ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

INFN
Istituto Nazionale
di Fisica Nucleare

Analisi Statistica dei dati nella Fisica Nucl. e Subnucl. [Laboratorio 💻]

Gabriele Sirri
Istituto Nazionale di Fisica Nucleare

2015.01.09

Sommario

- Ancora formalità…
- 💻 Discussione (Esercizio 1)
- Introduzione a RooFit
- 💻 Home work (Esercizio 2)

Ancora Formalità …

Calendario

- Lunedì 23 febbraio 2015 14-16 M. Sioli

MARZO

- Lunedì 2 marzo 2015 14-16 M. Sioli
  Giovedì 5 marzo 2015 11-13 T. Chiarusi
- Lunedì 9 marzo 2015 14-16 M. Sioli
  Giovedì 12 marzo 2015 11-13 M. Sioli
- Lunedì 16 marzo 2015 14-16 M. Sioli
  Giovedì 19 marzo 2014 11-13 T. Chiarusi
- Lunedì 23 marzo 2015 14-16 M. Sioli
  Giovedì 26 marzo 2015 11-13 M. Sioli
  Giovedì 26 marzo 2015 16-18 G. Sirri
- Lunedì 30 marzo 2015 14-16 M. Sioli

APRILE

- **Giovedì 2[illegible] 11-13 G. Sirri**
- Mercol. 8 aprile 2015 10-13 M. Sioli/T.Chiar.
- **Giovedì 9 aprile 2015 11-13 G. Sirri**
- Lunedì 13 aprile 2015 14-16 M. Sioli
  Giovedì 16 aprile 2015 11-13 M. Sioli
  Giovedì 16 aprile 2015 16-18 G. Sirri
- Lunedì 20 aprile 2015 14-16 M. Sioli
  Giovedì 23 aprile 2015 11-13 T. Chiarusi
  Giovedì 23 aprile 2015 16-18 G. Sirri
- **Giovedì 30 aprile 2015 11-13 G. Sirri**
  Giovedì 30 aprile 2015 16-18 G. Sirri

MAGGIO

- Lunedì 4 maggio 2015 14-16 M. Sioli
  Giovedì 7 maggio 2015 11-13 T. Chiarusi
- Lunedì 11 maggio 2015 14-16 M. Sioli
  Giovedì 14 maggio 2015 11-13 G. Sirri
  Giovedì 14 maggio 2015 16-18 G. Sirri
- Lunedì 18 maggio 2015 14-16 M. Sioli
  Giovedì 21 maggio 2015 11-13 T. Chiarusi
- Lunedì 25 maggio 2015 14-16 M. Sioli

Tutte le lezioni in Aula C, via Irnerio

Chiarimenti sulle modalità di esame

- L'**esame** del corso di Analisi Statistica dei Dati nella Fisica Nucleare e Subnucleare è solo **orale** e **unico per le 3 parti** (teoria, esercizi, laboratorio). Per la parte di laboratorio 💻, possono essere semplici domande sugli argomenti trattati. Non sarà chiesto di scrivere codice alla lavagna.

- L'**accesso** all'esame è **vincolato** dalla **spedizione** via mail della **soluzione degli esercizi** proposti alle lezioni di laboratorio 💻 a gabriele.sirri2@unibo.it . Gli esercizi possono essere svolti da soli o in coppia durante la lezione e/o completati a casa.

- Non è prevista la valutazione dei singoli esercizi (non c'è voto)

- Per il resto:

Informazioni sull'esame

L'esame è solo orale. Le date degli appelli sono da concordare con i docenti e vanno fissate all'interno delle sessioni d'esame. Per iscriversi è necessario spedire un'email a sioli@bo.infn.it. Per gli studenti della laurea magistrale l'esame verterà sull'intero programma del corso. Per i dottorandi, invece, si richiede la preparazione di un seminario (della durata di circa 30 minuti), in cui vengano mostrati con chiarezza gli aspetti della loro attività che hanno tratto profitto dalle tecniche apprese a lezione.

Esercitazioni (soluzioni)

- Le mie soluzioni saranno pubblicate sul sito, una volta raccolte le vostre, e saranno visibili alla **lista di distribuzione docenti-studenti**

  gabriele.sirri2.ASD-2015

  a cui vi invito ad iscrivervi
  (per il momento no password)

Stato

	Lista mail	ES. 1	
1		sì	0,074
2	sì	sì	0,076
3	sì		0,114
4	sì	sì	0,211
5		sì	0,324
6	sì	sì	0,386
7	sì		0,586
8		sì	0,598
9	sì	sì	0,604
10	sì	sì	0,634
11	sì	sì	0,696
12	sì	sì	0,796
13	sì	sì	0,801
14	sì		0,824
15	sì	sì	0,906
16		sì	0,968

→ Iscrivetevi alla Lista

→ Mandatemi le soluzioni dell'es.1
appena possibile
oppure
fissiamo un ricevimento

La lista ordinata secondo numeri casuali distribuiti uniformemente tra 0 e 1 per privacy.

💻 Discussione (Esercizio 1)

RECAP… Esercizio 1

Scaricate da http://www.bo.infn.it/~sirri/teaching/2013/ads/2_root/:
- uniform.C: Program to illustrate use of random number and histogram classes
- plotHist.C: Simple ROOT macro to plot the histogram.

[1]

Alcune variabili aleatorie r_i sono uniformemente distribuite nell'intervallo [0,1].
Si modifichi il programma uniform.C in modo da generare gli istogrammi di

(a) $x = r_1 + r_2 - 1$

(b) $x = r_1 + r_2 + r_3 + r_4 - 2$

(c) $x = \sum_{i=1}^{12} r_i - 6$

Si calcoli la media e la varianza delle variabili definite in (a)(b)(c) (sapendo che ciascun r_i ha media 1/2 e varianza 1/12) e si confrontino con i valori che ottenete dagli istogrammi dei numeri generati (quando visualizzate gli istogrammi con ROOT viene mostrata la media e lo scarto quadratico medio). Si ricordi di aggiustare i valori minimo e massimo dell'asse x dell'istogramma in modo che includa tutti i valori generati. Si commenti sulla connessione tra gli istogrammi e il teorema del limite centrale.

RECAP… Esercizio 1

[2]

Si incrementi *numValues* e si confronti il tempo di esecuzione della macro nella modalita' interpretata e nella modalita' compilata con ACLiC

[3]

Si modifichi **uniform.C** per generare random un istogramma distribuito secondo una gaussiana con media = 1 e sigma = 3 utilizzando *numValues* = 1000.
Ricordarsi di aggiustare i valori max e min dell'asse X dell'istogramma.

Si modifichi **plotHist.C** per visualizzare l'istogramma e sovrapporre un FIT gaussiano.

[4]

Si modifichi **plotHist.C** per creare una TCanvas divisa in due.
Nella prima metà : si disegni una p.d.f. Gaussiana con media 1 e sigma 1 e si sovrapponga una p.d.f. gaussiana con media 1 e sigma 3.

Nella seconda metà : si prenda l'esercizio 3 e si disegni l'istogramma utilizzando marker • e errori di misura. Si sovrapponga il fit.

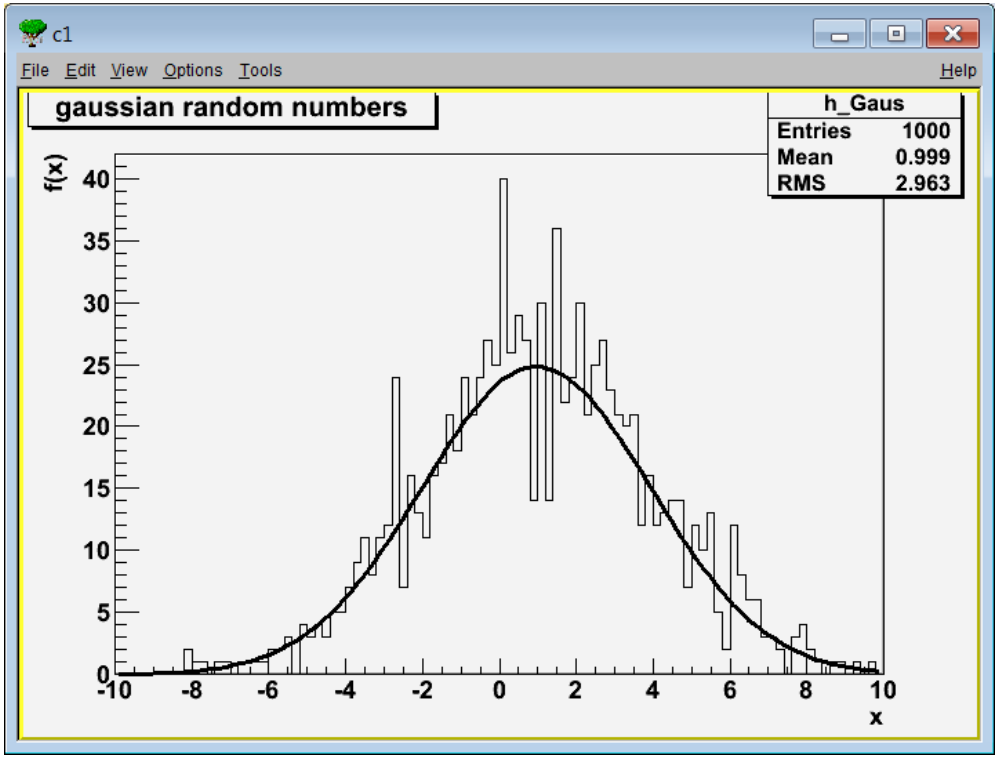
Esercizio 1 - Soluzione [3]

```
// Open output file
  TFile* file = new TFile("uniform.root", "recreate");

// Book histograms
TH1D* h = new TH1D("h", "random numbers“
                   ,  100,  0, 1.0);

// Create a TRandom3 object to generate random numbers
  int seed = 12345;
  TRandom3* ran = new TRandom3(seed);

// Generate some random numbers and fill histograms
  const int numValues = 10000;

  for (int i=0; i<numValues; ++i){
     double r = ran->Rndm();          // uniform in ]0,1]
     h->Fill(r);
  }

// Store all histograms in the output file and close up
  file->Write();
  file->Close();
```

```
// Open output file
  TFile* file = new TFile("gaussian.root", "recreate");

// Book histograms
TH1D* h_Gaus = new TH1D(" h_Gaus ", "random numbers“
                   ,  100,  -10, 10 );

// Create a TRandom3 object to generate random numbers
  int seed = 12345;
  TRandom3* ran = new TRandom3(seed);

// Generate some random numbers and fill histograms
  const int numValues = 1000;

  for (int i=0; i<numValues; ++i){
    double r = ran->Gaus(1,3);     // gaussian in mean = 1 , sigma = 3
    h_Gaus >Fill(r);
  }

// Store all histograms in the output file and close up
  file->Write();
  file->Close();
```

```
  TFile* f = new TFile("gaussian.root");
  f->ls();
  TH1D* h1 = (TH1D*)f->Get("h_Gaus");
  h1->SetXTitle("x");
  h1->SetYTitle("f(x)");
  h1->Fit("gaus") ;
  h1->Draw();
```

Esercizio 1 - Soluzione [3]



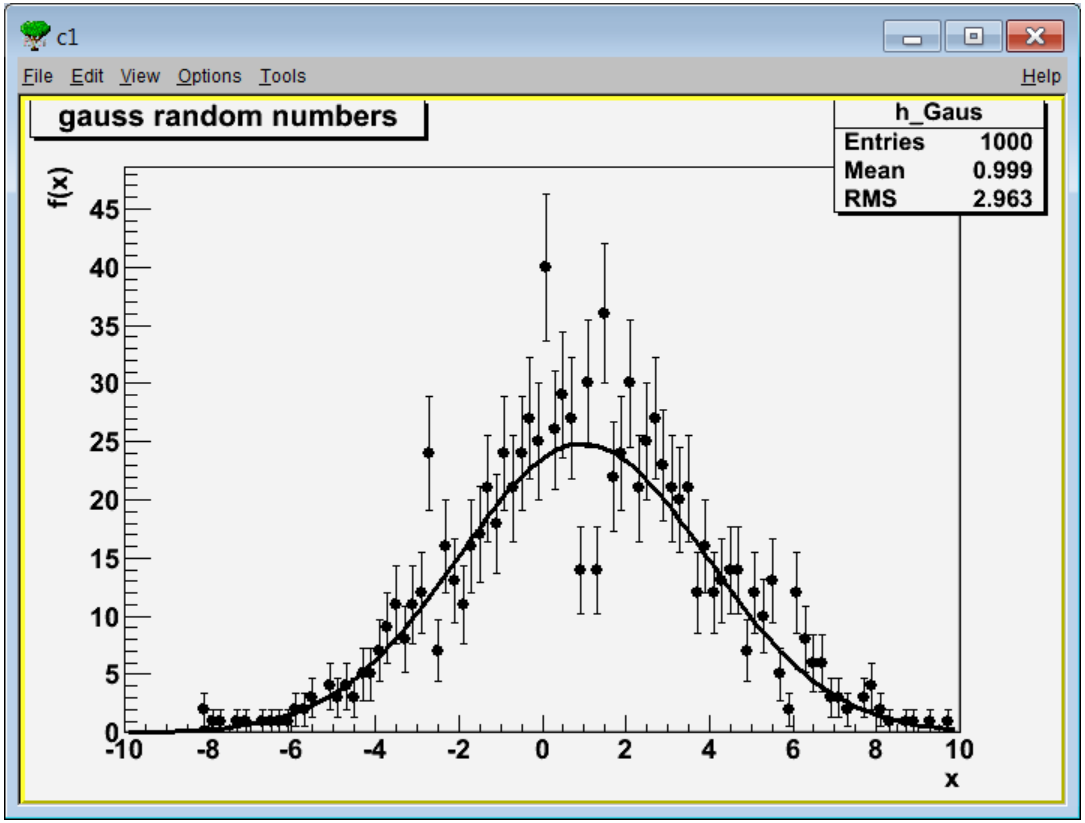
MINUIT is a physics analysis tool for function minimization.

```
FCN=81.0542 FROM MIGRAD     STATUS=CONVERGED       75 CALLS          76 TOTAL
                    EDM=3.71855e-010     STRATEGY= 1      ERROR MATRIX ACCURATE
 EXT PARAMETER                                          STEP         FIRST
 NO.   NAME      VALUE            ERROR          SIZE      DERIVATIVE
  1  Constant   2.48124e+001   1.05270e+000   3.62620e-003 -1.96918e-005
  2  Mean       9.66909e-001   1.01955e-001   4.50736e-004 -1.29068e-005
  3  Sigma      2.97928e+000   8.49430e-002   3.28220e-005 -3.66148e-003
```

Noi lo ignoriamo per questioni di tempo ma il significato di questo OUTPUT è da sapere !

Esercizio 1 - Soluzione [4]

```
const int numValues = 1000;

for (int i=0; i<numValues; ++i)
{
   double r = ran->Gaus(1,3);
   h_Gaus >Fill(r);
}
…


TFile* f = new TFile("gaussianMC.root");
 f->ls();

TH1D* h1 = (TH1D*) f->Get("h_Gaus");
 h1->SetXTitle("x");
 h1->SetYTitle("f(x)");
 h1->SetMarkerStyle(20);
 h1->Fit("gaus") ;
 h1->Draw("E1");
```



Commento : cosa c'è di strano nel plot?

Suggerimento: qual è il significato e come si calcolano le barre di errore?

Introduzione a RooFIT

How We Use Statistics

Broadly speaking, we use statistical techniques for a few main purposes:

- **Point estimation:** what is the best estimate of a particular parameter
  - eg. measurement of the Z boson mass
- **Confidence/Credible Intervals:** regions representing an range of a parameter compatible with the data (made precise in different ways)
  - eg. 95% contours, upper-limits, lower-limits
- **Hypothesis Testing:** test hypothesis against one or more alternatives
  - eg. Discover the Higgs, Discover SUSY, reject standard model
- **Goodness-of-fit:** quantify how well a model describes the data
  - eg. check if standard model describes data without a specific alternative

In a broader context, there are related issues:

- **Data reduction:** how to reduce the raw data while loosing minimal information that is useful for our ultimate goal (multivariate analysis is important here)
- **Decision making:** how do we make decisions in the face of uncertainty
- Where does the role of an experimentalist end? How does this impact how we **publish** our results?

Terminology

CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

1.3 Terminology used in this guide

model a probability density function that describes some observables. We use the term model for both parametric models (eg. a Gaussian is parametrized by a mean and standard deviation) and non-parametric models (eg. histograms or KEYS pdfs).

observable(s) quantities that are directly measured by an experiment and present in a data set. The distribution of the observables are predicted by the model. Models are normalized such that the integral of the model over the observables is 1.

auxiliary observable observables that come from an auxiliary experiment (eg. a control sample or a preceding experiment).

parameter of interest quantities used to parametrize a model that are 'interesting' in the sense that one wishes to estimate their values, place limits on them, etc (eg. masses, cross-sections, and the like).

nuisance parameter quantities used to parametrize a model that are uncertain but not 'interesting' in the above sense (eg. background normalization, shape parameters associated to systematic uncertainties, etc.)

control sample a data set independent of the main measurement (defining auxiliary observables) often used to constrain nuisance parameters by simultaneous considering it together with the main measurement.

Terminologia 2

sarebbe preferibile impiegare meno tempo dietro a :

puntatori,
assegnazione di variabili,
cicli for,
generatori di numeri casuali,
riempimento di istogrammi,
ecc…

… e concentrarsi sui modelli che descrivono il fenomeno che stiamo osservando.

RooFit & RooStats

CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

RooStats provides tools for high-level statistics questions in ROOT

- it builds on **RooFit** which provides basic building blocks for statistical questions

Today, I will start with an overview of RooFit and then move to RooStats

... tomorrow we will consider more advanced usage.

RooFit

variables functions

probability density functions

binned & unbinned datasets

fitting toyMC generation

minimization integration

RooStats

hypothesis prior

hypothesis tests

confidence intervals (limits)

combinations

test statistic sampling distribution

Note, excellent slides from Wouter Verkerke on RooFit at SoS '08 (I will borrow from them)

http://indico.in2p3.fr/materialDisplay.py?contribId=15&materialId=slides&confId=750

RooFit

Introduction to RooFit slides da 1 a 14 o 19

http://hadron.physics.fsu.edu/~skpark/document/ROOT/roofit-intro-roostats-v12a.pdf

💻 Home work (Esercizio 2)

Esercizio 2 - TESTO

Scaricate da http://www.unibo.it/docenti/gabriele.sirri2

→ Contenuti utili

→ Analisi statistica dei dati ... Calendario e Materiale

→ Lezione di oggi : “testo + code to start”

roofit_empty.C

[1] roofit_ex1.C

Editate la macro e seguendo lo schema costruire una p.d.f. gaussiana con media 0, sigma 1. Modificate la sigma a 3. Visualizzate la p.d.f. . Generate un dataset unbinned di 10000 eventi. Eseguite un Fit con Maximum Likelihood. Visualizzate i risultati.

Utilizzate le informazioni in Introduction to RooFit , nel manuale di roofit al paragrafo 2 (ftp://root.cern.ch/root/doc/RooFit_Users_Manual_2.91-33.pdf) e in http://root.cern.ch/drupal/content/roofit).

[2] roofit_ex2.C

Si modifichi lo script e generare un dataset binned (bin width = 0.5) .

The binning of the returned RooDataHist is controlled by the default binning associated with the observables generated. To set the number of bins in x to 200, do e.g. x.setBins(200) prior to the call to generateBinned()

Esercizio 2 - TESTO

[3] roofit_ex3.C

Rinominate la p.d.f. gaussiana «sig» e aggiungete al modello un fondo esponenziale «bkg» espresso in funzione di un parametro tau, exp(-x/tau) .

Il valore iniziale di tau =10.

Suggerimento: Si esprima -1./tau come RooFormulaVar

Definite un parametro «fsig» rapporto segnale/fondo.

Costruite un modello composito nella forma
model(x) = fsig*sig(x) + (1-fsig)*bkg(x)

Suggerimenti: usate la funzione RooAddPdf (paragrafo 3 del manuale)