The Logic of an EPP experiment

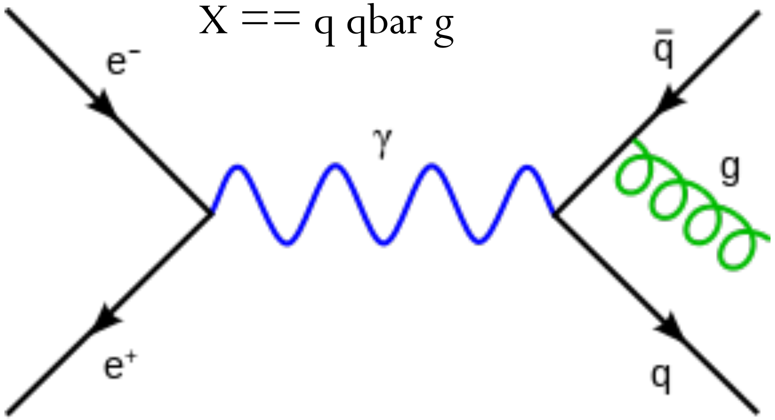
“Logic” of an EPP experiment - I

- 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 “**events”** with the final state he is searching, in such a way to count the number N_X of final states X.

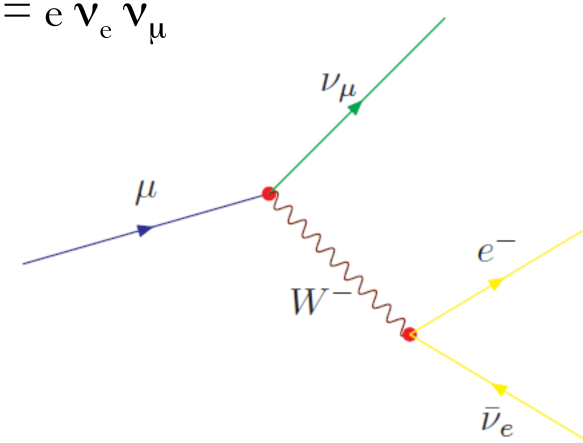
Why count events ?

- Why count events ?
- 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

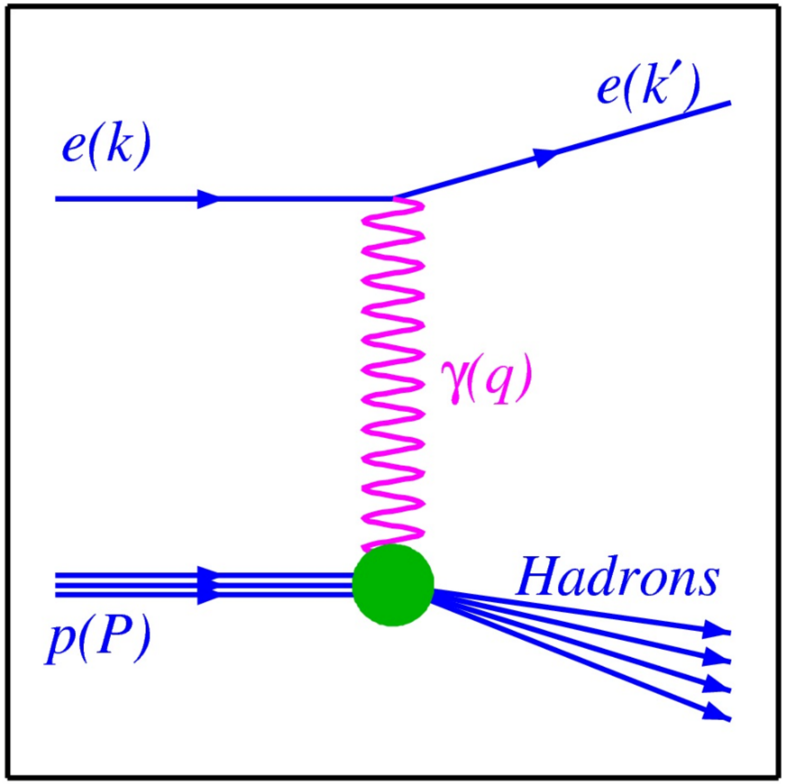


Example of decay:
X == e ν_e ν_μ

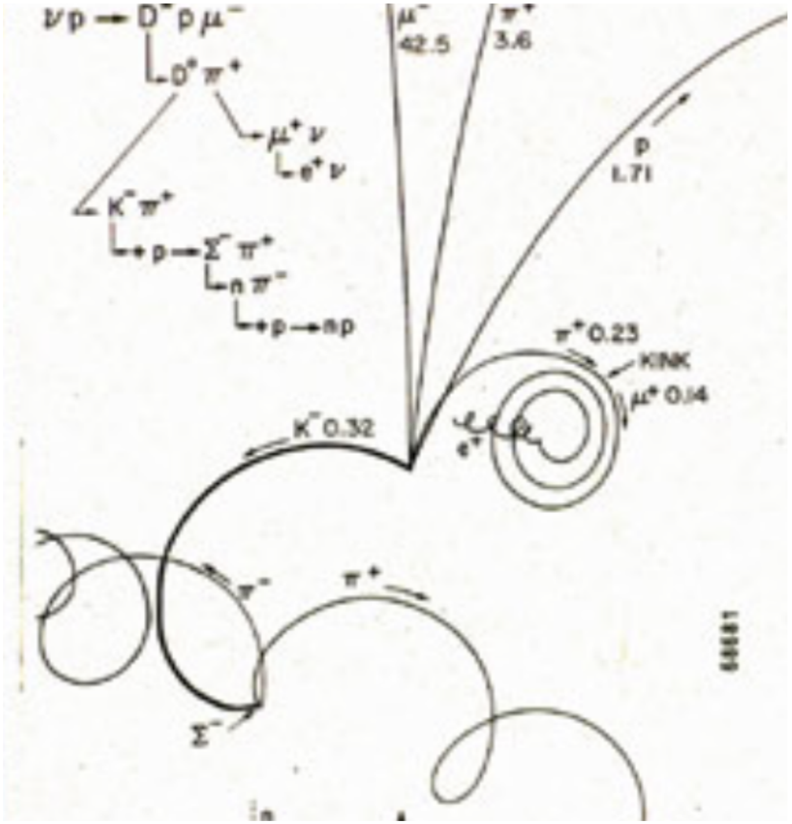


Event: a "photo" of a collision/decay

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 E and/or momentum modulus $|p|$;
  - Which particle is (e.g. from independent measurements of E and $|p|$ $m^2=E^2-|p|^2$) ➔ 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, **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 ok 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…

"Logic" of an EPP experiment - IV

2. OFFLINE selection: choice of a set of discriminating variables
   - cut-based selection
   - discriminating variables selection
   - multivariate classifier selection
3. 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 we have MC samples with the same format of a data sample. It 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.)
  - ***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 ($0<\varepsilon<1$)
  - 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) or using data themselves.

Event counting: a Poisson variable - I

- What is a Poisson process ? If we can divide the time in small intervals δt_i such that
  - $p(n=1, \delta t_i) << p(n=0, \delta t_i)$
  - $p(n=1, \delta t_i) = \alpha \, \delta t_i$
  - $p(n=1, \delta t_i)$ uncorrelated with $p(n=1, \delta t_{i+1})$
- If this happens, given a time interval Δt the number of counts *n* follows the Poisson distribution (λ is the only parameter)

$$p(n/\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

$$E[n] = \lambda$$

$$\mathrm{var}[n] = \lambda$$

Event counting: a Poisson variable - II

- 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 and the uncertainty is $\sqrt{N}$
- 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 tretment 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 ?

$$\varepsilon = \hat{p} = \frac{n}{N}$$

$$\sigma(\varepsilon) = \frac{\sigma(n)}{N} = \frac{1}{\sqrt{N}}\sqrt{\frac{n}{N}\left(1-\frac{n}{N}\right)}$$

Efficiency: a binomial variable - II

- So: I generate N_{gen} "signal" events. If I select N_{sel} of these events out of N_{gen}, the efficiency is:

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

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

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.

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

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

Summarizing

- N_{cand}: poissonian process ➔ the higher the better
- ε: binomial process ➔ high N_{gen} and high e
- N_b: normalized ≈poissonian process ➔ high R and high N_{gen}, low N_{exp}
- Moreover: unfortunately efficiency and background are correlated…

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***".

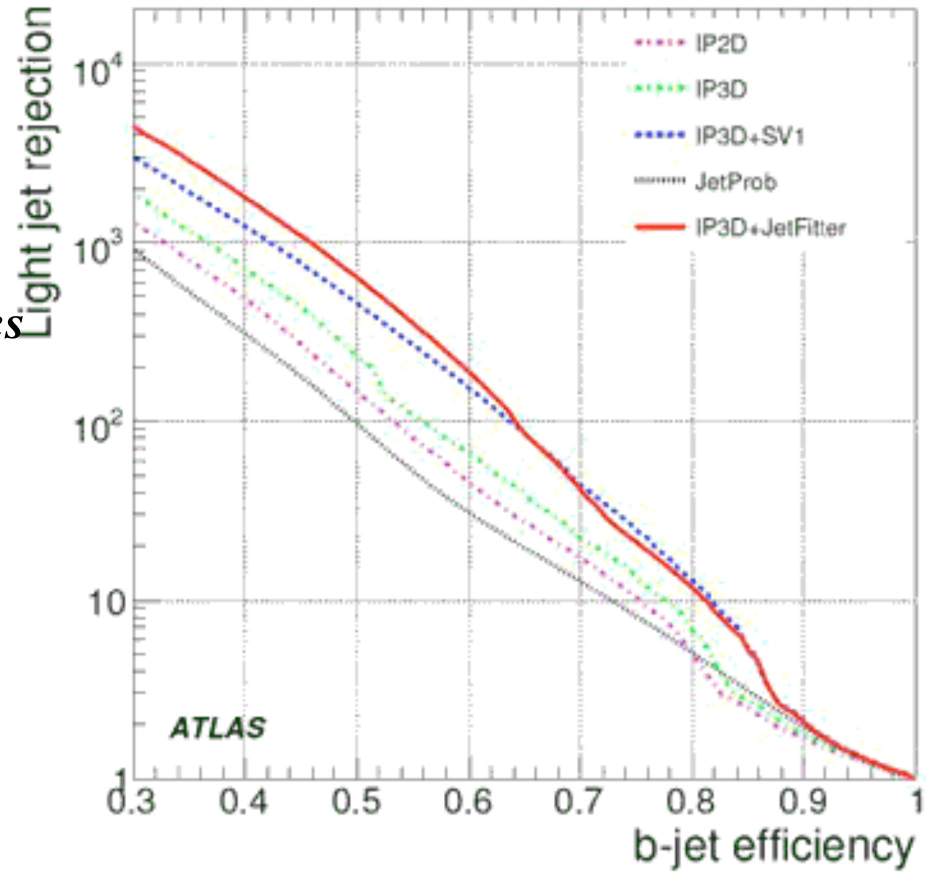
For each point I evaluate

- 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…

How to optimize a selection ? - I

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

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

- Which statistical uncertainty have I on N_X ?
  - Assume a Poisson statistics to describe N_{cand} negligible uncertainty on ε and on N_b. Minimize the uncertainty on N_{cand}-N_b. We call:
  - $N = N_{cand}$
  - $B = N_b$
  - $S=N-B$

$$\sigma^2(S) = \sigma^2(N) + \sigma^2(B) = N + \sigma^2(B) \approx N$$

$$\frac{S}{\sigma(S)} = \frac{S}{\sqrt{S+B}} \approx \frac{S}{\sqrt{B}}$$

How to optimize a selection ? - II

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

How to optimize a selection ? - III

- Let's define a more elaborated "score function" (widely used today).
  - Counting experiment: *S* and *B* are expected values of signal and background, *N* is my count. We evaluate the likelihoods in the hypothesys of *S*+*B* and only *B* and take the "likelihood ratio"

$$L(S+B) = \frac{e^{-(S+B)}(S+B)^N}{N!}$$

$$L(B) = \frac{e^{-B}B^N}{N!}$$

$$\Re = \frac{L(S+B)}{L(B)} = \frac{e^{-S}(S+B)^N}{B^N} = e^{-S}\left(1+\frac{S}{B}\right)^N$$

$$-2\log\Re = 2S - 2N\log\left(1+\frac{S}{B}\right)$$

  - Suppose now to count N=S+B and take the square root of the -2logR evaluated above:

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

  - This the so called "score function": significativity of the signal hypothesys.

Normalization

- In order to get quantities that can be compared with theory, once we have found a given final state and counted the number of events, 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.
- ***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 8again with no background). The absolute rate r is:

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

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

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

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

Not only event counting

- Many quantities are 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.
- 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:

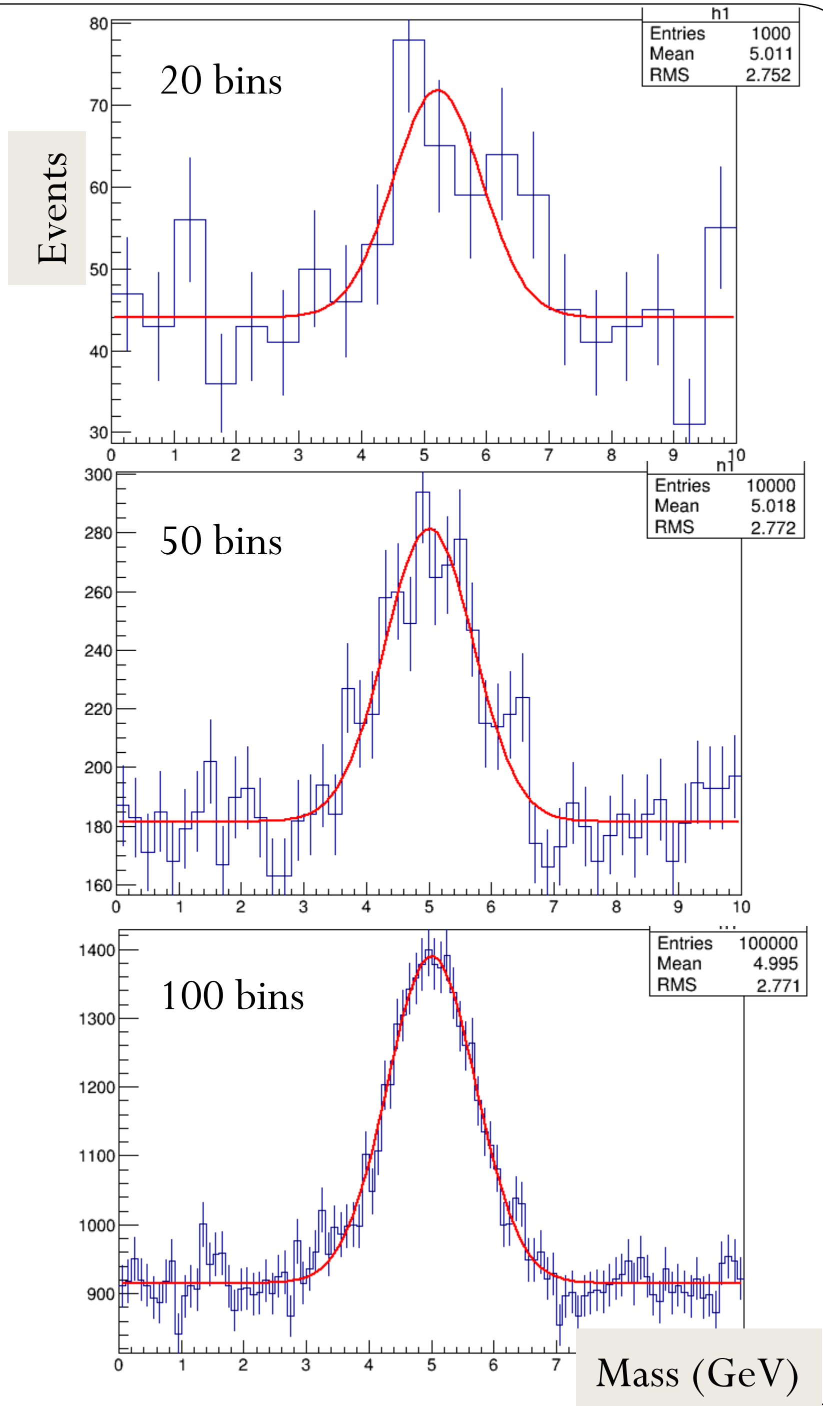
Mass = 5.22±0.22 GeV, χ^2 = 28 / 18 dof

$N=10^4$ 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



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.