

# Probabilità, funzioni di densità di probabilità. Istogrammi. Covarianza e correlazione.

## Laboratorio di Metodi Computazionali e Statistici (2023/2024)

6 Novembre 2023

# Perché un corso “avanzato” di probabilità e statistica

- Molte delle domande a cui dobbiamo rispondere giornalmente in laboratorio si devono confrontare con statistica limitata (errore statistico) ed incertezze sperimentali sistematiche. Strumenti statistici efficaci permettono di estrarre il massimo dell'informazione dai dati (e di darne un'interpretazione quantitativa chiara).
- Occorre quindi saper usare ed interpretare questi strumenti. È quindi essenziale approfondire le conoscenze di probabilità e statistica (il bagaglio di conoscenze acquisito al primo anno non è sufficiente).
- Organizzazione della seconda parte:
  - Richiami di probabilità. Covarianza e correlazione. Metodo di MonteCarlo. Propagazione degli errori.
  - Teoria della stima. Metodo di massima verosimiglianza. Intervalli di confidenza (S. Passaggio)
  - Test d'ipotesi, cenni a metodi multivariati
- Mini-esercitazioni (guidate): moduli da due ore al giovedì/venerdì pomeriggio (al posto delle esercitazioni)

# Misure di grandezze fisiche con risultati non riproducibili

Le misure fisiche danno spesso risultati non riproducibili. Questa tendenza stocastica potrebbe essere dovuta:

- La grandezza fisica potrebbe riferirsi ad **elementi di una popolazione** ed il suo valore esatto dipendere dall'elemento considerato
- La grandezza fisica potrebbe essere misurata con precisione ma risultare diversa ad ogni tentativo di misura a causa della sua **natura intrinsecamente stocastica** (decadimenti radioattivi, turbolenza,...)
- La grandezza fisica potrebbe essere definita, ma la lettura della “scala graduata” potrebbe essere spinta al di sotto del limite di riproducibilità dello strumento (**errore statistico**).

# Distribuzione delle misure affette da errore

Tutti le misure affette da incertezza sperimentale possono essere trattate come distribuzioni di variabili aleatorie ? Le raccomandazioni di “Guide to the expression of uncertainty in measurement (GUM)” vanno in questo senso.

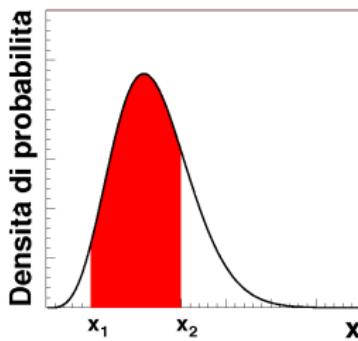
- Errore statistico: distribuzione gaussiana con centro media e deviazione standard l'errore standard.
- Errori di altra natura (massimi, etc...) per cui sia individuata un'intervalllo di variabilità massimo. GUM raccomanda di assumere per questo tipo di misure, espresse in genere come  $x \pm \Delta$ , una distribuzione rettangolare la cui varianza è  $\Delta^2/3$ . Questo riconduce, di fatto, a trattare queste misure come quelle di tipo a) ma con deviazione standard  $\Delta/\sqrt{3}$ .

Banalizzazione (o semplificazione) della distinzione tra errore statistico e massimo. Alla base ci sono molte ragioni: l'impossibilità di trattare intervalli non associati a distribuzioni, la necessità di stimare errori equivalenti a quelli statistici (vedi fit) e, non ultimo, il teorema del limite centrale (che assicura che in presenza di varie sorgente di errore la distribuzione risultante è ben approssimabile da una gaussiana).

# Distribuzioni di probabilità

Chiamiamo  $x$  la variabile che caratterizza la misura della grandezza che stiamo considerando. Descriviamo  $x$  con una variabile aleatoria, ovvero una variabile il cui valore non sia costante, ma vari seguendo una legge di probabilità nota.

$x$  continua assume tutti i valori



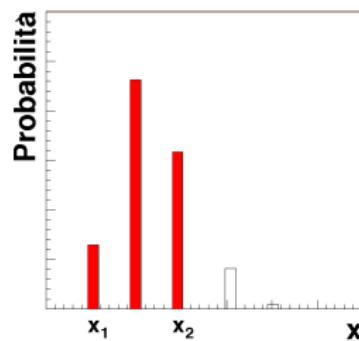
$$\int_{\Omega} p(x)dx = 1 \left( \int_{-\infty}^{\infty} p(x)dx = 1 \text{ se } \Omega = \mathbb{R} \right)$$

$$P(x \in [x, x + dx]) = p(x)dx$$

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} p(x)dx$$

↑  
Densità prob

$x$  discreta solo certi valori



$$\sum_i P(x_i) = 1$$

$$P(x_i) = p_i$$

$$P(x_1 \leq x \leq x_2) = \sum_{i=i_1}^{i_2} p_i$$

# Distribuzioni cumulative

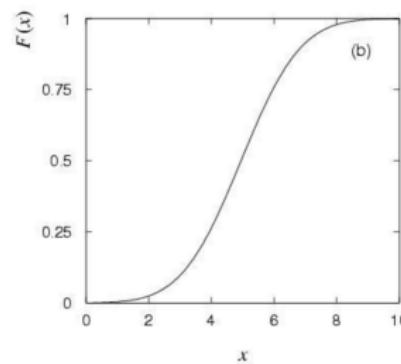
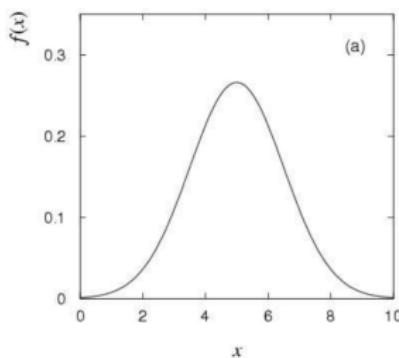
## Distribuzioni cumulative

Prob che  $x'$  assuma val minore di  $x$

$$\int_{-\infty}^x p(x') dx' = P(x) \rightarrow F(x) \text{ nel disegno}$$

Permette di definire (derivare) la pdf

$$p(x) = \frac{\partial P(x)}{\partial x}$$



# Caratterizzazione di una PDF: valore di aspettazione e momenti algebrici

Valore di aspettazione di  $g(x)$  sulla PDF  $p(x)$  è definito

$$E[g(x)] \equiv \int g(x)p(x)dx$$

ed è un operatore lineare.

Una distribuzione può essere caratterizzata sulla base dei **momenti algebrici** di ordine  $k$  definiti come

$$\mu'_k \equiv E[x^k] = \int x^k p(x)dx$$

Il momento di ordine 0 è la normalizzazione  $\rightarrow$  sempre definito *✓* PDF

$$\mu'_0 \equiv E[1] = \int p(x)dx = 1$$

il momento di ordine 1 è la media.

$$\mu \equiv \mu'_1 \equiv E[x] = \int xp(x)dx$$

# Caratterizzazione di una PDF: momenti

I momenti successivi sono convenientemente definiti rispetto alla media (**momenti "centrati"**)

$$\mu_k \equiv E[(x - \mu)^k] = \int (x - \mu)^k p(x) dx$$

la varianza è il momento di ordine 2

$$\mu_2 \equiv E[(x - \mu)^2] = \int (x - \mu)^2 p(x) dx = \sigma^2 = V[x]$$

Vale la pena ricordare che la varianza può essere scritta in termine dei primi due momenti algebrici.

$$\begin{aligned} E[(x - \mu)^2] &= E[(x^2 + \mu^2 - 2x\mu)] = E[x^2] - 2\mu E[x] + \mu^2 \\ &= E[x^2] - \mu^2 = E[x^2] - (E[x])^2 \end{aligned}$$

# Caratterizzazione di una PDF: momenti

Skewness  $\gamma_1$ :

- Se la distribuzione è simmetrica attorno alla sua posizione centrale, tutti i momenti centrali di ordine dispari sono nulli → qualunque momento centrale di ordine dispari diverso da zero indica asimmetria
- Il momento centrale di ordine 3 è la quantità più semplice per caratterizzare l'asimmetria di una distribuzione:

$$\mu_3 \equiv E[(x - \mu)^3] = \int (x - \mu)^3 p(x) dx$$

- Si definisce spesso il coefficiente di asimmetria (skewness) adimensionale

$$\gamma_1 \equiv \frac{\mu_3}{\sigma^3}$$

- Un coefficiente di asimmetria positivo (negativo) indica che la distribuzione presenta una coda più pronunciata a destra (a sinistra) della media

# Caratterizzazione di una PDF: momenti

La curtosi  $\gamma_2$  quantifica l'importanza delle code (rispetto al caso gaussiano)

- I momenti centrali di ordine pari (in virtù della maggiore importanza attribuita ai valori di  $x$  che si più si discostano dalla media) possono fornire informazioni sulle “code” della distribuzione.
- Anche in questo caso si usa il momento di ordine più basso:

$$\mu_4 \equiv E[(x - \mu)^4] = \int (x - \mu)^4 p(x) dx$$

- Il coefficiente adimensionale di curtosi è definito come:

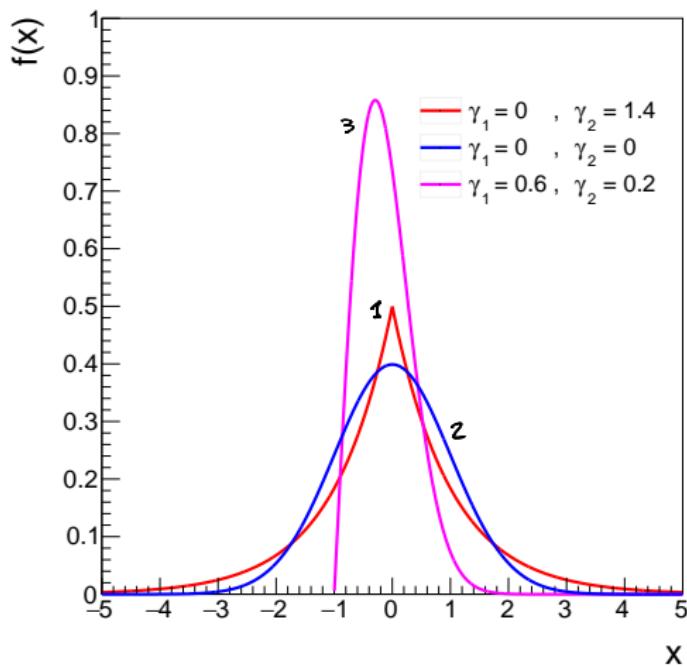
$$\gamma_2 \equiv \frac{\mu_4}{\sigma^4} - 3$$

*da info sui valori nelle code xk elevati allo  $^4$  hanno più importanza*

- $\gamma_2 = 0$  gaussiana
- $\gamma_2 > 0$  più piccata rispetto alla gaussiana
- $\gamma_2 < 0$  meno piccata rispetto alla gaussiana

# Skewness e Curtosi

$\gamma_1$        $\gamma_2$

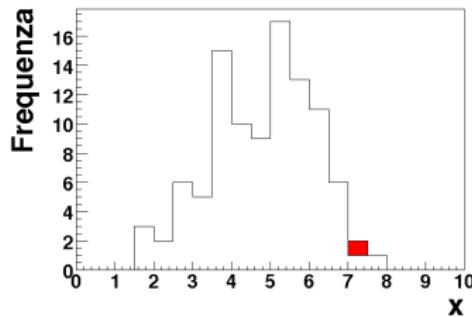
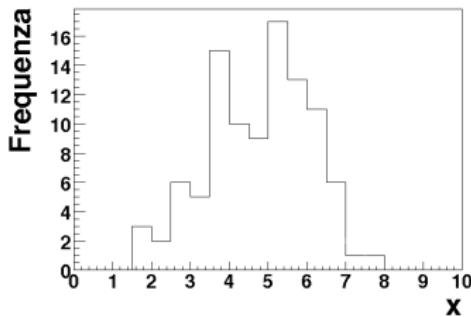


- 1 più piccata
- 2 meno piccata
- 3 non simmetrica

# Iistogrammi

Gli istogrammi sono grafici usati per visualizzare le distribuzioni di frequenza di eventi caratterizzati da variabili discrete o continue.

Per ogni intervallo  $\Delta x$  (detto “bin”) si disegna un rettangolo largo quanto l’intervallo e di altezza pari al numero di occorrenze dei valori appartenenti a tale intervallo.

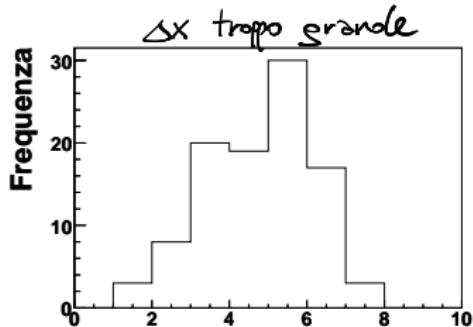
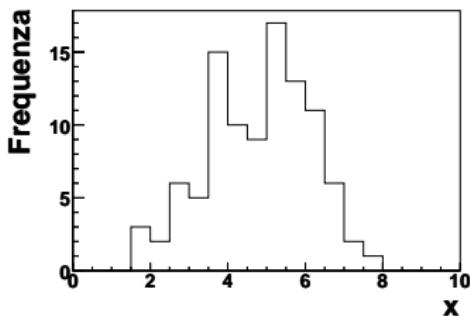
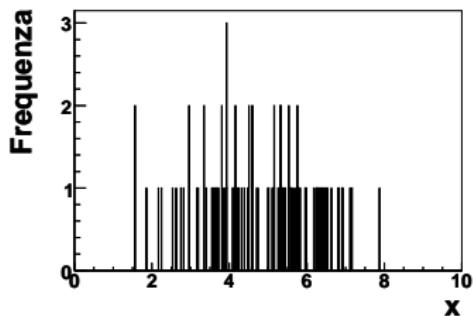


Nuovo evento  $x = 7.1$

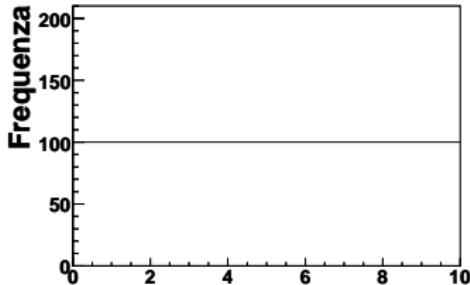
# Binning

Il numero di bin può influenzare in modo rilevante l'aspetto di un istogramma, e quindi va scelto caso per caso, cercando di evitare la presenza di molti bin poco popolati.

*bin vuoti o troppo piccolo*



*Δx troppo grande*



# PDF da istogramma

Un istogramma contenente le misure di una certa grandezza, se pensato come funzione

$$N(x)$$

può essere definito dalla relazione:

$$N(x) = N_i \quad (i \text{ indice del bin che contiene il valore } x)$$

Tale funzione non è normalizzata ma ha integrale (detti  $M$  il numero di bin dell'istogramma)

$$\int N(x)dx = \sum_{i=1}^M N_i \Delta x = N_{tot} \Delta x$$

# Classi di istogrammi in ROOT

In ROOT esistono classi di istogrammi 1,2,3-dimensionali (TH1X, TH2X, TH3X). X indica il tipo di variabile che possono contenere:

- **TH1,2,3C** istogrammi per variabili char (1 byte)
- **TH1,2,3S** istogrammi per variabili short int (2 bytes)
- **TH1,2,3I** istogrammi per variabili int (4 bytes)
- **TH1,2,3F** istogrammi per variabili float (4 bytes)
- **TH1,2,3D** istogrammi per variabili double (8 bytes)

# Classi di istogrammi in ROOT (I)

Costruttore:

```
TH1D(const char* name, const char* title, int nbinsx, double xlow, es
      double xup);
      name = identificatore
      title = titolo
      nbinsx = numero di bin
      xlow = estremo inferiore, xup = estremo superiore
      se xlow=xup=0 gli estremi sono calcolati automaticamente.
      Non so i range della var x → calcolati auto Ma molti metodi successivi non riconoscono gli estremi → auto calcolati
      TH2D(const char* name, const char* title, int nbinsx, double xlow,
            double xup, int nbinsy, double ylow, double yup);
            name = identificatore
            title = titolo
            nbinsx = numero di X bin
            xlow = estremo inferiore X, xup = estremo superiore X
            nbinsy = numero di Y bin
            ylow = estremo inferiore Y, yup = estremo superiore Y
            se xlow=xup=0 gli estremi della variabile X sono calcolati automaticamente
            se ylow=yup=0 gli estremi della variabile Y sono calcolati automaticamente.
```

*GetNbins  
non va  
se c'è 0,0*  
*METTERE SEMPRE  
ESTREMI*

# Classi di istogrammi in ROOT (II)

Per riempirli:

```
void TH1D::Fill(double x, double w);
```

*x: variabile aleatoria; w: peso, per il singolo evento w=1*

```
void TH1D::SetBinContent(int ibin, double val);
```

*pone il contenuto del bin ibin uguale a val*

incremento di 1 il bin  
altrimenti incremento di  
w il bin

*(N.B.): ibin parte da 1*  
*mette altezza del*  
*1 bin-esimo = val*

```
void TH2D::Fill(double x, double y, double w);
```

*x: variabile aleatoria; y: variabile aleatoria; w: peso, per il singolo evento w=1*

```
void TH2D::SetBinContent(int ibinX, int ibinY, double val);
```

*pone il contenuto del bin ibinX, ibinY uguale a val*

Per disegnarli:

```
void TH(1,2)D::Draw(const char *option);
```

*option: " " (o nulla) disegno "standard" con assi*

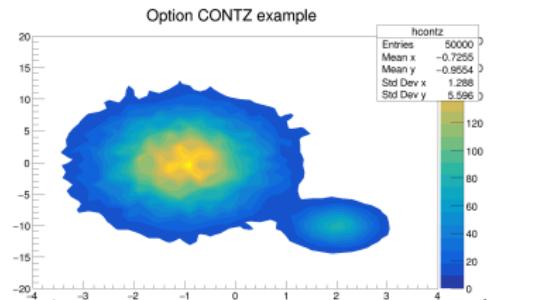
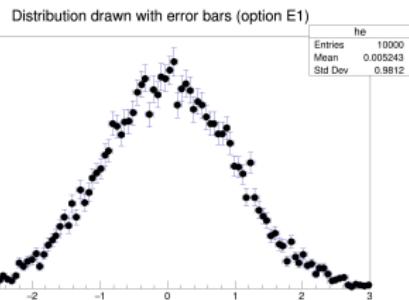
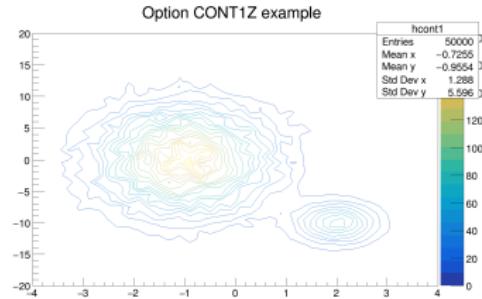
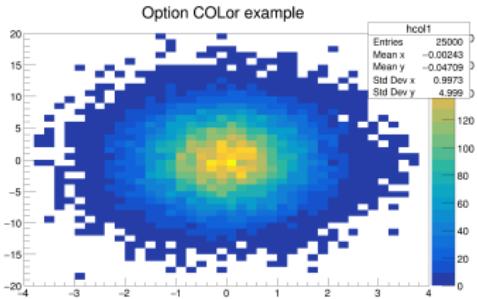
*"SAME" per sovrapporre*

*"COLZ" bin colorato dipendente dal contenuto (solo per 2D)*

*"CONT" per contour plot (solo per 2D)*

*"E" per disegnare le barre di errore*

# Classi di istogrammi in ROOT (III)



errore orizz è larghezza bin  
 errore sul bin è  $\sqrt{N}$  → frequenza  
 Da distib Poisson

conteggio di eventi in un intervallo  
 $\lambda$  n° medio eventi/intervalle (valore atteso)  
 $\lambda$  variabile  
 Funz generatrice dei momenti e  $\lambda(e^t - 1)$   
 $P_\lambda(n) = \frac{\lambda^n}{n!} e^{-\lambda}$   
 TOT eventi nell'intervalle

# Classi di istogrammi in ROOT (III)

Accesso alle informazioni: *Usato al posto del file di dati, come contenitore  
Per accedere ai dati:*

`double TH1D::GetBinContent(int ibin);`

*ritorna il contenuto del bin ibin  $\in [1, nbin]$*

`double TH1D::GetBinCenter(int ibin);`

*ritorna la coordinata del centro del bin ibin  $\in [1, nbin]$*

`double TH1D::GetNbinsX();`

*ritorna il numero di bin (nbin)*

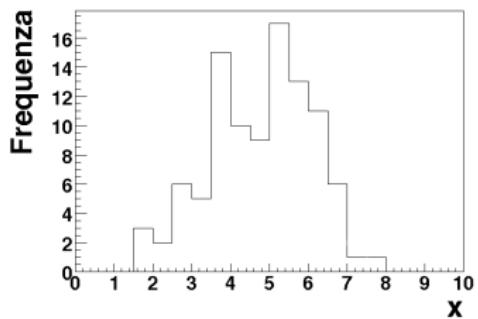
`double TH1D::GetEntries();`

*ritorna il numero di valori inseriti*

`double TH1D::GetBinWidth(int ibin);`

*ritorna la larghezza del bin ibin  $\in [1, nbin]$*

$h.GetBinContent(5)$  ritorna 2  $\rightarrow$  conteggio del bin 5  
 $h.GetBinCenter(3)$  ritorna 1.25  $\rightarrow$  centro del bin su x  
 $h.GetNbinsX()$  ritorna 20  $\rightarrow$  N° bin in esse x  
 $h.GetEntries()$  ritorna 100  $\rightarrow$  N tot conteggi



# Classi di istogrammi in ROOT (IV)

Accesso alle informazioni:

```
double TH2D::GetBinContent(int ibinX, int ibinY);  
    ritorna il contenuto del bin ibinX ∈ [1, nbinX], ibinY ∈ [1, nbinY]  
double TH2D::GetXaxis()::GetBinCenter(int ibin);  
    ritorna la coordinata del centro del bin X ibin ∈ [1, nbinX]  
double TH2D::GetYaxis()::GetBinCenter(int ibin);  
    ritorna la coordinata del centro del bin Y ibin ∈ [1, nbinY]  
double TH2D::GetXaxis()::GetBinWidth(int ibin);  
    ritorna la larghezza del bin X ibin ∈ [1, nbinX]  
double TH2D::GetYaxis()::GetBinWidth(int ibin);  
    ritorna la larghezza del bin Y ibin ∈ [1, nbinY]  
double TH2D::GetNbinsX();  
    ritorna il numero di bin X (nbinX)  
double TH2D::GetNbinsY();  
    ritorna il numero di bin Y (nbinY)  
double TH2D::GetEntries();  
    ritorna il numero di valori inseriti
```

# Uso degli istogrammi

A differenza di un normale grafico, che viene usato solo nella fase finale dei calcoli, un istogramma viene di solito usato, all'interno di un programma, come **elemento di memorizzazione dei dati**.

Un istogramma fornisce, in genere, anche la stima dei momenti della distribuzione

- `GetMean()`: media
- `GetRMS()` : deviazione standard
- `GetSkewness()`
- `GetKurtosis()`

# Esempio: C++

```
1 #include <iostream>
2 #include <fstream>
3 #include <TH1D.h>
4 #include <TF1.h>
5 #include <TMath.h>
6 #include <TApplication.h>
7 #include <cmath>
8 using namespace std;
9
10 int main(){
11     TApplication app("app",0,NULL);
12     ifstream file;
13     file.open("gaus.dat");
14     TH1D hexp("hexp","Histogramma sperimentale",20,0,20);
15     double tmp;
16     while(file >> tmp){
17         hexp.Fill(tmp);
18     }
19     hexp.SetMarkerStyle(20);
20     hexp.Draw("E");
21     app.Run(true);
22 }
```

# Esempio: macro ROOT

```
1 {  
2   ifstream file("gaus.dat");  
3   TH1D hexp("hexp","Histogramma sperimentale ",20,0,20);  
4   double tmp;  
5   while(file >> tmp){  
6     hexp.Fill(tmp);  
7   }  
8   hexp.SetMarkerStyle(20);  
9   hexp.Draw("E");  
10 }
```

# Esempio: macro ROOT con nome

```
1 void histo(){
2   TApplication app("app", NULL, NULL);
3   ifstream file("gaus.dat");
4   TH1D *hexp = new TH1D("hexp","Histogramma sperimentale ",20,0,20);
5   double tmp;
6   while(file >> tmp){
7     hexp->Fill(tmp);
8   }
9   hexp->SetMarkerStyle(20);
10  hexp->Draw("E");
11  app.Run(true);
12 }
```

N.B. Il file deve chiamarsi `histo.C`

# Esempio: Python

```
1 from ROOT import *
2
3 f = open(" gaus.dat")
4 hexp = TH1D(" hexp "," Histogramma sperimentale ",20,0,20)
5 for line in f:
6     val = float(line)
7     hexp.Fill(val)
8
9 hexp.SetMarkerStyle(20)
10 hexp.Draw("E")
11 gApplication.Run(True)
```

# Esempio: macro ROOT per TH2D

```
1 {
2     ifstream file("2D.dat");
3     TH2D *hexp = new TH2D("hexp", "Histogramma sperimentale", 20,
4                           0.,20.,20.,0.,20.);
5     double tmp1, tmp2;
6
7     while(file >>tmp1>>tmp2){
8         hexp->Fill(tmp1,tmp2);
9     }
10    hexp->Draw("COLZ");
11 }
```

# Confronto con le distribuzioni teoriche

Un modo semplice per confrontare graficamente un istogramma sperimentale con una distribuzione teorica è quello di **sovraporre il grafico della densità di probabilità**.

La cosa pare banale, ma si deve fare attenzione alla **corretta normalizzazione**: se i bin sono  $M$ , larghi  $\Delta x$  e si hanno in totale  $N_{tot}$  eventi l'integrale dell'istogramma vale

$$\sum_{i=1}^M N_i \Delta x = N_{tot} \Delta x$$

( $N_i$  contenuto del bin  $i$ -esimo).

Per avere una corretta normalizzazione si deve quindi moltiplicare la densità di probabilità  $p(x)$  per il fattore  $N_{tot} \Delta x$

`histo → GetEntries() * histo → GetBinWidth(1)`

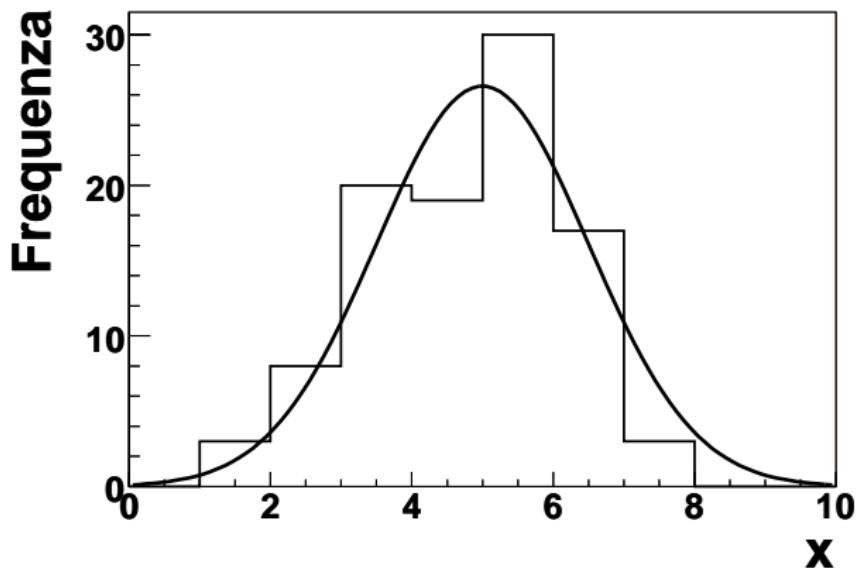
$N_{tot}$

$\Delta x$  bin 1 (uguale per tutti)

*NB aver fissato  
 $x_{min}$  e  $x_{max}$   
in TH1D*

# Confronto con le distribuzioni teoriche

Un modo semplice per confrontare graficamente un istogramma sperimentale con una distribuzione teorica è quello di **sovrapporre il grafico della densità di probabilità**.



# Esempio: sovrapposizione funzione a istogramma

```

1 #include <iostream>           Usato molto
2 #include <fstream>
3 #include <TH1D.h>
4 #include <TF1.h>
5 #include <TMath.h>
6 #include <TApplication.h>
7 #include <cmath>
8 using namespace std;
9
10 int main(){
11     TApplication app("app",0,NULL);
12     ifstream file;
13     file.open("gaus.dat");
14     TH1D hexp("hexp","Histogramma sperimentale ",20,0,20);
15     double tmp;
16     while(file >> tmp){
17         hexp.Fill(tmp);
18     }
19     hexp.SetMarkerStyle(20);
20     hexp.Draw();
21     TF1 f("f","[2]*TMath::Gaus(x,[0],[1],1)",0,20);
22     f.SetParameter(0,hexp.GetMean());
23     f.SetParameter(1,hexp.GetRMS());
24     f.SetParameter(2,hexp.GetEntries()*hexp.GetBinWidth(1)); ← NB
25     f.Draw("SAME");
26     app.Run(true);
27 }
```



# Confronto con le distribuzioni teoriche

Un secondo metodo di confronto consiste nel costruire un istogramma con struttura identica a quello sperimentale, e di **mettere in ogni bin il numero di eventi aspettati sulla base della densità di probabilità teorica**.

Il numero di eventi aspettati in ciascun bin è dato da

$$N_i^{\text{exp.}} = N_{\text{tot}} \int_{x_{\min}(i)}^{x_{\max}(i)} p(x) dx$$

*integrale metodo migliore per trovare valore aspettato della teoria*

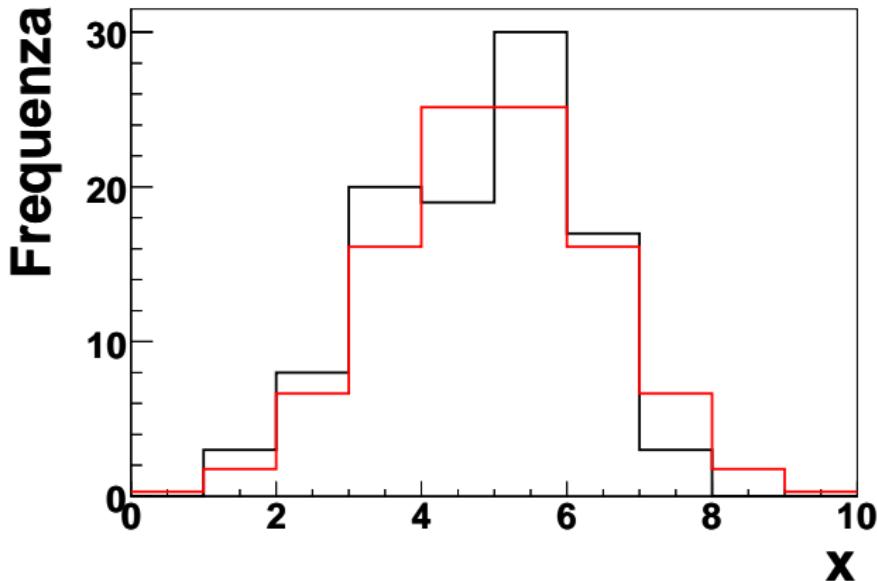
dove  $N_{\text{tot}}$  è il numero totale di eventi,  $x_{\min}(i)$  ed  $x_{\max}(i)$  sono l'estremo inferiore e superiore del bin  $i$ -esimo. Il calcolo dell'integrale può essere eseguito:

- analiticamente; ad esempio, per la gaussiana, la funzione `erf(x)` di `cmath` (o `TMath::Erf(x)` di ROOT)
- numericamente (attraverso il metodo `Integral` di `TF1`)

# Confronto con le distribuzioni teoriche

Un secondo metodo di confronto consiste nel costruire un istogramma con struttura identica a quello sperimentale, e di mettere in ogni bin il numero di eventi aspettati sulla base della densità di probabilità teorica.

Confronto  $N_i \leftrightarrow N_i^{\text{exp.}}$



# Esempio: sovrapposizione funzione a istogramma

```

1 from ROOT import *
2 import math as m
3
4 f = open("gaus.dat")
5 hexp = TH1D("hexp", "Histogramma sperimentale", 20, 0, 20)
6 for line in f:
7     val = float(line)
8     hexp.Fill(val)
9
10 hexp.SetMarkerStyle(20)
11 hexp.Draw()
12
13 hteo = TH1D(hexp)
14 for i in range(1, hteo.GetNbinsX()):
15     min = hteo.GetBinLowEdge(i)
16     max = hteo.GetBinLowEdge(i+1)
17     a1 = TMath.Erf((min-hexp.GetMean())/(m.sqrt(2)*hexp.GetRMS()))/2
18     a2 = TMath.Erf((max-hexp.GetMean())/(m.sqrt(2)*hexp.GetRMS()))/2
19     hteo.SetBinContent(i, (a2-a1)*hexp.GetEntries())
20
21 hteo.SetLineColor(kRed)
22 hteo.Draw("SAME")
23
24 gApplication.Run(True)

```

$$\int_{-\infty}^{x_{\text{min}}} G(x) dx \stackrel{\text{Gaussian}}{=} \frac{1}{2} \left( 1 + \operatorname{Erf} \frac{x - \mu}{\sqrt{2}\sigma} \right)$$

?? non ho capito scrifto

# Densità di probabilità multidimensionali

- Nel caso in cui si misurino più grandezze simultaneamente occorre ricorrere a distribuzioni di densità di probabilità multidimensionali.
- Nel caso N=2:

$$\int_{\Omega_x} \int_{\Omega_y} p(x, y) dx dy = 1$$

$$P(x \in [x, x + dx], y \in [y, y + dy]) = p(x, y) dx dy$$

$$P(x_1 < x < x_2, y_1 < y < y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} p(x, y) dx dy$$

# Caratterizzazione delle pdf multidimensionali

- Nel caso multidimensionale possiamo, ovviamente, considerare i momenti di ciascuna variabile...

Medie:

$$\mu_x \equiv \mu'_{x1} = \int_{\Omega_x} \int_{\Omega_y} x p(x, y) dx dy$$

$$\mu_y \equiv \mu'_{y1} = \int_{\Omega_x} \int_{\Omega_y} y p(x, y) dx dy$$

Varianze:

$$V[x] \equiv \sigma_x^2 \equiv \mu_{x2} = \int_{\Omega_x} \int_{\Omega_y} (x - \mu_x)^2 p(x, y) dx dy$$

$$V[y] \equiv \sigma_y^2 \equiv \mu_{y2} = \int_{\Omega_x} \int_{\Omega_y} (y - \mu_y)^2 p(x, y) dx dy$$

# Caratterizzazione delle pdf multidimensionali

- È tuttavia essenziale tenere conto del comportamento congiunto delle due variabili
- Il momento centrale, congiunto, di ordine più basso è la covarianza (valore atteso dei prodotti delle loro distanze dal valore medio)

$$\text{Cov}[x, y] \equiv E[(x - \mu_x)(y - \mu_y)] = \int_{\Omega_x} \int_{\Omega_y} (x - \mu_x)(y - \mu_y) p(x, y) dx dy$$

Si noti che:

$$\text{Cov}[x, x] = E[(x - \mu_x)(x - \mu_x)]$$

$$\sigma_x^2 = \text{Cov}[x, x] \quad \sigma_y^2 = \text{Cov}[y, y]$$

- Espressione della covarianza in termini dei momenti di ordine inferiore:

$$\begin{aligned} E[(x - \mu_x)(y - \mu_y)] &= \text{Cov}[x, y] = E[(xy - x\mu_y - \mu_x y + \mu_x \mu_y)] = \\ &= E[xy] - \mu_y E[x] - \mu_x E[y] + \mu_x \mu_y = \\ &= E[xy] - \cancel{\mu_y \mu_x} - \cancel{\mu_x \mu_y} + \cancel{\mu_x \mu_y} = E[xy] - E[x]E[y] \end{aligned}$$

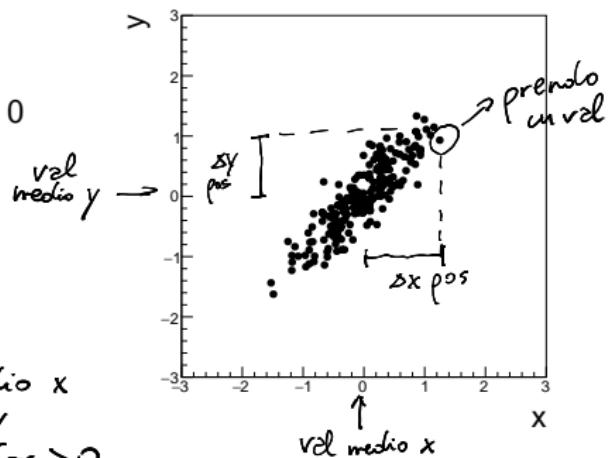


# Significato della covarianza

- La covarianza può assumere valori positivi o negativi (in contrasto con la varianza sempre positiva):

$$\bullet \text{Cov}[x, y] \equiv E[(x - \mu_x)(y - \mu_y)] > 0$$

La variazione della grandezza  $y$  rispetto a  $\mu_y$  tende ad avere lo stesso segno della variazione della grandezza  $x$  rispetto a  $\mu_x$



Ogni punto una misura nel piano  $xy$

Preso un punto: per uno spost pos del val medio  $x$   
ho spost pos del val medio  $y$

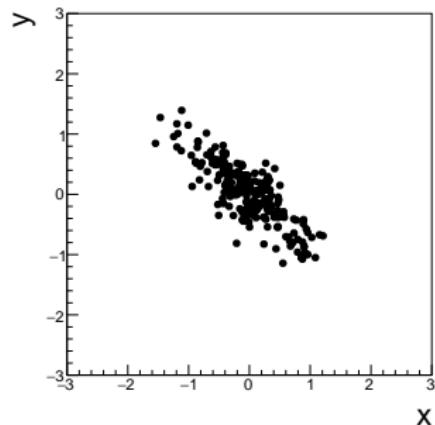
Uguali per spost neg rispetto  $\mu_x$  e  $\mu_y \rightarrow \text{Cor} > 0$

# Significato della covarianza

- La covarianza può assumere valori positivi o negativi (in contrasto con la varianza sempre positiva):

- $\text{Cov}[x, y] < 0$

La variazione della grandezza  $y$  rispetto a  $\mu_y$  tende ad avere segno opposto alla variazione della grandezza  $x$  rispetto a  $\mu_x$

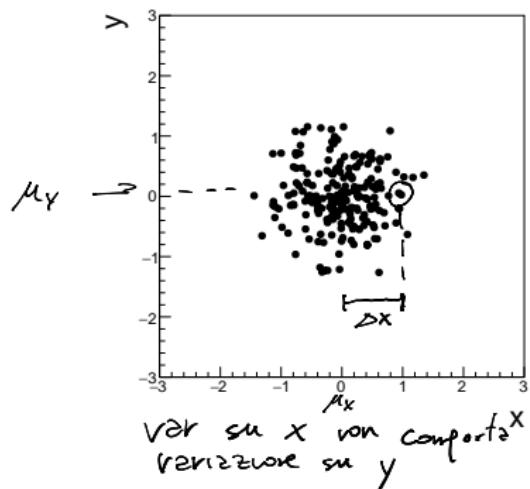


# Significato della covarianza

- La covarianza può assumere valori positivi o negativi (in contrasto con la varianza sempre positiva):

- $\text{Cov}[x, y] = 0$

La variazione della grandezza  $y$  rispetto a  $\mu_y$  non ha nessuna relazione di segno con la variazione della grandezza  $x$  rispetto a  $\mu_x$



# Covarianza e coefficiente di correlazione

- Si dice quindi che la covarianza misura la correlazione tra due variabili aleatorie.
- È utile introdurre una variabile adimensionale che quantifichi la correlazione: il **coefficiente di correlazione**

$$\rho(x, y) \equiv \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

che ha la proprietà  $|\rho| < 1$

- Per dimostrarlo si può considerare una variabile  $z = ax + y$

$$V(ax + y) =$$

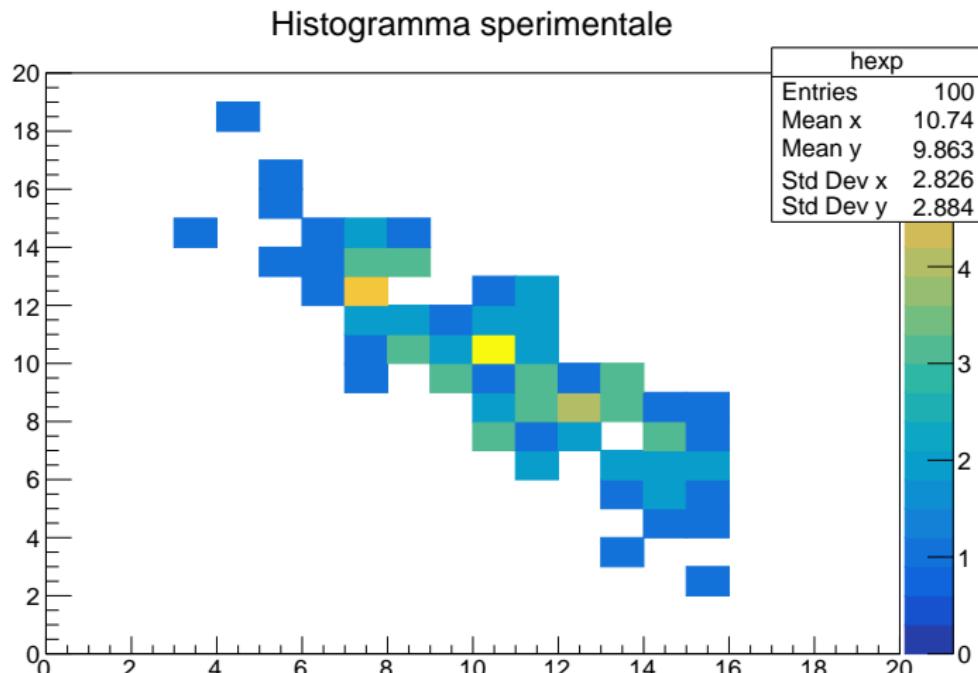
$$\begin{aligned} E[(z - \mu_z)^2] &= a^2 E[(x - \mu_x)^2] + E[(y - \mu_y)^2] + 2aE[(x - \mu_x)(y - \mu_y)] \\ &= a^2 \sigma_x^2 + 6y^2 + 2a \text{cov}(x, y) \geq 0 \quad \text{eq 2° grado} \rightarrow \text{se non voglio sol} \\ \Delta < 0 \quad \Delta_2 &= 4 \text{cov}^2(x, y) - 4\sigma_x^2\sigma_y^2 \leq 0 \quad \frac{\text{cov}^2[x, y]}{\sigma_x^2\sigma_y^2} = \rho^2 \leq 1 \end{aligned}$$

# TH2D::GetCorrelation

Esiste un metodo TH2D::GetCorrelationFactor che vi fornisce il coefficiente di correlazione  $\rho$

```
1 {  
2     ifstream file("2D.dat");  
3     TH2D *hexp = new TH2D("hexp", "Histogramma sperimentale", 20,  
4                             0., 20., 0., 20.);  
5  
6     double tmp1, tmp2;  
7  
8     while(file >> tmp1 >> tmp2){  
9         hexp->Fill(tmp1, tmp2);  
10    }  
11    hexp->Draw("COLZ");  
12    cout << "coeff correlazione " << hexp->GetCorrelationFactor() << endl;  
13 }
```

# TH2D::GetCorrelation



*anticonelate  $\rightarrow$  abbastanza vicine a 1*  
 coeff correlazione = -0.84

# Correlazione ed indipendenza

- Due variabili casuali  $x$  e  $y$  si dicono correlate positivamente, negativamente o scorrelate quando il coefficiente di correlazione è, rispettivamente, positivo, negativo o nullo.
- Due variabili si dicono indipendenti se la pdf congiunta  $p(x, y)$  fattorizza nel prodotto delle due pdf  $p_x(x)$  e  $p_y(y)$ :

$$p(x, y) = p_x(x)p_y(y)$$

- Se due variabili sono indipendenti allora sono scorrelate. Prova:

$$\text{Cov}[x, y] = E[xy] - E[x]E[y] = 0?$$

Calcoliamo  $E[xy] =$

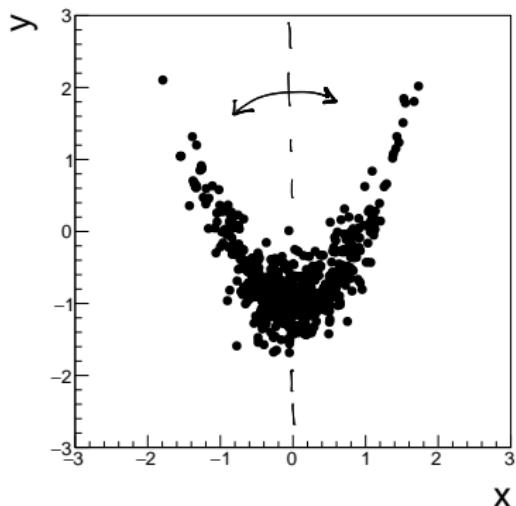
$$\begin{aligned} E[xy] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} xy p(x, y) dx dy = \\ &= \int_{\mathcal{X}} x p(x) dx \cdot \int_{\mathcal{Y}} y p(y) dy = \\ &= E[x] E[y] \end{aligned}$$

- Non è in generale vero invece che due variabili scorrelate siano indipendenti

# Correlazione nulla non implica indipendenza

Se var scorrelate  ~~$\Rightarrow$~~  indipendenti

- Supponiamo di avere una pdf con questa simmetria:  
 $p(-x, y) = p(x, y)$  (cioè la pdf è simmetrica rispetto all'asse y)
- $\text{Cov}[x, y] = E[xy] - E[x]E[y]$
- Per simmetria:  $E[x] = 0$



$$\begin{aligned} E[xy] &= \int_{\Omega_x} \int_{\Omega_y} xyp(x, y) dxdy = \\ &\int_{-\infty}^{\infty} \int_{-\infty}^0 xyp(x, y) dxdy + \int_{-\infty}^{\infty} \int_0^{\infty} xyp(x, y) dxdy = 0 \end{aligned}$$

$\text{Cov}[x, y] = 0$ . Le variabili sono scorrelate. Ma le variabili sono dipendenti.

# Matrice di covarianza

- Supponiamo di avere  $n$  variabili aleatorie:  $\vec{x} = \{x_1, x_2, \dots, x_n\}$
- Si può scrivere la covarianza di ogni coppia come una matrice  $n \times n$

$$V_{ij} = \text{cov}[x_i x_j] = E[(x_i - \mu_{x_i})(x_j - \mu_{x_j})] = \rho_{ij} \sigma_i \sigma_j$$

$$V = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{pmatrix}$$

- Simmetrica
- Diagonale: varianze

# Matrice di correlazione

- Legata alla matrice di covarianza è la matrice  $n \times n$  dei coefficienti di correlazione:

$$\rho_{ij} = \frac{\text{cov}[x_i x_j]}{\sigma_i \sigma_j}$$

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}$$

- Per costruzione, gli elementi della diagonale sono  $\rho = 1$

# Lezione in guscio di noce

- Le misure sperimentali sono descritte da distribuzioni di probabilità.
- Le distribuzioni di probabilità sono caratterizzate dai loro momenti (media, varianza, skewness, curtosi).
- Gli histogrammi sono strumenti utili per “rappresentare” le distribuzioni sperimentali.
- Il momento centrale congiunto di ordine minimo per distribuzioni di due variabili è la covarianza. Il coefficiente di correlazione

$$\rho = \frac{\text{Cov}[x, y]}{\sqrt{\text{Cov}[x, x] \text{Cov}[y, y]}}$$

definisce se le due variabili sono correlate positivamente, negativamente o sono scorrelate ( $\rho > 0$ ,  $\rho < 0$ ,  $\rho = 0$ ).

- Due variabili indipendenti sono sempre scorrelate, non è vero il viceversa.