The Logic of an EPP experiment

Go back to Rutherford and the logical steps of his experiment (slide 8)

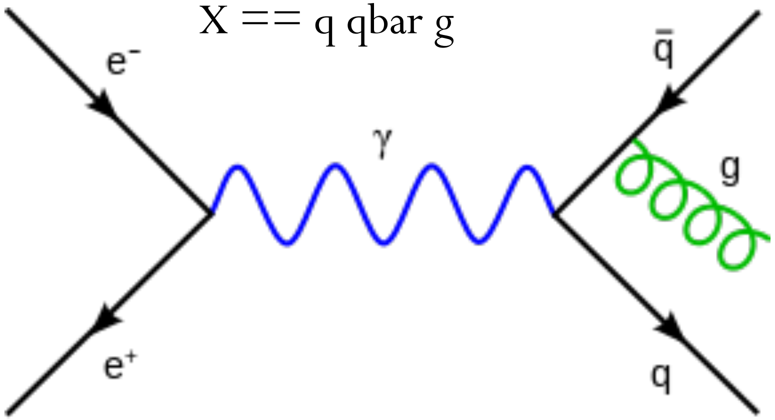
“Logic” of an EPP experiment - I

- Collision or decay: → **process to look at**
  - Initial state (proj. + target) OR (decaying particle);
  - Final state X = all particles produced
- Quadri-momentum conservation should always be at work
- In principle there is no need to measure ALL final state particles: a final state could be: → $\mu^+\mu^-$ + X (“inclusive” search)
- Possible final states:
  - a + b → a + b : elastic collision (e.g. pp→ pp)
  - a + b → X : inelastic collision (e.g. pp→ppπ^0)
- The experimentalist should set-up an experimental procedure to select the final state he/she searches. First of all he should be able **to count the number N_X of final states X**.

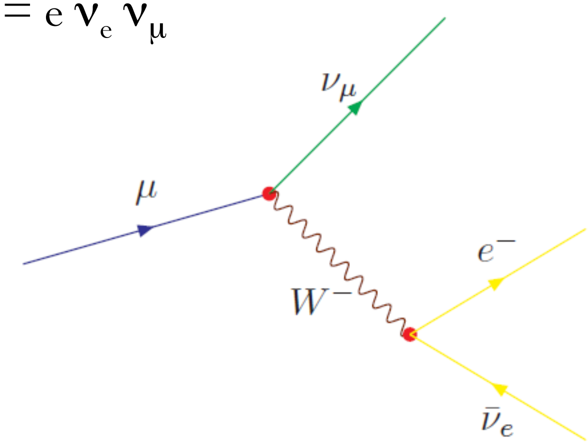
Why count ? – I

- Why count ?
- Because QFT based models allow to predict quantities (like ***cross-sections***, ***decay widths*** and ***branching ratios***, see later) that are proportional to "***how probable is***" a given final state.

Example of collision:
X == q qbar g



Example of decay:
X == e ν_e ν_μ



Why count ? – II

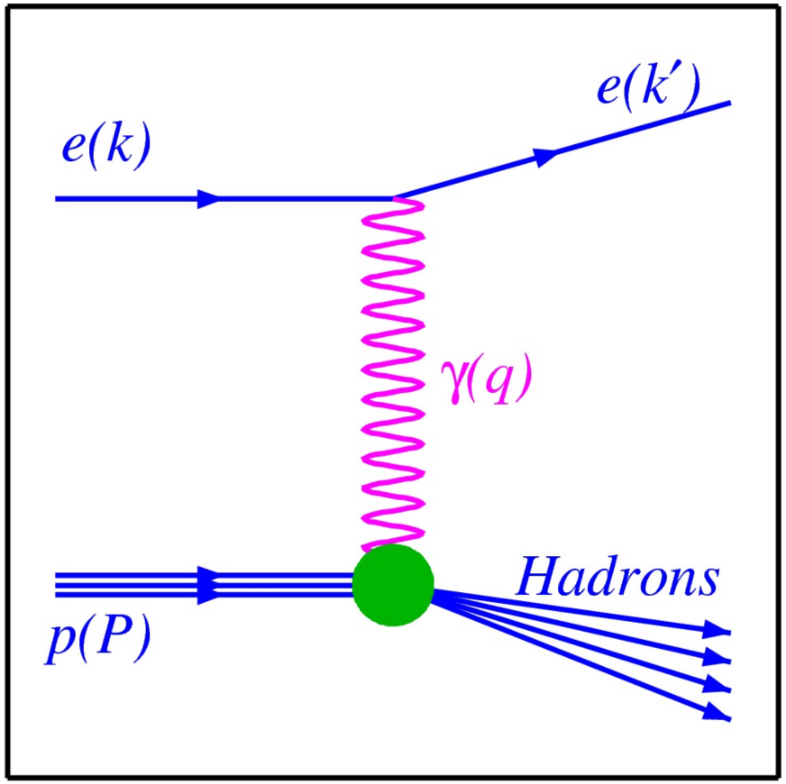
- Given a collision or a decaying particle you have several possibilities, several different final states.
- So: if I have produced N initial states (either a+b collisions or decaying particles), and out of them n times I observe the final state I am looking for, I can access this probability that should be $\approx n/N$
- Let me introduce the concept of **Event:**
  - The collection of all the particles of the final state from a single collision.
  - It is a collection of particles with their quadri-momenta.
  - Be careful not to overlap particles from different collisions.

Binomial or Poissonian ?

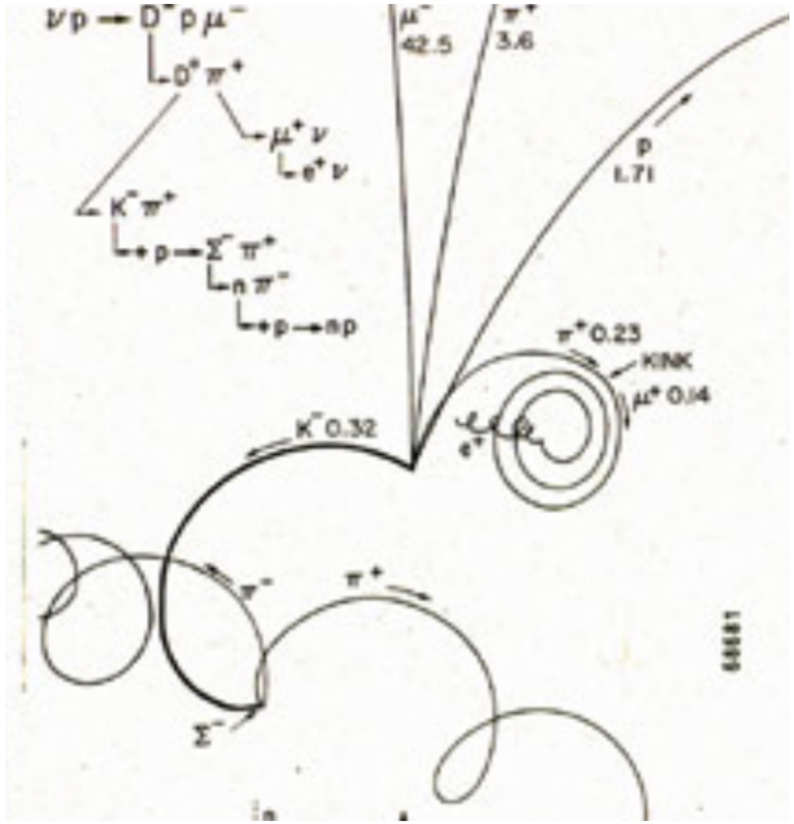
- ***N*** initial states prepared ***n*** final states observed → inference on ***p***. So binomial ? Yes BUT:
- ***N*** is not known exactly
- If $N \rightarrow \infty$ and $p \rightarrow 0$ → **n follows a poissonian distribution** (easy to prove)

Event: a "photo" of a collision/decay

Inclusive Event: measure the electron only



Exclusive Event: measure all particles to "close" the kinematics



"Logic" of an EPP experiment - II

- An ***ideal detector*** allows to measure the quadri-momentum of each particle involved in the reaction.
  - Direction of flight;
  - Energy $\boldsymbol{E}$ and/or momentum modulus $|\boldsymbol{p}|$;
  - Which particle is (e.g. from independent measurements of $\boldsymbol{E}$ and $|\boldsymbol{p}|$, $\boldsymbol{m}^2=\boldsymbol{E}^2-|\boldsymbol{p}|^2$) ➔ Particle ID
- BUT for a ***real detector***:
  - Not all quadri-momenta are measured: some particles are out of acceptance, or only some quantities are accessible, there are unavoidable **inefficiencies**;
  - Measurements are affected by **resolution**
  - Sometimes the particle nature is "confused"

"Logic" of an EPP experiment - III

- Selection steps:
  1. **TRIGGER SELECTION**
     - Retain only "interesting events": from bubble chambers to electronic detectors
     - ➔ "logic-electronic" eye: decides in a short time O(μs) if the event is interesting or not.
     - In some cases (e.g. pp), it is crucial since interactions are so probable…
     - LHC: every 25 ns is a bunch crossing giving rise to interactions: can I write 40 MHz on "tape" ? A tipical event has a size of 1 MB ➔ 40 TB/s. Is it conceivable ? And how many CPU will be needed to analyze these data ? At LHC from 40 MHz to 200 Hz ! Only one bunch crossing every 200000 !
     - "pre-scale" is an option
     - e^+e^-: the situation is less severe but a trigger is in any case necessary.

"Logic" of an EPP experiment - IV

2. **EVENT RECONSTRUCTION**: Once you have the final event sample, for each trigger you need to reconstruct at your best the kinematic variables.
3. **OFFLINE SELECTION**: choice of a set of discriminating variables on which apply one of the following:
   - cut-based selection
   - discriminating variables selection
   - multivariate classifier selection
4. **PHYSICS ANALYSIS**: analysis of the sample of ***CANDIDATES***

The selection strategy is a crucial part of the experimentalist work: defined and optimized using *simulated data samples*.

“Logic” of an EPP experiment - V

- Simulated samples of events: the Montecarlo.
  - “Physics” simulation: final state with correct kinematic distributions; also dynamics in some cases is relevant.
  - “Detector” simulation: the particles are traced through the detector, interactions, decays, are simulated.
  - “Digitization”: based on the particle interactions with the detector, signals are simulated with the same features of the data.
- ➔ For every interesting final state MC samples with the same format of a data sample are built. These samples can be analyzed with the same program. In principle one could run on a sample without knowing if it is data or MC.
- To design a “selection” strategy for a given searched signal one needs: ***signal MC samples*** and ***background MC samples***.

"Logic" of an EPP experiment - VI

- End of the selection: CANDIDATES sample N_{cand}
- Which relation is there between N_{cand} and N_X ?
  - ***Efficiency***: not all searched final states are selected and go to the candidates sample.(Trigger efficiencies are particularly delicate to treat.) Efficiency includes also the **acceptance.**
  - ***Background***: few other final states are faking good ones and go in the candidates sample.

$$\varepsilon N_X = N_{cand} - N_b$$

- where:
  - ε = efficiency ($0<\varepsilon<1$); $\boldsymbol{\varepsilon} = \boldsymbol{A} \times \boldsymbol{\varepsilon_d}$
  - N_b = number of background events
- Estimate ε and N_b is a crucial work for the experimentalist and can be done either using simulation (this is tipically done before the experiment and updated later) or using data themselves.

Quantities to measure

- In order to estimate N_X we need to measure:
  - N_{cand}
  - ε
  - N_b
- We already know that each of these variables have a fluctuation model:
  - N_{cand} is described by a Poisson process
  - ε is described by a Bernoulli process
  - N_b ??

N_{cand}: a Poisson variable

- If events come in a random way (without any time structure) the event count N is a Poisson variable.
- → if I count N, the best estimate of λ is N itself (or better $N+1$) and the uncertainty is $\sqrt{N}$ *(see previous lectures)*
- If N is large enough (N>20) Poisson → Gaussian. → $N \pm \sqrt{N}$ is a 68% probability interval for N.
- If N is small (close to 0) the Gaussian limit is not ok, a specific treatment is required (see later in the course).

Efficiency: a binomial variable - I

- Bernoulli process: success/failure ***N*** proofs, 0<***n***<***N***, ***p*** = success probability. ***p*** == ***ε***

$$P(n/N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

$$E[n] = Np$$

$$\mathrm{var}[n] = Np(1-p)$$

- Inference: given ***n*** and ***N*** which is the best estimate of ***p*** ?

And its uncertainty ? *(see previous lectures)*

$$\varepsilon = \hat{p} = \frac{n+1}{N+2}$$

$$\sigma(\varepsilon) = \frac{\sigma(n)}{N} = \frac{1}{\sqrt{N+2}}\sqrt{\hat{p}(1-\hat{p})}$$

Efficiency: a binomial variable - II

- How measure it ?
  - From data: Sample of N true particles and I measure how many, out of these give rise to a signal in my detector
  - From MC: I generate N_{gen} "signal" events. If I select N_{sel} of these events out of N_{gen}, the efficiency is (assume N_{gen} and N_{sel} large numbers):

$$\varepsilon = \frac{N_{sel}}{N_{gen}}$$

$$\sigma(\varepsilon) = \frac{\sigma(N_{sel})}{N_{gen}} = \frac{1}{\sqrt{N_{gen}}}\sqrt{\frac{N_{sel}}{N_{gen}}\left(1-\frac{N_{sel}}{N_{gen}}\right)}$$

Background N_b

- Simulation of N_{gen} "bad final states"; N_{sel} are selected. What about N_b ?
- We define the "rejection factor" $R = N_{gen}/N_{sel} > 1$
- We also need a correct normalization in this case: we need to know N_{exp} = total number of expected "bad final states" in our sample (N_{exp} related to luminosity and cross-section).

$$N_b = N_{sel}\frac{N_{\exp}}{N_{gen}} = \frac{N_{\exp}}{R}$$

$$\sigma(N_b) = \sigma(N_{sel})\frac{N_{\exp}}{N_{gen}} = \sqrt{N_{sel}}\frac{N_{\exp}}{N_{gen}} = \frac{N_{\exp}}{\sqrt{RN_{gen}}}$$

Statistical Errors

- In alla cases there is an unreducible error on N_X given by limited statistics. It is a random error, coming from the procedure of "sampling" that is intrinsic in our experiments.
- In all cases increasing the statistics, the error decreases

$$\frac{\sigma(N_{cand})}{N_{cand}} = \frac{1}{\sqrt{N_{cand}}}$$

$$\sigma(\varepsilon) \approx \frac{1}{\sqrt{N_{gen}}}$$

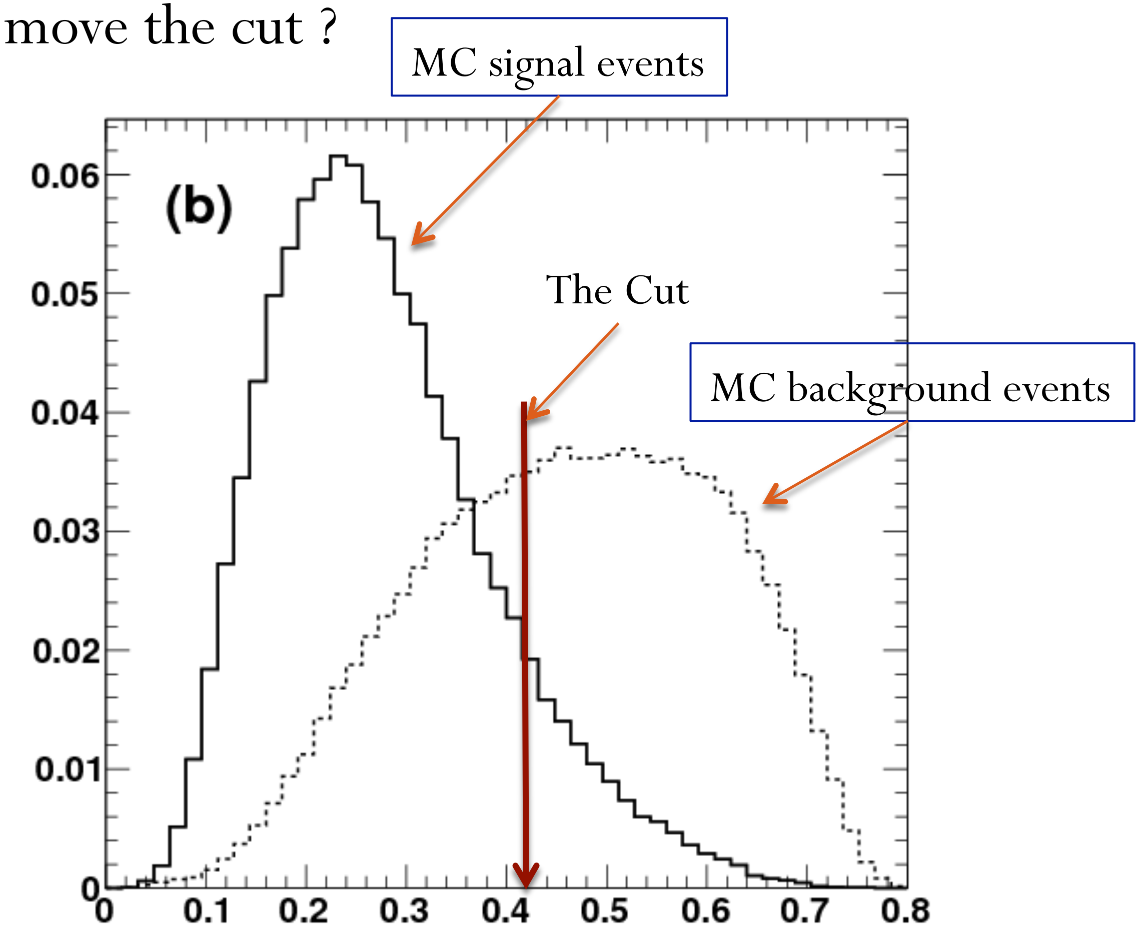
$$\sigma(N_b) \approx \frac{1}{\sqrt{N_{gen}}}$$

Summarizing

- N_{cand}: poissonian process ➔ the higher the better
- ε: binomial process ➔ high N_{gen} and high $\boldsymbol{\varepsilon}$
- N_b: normalized ≈poissonian process ➔ high R and high N_{gen}, low N_{exp}
- Moreover: unfortunately efficiency and background cannot be both improved simultaneously…

Efficiency vs. background

What happens if I move the cut ?



Efficiency-background relation

Example: selection of b-jets in ATLAS.

"b-jet" is the signal;

"light jet" is the background.

MC samples of ***b-jets*** and ***light-jets***

Application of 5 different ***selection recipes***

each with a "***free-parameter***".

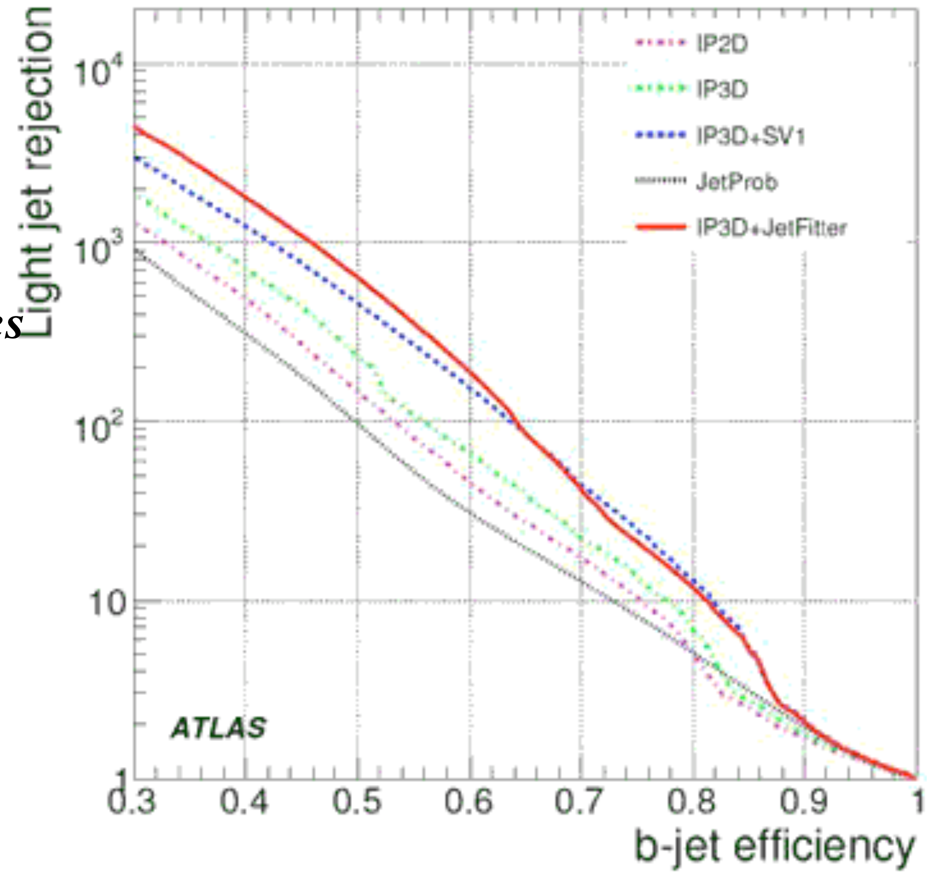
For each point I evaluate

- b-jet efficiency

  $= N_{sel}/N_{gen}$ (b-jet sample)

- light-jet rejection

  $= N_{gen}/N_{sel}$ (light-jet sample)



Choice of a working point, "compromise".

Unlucky situation: if you gain in efficiency you increase your bckg and viceversa…

Combining uncertainties

- Given the uncertainties on N_{cand}, ε and N_b, how can we estimate the uncertainty on N_X ?
- ➔ Uncertainty Propagation. General formulation *(see proof on blackboard)*

$$\left(\frac{\sigma(N_X)}{N_X}\right)^2 = \left(\frac{\sigma(\varepsilon)}{\varepsilon}\right)^2 + \frac{\sigma^2(N_{cand}) + \sigma^2(N_b)}{(N_{cand} - N_b)^2}$$

Assumption: three indipendent contributions

NB: if $N_{cand} \approx N_b$ the relative uncertainty becomes very large (the Formula cannot be applied anymore…

Can we say we have really observed a signal ???

Or we are simply observing some fluctuation of the background ?

Have we really observed the final state X ? - I

- We need a criterium to say ok, we have seen the signal or our data are compatible with the background.
- Which statistical uncertainty have we on N_X ?
  - Assume a Poisson statistics to describe N_{cand} negligible uncertainty on ε. We call (using more "popular" symbols):
  - $N = N_{cand}$
  - $B = N_b$
  - $S = N - B$

$$\left(\frac{\sigma(N_X)}{N_X}\right)^2 = \frac{\sigma^2(N) + \sigma^2(B)}{S^2} = \frac{N + \sigma^2(B)}{S^2}$$

$$\frac{N_X}{\sigma(N_X)} = \frac{S}{\sigma(S)} = \frac{S}{\sqrt{N + \sigma^2(B)}} = \frac{S}{\sqrt{S + B}}$$

Additional assumption: $\sigma^2(B) << N$

$\sigma(S)/S$ is the relative uncertainty on S, its inverse is "how many st.devs. away from 0" → $S/\sqrt{B}$ when low signals on top of large bck

Have we really observed the final state X ? - II

- This quantity is the "significativity" of the signal. The higher is $S/\sigma(S) = S/\sqrt{S+B}$, the larger is the number of std.dev. away from 0 of my measurement of S
  - $S/\sqrt{S+B} < 3$ probably I have not osserved any signal (my candidates can be simply a fluctuation of the background)
  - $3 < S/\sqrt{S+B} < 5$ probably I have observed a signal (probability of a background fluctuation very small), however no definite conclusion, more data needed. ➔ ***evidence***
  - $S/\sqrt{S+B} > 5$ observation is accepted. ➔ ***observation***
- NB1: All this is "conventional" it can be discussed
- NB2: $S/\sqrt{S+B}$ is an approximate figure, it relies on some assumptions *(see previous slide)*.

How to optimize a selection ? - I

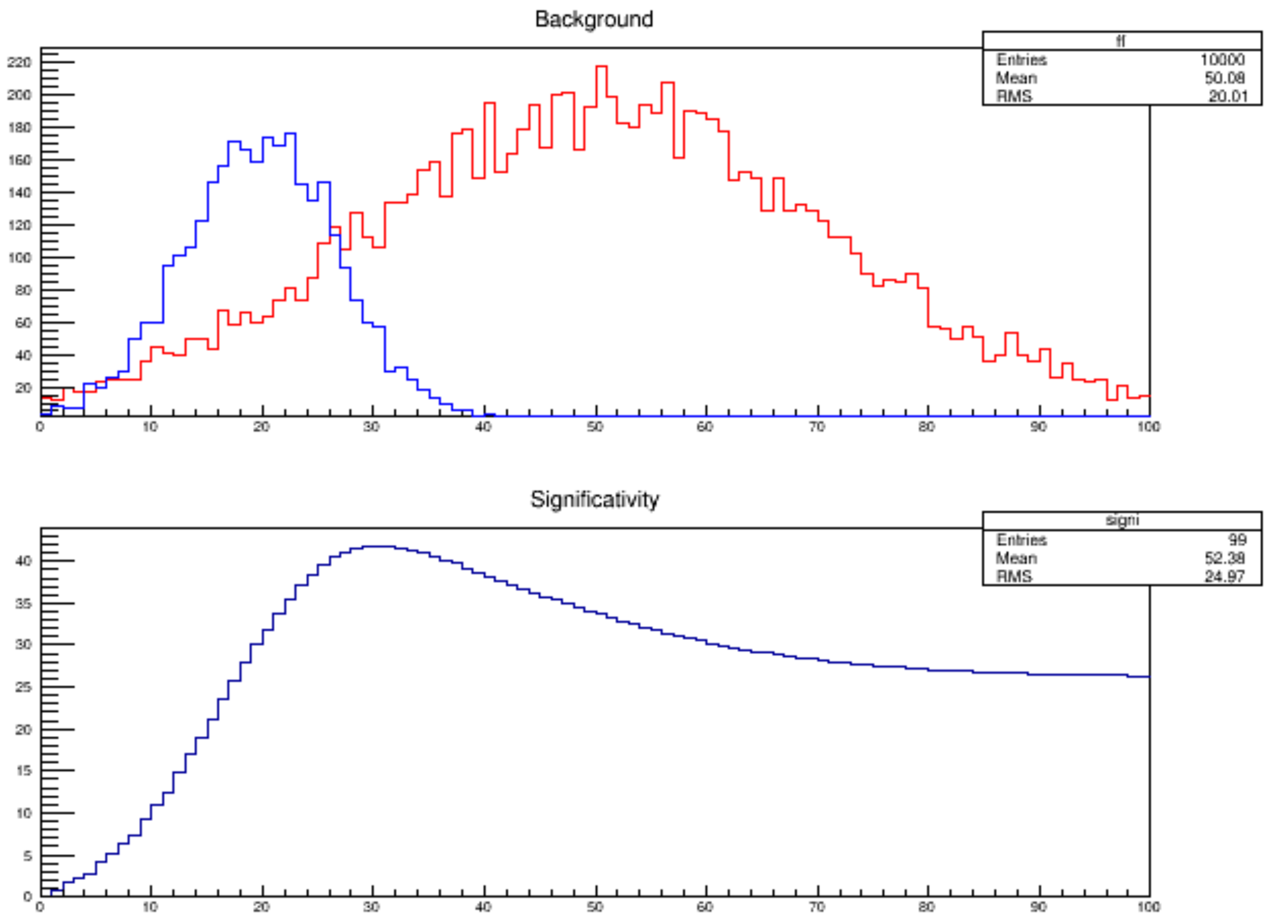
- The perfect selection is the one with
  - $\varepsilon = 1$
  - $N_b = 0$
- Intermediate situations ? Assume a given ε and a given N_b.

$$N_X = \frac{N_{cand} - N_b}{\varepsilon}$$

- By moving the cut we change each single ingredient. We want to see for which choice of the cut we get the lower statistical error on $\boldsymbol{N_X}$.
  - Again: if we assume a Poisson statistics to describe N_{cand} , negligible uncertainty on ε and on N_b we have to minimize the uncertainty on S=N_{cand}-N_b
  - S/sqrt(S+B) ≈ S/sqrt(B) is the good choice: the higher it is the higher is our sensitivity to the final state X. It is the "score function".

Example - I

B=10000
σ_x (B) = 15
S=3000
σ_x(S) = 5



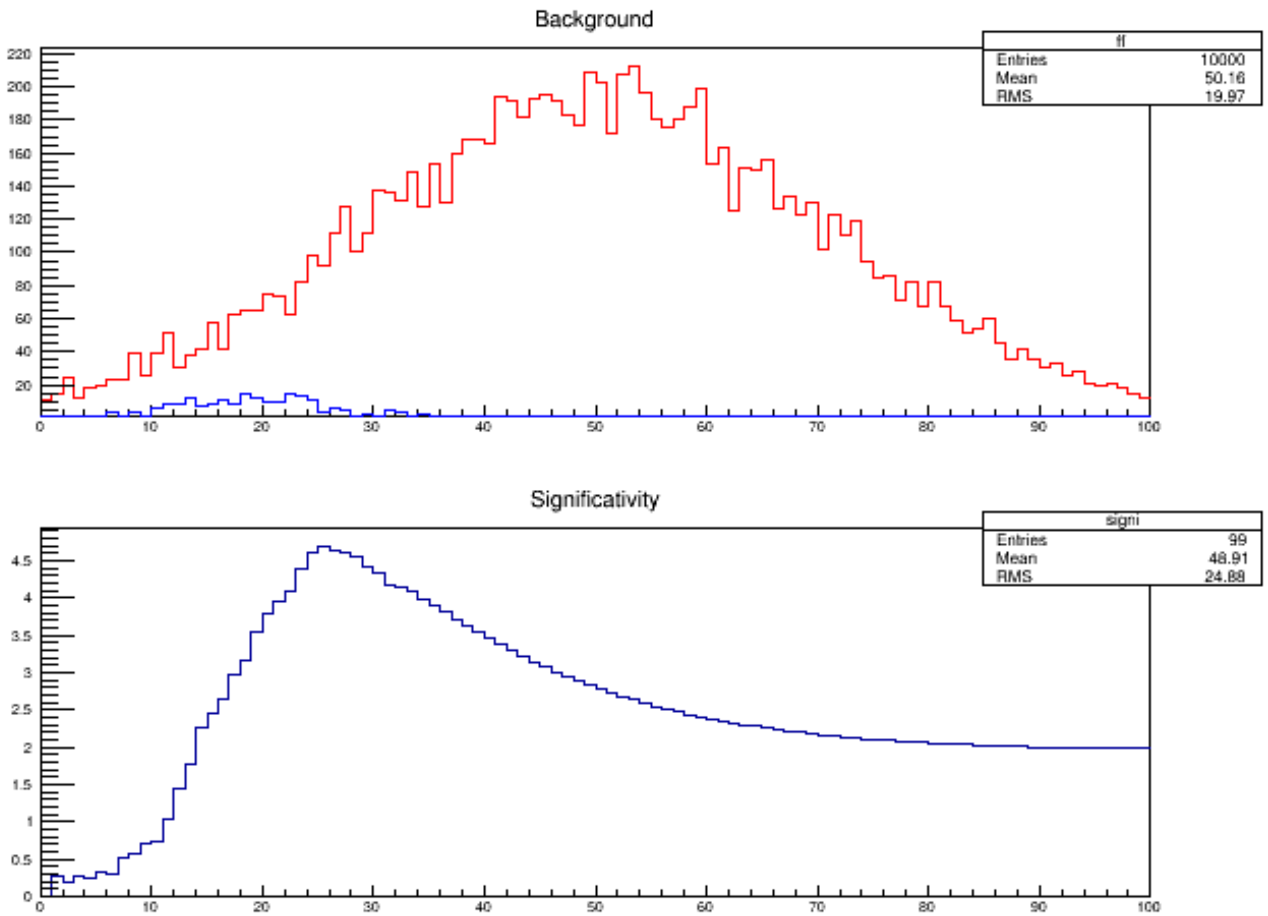
Example - II

B=10000

$\sigma_x\,(B) = 15$

S=200

$\sigma_x(S) = 5$



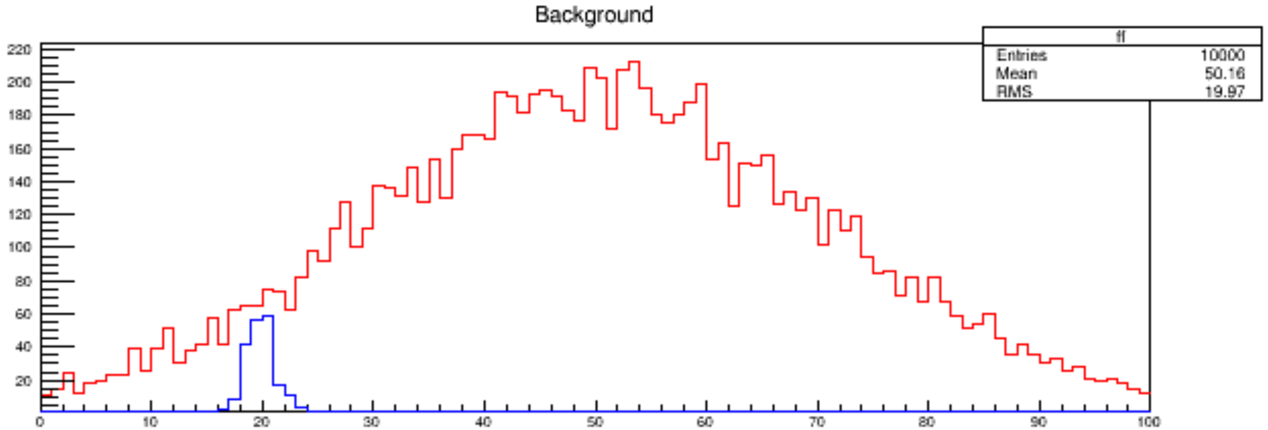
Example - III

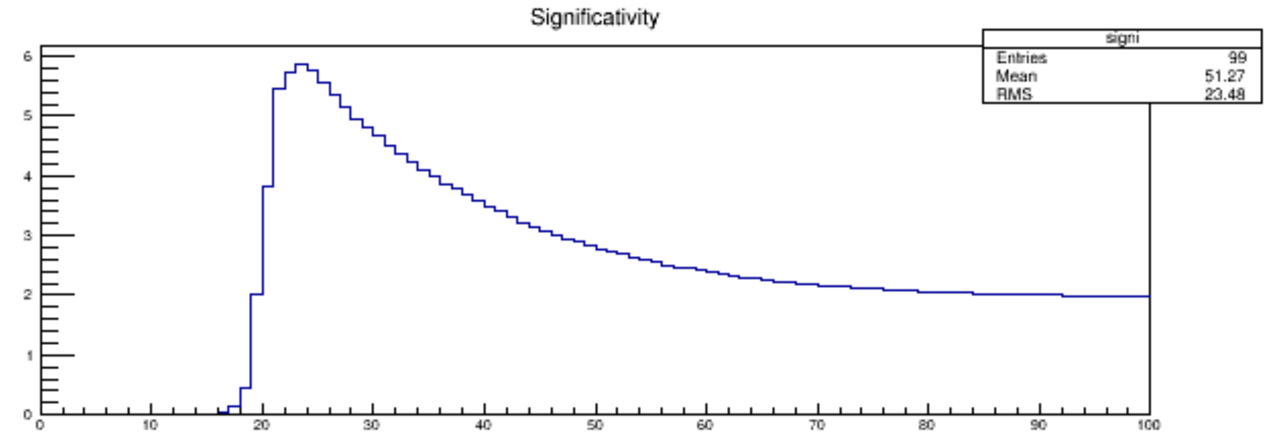
$B = 10000$

$\sigma_x(B) = 15$

$S = 200$

$\sigma_x(S) = 1$





Normalization

- In order to get quantities that can be compared with theory, once we have found a given final state and estimated N_X with its uncertainty we need to normalize to "how many collisions" took place.
- Measurement of:
  - Luminosity (in case of colliding beam experiments);
  - Number of decaying particles (in case I want to study a decay);
  - Projectile rate and target densities (in case of a fixed target experiements).
- Several techniques to do that, all introducing additional uncertainties (discussed later in the course).
- ***Absolute*** vs. ***Relative*** measurements.

The simplest case: rate measurement

- Rate: r = counts /unit time (normally given in Hz). We count N in a time Δt (neglect any possible background) and assume a Poisson process with mean λ

$$r = \frac{\lambda}{\Delta t} = \frac{N}{\Delta t} \pm \frac{\sqrt{N}}{\Delta t}$$

- NB: the higher is N, the larger is the absolute uncertainty on r but the lower the relative uncertainty.

$$\frac{\sigma(r)}{r} = \frac{1}{\sqrt{N}}$$

- Only for large N (N>20) it is a 68% probability interval.

Cosmic ray "absolute" flux

- Rate in events/unit surface and time
- My detector has a surface S, I take data for a time Δt with a detector that has an efficiency ε and I count N events (again with no background). The absolute rate r is:

$$r = \frac{N}{\varepsilon \Delta t S}$$

- Uncertainty: I combine "in quadrature" all the potential uncertainties.

$$\frac{\sigma(r)}{r} = \sqrt{\frac{1}{N} + \left(\frac{\sigma(\varepsilon)}{\varepsilon}\right)^2 + \left(\frac{\sigma(\Delta t)}{\Delta t}\right)^2 + \left(\frac{\sigma(S)}{S}\right)^2}$$

- Distinction between "***statistical***" and "***systematic***" uncertainty

Combination of uncertainties

- Back to the previous formula.

$$\frac{\sigma(r)}{r} = \sqrt{\frac{1}{N} + \left(\frac{\sigma(\varepsilon)}{\varepsilon}\right)^2 + \left(\frac{\sigma(\Delta t)}{\Delta t}\right)^2 + \left(\frac{\sigma(S)}{S}\right)^2}$$

1. Suppose we have a certain "unreducible" uncertainty on S and/or on $\boldsymbol{\varepsilon}$ (the uncertainty on Δt we assume is anyhow negligible..). Is it useful to go on to take data ? Or there is a limit above which it is no more useful to go on ?
2. Suppose that we have a limited amount of time to take data N is fixed: is it useful to improve our knowledge on $\boldsymbol{\varepsilon}$?

Not only event counting

- Once the candidate sample is obtained many quantities can be measured (particle properties, e.g. particle mass).
- BUT in most cases they are obtained from a **FIT** to a data distribution. So, you divide events in bins and extract the quantity as a ***fit parameter*** ➔ the event counting is still one major source of uncertainty ➔ the uncertainty on the parameter depends on the statistics $\approx \sqrt{N_i}$.
- Example:
  - Measure the mass of a "imaginary" particle of M=5 GeV.
  - Mass spectrum, gaussian peak over a uniform background
  - FIT in three different cases: 10^3, 10^4 and 10^5 events selected

Mass uncertainty due to statistics

Observations:

- → Poissonian uncertainty on each bin
- → Reduce bin size for higher statistics
- → Fit function = A+B*Gauss(M)
- → Free parameters: A,B,M (fixed width)
- → The fit is good for each statistics

Results

$N=10^3$ events:

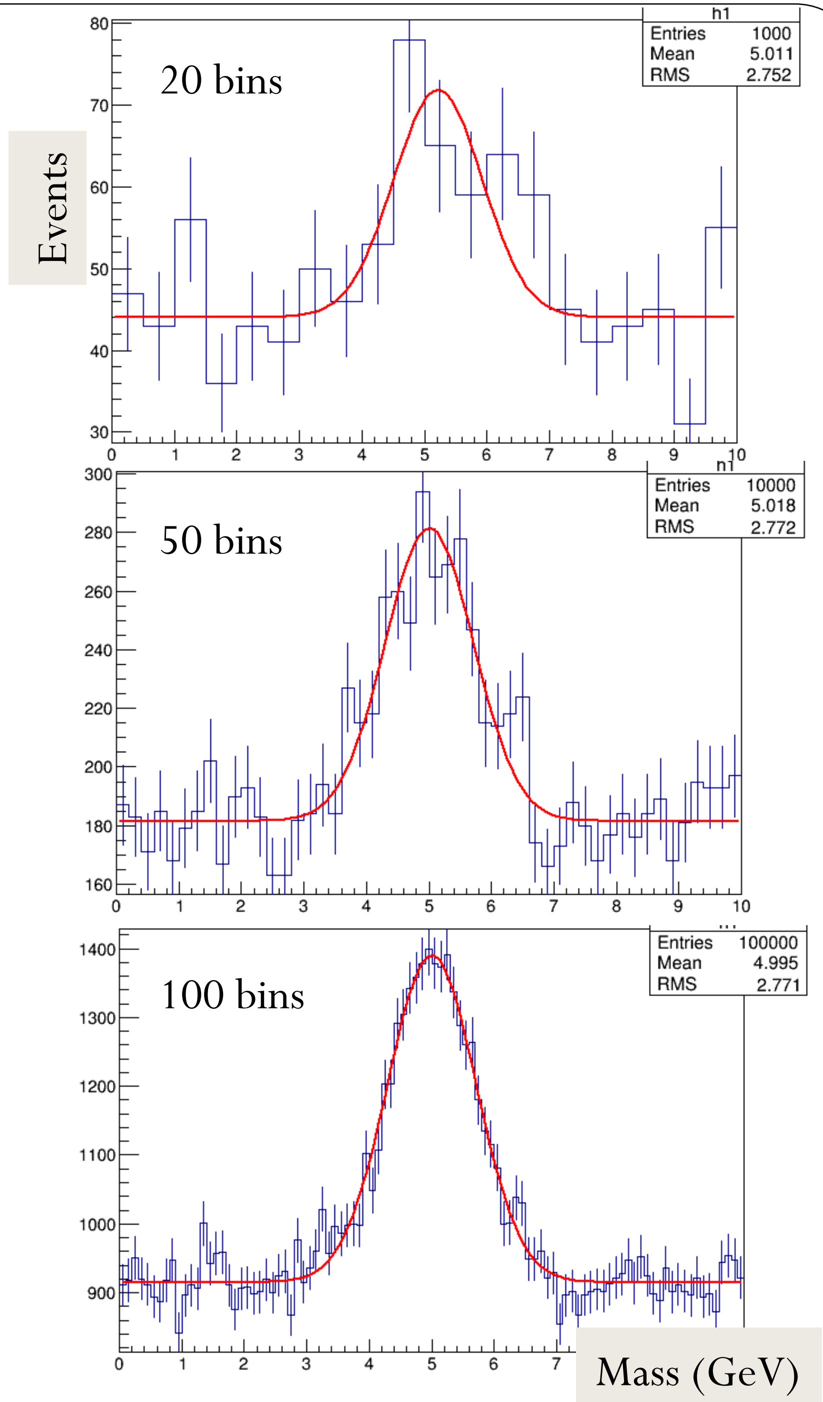
Mass = 5.22±0.22 GeV, χ^2 = 28 / 18 dof

$N=10^4$ events:

Mass = 5.01±0.06 GeV, χ^2 = 38 / 48 dof

$N=10^5$ events:

Mass = 5.02±0.02 GeV, χ^2 = 83 / 98 dof



Where could be a systematic uncertainty here ?

- Absolute mass scale: this can be measured using a candle of known mass. Not always it is available. e.g. Z for the Higgs mass at the LHC.
- Mass resolution: in most cases the width of the peak is given by the experimental resolution that sometimes is not perfectly gaussian, giving rise to possible distortion to the curve.
- Physics effects: knowledge of the line-shape, interference with the background…
- In general: M = central value ± stat.uncert. ± syst.uncert.

Uncertainty combination

central value ± **stat.uncert.** ± **syst.uncert.**

Can we combine stat. and syst. ? If yes how ?

The two uncertainties might have different probability meaning: typically one is a gaussian 68% C.L., the other is a “maximum” uncertainty, so in general it is better to hold them separate.

If needed better to add in quadrature rather than linearly.

Summarizing

- Steps of an EPP experiment (assuming the accelerator and the detector are there):
  - Design of a **trigger**
  - Definition of an offline **selection**
  - **Event counting** and **normalization** (including **efficiency** and **background** evaluation)
  - **Fit** of "candidate" distributions
- Uncertainties
  - Statistical due to Poisson fluctuations of the event counting
  - Statistical due to binomial fluctuations in the efficiency measurement
  - Systematic due to non perfect knowledge of detector effects.