Introduction to statistics

Summer internship
LIP – 19/07/2017

Pedrame Bargassa
CMS - LIP

Disclaimer & Goal:

- 30 minutes: Way too short to give a course on statistics
  - Covering important notions only, without demonstration
- Goal: Cover basic concepts that you might come across during the internship
- Courses on statistics:
  - “Statistics for Nuclear and Particle Physicists” Pr. Louis Lyons

Outline:

- Experimental error
- Distribution & Probability / Sample of events
- Important distributions – Central Theorem
- Error propagation
- Hypothesis testing: The χ^2 example

Experimental error

<u>Any</u> experimentally measured quantity has an uncertainty

Reflecting the precision of the measure

<u>Example</u>:

The speed of light is c = 2.99792458×10^8 m/s

A new experiment gives c = $(2.9900 \pm \sigma) \times 10^8$ m/s

- σ = 0.01: New result is consistent with previous result
- σ = 0.001: New result is inconsistent with previous result
  - Either: We have made a new discovery
  - Or: Either the new value or error is wrong
- σ = 1.0: The new result is irrelevant

Experimental error

2 types of experimental errors:

- **Random/Statistical**: Inability to measure w infinite accuracy
  - Opinion polls, counting radioactive decays
- **Systematic**: "In the nature of our measure": Often points to mis-calibration of device, mis-calibration that we must measure & include in our final result
  - We know that 100 atoms of Cesium decayed. We measure only 98 decay products: Systematic uncertainty of 2% specific to measuring device

Distribution & Probability

Distribution n(x): Describes how often a value of the variable x occurs in a definite sample of Data

Range	x variable	n(x)
[0,7]	Number of days in 1 week	N(Sunny days)
[-13.6,0] eV	Energy states of H atom	N(Atoms w electrons w E=x @ 10K)
[0,∞[	Hours to understand stats	N(Person having understood after x hours)

Probability p(x): That with sample of N measurements, the value x is obtained n_x times

$p(x) = \lim_{N\to\infty} (n_x/N)$ $\qquad p(x) \in [0,1]$

Distributions/Probabilities are characterized mainly by 2 quantities:

Mean/Expectation value: $E(x) = \int x\, p(x)\, dx$

Variance: $\sigma^2(x) = \int (x - E(x))^2\, p(x)\, dx$

During estagio:

- Most of the time
  - p_T, E, (r,ϕ,η) of a reconstructed particle
- If you use root: "E(x)" & σ will be given to you
  - But now you know what they correspond too :-)

Sample of events

For a set of N separate measurements of $x=\{x_1,\ldots x_n\}$, how can we estimate the expectation value & variance ?

$\bar{x} = (1/N)\, \Sigma_i\, x(i)$: Nothing else than $E(x) = \int x\, p(x)\, dx$ for a discrete case

where: $p(x)=p=1/N$

$\sigma^2(\bar{x}) = \sigma^2(x) / N$:

Each of the measurements x has an uncertainty, but...

The more measurement we will have, the preciser the mean of all measurements will be

During estagio:

- The more your sample has events, the preciser your relative precision will be
- Relative uncertainty: $r = \sigma(\bar{x}) / N$
  - Let's assume that: $\sigma(\bar{x}) = \sqrt{N}$ (Poisson, see next slides)

    Then: $r = 1/\sqrt{N}$

Important distributions

Binomial:
When we have 2 possible outcomes of the experience

Probability: of having
m success
out of n trials
with p: Probability of success
q: Probability of failure p+q=1

$$p_n(m) = C^m_n \, p^m \, q^{m-n}$$

$$C^m_n = n!/m!(n-m)!$$

Expectation value: $E(m) = n\,p$

Variance: $\sigma^2 = n\,p\,q$

Examples: Tossing a coin & looking for e.g. heads
Detecting a produced particle or not: Can be used for calculating efficiency & its uncertainty

Important distributions

Poisson: When the probability of observing an event is small

Probability: $\lim_{N\to\infty} P_n(m) = \mu^m/m!$

μ = Mean counted events

Expectation value: $E(m) = \mu$

Variance: $\sigma^2 = \mu$

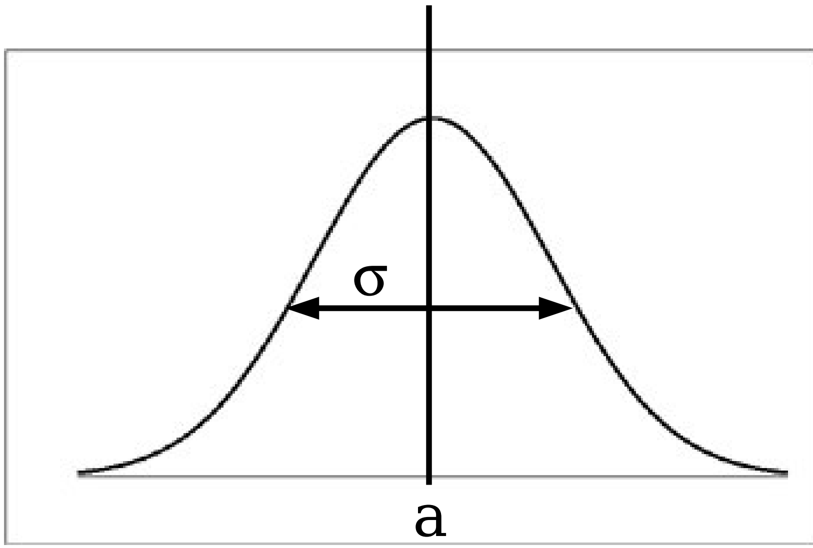
Example: Consider the very large number of radioactive nuclei. The probability that one of the nuclei decays within time interval Δt follows the Poisson distribution

During estagio:

- When you count N events in a sample: The uncertainty on this number is $\sqrt{N}$

Important distributions

Gaussian:



Distribution:
$G(x) = [1 / \sqrt{2\pi}\, \sigma] \exp[- (x-a)^2 / 2\sigma^2]$

Expectation value: By definition: a

Variance: By definition: σ^2

Example: If the measure of a variable x (mean & variance) is well done, then the pull = [x(measured) – x(true) / σ] follows a gaussian of $a=0$ & $\sigma=1$

During estagio:

- You might come across this distribution/probability when, i.e. fitting the resolution of a detector
- See next slide ;-)

Central limit theorem

If $x=\{x_1,\ldots x_n\}$ are a set of n independent variables all following an arbitrary distribution with mean a and variance σ^2, then in the limit $n\rightarrow\infty$, their arithmetic mean $\bar{x} = (1/n)\ \Sigma_i\ x(i)$ follows a Gaussian distribution with mean a and variance σ^2/n

Error propagation

Imagine you are calculating an experimental quantity f which is a function of 2 numbers l & j
l: Counted number of reconstructed leptons
j: Counted number of reconstructed jets

What is the uncertainty on f ?

$$\sigma^2(f) = (\delta l/\delta f)^2\, \sigma^2(l) + (\delta j/\delta f)^2\, \sigma^2(j) + 2(\delta l/\delta f)(\delta j/\delta f)\, \mathrm{cov}(l,j)$$

We are counting: Poisson is the uncertainty to take into account: $\sigma^2(l) = l$; $\sigma^2(j) = j$

cov(l,j): Covariance between l & j:
Measure of “how much l/j moves when j/l moves”
$\mathrm{cov}(l,j) = \rho(l,j)\ .\ \sigma(l)\ .\ \sigma(j)$ with $\rho(l,j) \in [0,100]\%$

Hypothesis testing: The χ^2 example

Imagine that we have $i \in [1,N]$ measurements $(O(i),\sigma(i))$ as a function of a variable x
We want to know the best hypothesis representing these observations, i.e. the function best fitting N observed Data

Finding a function f to test isn't a problem. The real question:
How can-I quantitatively test the goodness of my hypothesis ?

$$\chi^2 = \Sigma_i \left[(f(i) - O(i))^2 / \sigma(i)^2\right]$$

Accounts for: Each & full observation point $(O(i),\sigma(i))$
The tested hypothesis f

If the hypothesis is reasonably good:
$\forall\, i \quad f(i) - O(i) \sim 1\, \sigma(i) \rightarrow$ **χ^2/N(Degrees freedom) ~ 1**

During estagio:
- root does this for you but now you know what does it correspond to :-)

Closing words

Always keep in mind that any quantity you measure, whatever the method of the measure, has an uncertainty

If you have stat problems and/or there are related points you would like to discuss: Let's discuss together, with your supervisor(s) bargassa@cern.ch

Wish you an interesting internship !