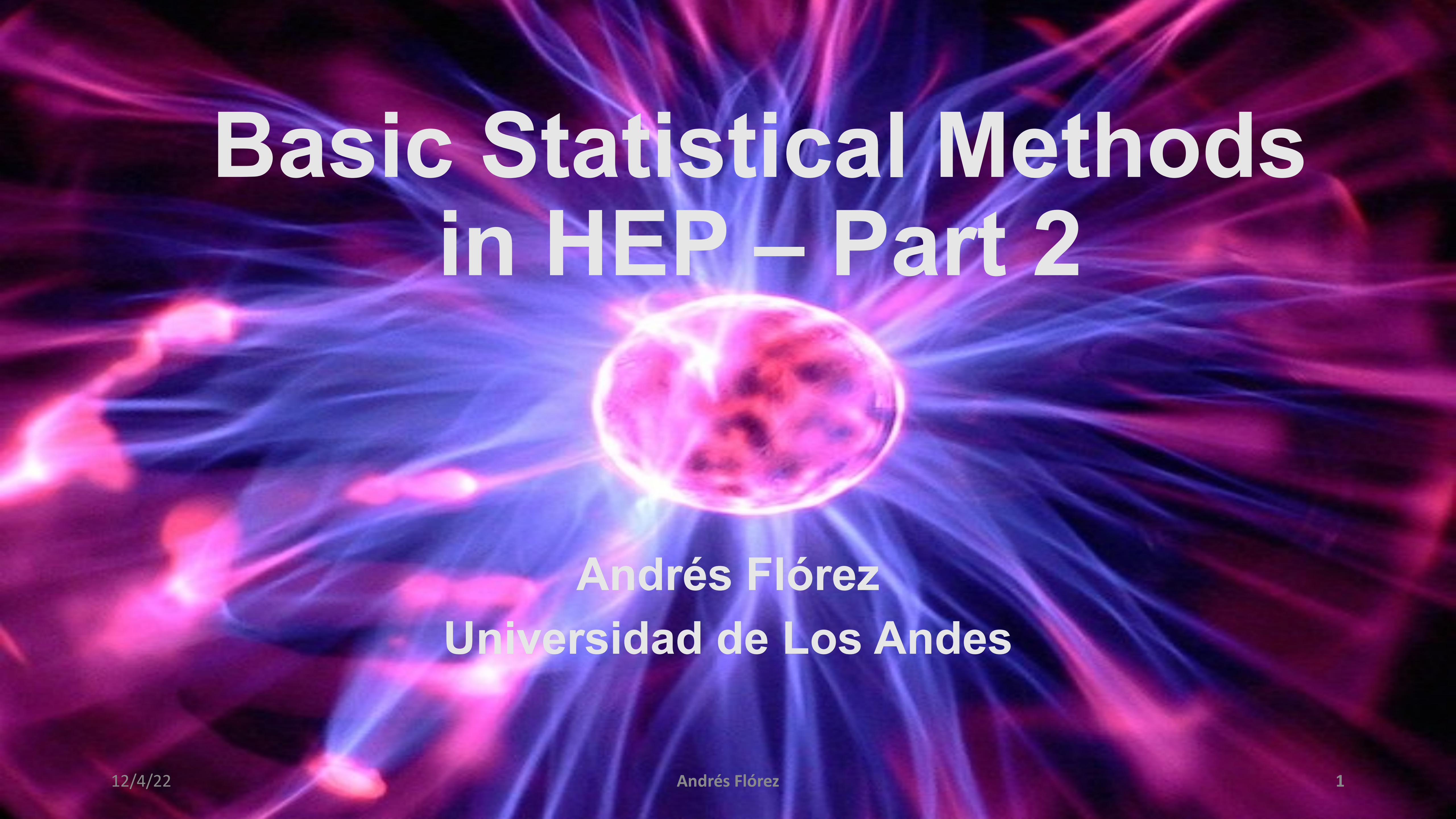


Basic Statistical Methods in HEP – Part 2

Andrés Flórez
Universidad de Los Andes

Bayesian Statistics: In a

- Note that when the efficiency is 1 or 0 $\sigma_k = \sqrt{n\epsilon\,(1-\epsilon)}$ becomes 0. This is problematic!
- In order to deal with these cases or cases where the efficiency is on the boundary of 0 or 1, we generally use Bayesian statistics.
- **Bayesian statistics considers priors, understood as believes or educated hypotheses based upon evidences.**
- This is approach is contrary with another statistical inference, known as frequentist statistics, where probabilities are understood as a frequency of random events, that emerge after many repeated trials of the test.

Bayesian Statistics: In a

- To determine the probability that ϵ is the true efficiency given the measurement of k positive results with respect to n measurements:

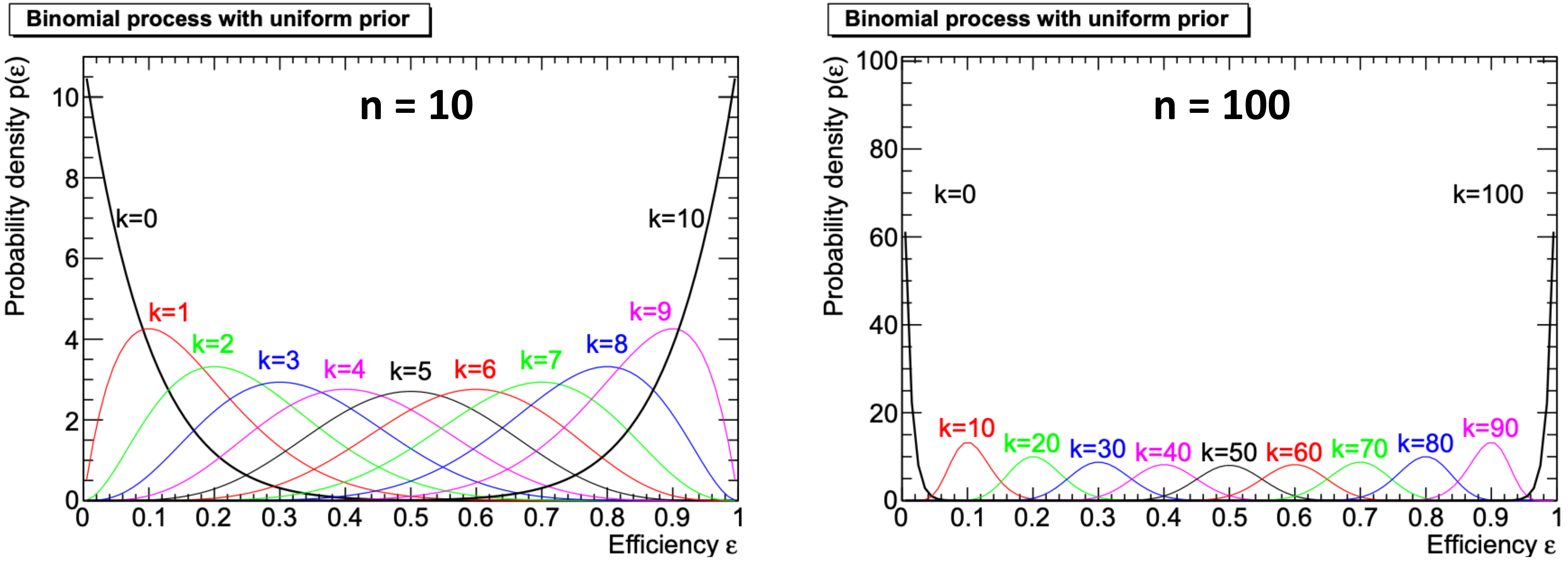
$$P(\epsilon; k, n) = \frac{P(k; \epsilon, n)P(\epsilon; n)}{C}$$

- Note that $P(k; \epsilon, n)$ is the binomial probability we just explained.
- $P(\epsilon; n)$ is the probability to measure ϵ for a given value of n. Before any measurement we assume that $P(\epsilon; n)$ is uniform in the range between 0 and 1, as ϵ can also take any value between 0 and 1.
- The constant is $C = \frac{1}{n+1}$ → Please check Eq. 10 in this paper (Ullrich and Xu).

Bayesian Statistics: In a

- Putting together all the parts:

$$P(\epsilon; k, n) = \frac{(n+1)!}{k!\,(n-k)!}\epsilon^{k}(1-\epsilon)^{n-k}$$



Figures from Diego Casadei´s paper (CLICK)

Bayesian Statistics: In a

- Using the probability, we can calculate the moments.
- For the efficiency, we find that the mean is:

$$\bar{\epsilon} = \frac{k+1}{n+2}$$

- For the variance (proof in Ullrich and Xu paper):

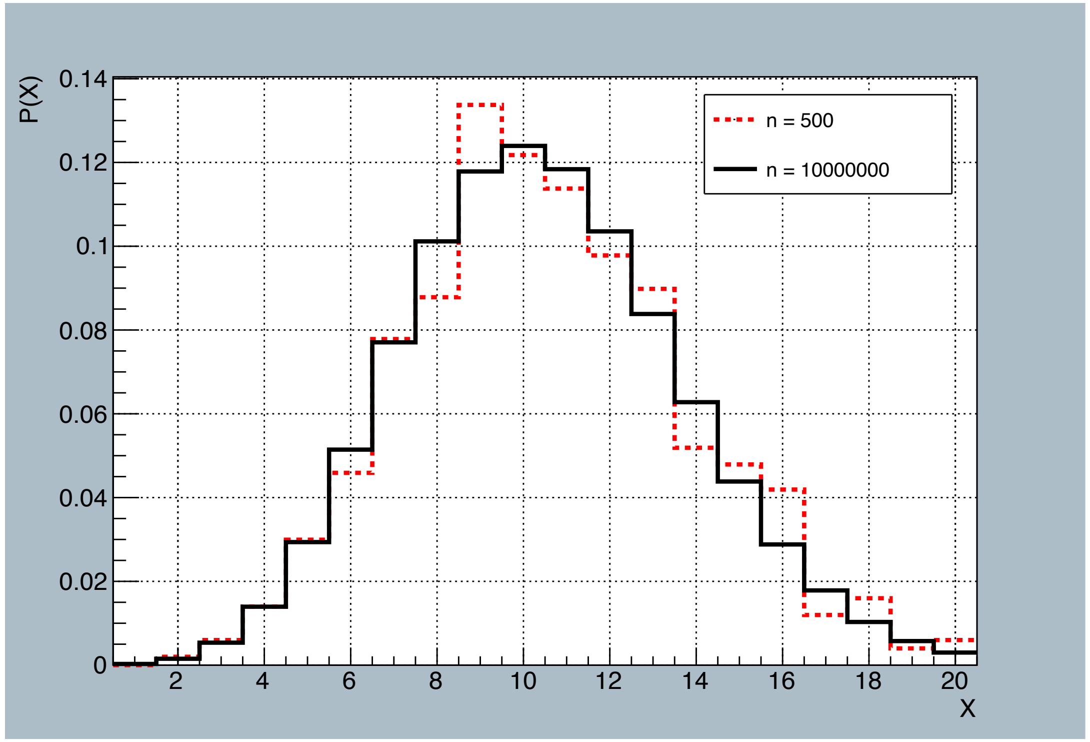
$$\sigma_\epsilon^2 = \frac{(k+1)(k+2)}{(n+1)(n+2)} - \frac{(k+1)^2}{(n+2)^2}$$

- Note that the expression for the variance allows to handle the extreme case for k = 0 and k = 1.

$$\sigma_\epsilon(k=0, n=10) = 0.09 \ , \sigma_\epsilon(k=10, n=10) = 0.4$$

Poissonian Statistics

- Perhaps one of the most important distributions in particles physics.
- Very useful in experiments with discrete counts, at a fixed rate.
- For the distribution on the right, we have the exact same Poissonian random generator, with histograms with the same binning. The only difference is the number of counts.
- **Do you note something different?**



Poissonian Statistics

- Imagine that for a given time “t” we expect μ events.
- We can divide “t” into “N” intervals: $\delta t = t/N$.
- So, the probability of one event in a δt is $\delta p = \mu \frac{\delta t}{t} = \frac{\mu}{N}$.
- Note that the problem as circled back to a sort of binomial distribution: N trials with μ positive (discreate) outcomes:

$$P(n;\mu) = \lim_{N\to\infty} \delta p^n (1-\delta p)^{N-n} \frac{N!}{n!\,(N-n)!}$$

- Taking the natural log, expanding, using Stirlings approximation, and
- calculating the limit, we can proof: $\boxed{P(n;\mu) = \frac{\mu^n e^{-\mu}}{n!}}$

Poissonian Statistics

For the Mean

$$\langle n\rangle = \sum_{n=0}^{\infty} nP(n;\mu) = \sum_{n=0}^{\infty} n\frac{\mu^n e^{-\mu}}{n!}$$

$$\langle n\rangle = \sum_{n=1}^{\infty} n\frac{\mu^n e^{-\mu}}{n!} = \mu\sum_{n=1}^{\infty} \frac{\mu^{n-1} e^{-\mu}}{(n-1)!}$$

$$\langle n\rangle = \mu\sum_{n'=0}^{\infty} \frac{\mu^{n'} e^{-\mu}}{n'!}$$

$$\boxed{\langle n\rangle = \mu}$$

For the Variance

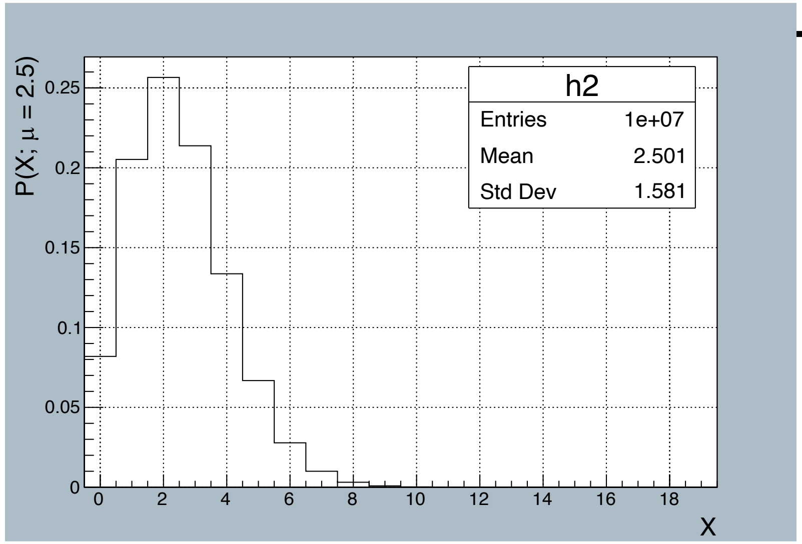
$$\langle n^2\rangle = \sum_{n=0}^{\infty} n^2P(n;\mu) = \sum_{n=0}^{\infty} n^2\frac{\mu^n e^{-\mu}}{n!}$$

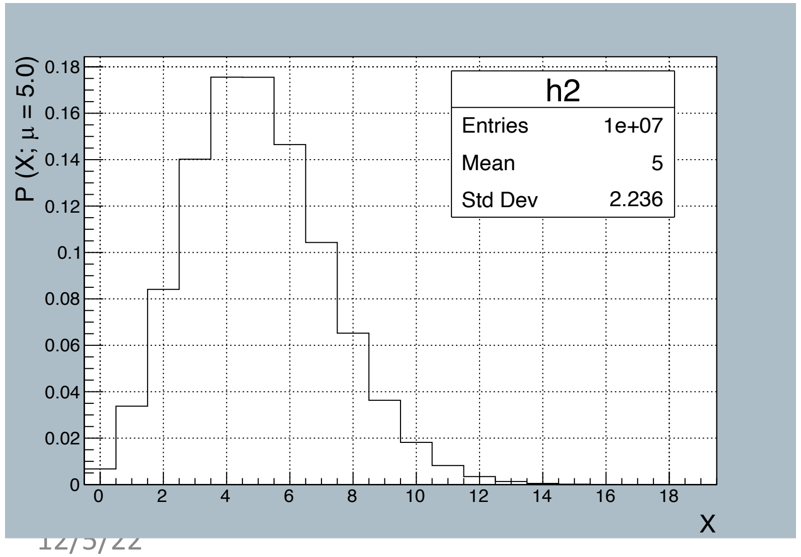
$$\langle n^2\rangle = \sum_{n=1}^{\infty} n^2\frac{\mu^n e^{-\mu}}{n!} = \mu\sum_{n=1}^{\infty} n\frac{\mu^{n-1} e^{-\mu}}{(n-1)!}$$

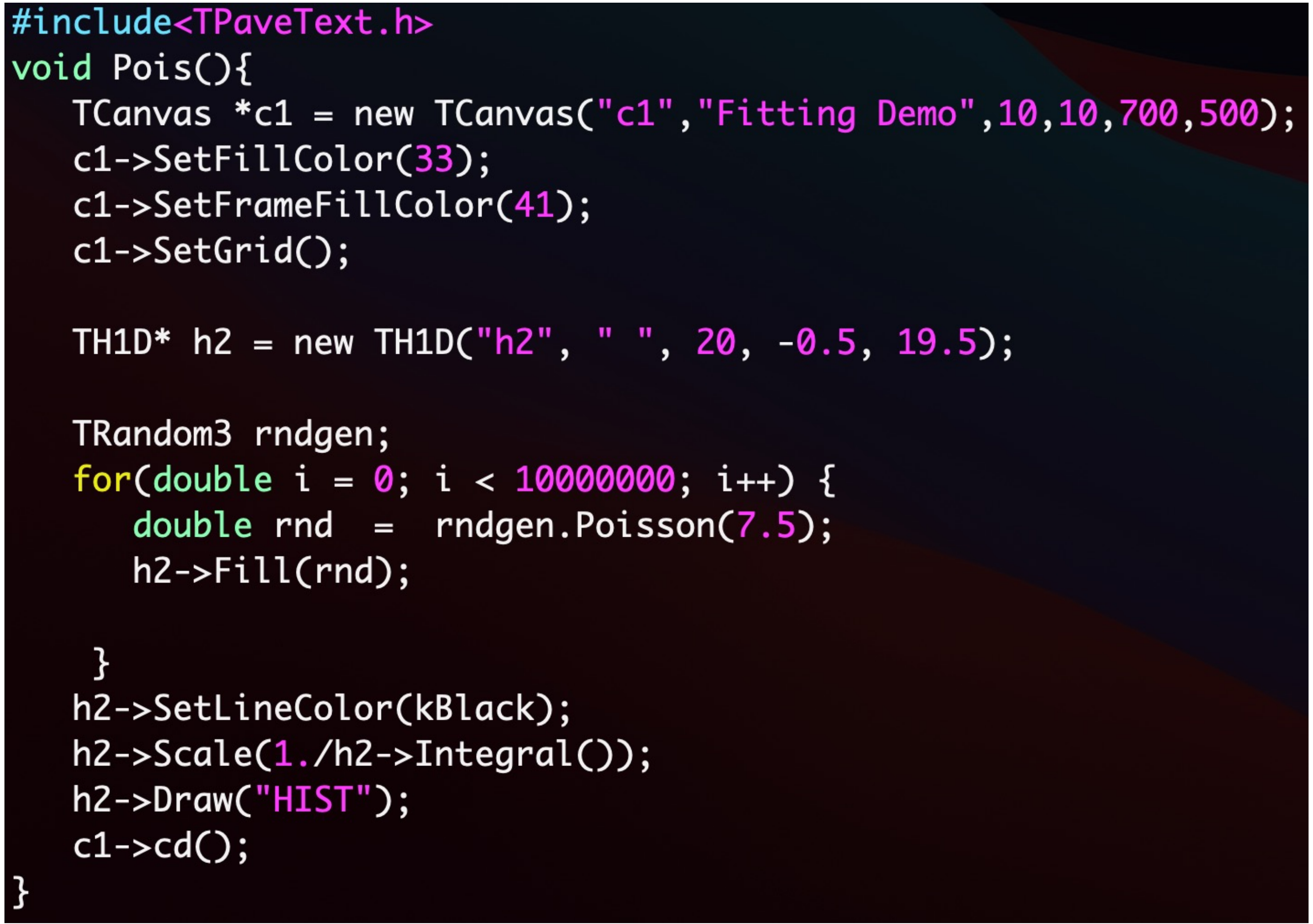
$$\langle n^2\rangle = \mu\sum_{n'=0}^{\infty} (n'+1)\frac{\mu^{n'} e^{-\mu}}{n'!} = \mu^2 + \mu$$

$$\boxed{\sigma^2 = \langle n^2\rangle - \langle n\rangle^2 = \mu}$$

Poissonian Statisitics







```
#include<TPaveText.h>
void Pois(){
   TCanvas *c1 = new TCanvas("c1","Fitting Demo",10,10,700,500);
   c1->SetFillColor(33);
   c1->SetFrameFillColor(41);
   c1->SetGrid();

   TH1D* h2 = new TH1D("h2", " ", 20, -0.5, 19.5);

   TRandom3 rndgen;
   for(double i = 0; i < 10000000; i++) {
      double rnd  =  rndgen.Poisson(7.5);
      h2->Fill(rnd);

    }
   h2->SetLineColor(kBlack);
   h2->Scale(1./h2->Integral());
   h2->Draw("HIST");
   c1->cd();
}
```

Simple Example

- Imagine we want to measure experimentally the production cross section for a specific process of interest (for example: $Z \rightarrow \mu^{+}\mu^{-}$).
- We are informed that the collected luminosity by our experiment is $L = 100.0\ fb^{-1}$ and for this luminosity we expect a number of events

$$\mu = \sigma L$$

- We select exactly two well reconstructed muons with opposite charge, which pass tight identification criteria and isolation. We veto other lepton flavors and b-jets.
- With the selection criteria described above, we obtain 63,900,000 events that pass the "cuts".

Simple Example

- Note that the number of events is our best unbiased estimate of μ:

$$\mu_e = N$$

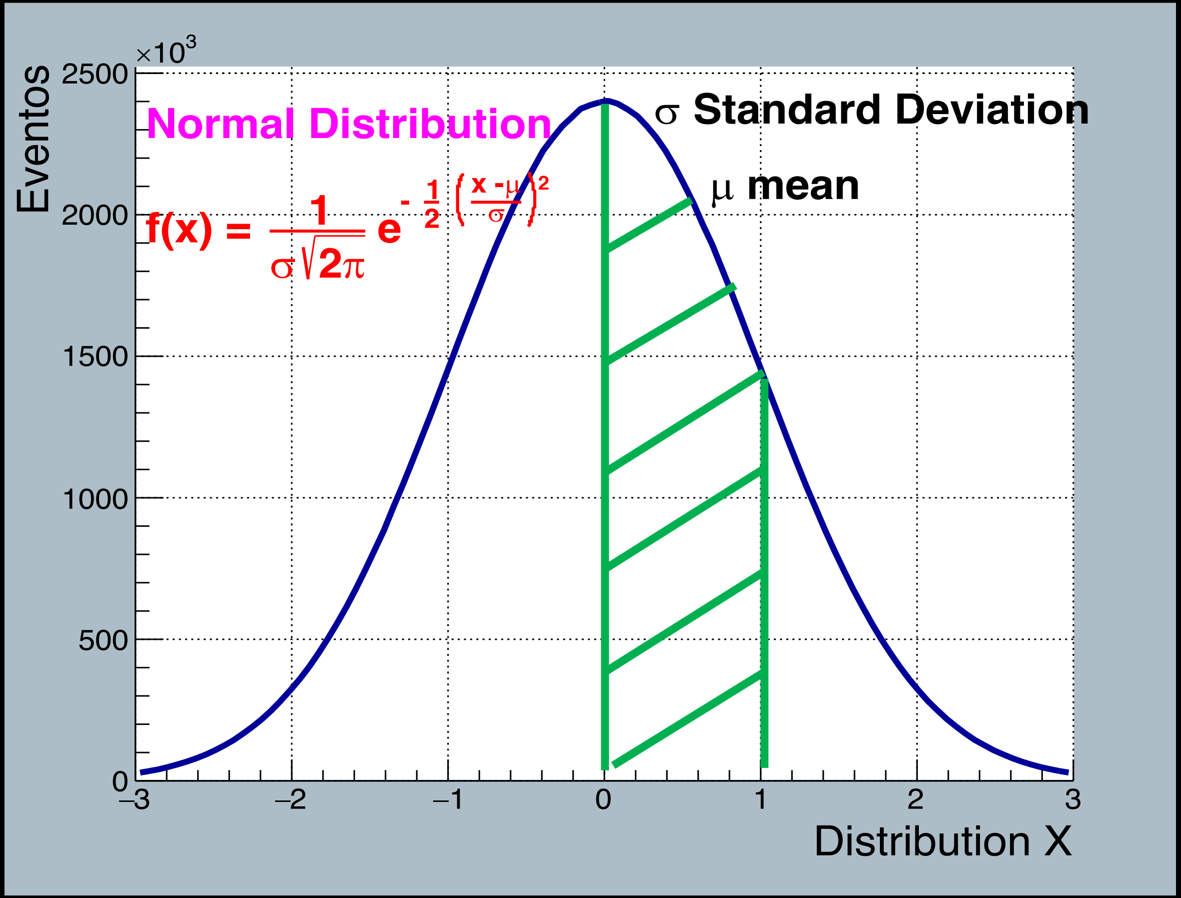
- The uncertainty on the mean of the underlaying Poisson distribution is $\sqrt{N}$

$$\mu_e = N \pm \sqrt{N}$$

- VERY IMPORTANT: This is not the error on N – there is no uncertainty on what you counted.
- Putting together all the pieces together:

$$\sigma = (639 \ \pm 0.08)\times 10^3 \ fb$$

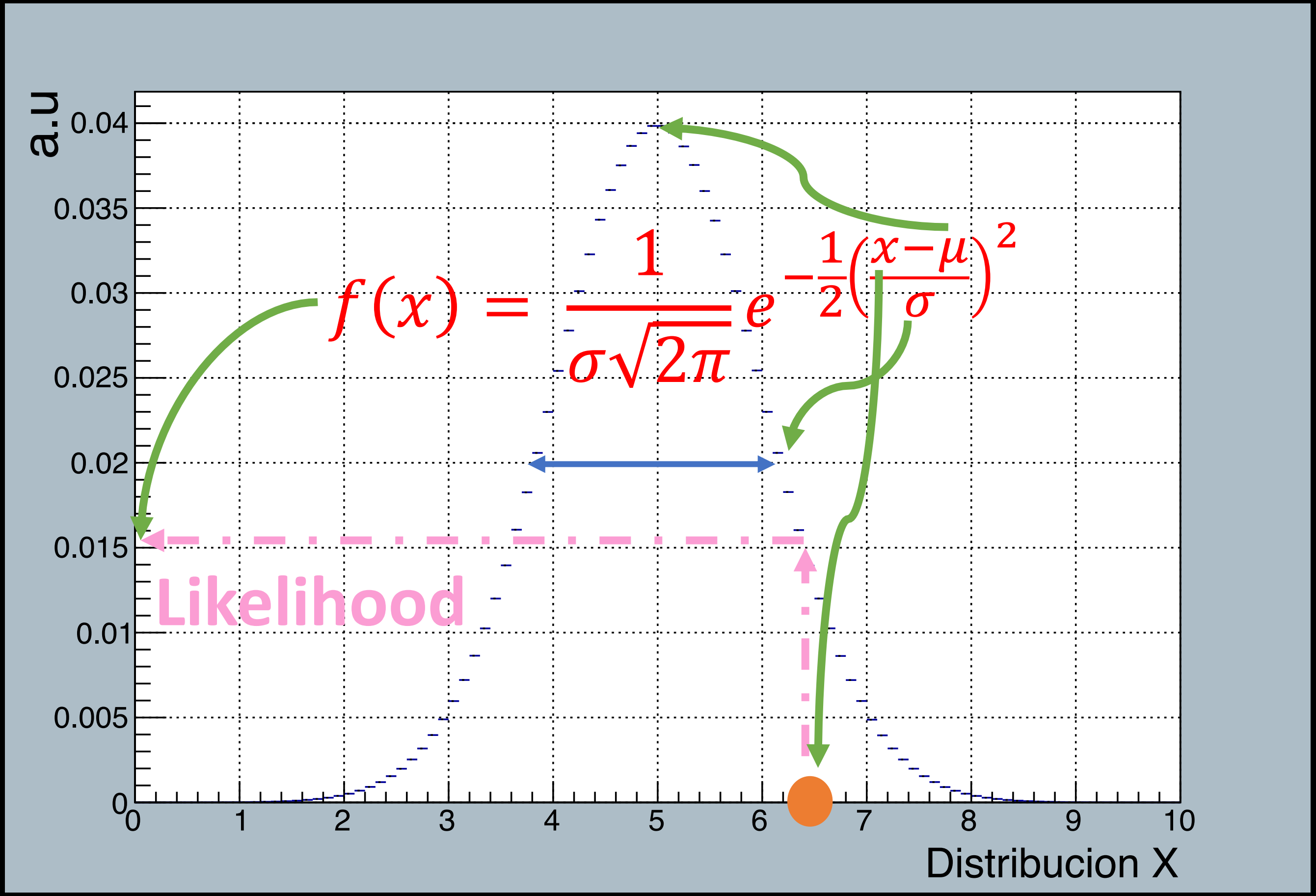
Probability



- We understand probability as the area under the curve in a fixed distribution, with respect to the total area.
- We interpret it as the feasibility of a process to occur, given a known PDF.

$P(data \mid distribution)$ = **Probability to get some data, given a distribution**

Likelihood



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Distribution normalized to unity → area under the curve = 1
- a.u: “arbitrary units”
- Likelihood represents the value that a given point can have with respect to a probability distribution that can be moved.

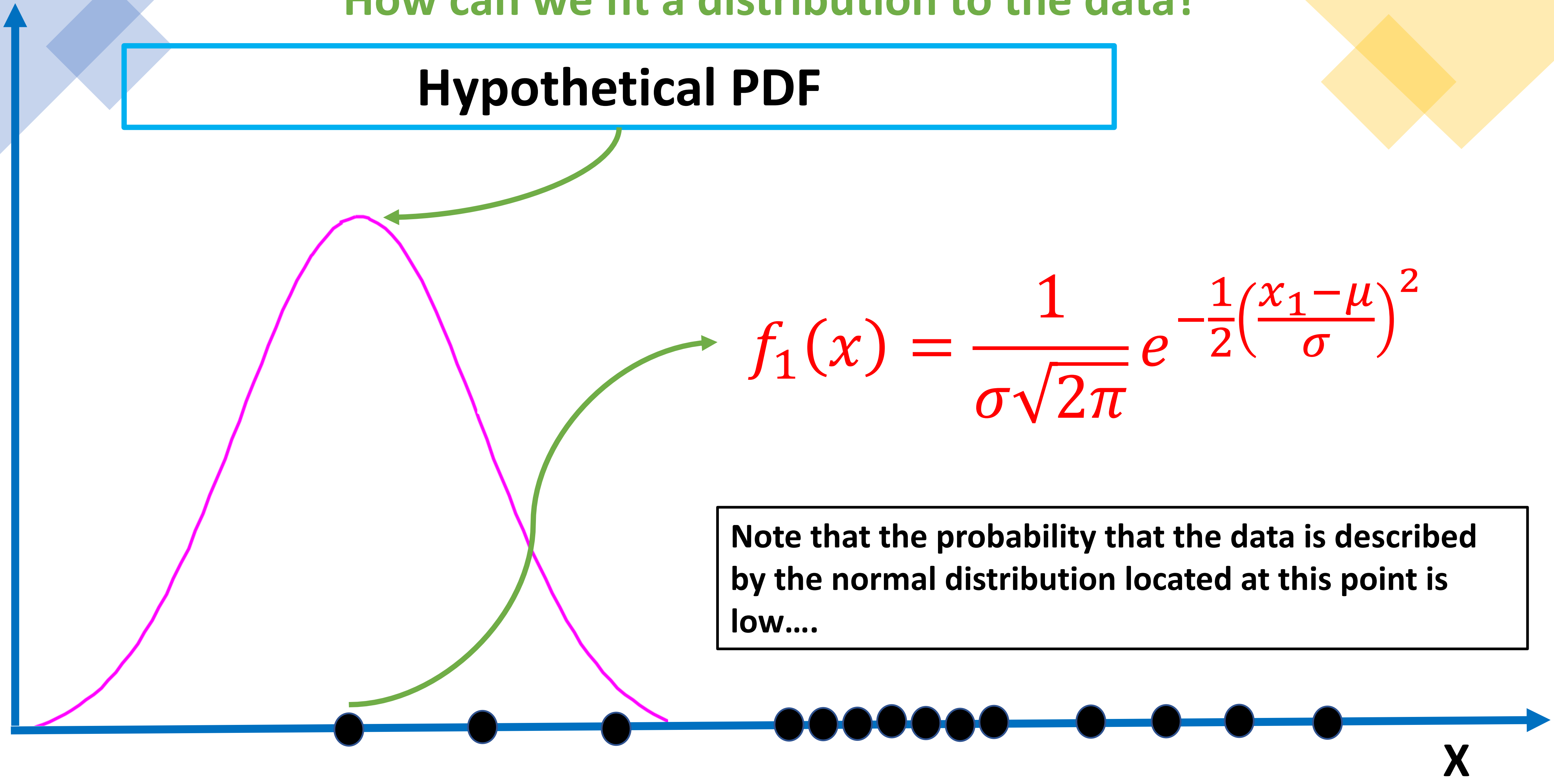
L(distribution | $data$) = Likelihood of a distribution, given data

Introduction to Likelihoods

- In different type of experiments, we can have data of some kind, with different information.
- If we have “N” events collected at the LHC:
  - How many correspond to background and how they distribute across the different observables?
  - Are there events from particles beyond the SM (signal)?
  - **How likely is that a difference between the observed data and the predicted background yields, come from a new signal?**
  - **If so, how those signal events are distributed? What is shape that characterize them?**
- **As an example, lets assume that we have a random set of measurements of a variable X. Lets use a hypothesis that those events are normally distributed.**

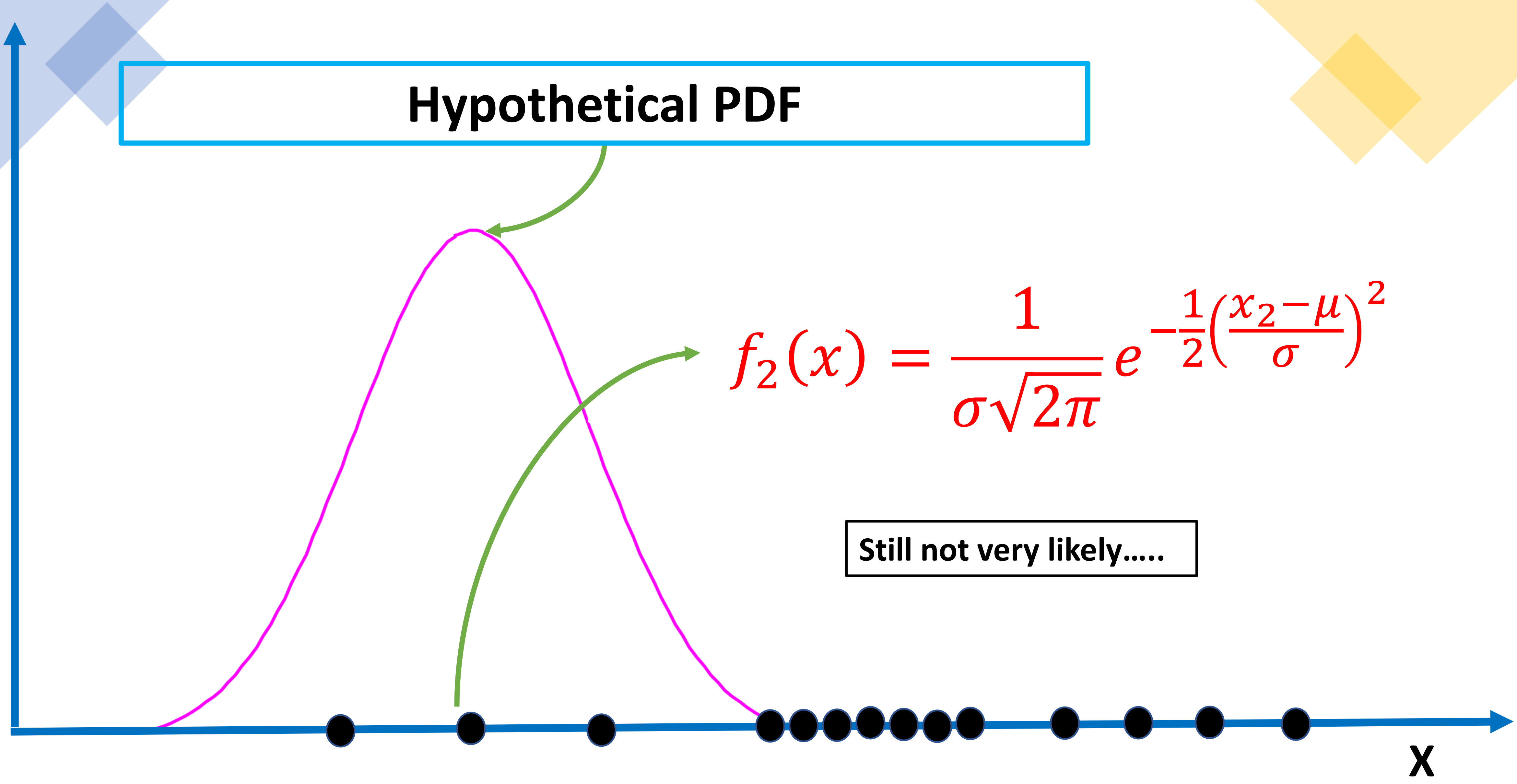
How can we fit a distribution to the data?

Hypothetical PDF



$$f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2}$$

Note that the probability that the data is described by the normal distribution located at this point is low....



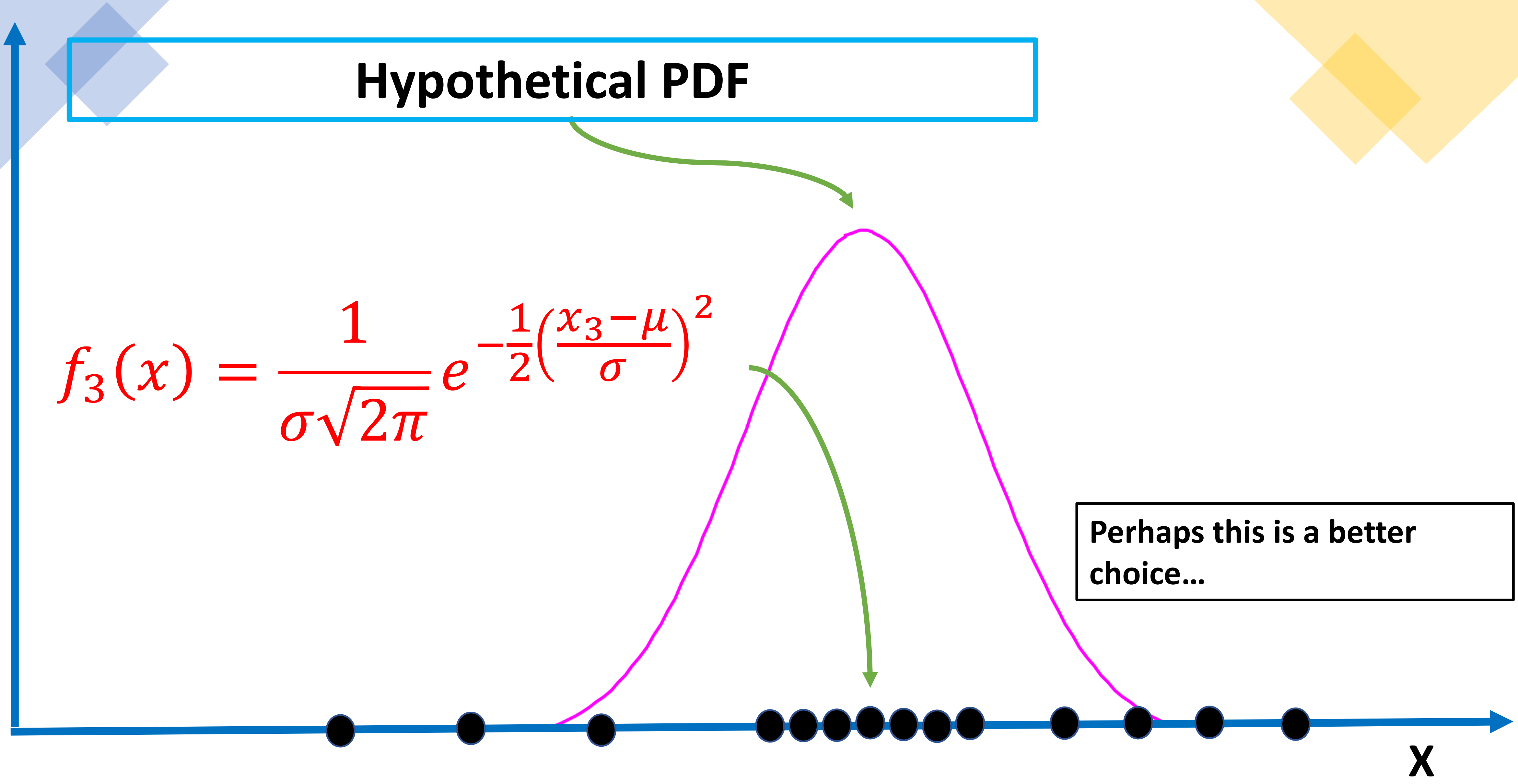
Hypothetical PDF

$$f_2(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2}$$

Still not very likely.....

Hypothetical PDF

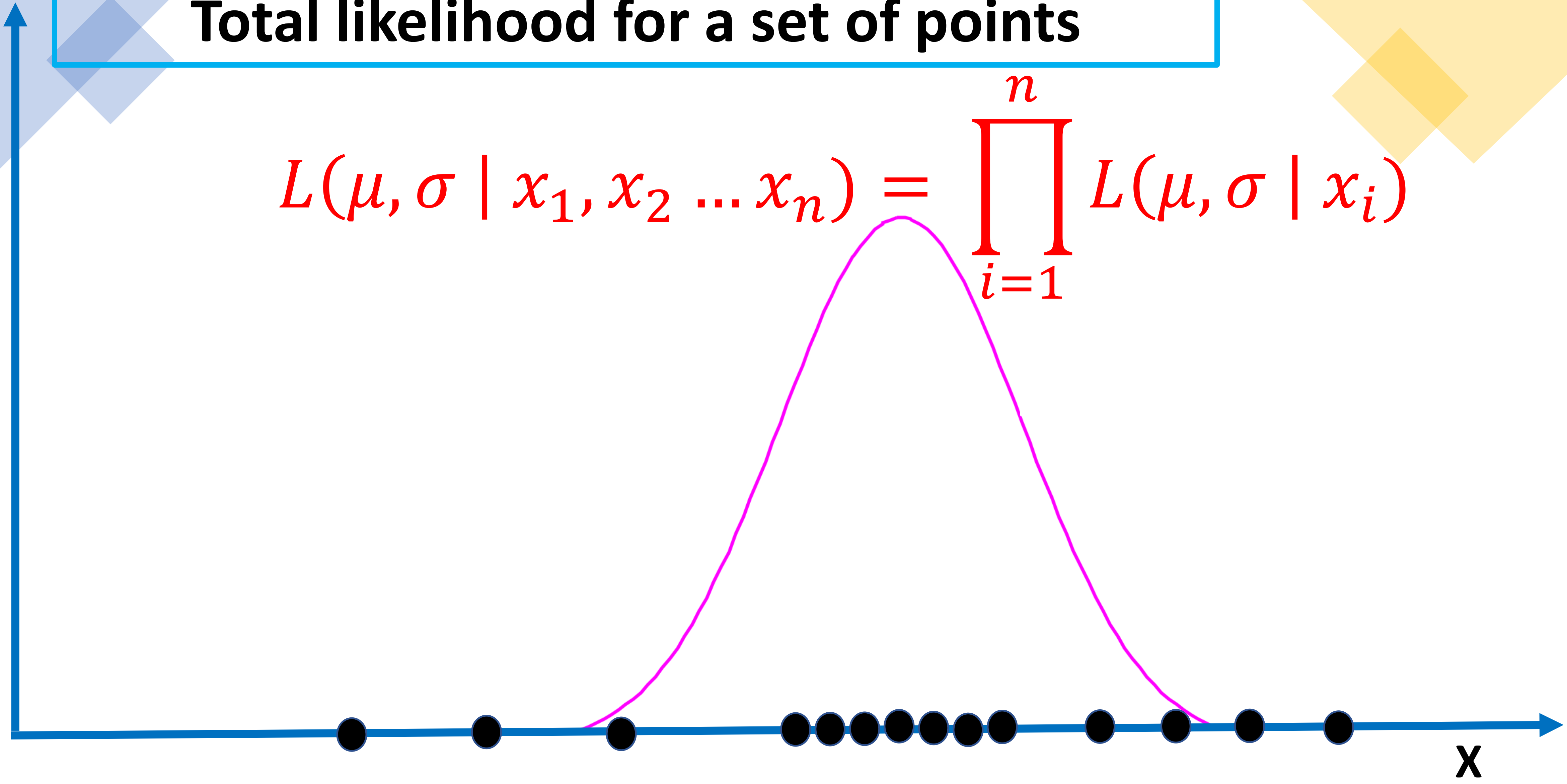
$$f_3(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_3-\mu}{\sigma}\right)^2}$$



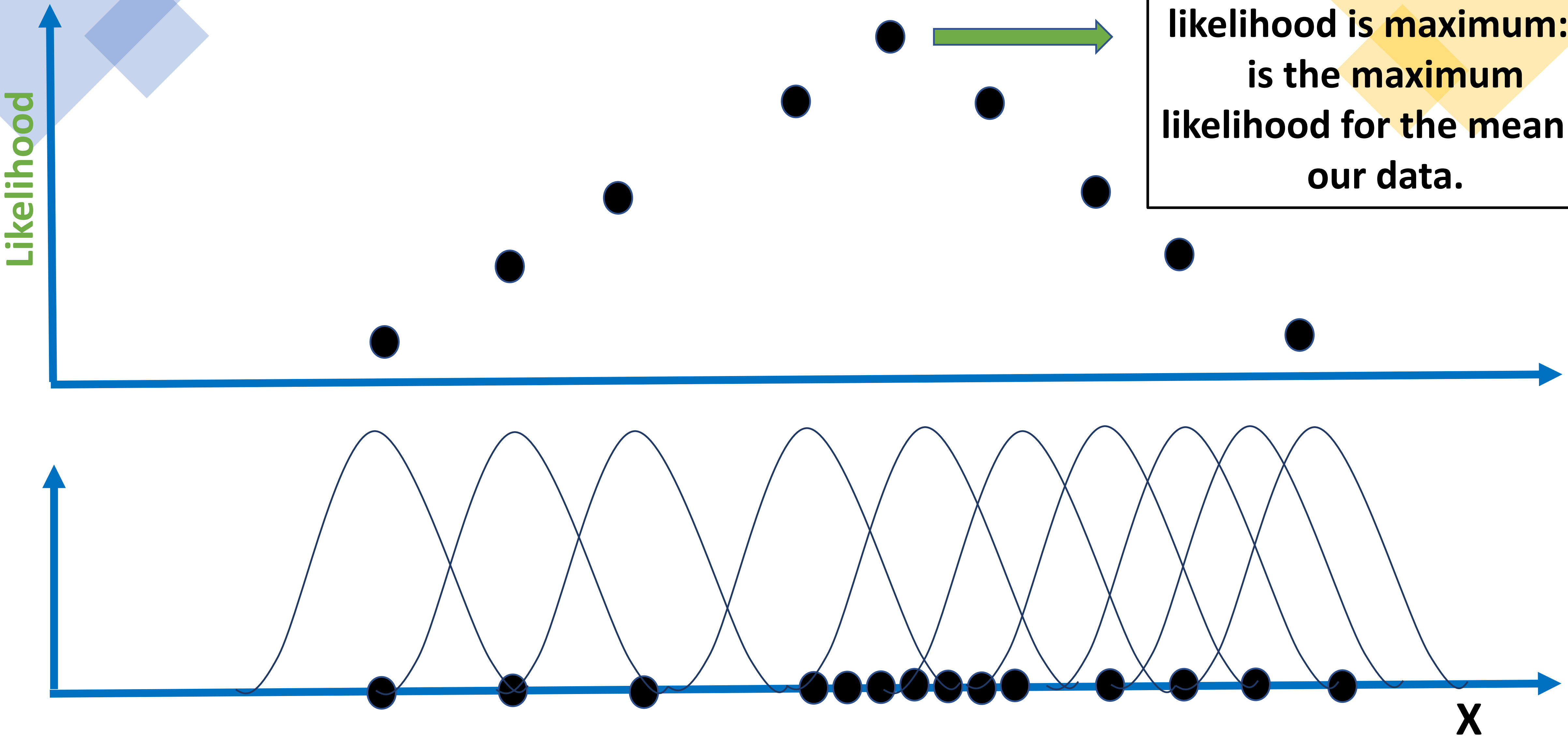
Perhaps this is a better choice…

Total likelihood for a set of points

$$L(\mu,\sigma \mid x_1, x_2 \ldots x_n) = \prod_{i=1}^{n} L(\mu,\sigma \mid x_i)$$



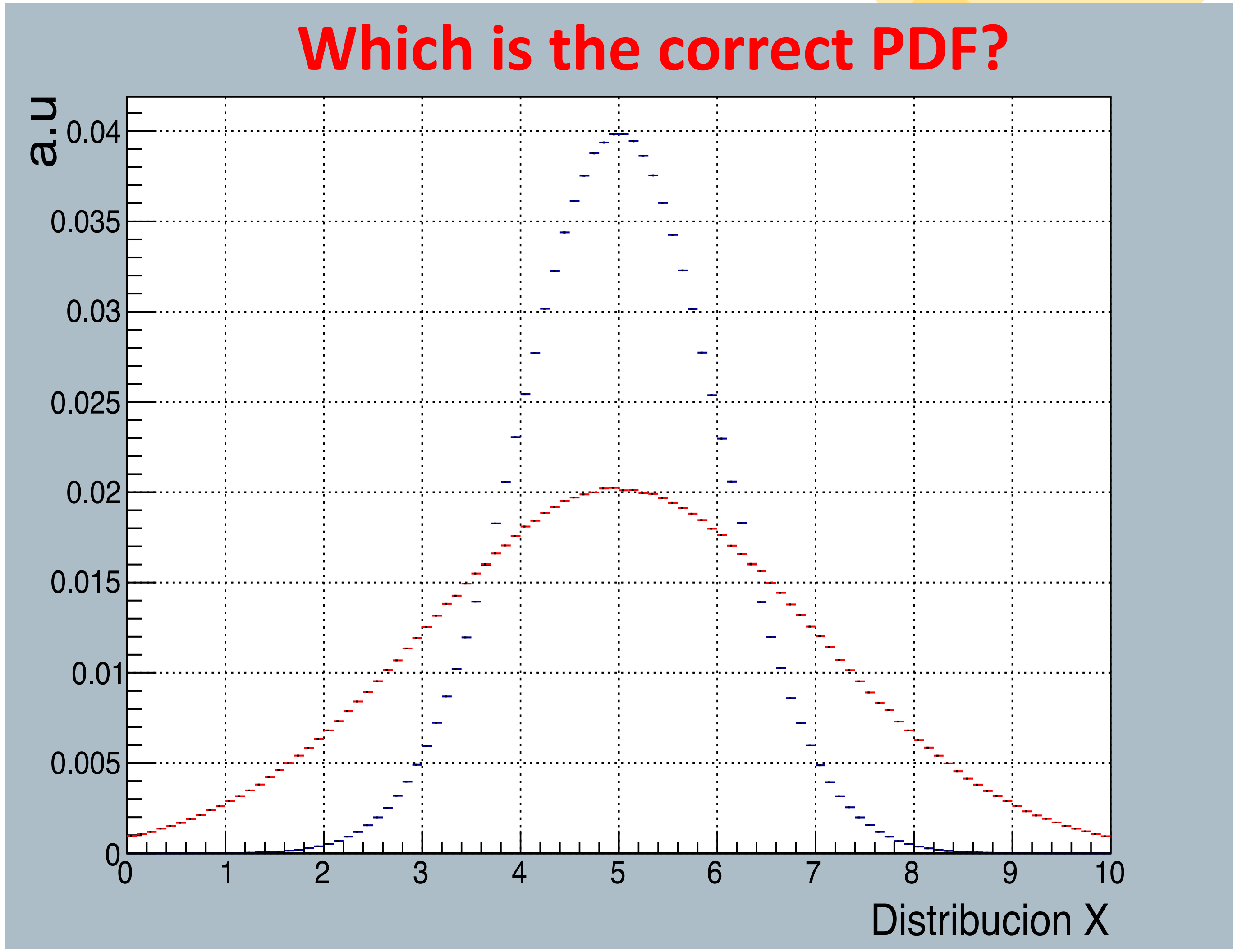
Note we started with a normal distribution with random mean and s.t.d

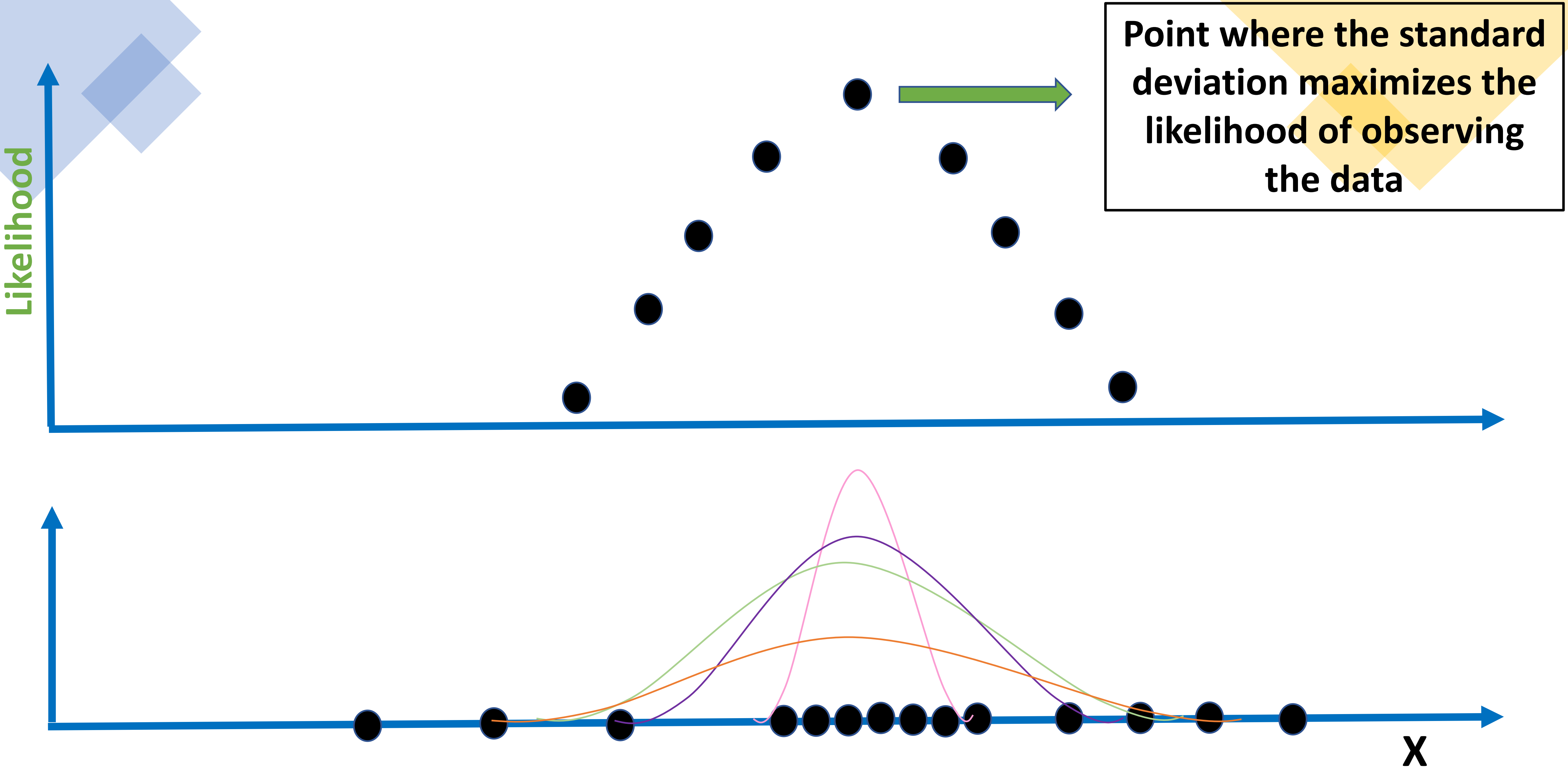


Point where the likelihood is maximum: It is the maximum likelihood for the mean of our data.

Note we started with a normal distribution with random mean and s.t.d

- So far, we just estimated the maximum likelihood for the mean, but what about the maximum likelihood for the s.t.d?
- We follow the same procedure!





Point where the standard deviation maximizes the likelihood of observing the data

Note we started with a normal distribution with random mean and s.t.d

Introduction to Likelihoods

- We can do this mathematically! We will use the Gaussian distribution as an example. Remember our expression for the total likelihood:

$$L(\mu, \sigma \mid x_1, x_2 \ldots x_n) = \prod_{i=1}^{n} L((\mu, \sigma \mid x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} \times \cdots .. \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2}$$

- Now, we need to take the derived with respect to μ (σ), keeping σ (μ) constant, and equal the expression to zero to find the maximum for μ (σ).
- To facilitate the process, we take the "ln" of the likelihood:

$$\ln(L(\mu, \sigma \mid x_1, x_2 \ldots x_n)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} \times \cdots .. \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2}\right)$$

Introduction to Likelihoods

- Using the properties of ln:

$$\ln(L(\mu,\sigma \mid x_1, x_2 \ldots x_n)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2}\right) + \cdots + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2}\right)$$

- Note that

$$\ln\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2}\right) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \ln\left(e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2}\right) = \ln\left[(2\pi\sigma^2)^{-1/2}\right] - \frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 \ln(e)$$

$$\ln\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2}\right) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 = -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2$$

Introduction to Likelihoods

- Therefore:

$$\ln(L(\mu,\sigma \mid x_1,x_2 \ldots x_n)) = -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 - \cdots - -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2$$

- Now we can combine the terms:

$$\ln(L(\mu,\sigma \mid x_1,x_2 \ldots x_n)) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 - \cdots - \frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2$$

- Taking the derivative with respect μ:

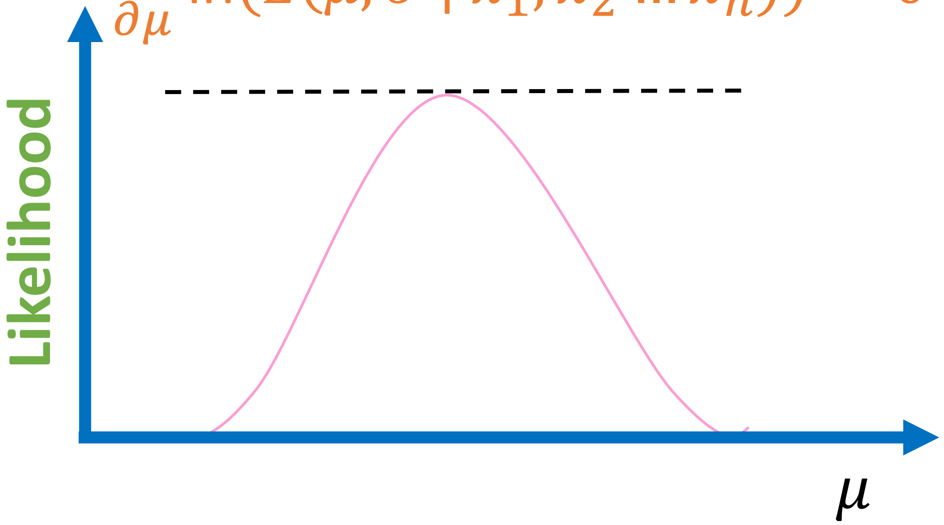
$$\frac{\partial}{\partial\mu}\ln(L(\mu,\sigma \mid x_1,x_2 \ldots x_n)) = \left(\frac{x_1-\mu}{\sigma^2}\right) + \cdots + \left(\frac{x_n-\mu}{\sigma^2}\right) = \frac{1}{\sigma^2}[(x_1 + \cdots + x_n) - n\mu]$$

- Likewise, we take the derivative with respect σ:

$$\frac{\partial}{\partial\sigma}\ln(L(\mu,\sigma \mid x_1,x_2 \ldots x_n)) = -\frac{n}{\sigma} + \frac{(x_1-\mu)^2}{\sigma^3} + \cdots + \frac{(x_n-\mu)^2}{\sigma^3}$$

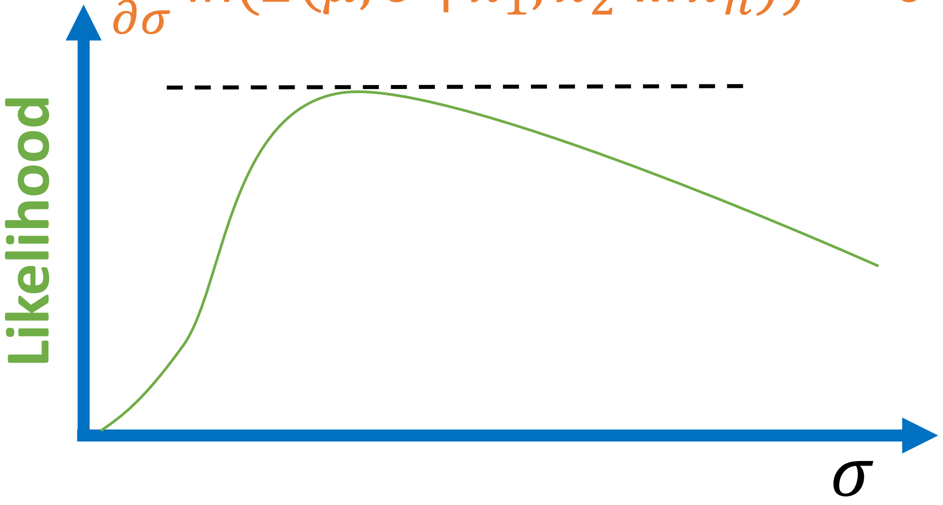
Introduction to Likelihoods

$$\frac{\partial}{\partial\mu}\ln(L(\mu,\sigma\mid x_1,x_2\ldots x_n))=0$$



$$\boxed{\boldsymbol{\mu}=\frac{(\boldsymbol{x_1}+\cdots.+\boldsymbol{x_n})}{\boldsymbol{n}}}$$

$$\frac{\partial}{\partial\sigma}\ln(L(\mu,\sigma\mid x_1,x_2\ldots x_n))=0$$



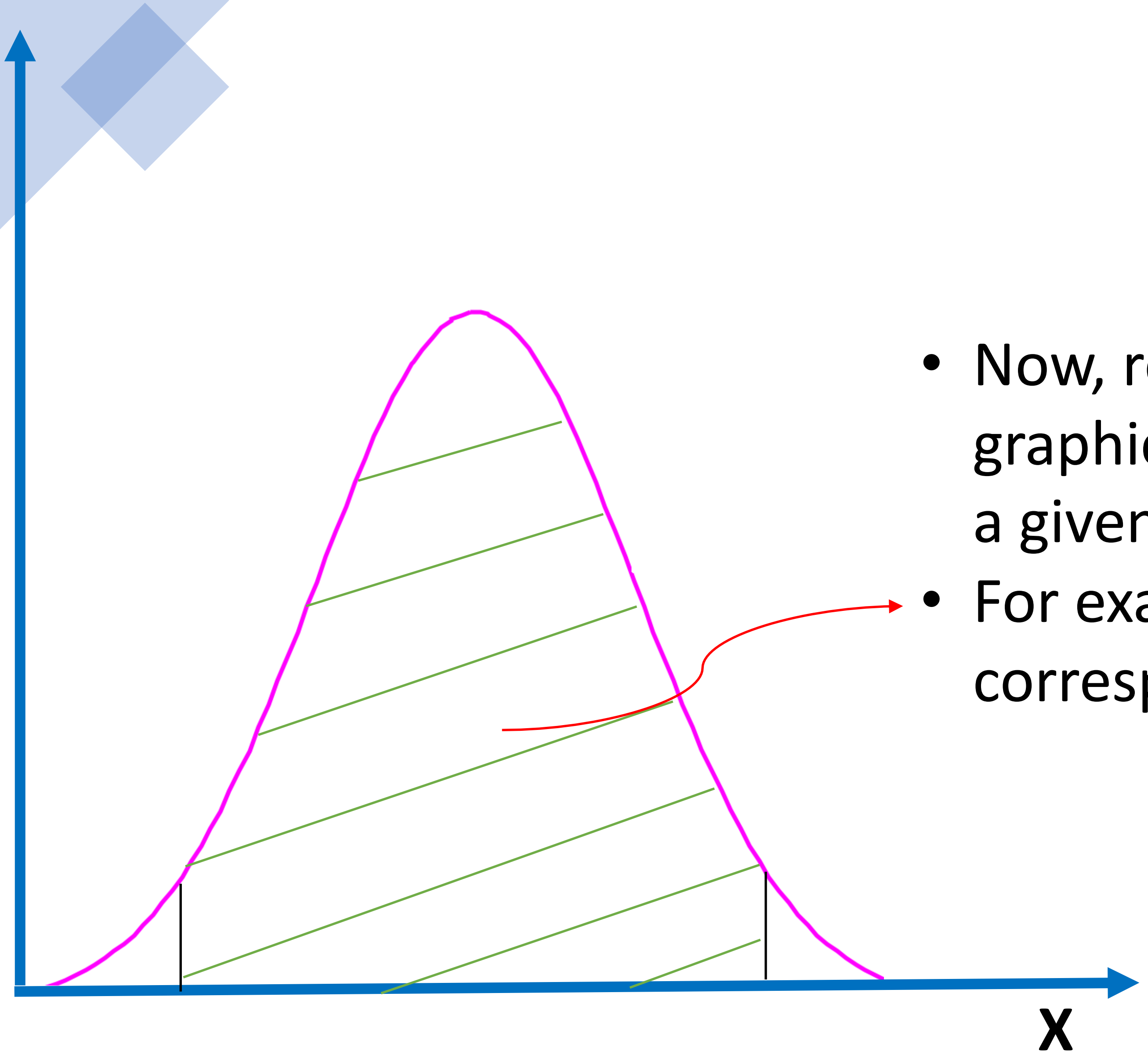
$$\boxed{\boldsymbol{\sigma}=\sqrt{\frac{(\boldsymbol{x_1}-\boldsymbol{\mu})^2+\cdots+(\boldsymbol{x_n}-\boldsymbol{\mu})^2}{\boldsymbol{n}}}}$$

Introduction to Likelihoods

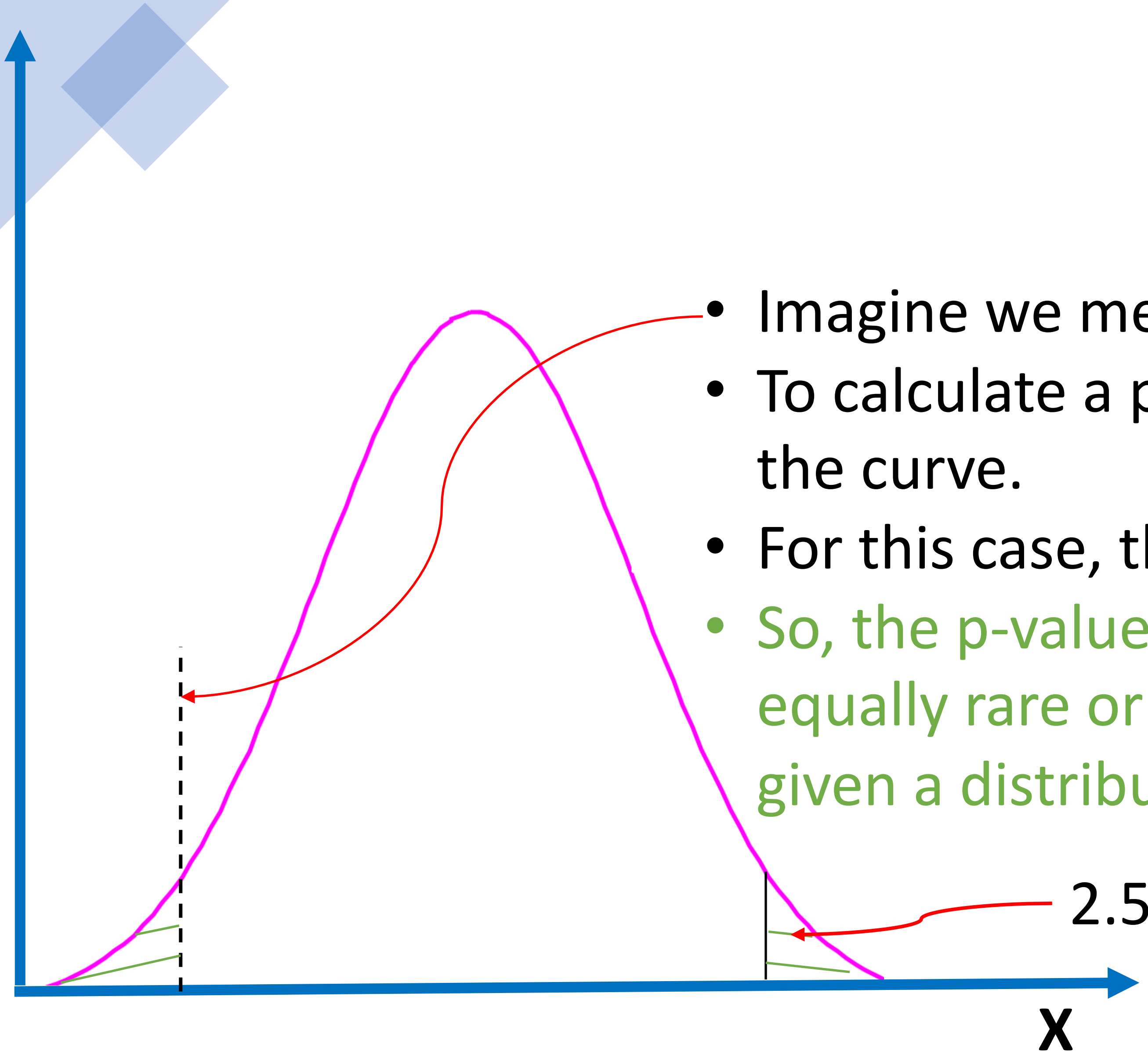
- With this information we conclude:
  - The mean of the data is the maximum likelihood estimate for where the center of the Gaussian distribution must be located.
  - We use the formula for the s.t.d to determine the width of the Gaussian curve that, according to the data, give the maximum likelihood.
- **Although this example was performed with the Gaussian distribution, the logic is the same for other types of PDFs!**

p-values

- If I have a measurement where there is a difference between the BKG and the observed data, **the null-hypothesis test** states that the difference comes from random chance alone.
- The p-value is defined as the probability that random chance generates an extreme result that could explain the difference between the observed and the expected data, assuming that the null hypothesis is true.
- P-values are composed by three parts:
  1. The probability that random change would yields the observed data.
  2. The probability of observing something else that is equally rare.
  3. The probability of observing something more extreme.

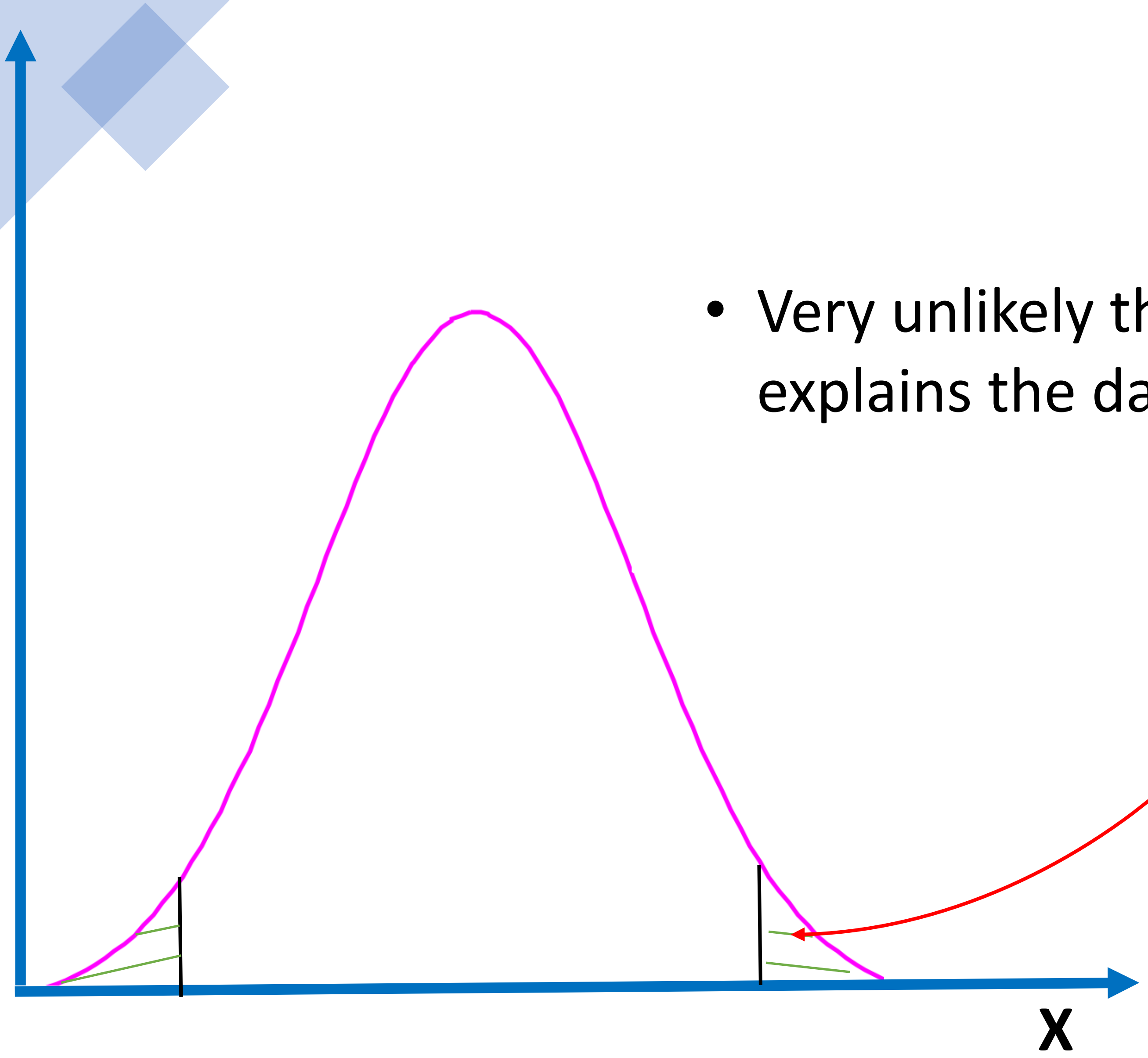


- Now, remember that we understand, graphically, the area under the curve for a given range with respect the total area.
- For example, the shaded area corresponds to 95% of the total area.



- Imagine we measure the extreme value
- To calculate a p-value, we basically add the areas un the curve.
- For this case, the p-value is = 0.025 + 0.025 = 0.05.
- So, the p-value is the probability that something equally rare or rarer explains the observed data, given a distribution.

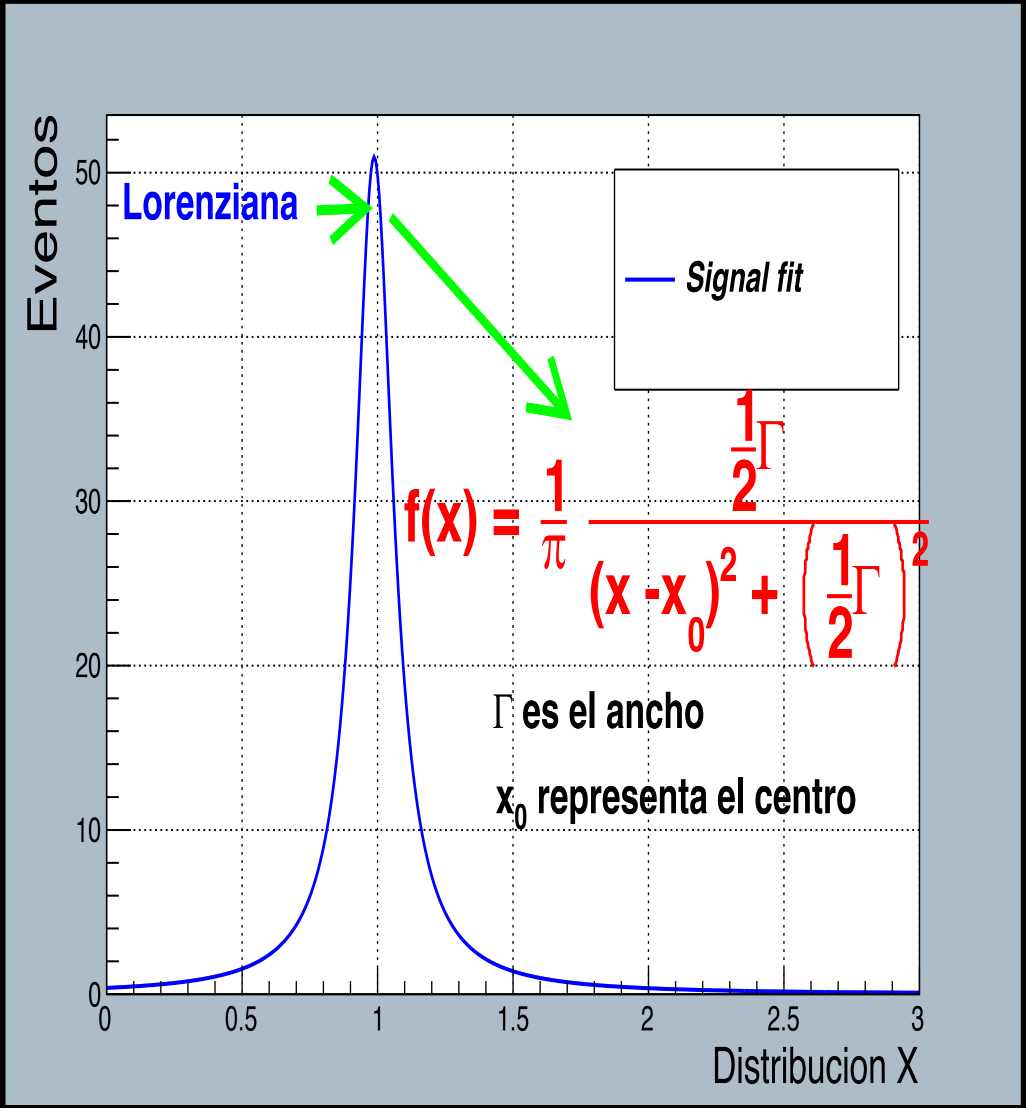
2.5% probability.



- Very unlikely that any event falling here, explains the data with this given distribution

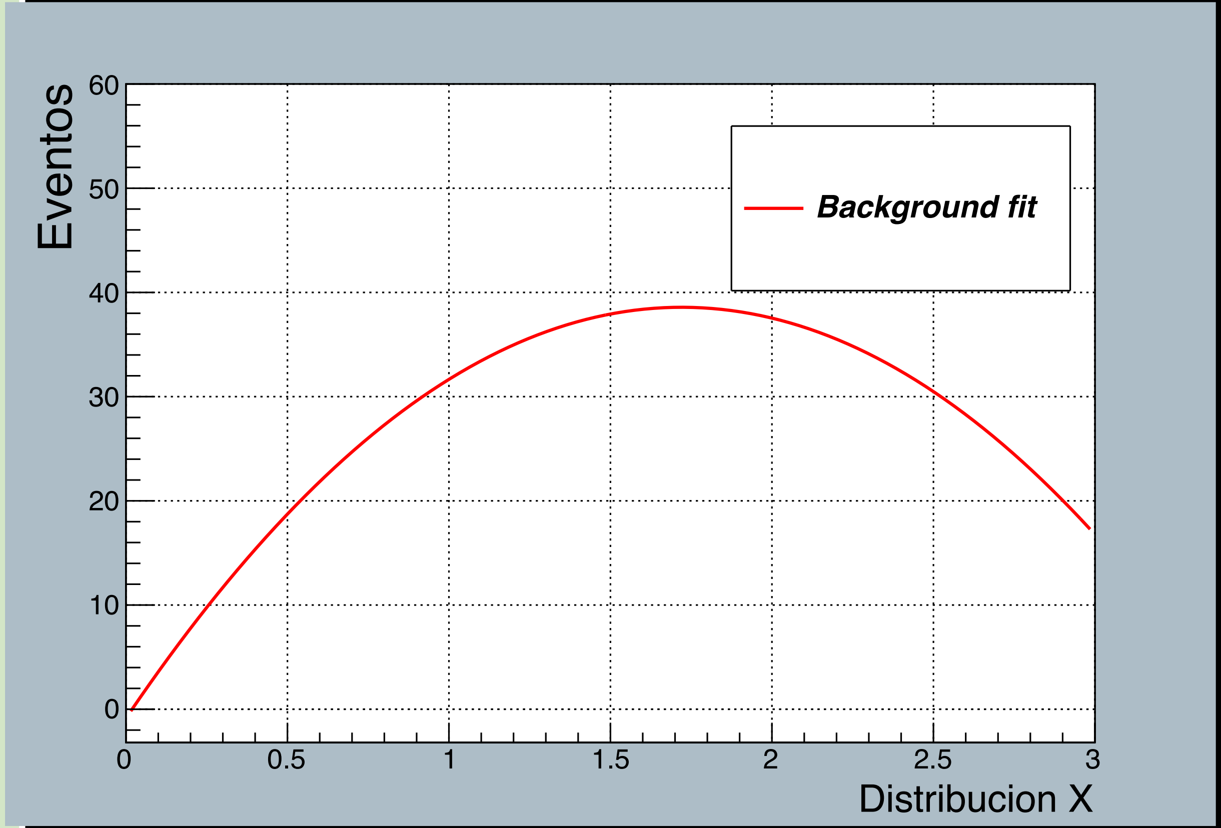
Example

Hypothetical distribution for signal



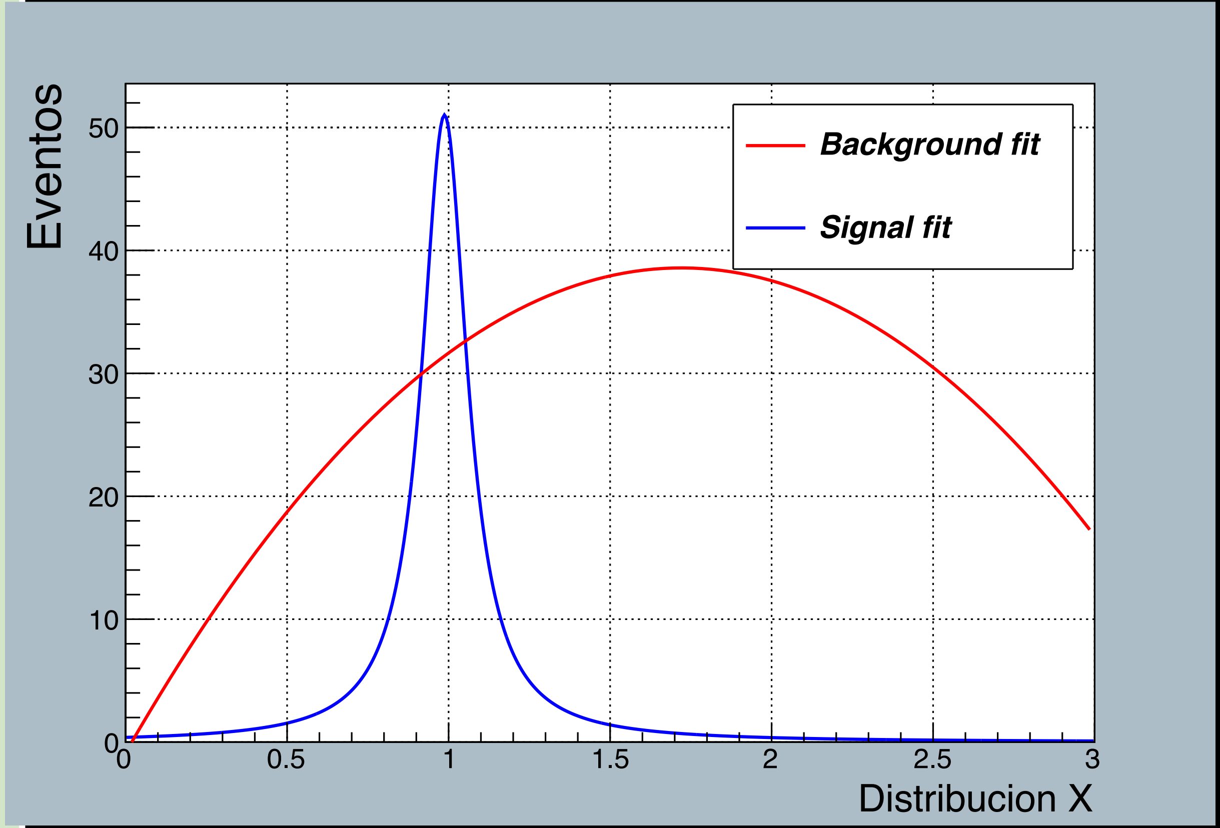
Example

Background distribution: shape and rate



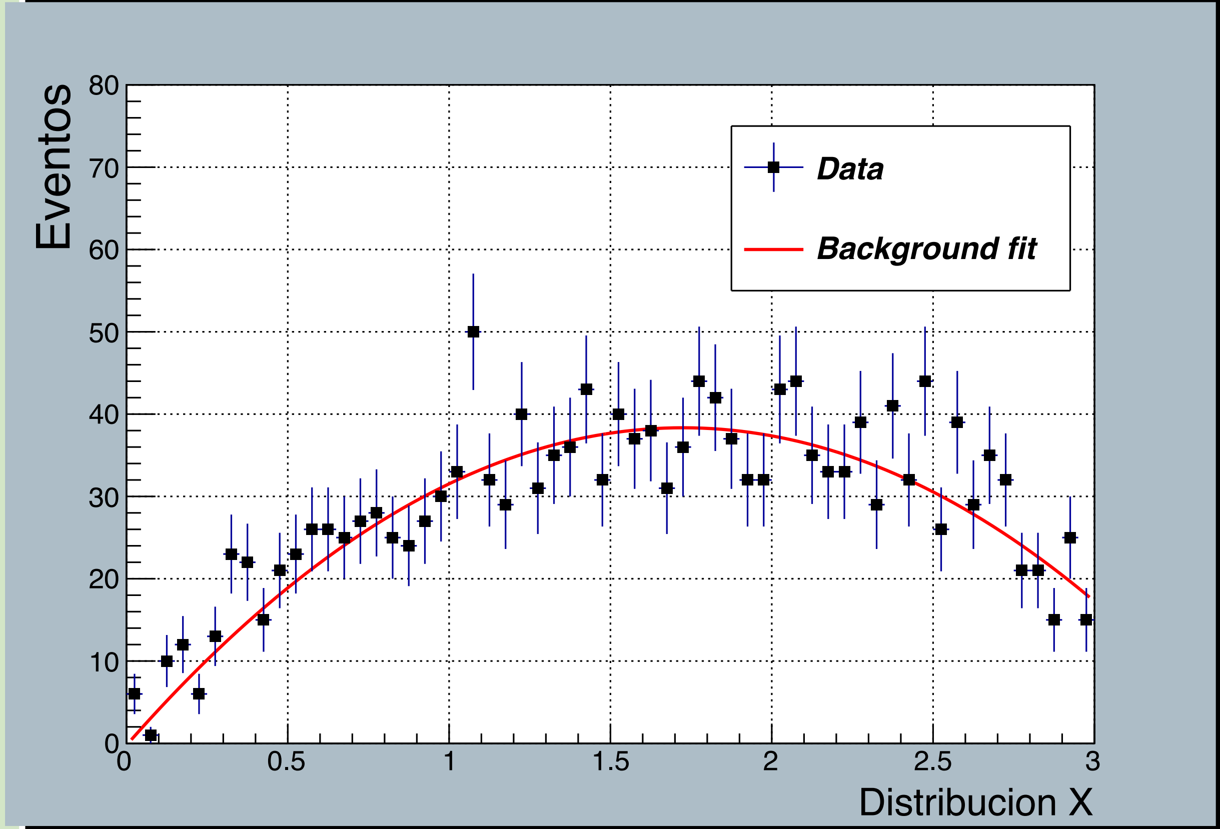
Example

Overlaying the two distributions



Example

First case: there is agreement between the observed data and the predicted background



Likelihood ratio as test statistic…

- So, we have some number of events for background distributed according to some shape, which depends on the topology of the analysis. We will represent the distribution of background events as $f(x|H_0)$, where H_0 represents all the background events.
- **Note that known precisely $f(x|H_0)$ is an idealized situation.**
- Similarly, $f(x|H_1)$ describes the distribution of signal and background events mixed.
- Now, imagine that for a given luminosity, we have “s” number of signal events and “b” number of background events.

Likelihood ratio as test statistic…

- As we discussed yesterday, we expect that the observed number of events follow a Poissonian distribution:

$$P(n|\ b) = \frac{b^n}{n!} e^{-b} \quad , P(n|b+s) = \frac{(b+s)^n}{n!} e^{-(b+s)}$$

- The likelihood function for the entire experiment assuming the background-only hypothesis (H_0):

$$L_b = \frac{b^n}{n!} e^{-b} \prod_{i=1}^{n} f(x_i|b)$$

- Similarly, for the signal plus background hypothesis:

priors

$$L_{b+s} = \frac{(b+s)^n}{n!} e^{-(b+s)} \prod_{i=1}^{n} (\ \alpha_b f(x_i|b) + \alpha_s f(x_i|s)\ )$$

Likelihood ratio as test statistic…

- To test if the observed data, given the certain predicted background, might contain signal events of interest, we use a monotonic set test statistic, defined as Q:

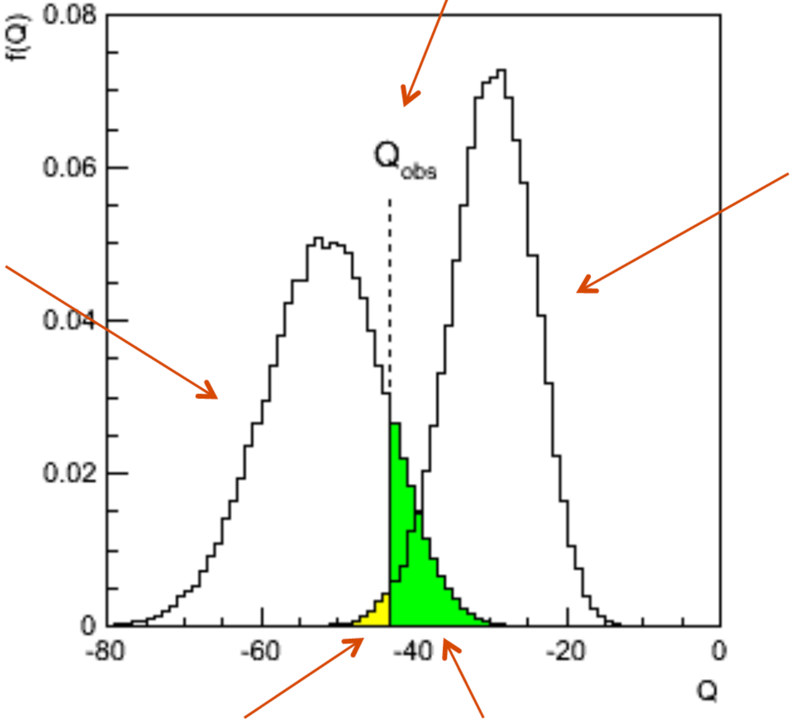
$$Q = -2\ln\frac{L_{b+s}}{L_b} = -s + \sum_{i=1}^{n} \ln\left(1 + \frac{s}{b}\frac{f(x_i|s)}{f(x_i|b)}\right)$$

- Note that we would need to know $f(x|b)$ and $f(x|s)$ with relatively good precision.
- To compute p-values for the b and s+b hypotheses given an observed value of Q we need the distributions $f(\mathrm{Q}|b)$ and $f(\mathrm{Q}|b+s)$.

Distribution of Q (From Glen Cowan)

Take e.g. b = 100, s = 20.

Suppose in real experiment Q is observed here.



$f(Q|\mathrm{s+b})$

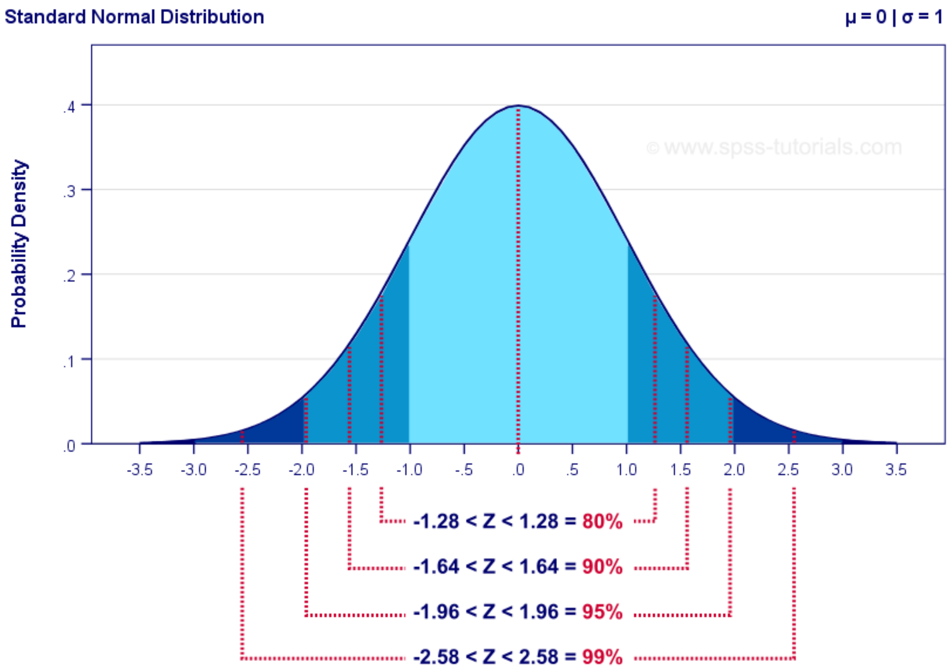
$f(Q|\mathrm{b})$

p-value of b only

p-value of s+b

Practical experimental example…

- Suppose we have a prediction of $b = 9.00 \pm 3.00$, an observation of $n = 10$, and expected signal yield of $s = 1.00$, considering a production cross section for signal $\sigma_{xs} = 1.00$ pb.
- The main question is: Can we exclude this signal model, given the observed data and the predicted BKG rate?
- Note that for a normal distribution the Z-score ($Z = \frac{x-\mu}{\sigma}$) of 1.64, corresponds to 90%.



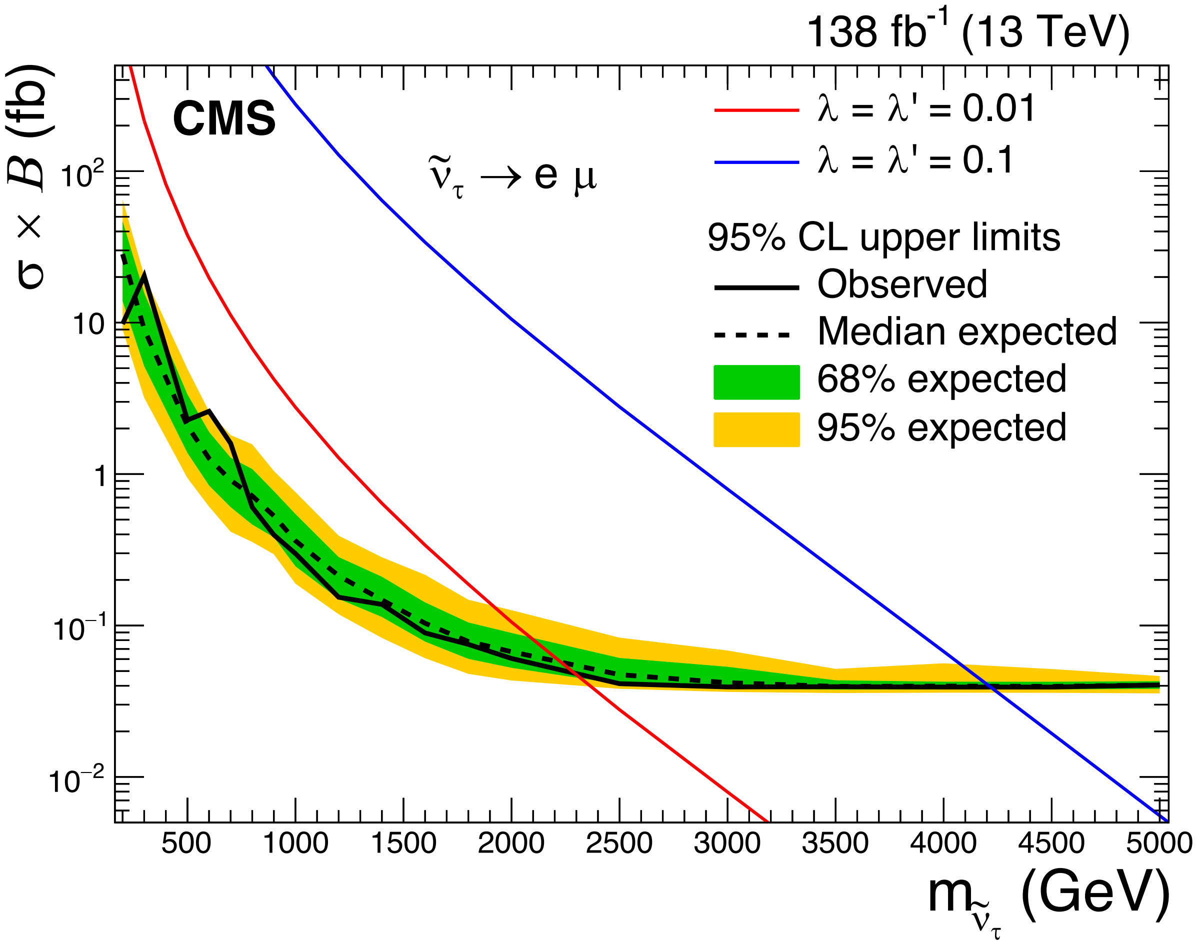
Practical experimental example…

- Therefore, to be able to exclude a signal at 90% C.L, given the observed data and predicted BKG, we can make the following **rough** calculation.
- We will need at least $s_{need} = 1.64 \times \delta b = 1.64 \times 3.00 = 4.92$ events in order to be able to exclude the signal model.
- Nevertheless, we have only 1.00 expected events in signal, weighted with a cross section of 1.0 pb.
- We can parametrize this as $s_{need} = s \times r = 4.92$, so $r = 4.92$.
- Now, we can estimate the experimental cross section is

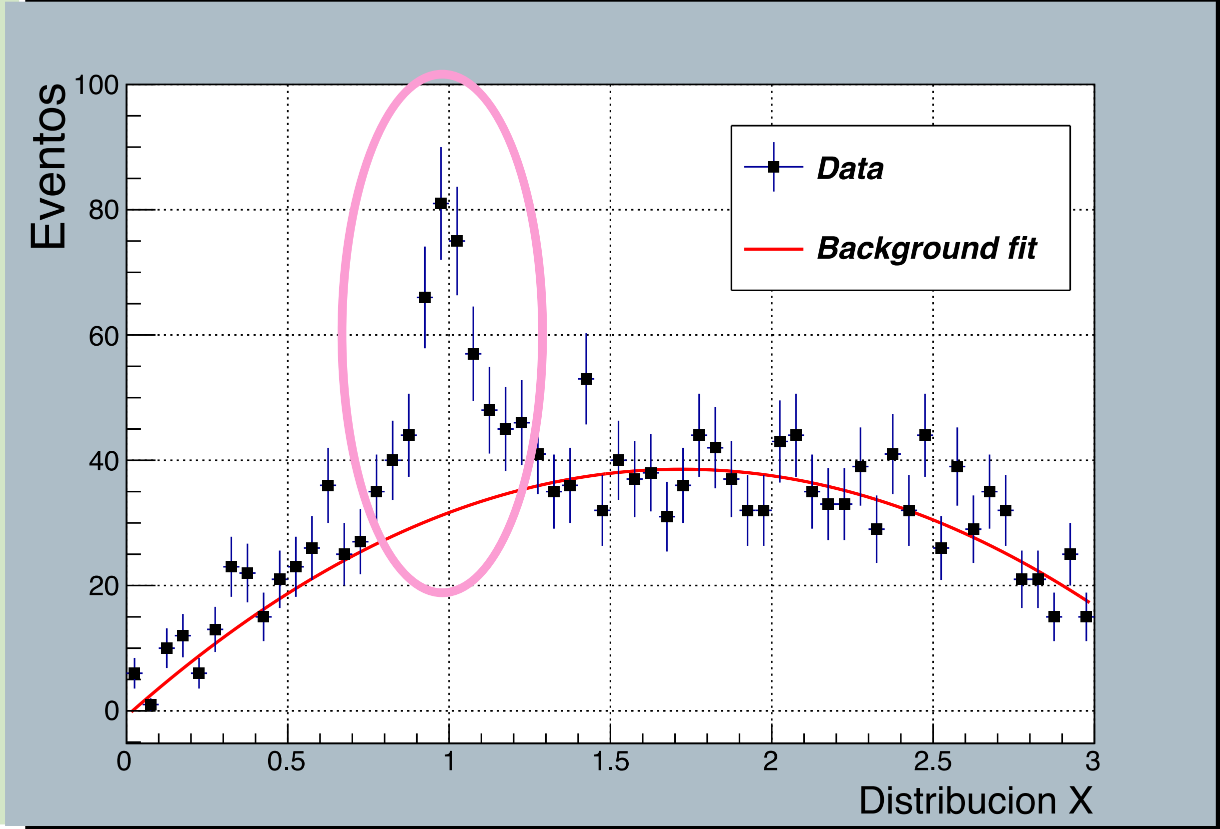
$$\sigma_{xs-need} = \sigma_{xs} \times r = 4.92$$

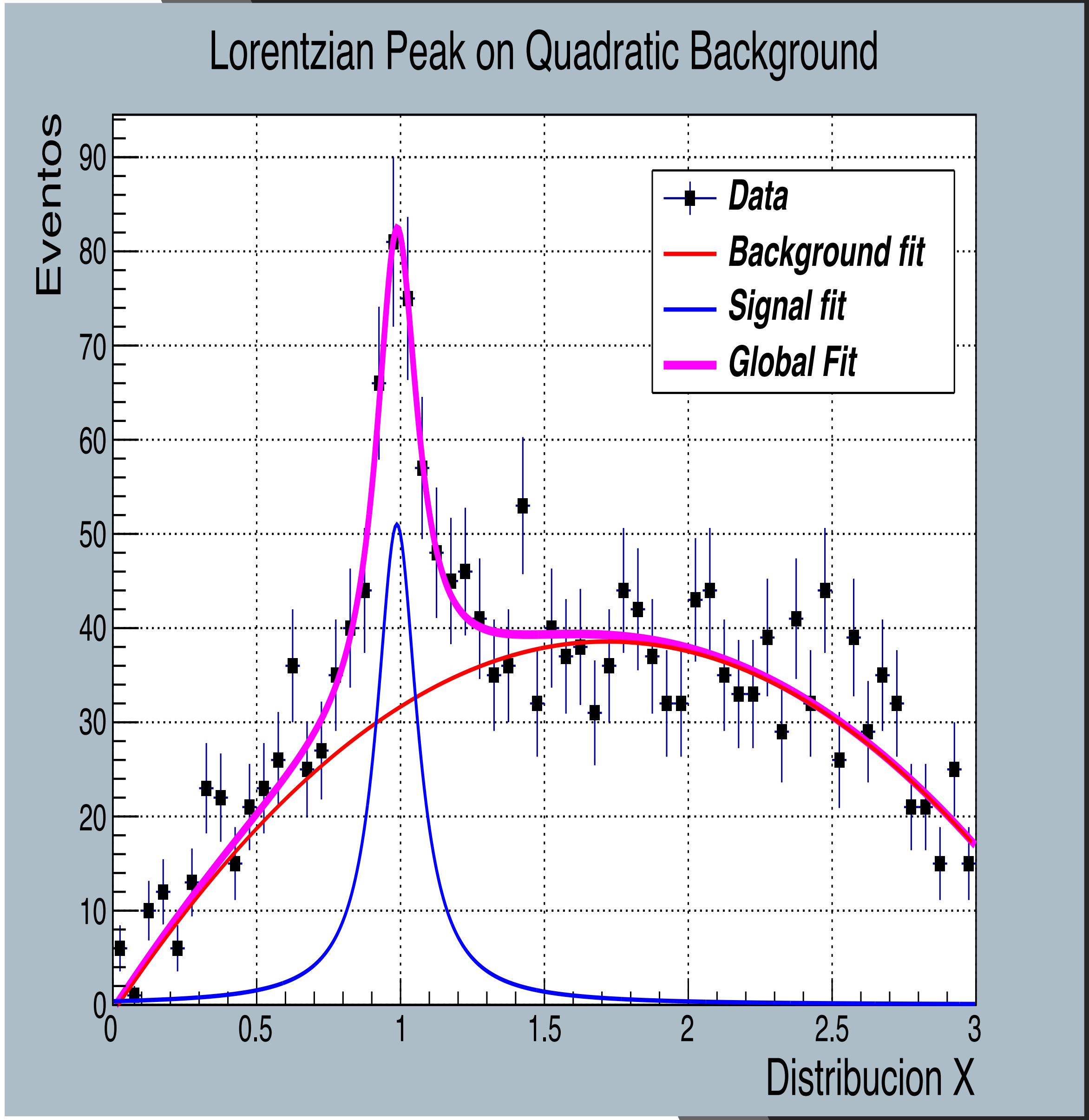
- Since the experimental cross section is almost 5 times the theoretical cross section, we cannot exclude this signal hypothesis.

Example from CMS



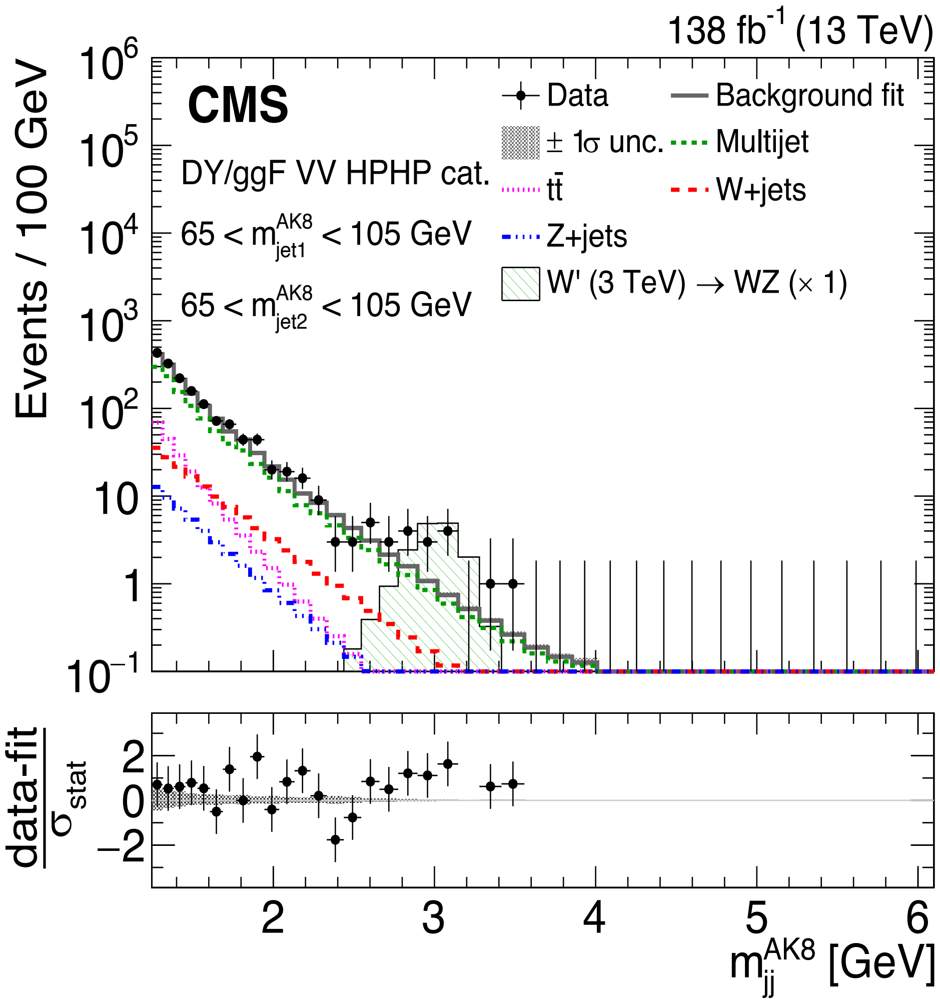
Second case: There is a significant excess of events with respect to the background prediction.

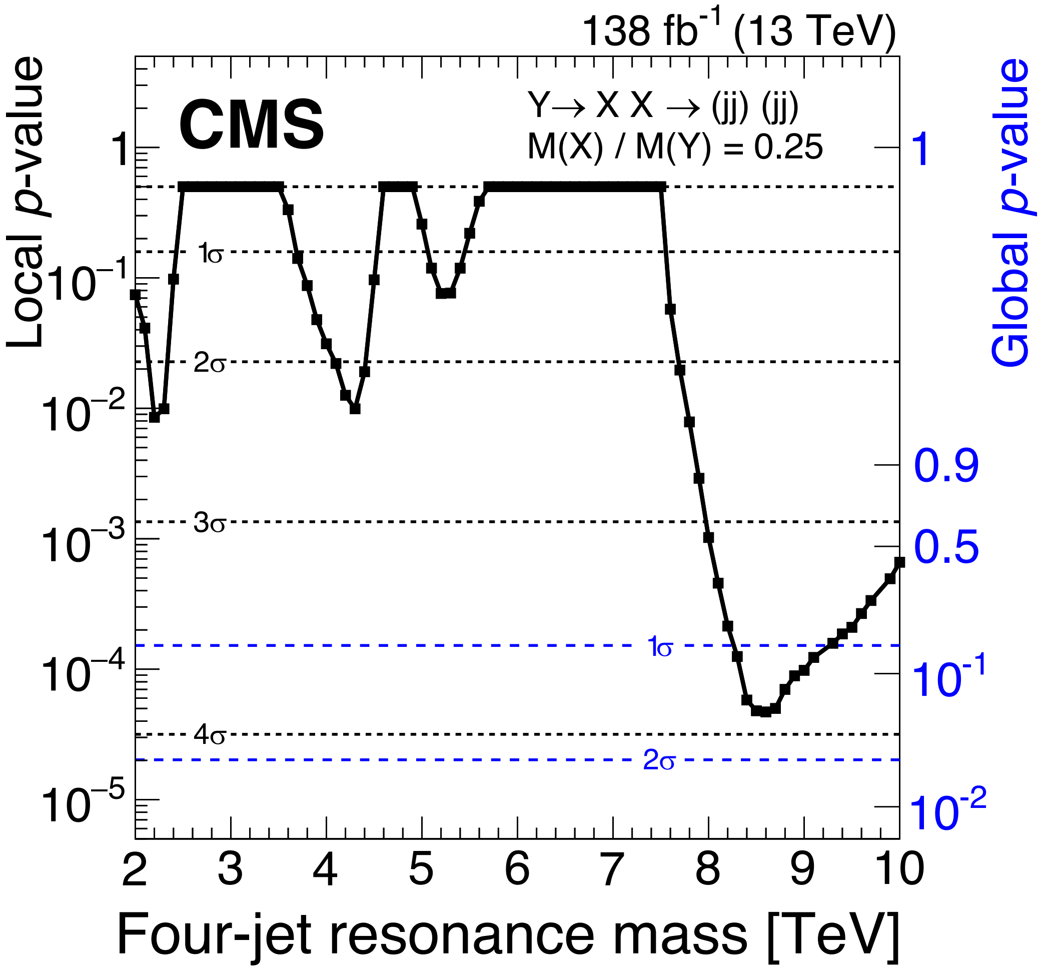




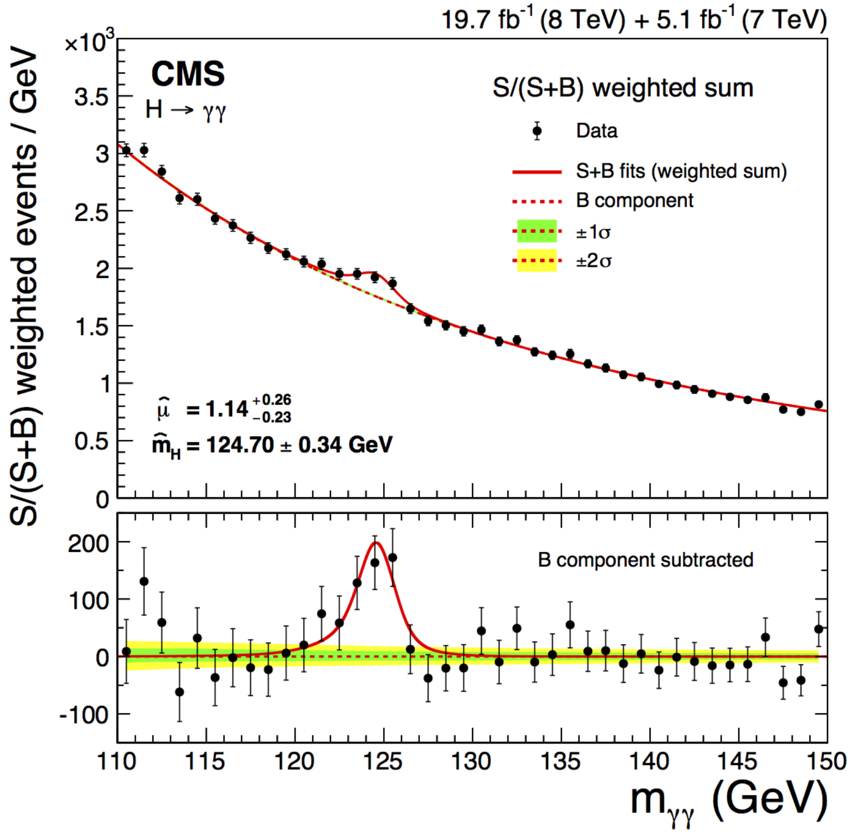
Now, we perform a fit in order to determine if our hypothetical signal agrees, both shape and rate, with the excess of events.

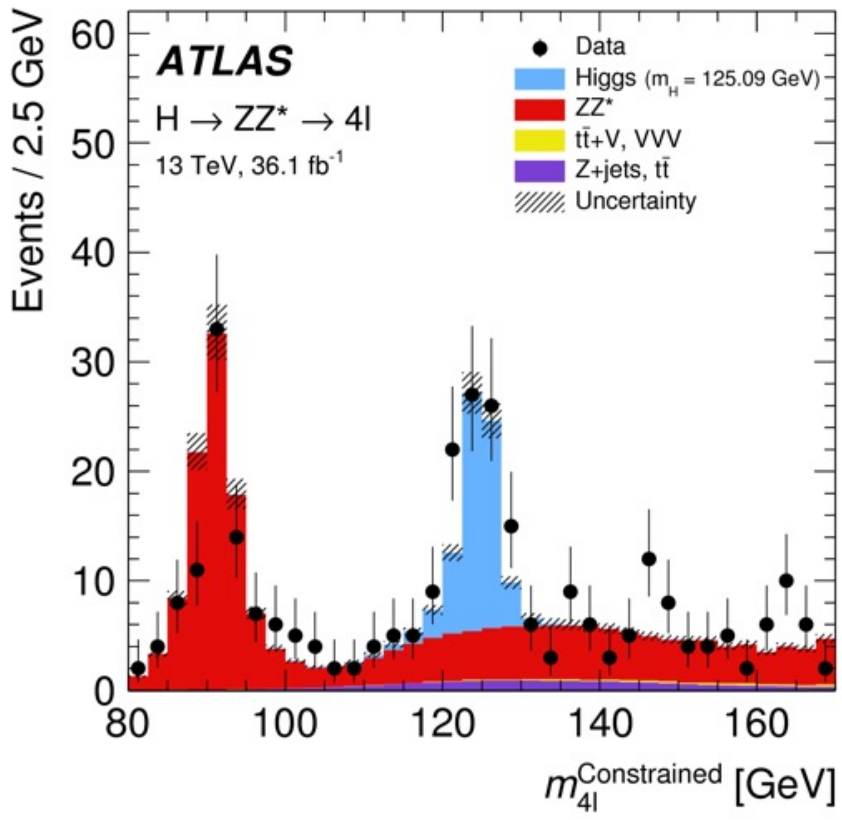
Example from CMS





The Higgs Boson!





Summary

- Understanding some fundamental statistical concepts is fundamental is particles physics: mean, variance, PDF….
- Some PDF are widely used in our field, and many others, and it is important to understand their similarities and differences: Gaussian, Poissonian, Lorentzian, etc.
- Understanding the difference between probability and likelihood is very important.
- **There are other important concepts I could not cover because of time, but I hope this material could be useful somehow for your careers….it was prepared with love and dedication for all of you!**

12/5/22

Thank you!

Andrés Flórez