METODI STATISTICI E COMPUTAZIONALI

Stefania Spagnolo

Dipartimento di Matematica e Fisica, Univ. del Salento

LEZIONE 9-BIS

REGRESSIONE

Testi:
Dispense M. Marchetti Università Di Firenze “**Introduzione all’analisi dei dati multivariati**” (http://local.disia.unifi.it/rampichini/intro.pdf)
H.J.C. Berendens “Data and Error Analysis”

ANALISI MULTIVARIATA

- Col termine ***analisi multivariata*** si indica quell'insieme di metodi statistici usati per analizzare simultaneamente più caratteri.

- In tutte le analisi statistiche multivariate il materiale grezzo e' costituito da un certo numero di eventi (osservazioni) che si vogliono studiare simultaneamente.

- In alcuni casi si cercano correlazioni relazioni funzionali tra gruppi di variabili (***regressione***)

- In altri casi ancora si può essere interessati a ridurre le dimensioni della variabile multipla considerata (identificazione delle ***componenti principali***).

- In alcuni casi l'obiettivo dell'analisi e' semplicemente quello di classificare gli eventi sulla base di tutte le variabili considerate (***classificazione***).

- L'analisi multivariata è comunemente usata in fisica e inconsciamente ne avete già fatto uso. La regressione è un esempio di analisi multivariata

REGRESSIONE

- Supponiamo di avere n coppie di dati xi (note con errori trascurabili) e yi le quali hanno errori σi e supponiamo che le le variabili x e y siano legate dalla relazione: y=f(x,λ) dove λ è un parametro incognito.
- La probabilità che dato un certo x si ottenga un certo y è una probabilità condizionata ad una particolare scelta di λ. Si può valutare cioè per un fissato λ la probabilità di ottenere una particolare coppia di x e y:
  - $P(x_i, y_i ; \lambda)$
- Se ipotizzo che le misure fluttuino in maniera gaussiana allora:
  - $$P(x_i, y_i; \lambda) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - f(x_i;\lambda))^2}{2\sigma^2}}$$
- Il che equivale a supporre che: $y_i = f(x_i; \lambda) + \epsilon_i$ dove gli ϵ_i sono gli errori di misura distribuiti gaussianamente attorno a valore medio 0

REGRESSIONE

- Applicando il metodo di massima verosimiglianza si arriva al metodo dei minimi quadrati.

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n}\left(\frac{1}{\sigma_i}\right)\prod_{i=1}^{n} e^{-\frac{(y_i - f(x_i,\lambda))^2}{2\sigma_i^2}}$$

Calcolando il logaritmo $\ln(L) = \cos t. + \sum_{i=1}^{n} -\frac{(y_i - f(x_i,\lambda))^2}{2\sigma_i^2}$

- ***Massimizzare la likelihood equivale a minimizzare il χ^2***

$$\sum_{i=1}^{n}\frac{(y_i - f(x_i,\lambda))^2}{2\sigma_i^2} = \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - f(x_i,\lambda))^2}{\sigma_i^2} = \frac{1}{2}\chi^2$$

Variabile aleatoria che si distribuisce come un χ^2

- Se interpretiamo L come la probabilità che il particolare set di misure (***x,y***) per un dato valore di λ si verifichi, questa è una $P(\lambda \mid x, y)$ per quel particolare set e ha una distribuzione del tipo:

$$P(\lambda \mid \mathbf{x}, \mathbf{y}) \sim exp\left[-\frac{1}{2}\chi^2(\lambda)\right]$$

REGRESSIONE

- Se gli errori sulle misure x_i non sono trascurabili posso usare lo stesso formalismo, ma le σ_i che compaiono nell'espressione non saranno coincidenti con le $\sigma(y_i)$ (errori della sola variabile y) ma dovranno rappresentare anche l'incertezza sulle x_i

$$\varepsilon_i = y_i - f(x_i, \lambda)$$

Sviluppando fino al primo ordine

$$\sigma_i^2 = \sigma_{yi}^2 + \sigma_{xi}^2 \left(\frac{\partial f}{\partial x} \right)^2_{x=x_i}$$

- Il caso più semplice da trattare è il caso in cui la dipendenza dai parametri della
- funzione f(x) è di tipo lineare $y_i = \sum_{i=0}^{k} \lambda_i x_i$
- In questo caso è possibile risolvere analiticamente il problema della determinazione del minimo della funzione S introdotta in precedenza.
  Nel caso di dipendenze non lineari si cerca di linearizzarla altrimenti la soluzione numerica è spesso l'unica possibile.

REGRESSIONE - TOY MC

- Ipotizziamo che tra due variabili esista una relazione lineare del tipo ***y=5x+2*** che rappresenta un certo processo fisico che stiamo studiando.
  Assumiamo di disporre di 10 misure sperimentali.

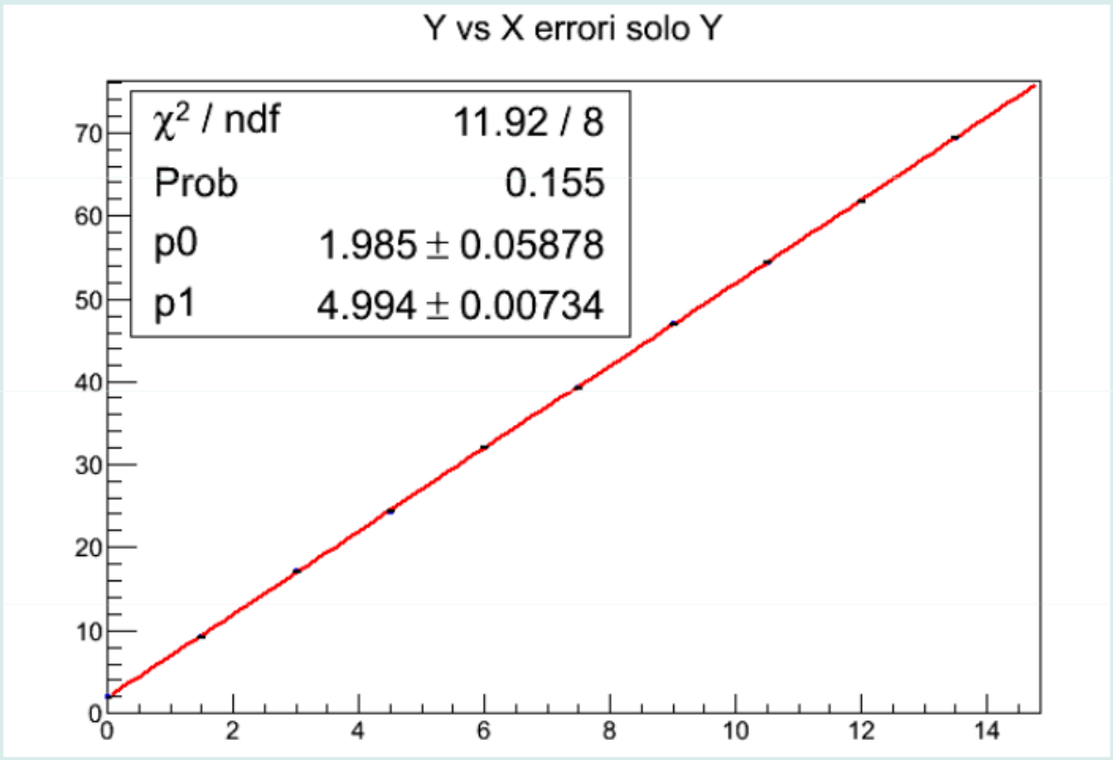
- Predisponiamo un Toy Monte Carlo che ci permetta di simulare l'esperimento. Il Toy Monte Carlo ci permetterà di valutare le varie situazioni possibili e come queste influenzano l'analisi e il processo di adattamento dei dati alla f
  - Errori su x trascurabili, errori su y importanti
  - Errori su y trascurabili, errori su x importanti
  - Errori importanti su x e su y
  - Incertezza sperimentali
    - stimati correttamente
    - sottostimati
    - sovrastimata

REGRESSIONE - TOY MC

```
    TRandom R(1234567);
    double xTrue[NMISURE],x[NMISURE];
    double yTrue[NMISURE],y[NMISURE];
    double ex[NMISURE],ey[NMISURE];
    double stepX=1.5;
    for (int i=0; i<NMISURE; i++) {
// questi sono i valori "veri" del processo fisico
        xTrue[i]=i*stepX;
        yTrue[i]=5*xTrue[i]+2;
// assume che gli errori di misura sulla y
// siano distribuiti gaussianamente con sigma=0.1
        double eps=R.Gaus(0,0.1);
// questo e' il valore misurato.
        y[i]=yTrue[i]+eps;
// Errori sulle x nulli e sulle y sempre uguali a 1, 0.1, 0.01
        ex[i]=0;
        ey[i]=1;
    }
```

- NMISURE è il numero di coppie di punti di cui si dispone (10 nel nostro caso).
- I vettori xTrue e yTrue contengono i valori “veri” del processo sotto analisi.
- I vettori x e y contengono i valori “misurati” delle due grandezze.
- I vettori ex e ey contengono la nostra stima delle incertezze statistiche su ogni misura

REGRESSIONE - TOY MC



Configurazione I

X "esatti" e Y soggette ad errore sperimentale.

Errore assunto (ey) e fluttuazione dei punti coincidono

- p_0, p_1 valori dei parametri di best fit, ossia valori che minimizzano il χ^2
- χ^2 minimo =>>> $S_0 = \sum_{i=1}^{N} \frac{(y_i - f(x_i; \vec{\lambda}))^2}{\sigma_i^2}$, con $\vec{\lambda} = (p_0, p_1)$
- n.d.f. = number of degrees of freedom = N - dimensione di $\vec{\lambda}$

REGRESSIONE

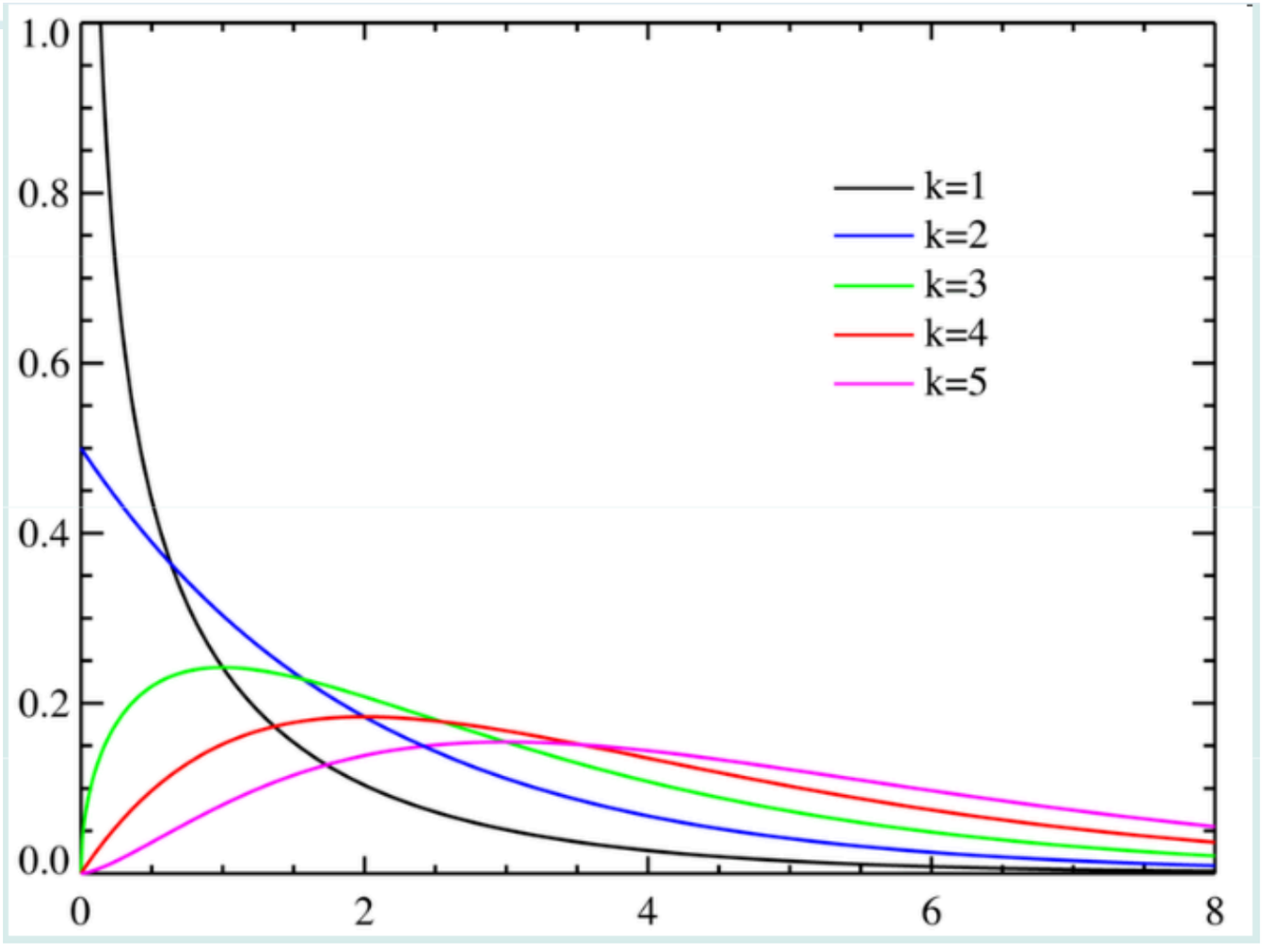
- Per valutare la bontà di un fit si può ricorrere al test del χ^2.
  Il test del χ^2 è affidabile solo nel caso in cui si ha una ragionevole accuratezza nella stima delle incertezze σ_i dei singoli punti.

- In questo caso il modo di procedere vi è già noto:
  - 1) Si determina il valore del χ^2 in corrispondenza del valore dei parametri che lo minimizzano
  - 2) Si identificano il numero di gradi di libertà che in questo caso saranno pari al numero di punti meno il numero di parametri
  - 3) Si assume che la grandezza S_0 sia distribuita come una χ^2
  - 4) Si fissa un intervallo di confidenza e si valuta se il valore ottenuto e' contenuto o meno nell'intervallo di confidenza.
- E' preferibile fare un test a due code e escludere le regioni in cui S_0 assume valori troppo grandi o troppo piccoli.

REGRESSIONE

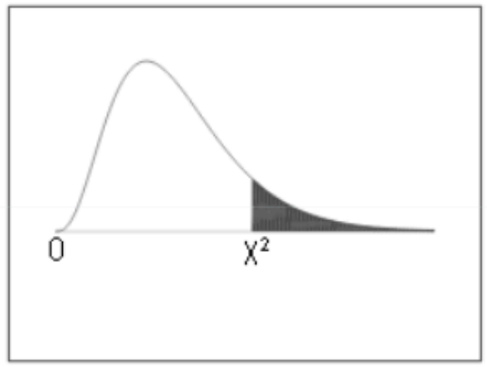
$$p_\nu(\chi^2) = \frac{1}{2^{\frac{\nu}{2}} \cdot (\frac{\nu}{2}-1)!}(\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}}$$

- Media v, Varianza 2v
- Per v -> inf. tende a una gaussiana



REGRESSIONE

Chi-Square Distribution Table

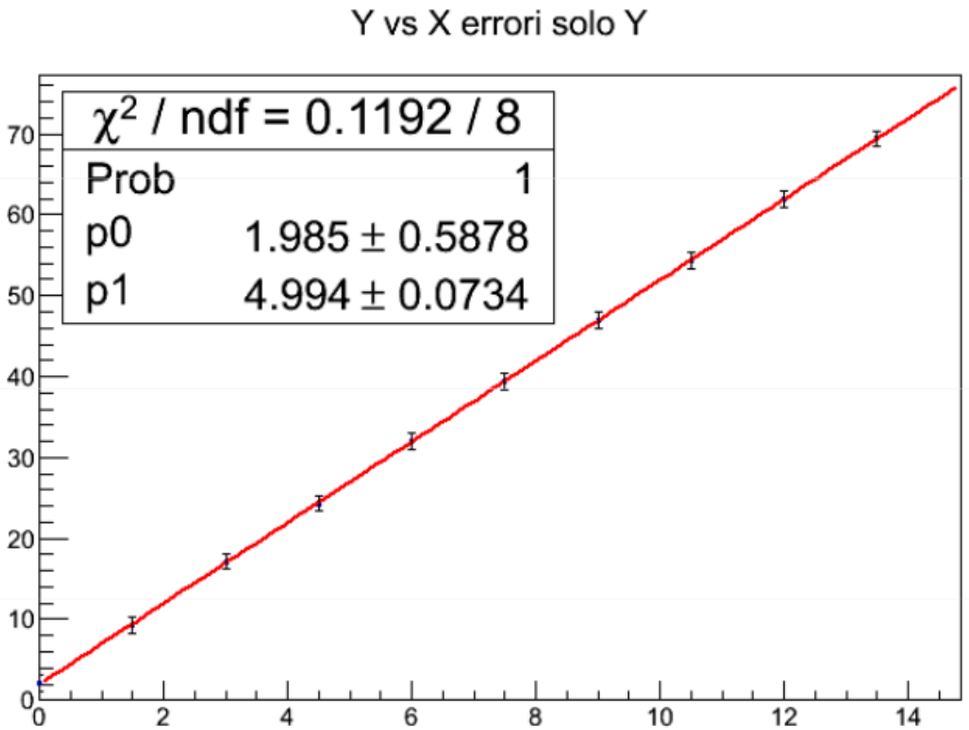


The shaded area is equal to α for $\chi^2 - \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

Esempio di tabella di χ^2

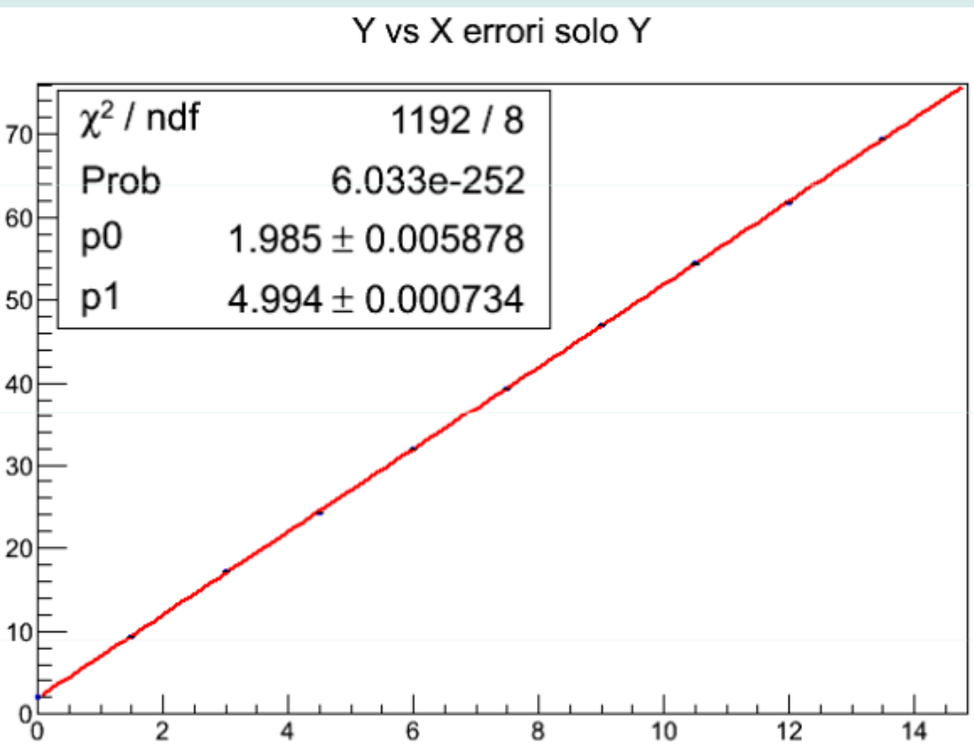
REGRESSIONE



Configurazione II

X "esatti" e Y soggette ad errore sperimentale.

Errore assunto (ey) e molto più grande delle fluttuazione dei punti.



Configurazione III

X "esatti" e Y soggette ad errore sperimentale.

Errore assunto (ey) e molto più piccolo delle fluttuazione dei punti.

ERRORI SUI PARAMETRI DETERMINATI DAL FIT E LORO UTILIZZO

- Nello stimare il valore dei parametri che meglio si adatta ad un insieme di misure è rilevante la determinazione dell'incertezza associata ai parametri stessi.
- Nel processo di identificazione dei parametri ottimali viene fornita una matrice di covarianza che tiene conto e dell'incertezza sul singolo parametro e del gradi di correlazione dei parametri tra di loro. La matrice di covarianza è una matrice simmetrica sulla cui diagonale compaiono le varianze dei singoli parametri.
- Nel caso semplice di due parametri ***a*** e ***b***, come nel nostro esempio, la matrice di covarianza assume la forma:

$$\begin{pmatrix} \sigma_a^2 & \mathrm{cov}(a,b) \\ \mathrm{cov}(a,b) & \sigma_b^2 \end{pmatrix}$$

- I termini diagonali sono le varianze associate ai singoli parametri. I termini non diagonali rappresentano le rispettive covarianze che quantificano il grado di correlazione tra i parametri.

ERRORI SUI PARAMETRI DETERMINATI DAL FIT E LORO UTILIZZO

- Le classi che gestiscono il fit in Root ci danno la stima della matrice di covarianza

```
44  // definisco una funzione parametrica
45      TF1 * fun= new TF1("fun","[0]+[1]*x",0,20);
46  // La uso per fare un fit
47      gr->Fit("fun");
48  // estraggo i risultati della regressione
49      TVirtualFitter * fitter = TVirtualFitter::GetFitter();
50  // Estraggo e stampo il valore dei parametri.
51      double p1=fitter->GetParameter(1);
52      double p0=fitter->GetParameter(0);
53      printf("\n p1=%f p0=%f\n",p1,p0);
54  // estraggo e stampo la matrice di covarianza
55      double * covMatrix = fitter->GetCovarianceMatrix();
56      printf("\n c11=%f   c12=%f\n ",covMatrix[0],covMatrix[1]);
57      printf("\n c21=%f   c22=%f\n ",covMatrix[2],covMatrix[3]);
```

```
FCN=0.417844 FROM MIGRAD    STATUS=CONVERGED      30 CALLS          31 TOTAL
                    EDM=1.9285e-014    STRATEGY= 1      ERROR MATRIX ACCURATE
 EXT PARAMETER                                   STEP         FIRST
 NO.   NAME      VALUE            ERROR          SIZE      DERIVATIVE
  1  p0           2.01184e+000  2.99125e-001  9.35706e-005  1.21548e-006
  2  p1           4.99013e+000  3.73543e-002  1.16850e-005  8.67053e-006

p1=4.990132 p0=2.011839

c11=0.089476   c12=-0.009419

c21=-0.009419   c22=0.001395
```

Stampa aggiunta manualmente

ERRORI SUI PARAMETRI DETERMINATI DAL FIT E LORO UTILIZZO

```
FCN=0.417844 FROM MIGRAD    STATUS=CONVERGED      30 CALLS          31 TOTAL
                    EDM=1.9285e-014    STRATEGY= 1      ERROR MATRIX ACCURATE
  EXT PARAMETER                                   STEP         FIRST
  NO.   NAME      VALUE            ERROR          SIZE      DERIVATIVE
   1  p0           2.01184e+000   2.99125e-001   9.35706e-005   1.21548e-006
   2  p1           4.99013e+000   3.73543e-002   1.16850e-005   8.67053e-006

p1=4.990132  p0=2.011839

c11=0.089476    c12=-0.009419

c21=-0.009419    c22=0.001395
```

Errore del parametro p0: σ_{p0}= 0.299125
c11= σ_{p0}^2=0.089476

Analogamente per p_1

- A cosa serve la covarianza?
  - Ipotizziamo che una volta che si sono identificati i parametri della funzione incognita vogliamo calcolare il valore atteso della variabile y in un punto x diverso dai valori sperimentali disponibili. Qual e' il giusto errore da associare a questa stima?

ERRORI SUI PARAMETRI DETERMINATI DAL FIT E LORO UTILIZZO

$$y = ax + b$$

$$\mathrm{var}(y) = \mathrm{var}(ax) + \mathrm{var}(b) + 2\,\mathrm{cov}(ax, b)$$

$$\mathrm{cov}(ax, b) = x\,\mathrm{cov}(a, b)$$

$$\mathrm{var}(ax) = x^2\,\mathrm{var}(a)$$

$$\mathrm{var}(a) = 1.497E-3$$

$$\mathrm{var}(b) = 9.257E-2$$

$$\mathrm{cov}(a, b) = -9.919E-3$$

Senza tener conto del termine di covarianza

$$\mathrm{var}(y) = x^2\,\mathrm{var}(a) + \mathrm{var}(b) = 0.241$$

$$\mathrm{var}(y) = 10^2\,\mathrm{var}(a) + \mathrm{var}(b) + 20\,\mathrm{cov}(a, b) = 0.1497 + 9.256 \cdot 10^{-2} - 0.19838 = 0.04388$$

Y(10)=42.6±0.2

- Come verifichiamo che sia questa la stima corretta dell'errore ?

COME SONO DETERMINATI GLI ERRORI NEL FIT

- Come sono determinati gli errori sui parametri ottenuti con il metodo dei minimi quadrati o massima verosimiglianza ?
- La likelihood rappresenta la probabilità con cui posso osservare il mio campione di punti sperimentali (y_i, x_i) data la legge oggetto di studio $y = f(x; \vec{\lambda})$
  - I valori di $\vec{\lambda}$ determinati cercando il minimo del χ^2 (o il massimo della Likelihood) sono variabili aleatorie (funzioni di variabili aleatorie)
  - Posso interpretare la likelihood come la distribuzione di probabilità congiunta dei parametri in studio $\vec{\lambda}$
- Abbiamo visto che vale la relazione: $P(\vec{\lambda} \,|\, \mathbf{x}, \mathbf{y}) \sim exp\left[-\frac{1}{2}\chi^2(\vec{\lambda}) \right]$
- Per un solo parametro $P(\lambda \,|\, \mathbf{x}, \mathbf{y}) \sim exp\left[-\frac{1}{2}\chi^2(\lambda) \right]$
- Per valutare l'errore sul parametro λ dobbiamo capire come varia la sua distribuzione di probabilità attorno al valore di best fit (valore di minimo del χ^2)

COME SONO DETERMINATI GLI ERRORI NEL FIT

- Consideriamo il caso più semplice di tutti:
  - Una sola misura ***y***, in corrispondenza di un valore x, ***affetta da errore σ gaussiano***
  - λ_0 =valore di best fit
- Dalla L(y,λ) si vede che
  - se $y = y_{\pm 1\sigma} = f(x;\lambda_0) \pm \sigma$
    - si ha $\Delta\chi^2 = 1$, $\Delta \ln L = -1/2$
    - $P(y_{-1\sigma} < y < y_{+1\sigma})$ = 68.3%
  - $y = y_{\pm 2\sigma} = f(x;\lambda_0) \pm 2\sigma$
    - si ha $\Delta\chi^2 = 4, \Delta \ln L = -2$
    - $P(y_{-2\sigma} < y < y_{+2\sigma})$ = 95.5%

$$L(y;\lambda) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y - f(x;\lambda))^2}{2\sigma^2}\right)$$

$$lnL(y;\lambda) \sim \left(-\frac{(y - f(x;\lambda))^2}{2\sigma^2}\right) = \left(-\frac{\chi^2}{2}\right)$$

$$\chi^2 = \frac{(y - f(x;\lambda))^2}{\sigma^2}$$

- Definiamo come errore su λ l'intervallo Δλ tale per cui

$$\Delta\chi^2 = \frac{(y - f(x;\lambda_0 + \Delta\lambda))^2}{\sigma^2} = 1$$

Infatti la probabilità che $\lambda_0 - \Delta\lambda < \lambda < \lambda_0 + \Delta\lambda$ corrisponde alla probabilità che il valore del χ^2 sia compreso tra il suo valore minimo S_0 e S_0+1, e questa corrisponde a

$P(y_{-1\sigma} < y < y_{+1\sigma})$ = 68.3%

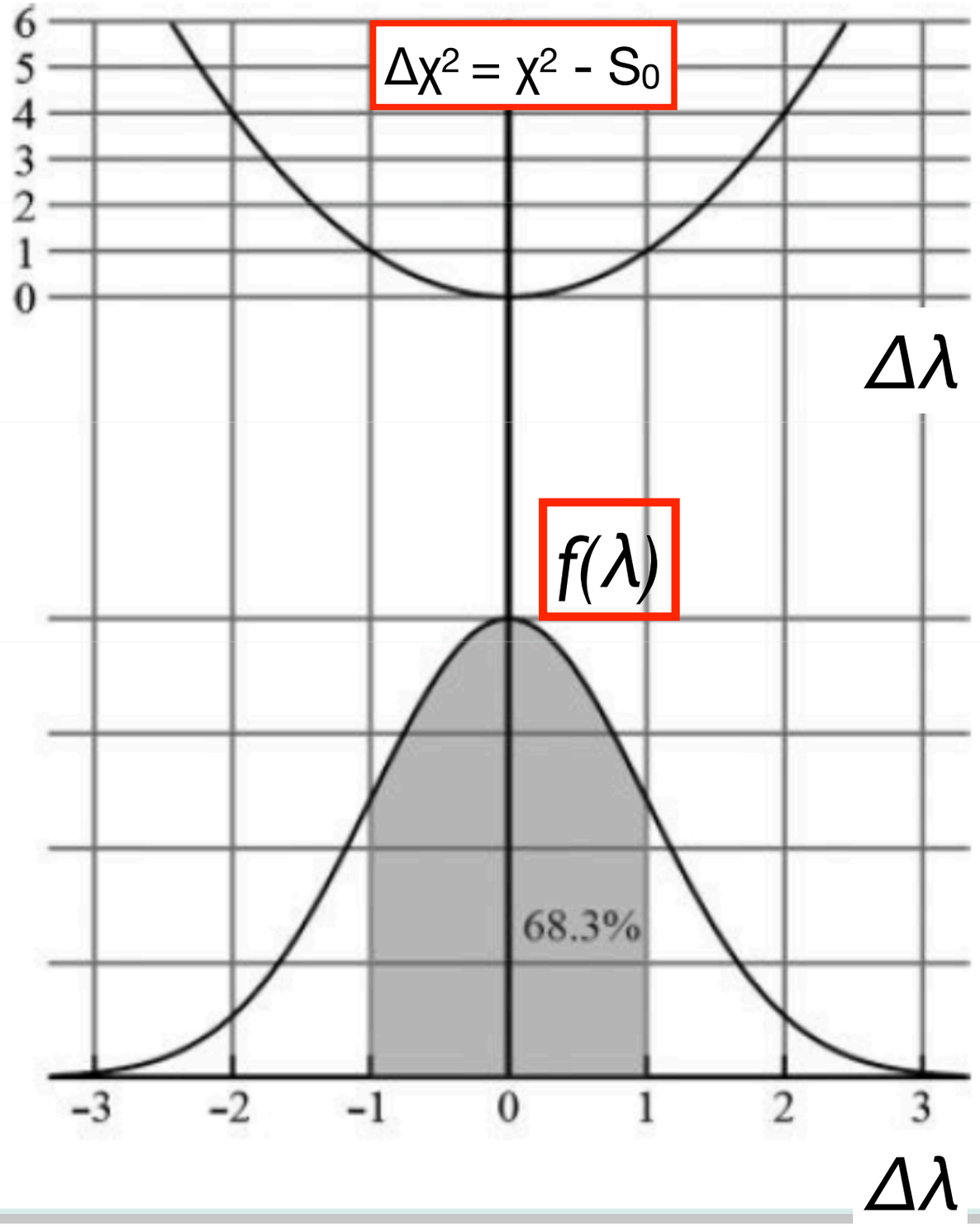
COME SONO DETERMINATI GLI ERRORI NEL FIT

- ***Nel caso generale di N misure e 1 parametro***
  - Se abbiamo N misura y_i, in corrispondenza di valori x_i, ***affette da errore σi*** *(rappresentativi eventualmente anche delle incertezze alle x)* ***non necessariamente gaussiane,*** nel limite di N grande, la $P(\lambda \,|\, \mathbf{x}, \mathbf{y}) \sim exp\left[-\frac{1}{2}\chi^2(\lambda) \right]$ rappresenta una distribuzione di probabilità approssimativamente gaussiana per λ.

NOTA: questo significa che il χ2 = -ln *L* è una funzione quadratica di λ

  - Se $\boldsymbol{\lambda_0}$ =valore di best fit e **Δλ** è l'ampiezza dell'intervallo in corrispondenza del quale si ha $\Delta\chi^2 = 1$ (e allo stesso tempo $\Delta \ln L = -1/2$) la probabilità $P(\lambda_0 - \Delta\lambda < \lambda < \lambda_0 + \Delta\lambda) = 68.3\,\%$
    - Quindi possiamo dire che $\lambda = \lambda_0 \pm \Delta\lambda$ con il 68.3% di C.L.
  - Se **Δλ** è l'ampiezza dell'intervallo in corrispondenza del quale si ha $\Delta\chi^2 = 4$ (e allo stesso tempo $\Delta \ln L = -2$) la probabilità $P(\lambda_0 - \Delta\lambda < \lambda < \lambda_0 + \Delta\lambda) = 95.5\,\%$ —>> $\lambda = \lambda_0 \pm \Delta\lambda$ con il 95.5% di C.L.

COME SONO DETERMINATI GLI ERRORI NEL FIT

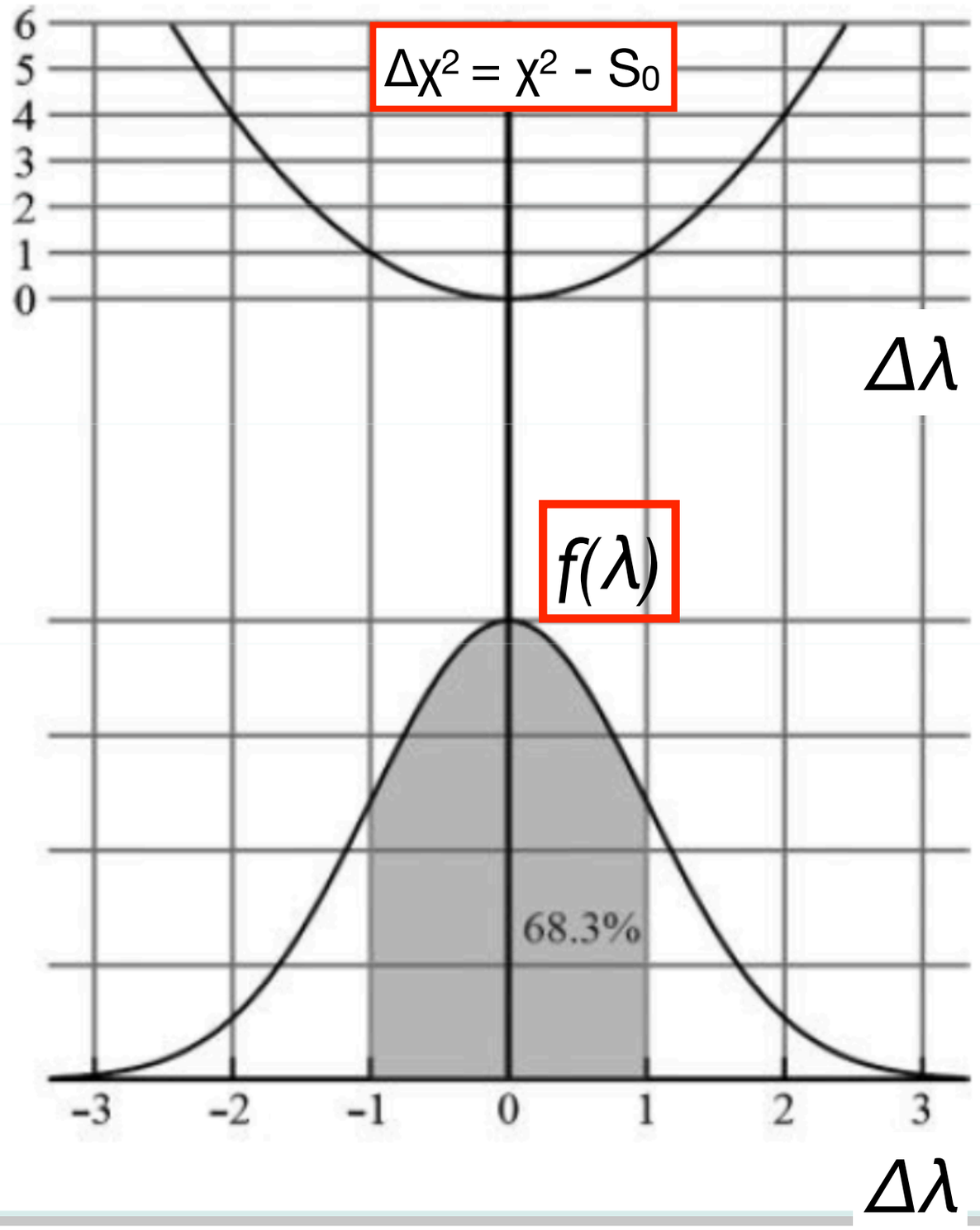


- In pratica, l'intervallo di y attorno al valore "vero" [$f(x,\lambda_0)$] che corrisponde a un incremento del χ^2 pari a 1 rappresenta un intervallo in cui y ricade con probabilita' del 68.3% Dao il valore λ del parametro

- Analogamente

- l'intervallo di λ, attorno al valore "vero" [λ_0] che corrisponde a un incremento del χ^2 pari a 1, rappresenta un intervallo in cui λ ricade con probabilita' del 68.3% dato il risultato y della misura

COME SONO DETERMINATI GLI ERRORI NEL FIT



- Quindi operativamente determinare con un fit il valore di un parametro, con errore, significa:
  - Trovare il valore di λ che minimizza il χ2 (o massimizza la log-likelihood), λ_0
  - Trovare l'intervallo Δλ attorno a λ_0 tale che per $\lambda = \lambda_0 +/- \Delta\lambda$ il $\chi^2 = S_0 + 1$

DOMANDA

Come si confronta questo metodo per ricavare gli errori su uno o piu' parametri, con l'espressione approssimata degli errori sui parametri che abbiamo quotato per il metodo di massima verosimiglianza ???

METODO DELLA MASSIMA VEROSIMIGLIANZA

RICORDIAMO (dalla lezione sul metodo di massima verosimiglianza)

- Per definizione la varianza di $\hat{\lambda}$ e'

$$\sigma^2(\hat{\lambda}) = \int (\hat{\lambda}(x') - \hat{\lambda}_{medio})^2 L(x'_1, x'_2, \ldots, x'_n;\ \lambda)\ dx'_1 dx'_2 \ldots dx'_n$$

- Difficile da calcolare, tipicamente si affronta numericamente
- Una relazione approssimata, valida per ***n*** grande, e'

$$\sigma^2(\hat{\lambda}) = -\left(\frac{d^2 \ln L}{d\lambda^2}\right)^{-1}$$

- Con lo stesso livello di approssimazione se la distribuzione di probabilita' dipende da più' parametri, l'inverso della matrice i cui elementi sono

$$U_{ij} = -\frac{d^2 \ln L}{d\lambda_i\ d\lambda_j}$$

rappresenta la matrice di covarianza, i cui elementi diagonali sono le varianze e quelli fuori diagonali le covarianze di coppie di variabili

METODO DELLA MASSIMA VEROSIMIGLIANZA

NOTA: Se abbiamo N misura y_i, in corrispondenza di valori x_i, affette da errore σi **n**on necessariamente gaussiane, la Likelihood rappresenta un pdf approssimativamente gaussiana per λ. Questo significa che il χ2 = -ln *L* è una funzione quadratica di λ, per esempio

$$L^\star = -\ln L = A\lambda_1^2 + B\lambda_2^2$$

RICORDIAMO (dalla lezione sul metodo di massima verosimiglianza)

$$-\frac{d^2 \ln L}{d\lambda_1^2} = 2A \qquad -\frac{d^2 \ln L}{d\lambda_2^2} = 2B \qquad -\frac{d^2 \ln L}{d\lambda_1\, d\lambda_2} = 0$$

$$\begin{pmatrix} \frac{1}{2A} & 0 \\ 0 & \frac{1}{2B} \end{pmatrix} \begin{pmatrix} 2A & 0 \\ 0 & 2B \end{pmatrix} = \mathbb{I}$$

Matrice **inversa** della matrice di derivate seconde della logL **= matrice di covarianza**

$$\sigma(\lambda_1) = \sqrt{\frac{1}{2A}}$$

$$\sigma(\lambda_2) = \sqrt{\frac{1}{2B}}$$

$\hat{\lambda}_1 = 0 \quad \hat{\lambda}_2 = 0$ Valori di best fit, danno $L^\star = 0$

$$\Delta L^\star = L^\star(\hat{\lambda}_1 + \sigma(\lambda_1), \hat{\lambda}_2) - L^\star(\hat{\lambda}_1, \hat{\lambda}_2) = \frac{1}{2}$$

$$\Delta L^\star = L^\star(\hat{\lambda}_1, \hat{\lambda}_2 + \sigma(\lambda_2)) - L^\star(\hat{\lambda}_1, \hat{\lambda}_2) = \frac{1}{2}$$

DOMANDA

Come si confronta questo metodo per ricavare gli errori su uno o piu' parametri, con l'espressione approssimata degli errori sui parametri che abbiamo quotato per il metodo di massima verosimiglianza ???

Bene: **coincidono** per N grande, perche' la L rappresenta è una funzione gaussiana di λ e quindi il la $\log L = -(1/2)\chi^2$ è quadratica in λ

COME SONO DETERMINATI GLI ERRORI NEL FIT

- ***Nel caso generale di N misure e m parametri***
  - Se abbiamo N misura ***yi***, in corrispondenza di valori xi, ***affette da errore σi*** *(rappresentativi eventualmente anche delle incertezze alle x)* ***non necessariamente gaussiane,*** nel limite di N grande, la $P(\vec{\lambda}\,|\,\mathbf{x},\mathbf{y}) \sim exp\left[-\frac{1}{2}\chi^2(\vec{\lambda})\right]$ rappresenta una distribuzione di probabilità congiunta approssimativamente gaussiana per i parametri $(\lambda_1, \lambda_2, \ldots, \lambda_m)$
  - **NOTA**: questo significa che il χ2 = -ln *L* è una funzione approssimativamente quadratica di λi
  - Se $\vec{\lambda}_0$ = valore di best fit, gli errori $\Delta\vec{\lambda}_i$ ad un certo C.L. sono determinati in corrispondenza di incrementi del valore di χ^2 che garantiscano la copertura richiesta dal livello di confidenza
  - Quali sono in questo caso i valori $\Delta\chi^2$ che corrispondono a intervalli di probabilità tipici ??
    - Chiameremo intervallo a 1 sigma quello a cui corrisponde una prob -> 68.3%
    - 2 sigma -> 95.5%
    - ... *Dipende da m*

COME SONO DETERMINATI GLI ERRORI NEL FIT

Table 40.2: Values of $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of M parameters.

$(1-\alpha)$ (%)	$M=1$	$M=2$	$M=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

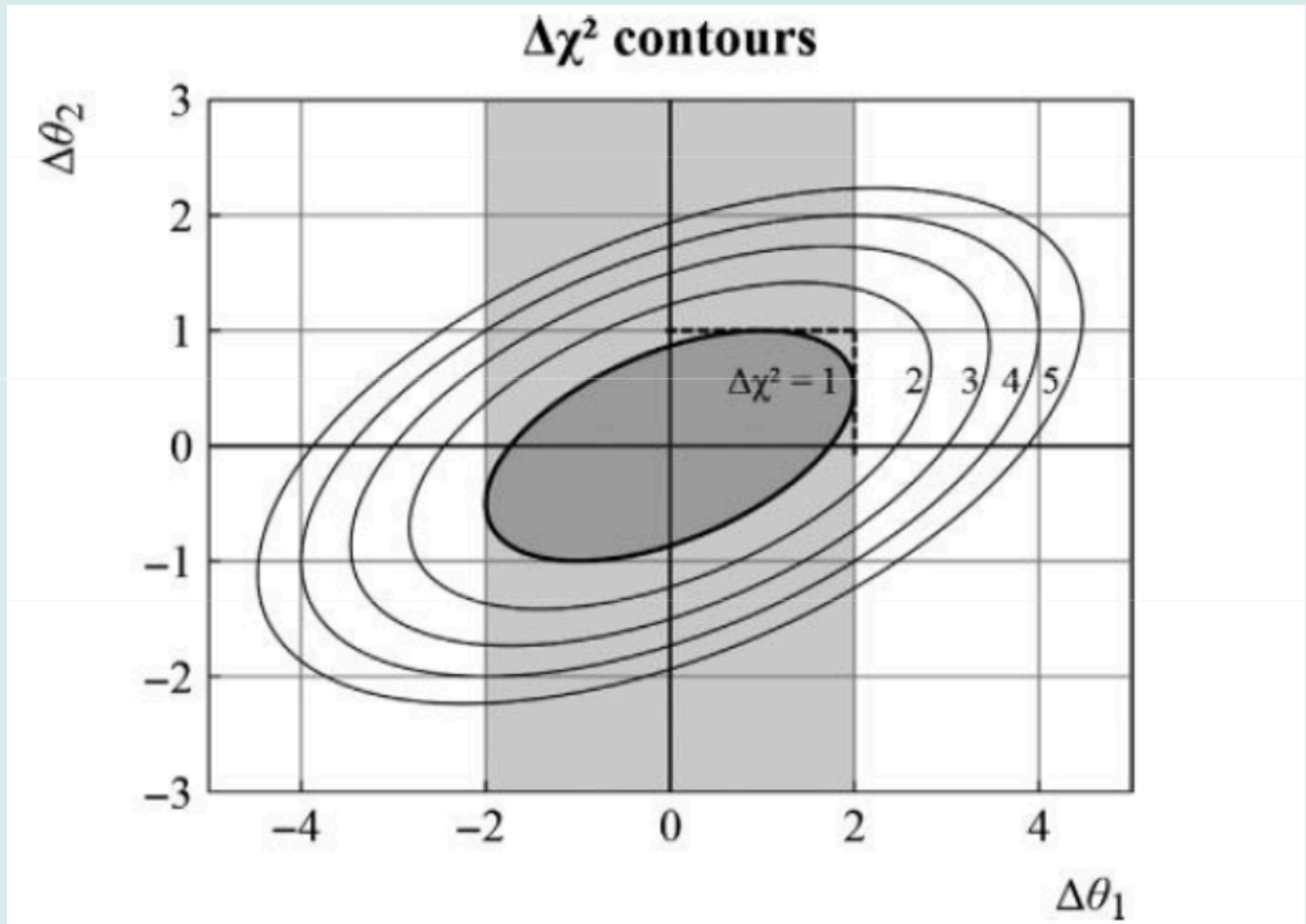
1 sigma

2 sigma

Numero di parametri determinati contemporaneamente nel fit

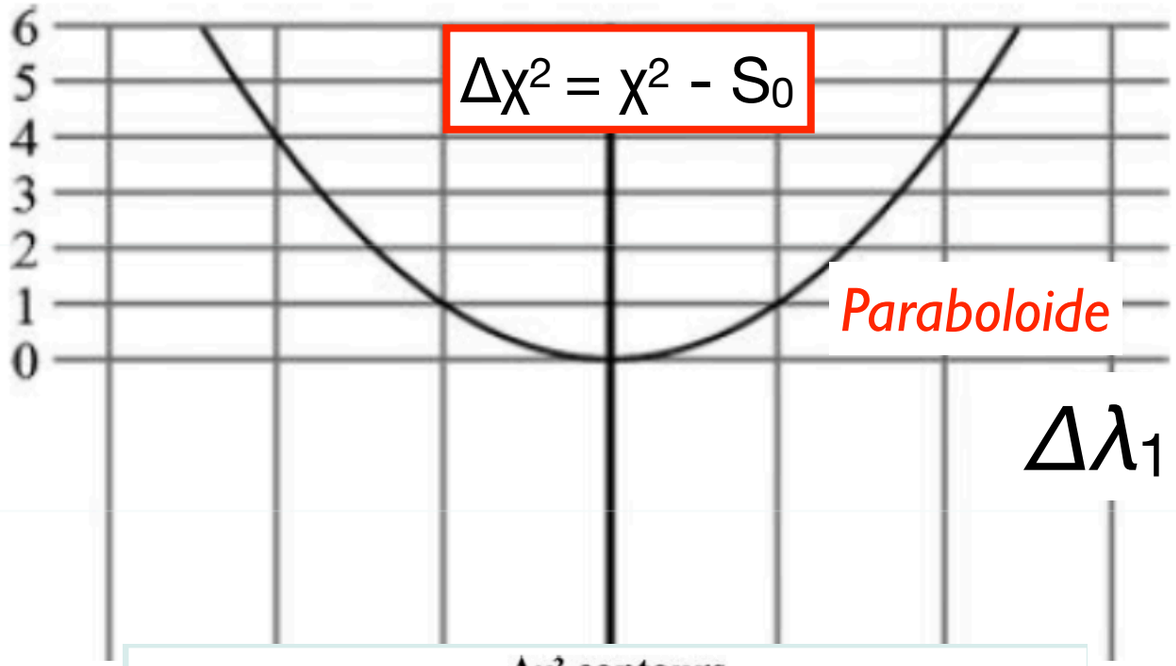
https://pdg.lbl.gov/2022/reviews/contents_sports.html

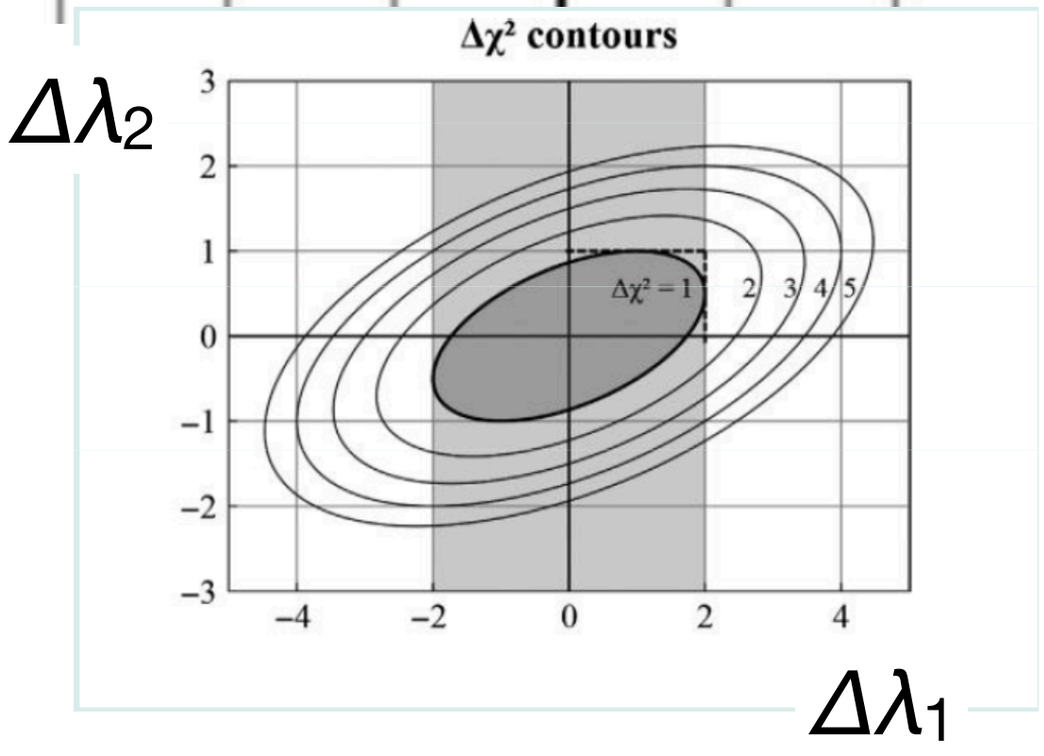
COME SONO DETERMINATI GLI ERRORI NEL FIT



- Curve di livello $\Delta\chi^2$ = costante

COME SONO DETERMINATI GLI ERRORI NEL FIT





- Quindi operativamente determinare con un fit contemporaneamente il valore di piu' di un parametro, con relativi errore, significa:
  - Trovare i valori λ_i che minimizzano il χ^2 (o massimizzano la log-likelihood)
  - Trovare la regione (un ellissoide, se la f è una funzione lineare dei λ_i) nello spazio dei parametri attorno al punto di minimo $\vec{\lambda}_0$ entro la quale il χ^2 varia rispetto al minimo meno di un certo $\Delta\chi^2$ (determinato dal C.L. fissato)
  - Stimare i limiti di questa regione per ciascuno dei parametri e determinare le correlazioni tra i parametri

COME SONO DETERMINATI GLI ERRORI NEL FIT

Stimare i limiti di questa regione per ciascuno dei parametri e determinare le correlazioni tra i parametri

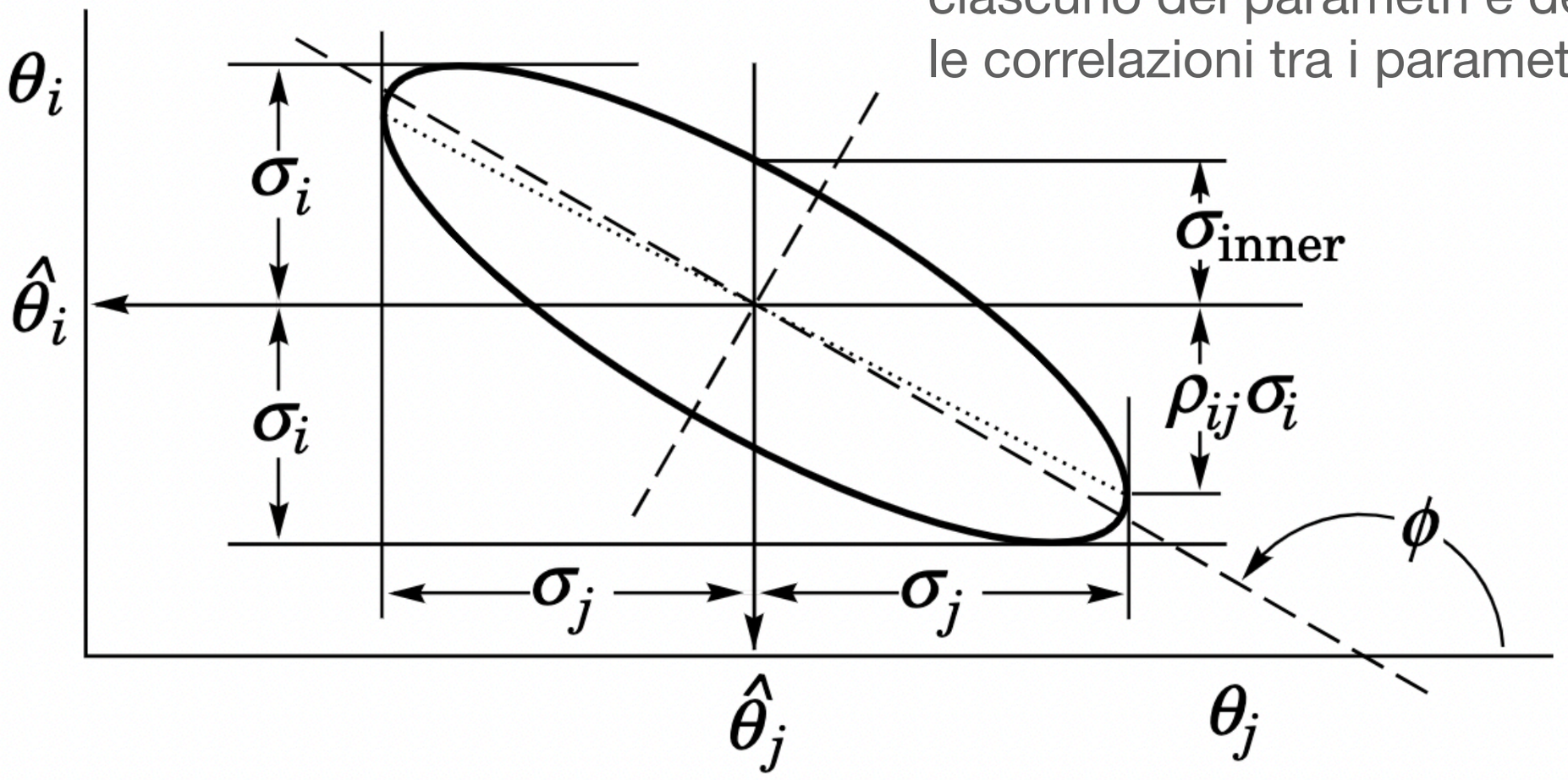
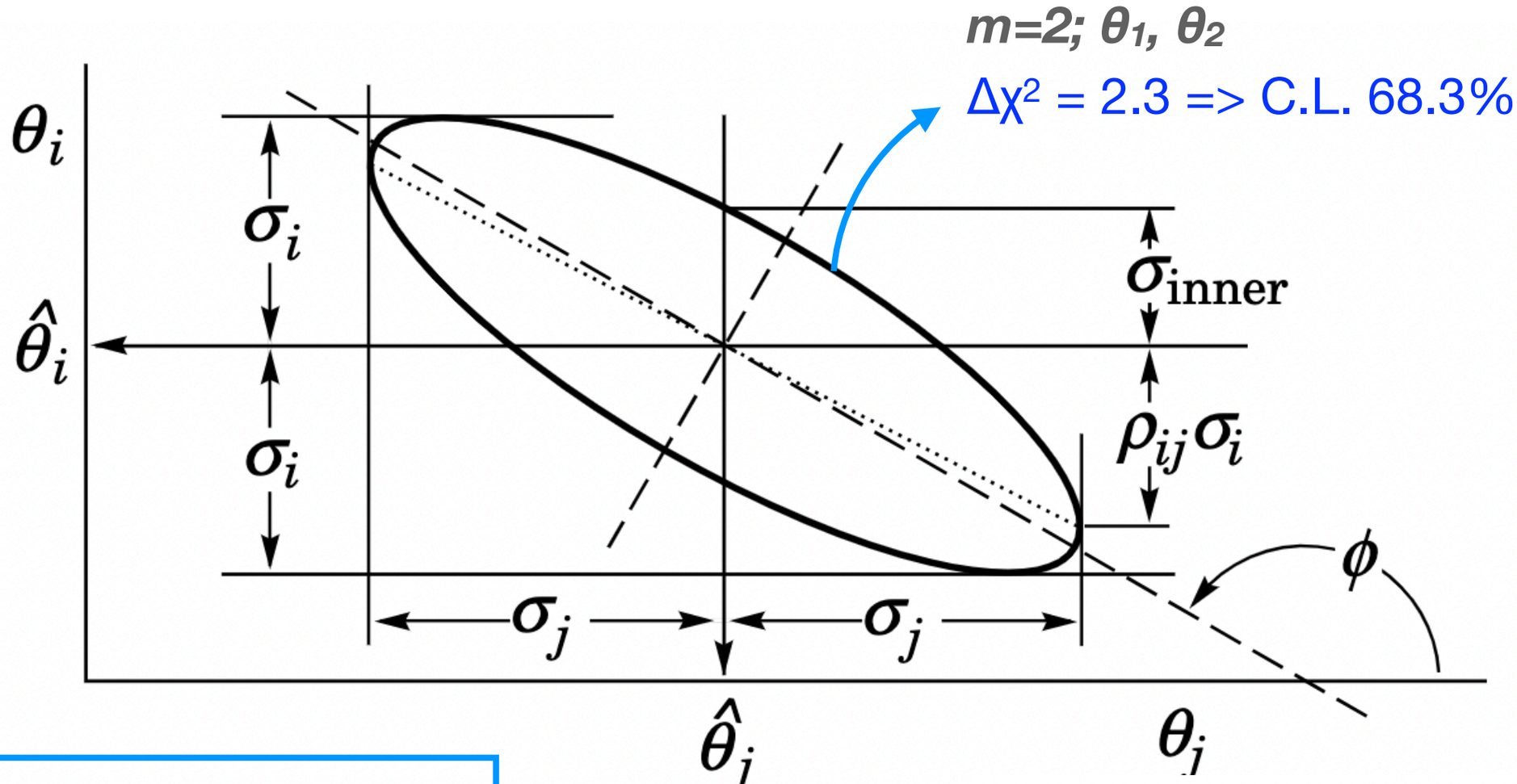


Figure 40.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In the case shown the correlation is negative.

https://pdg.lbl.gov/2022/reviews/contents_sports.html

COME SONO DETERMINATI GLI ERRORI NEL FIT



al CL del 68%

$$\hat{\theta}_i - \sigma_i < \theta_i < \hat{\theta}_i + \sigma_i$$
$$\hat{\theta}_j - \sigma_j < \theta_j < \hat{\theta}_j + \sigma_j$$

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2}$$

$$\rho_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]/\sigma_i\sigma_j$$

COME SONO DETERMINATI GLI ERRORI NEL FIT

- Spesso in questo genere di rappresentazioni essendo i parametri correlati ha poco senso parlare di 1 o 2 sigma ma quello che si fa è di identificare regioni entro le quali i parametri sono contenuti con una certa probabilità.
- Il concetto è molto simile a quello che abbiamo affrontato nel caso di intervalli di confidenza.
  - 1. Si stabilisce un livello di confidenza (es.:90%)
  - 2. Si traccia la curva del χ^2 che identifica la regione entro la quale si possono muovere i parametri affinché la probabilità congiunta dei due sia del 90%.
- L'identificazione analitica della curva è possibile solo nel caso di dipendenza lineare dai parametri. In questo caso, infatti, è possibile dimostrare che regioni che assumono lo stesso valore di χ^2 nello spazio dei due (n) parametri sono delle ellissi (ellissoide n-dimensionali). In tutti gli altri casi occorre ricorrere a metodi numerici.

ESEMPIO

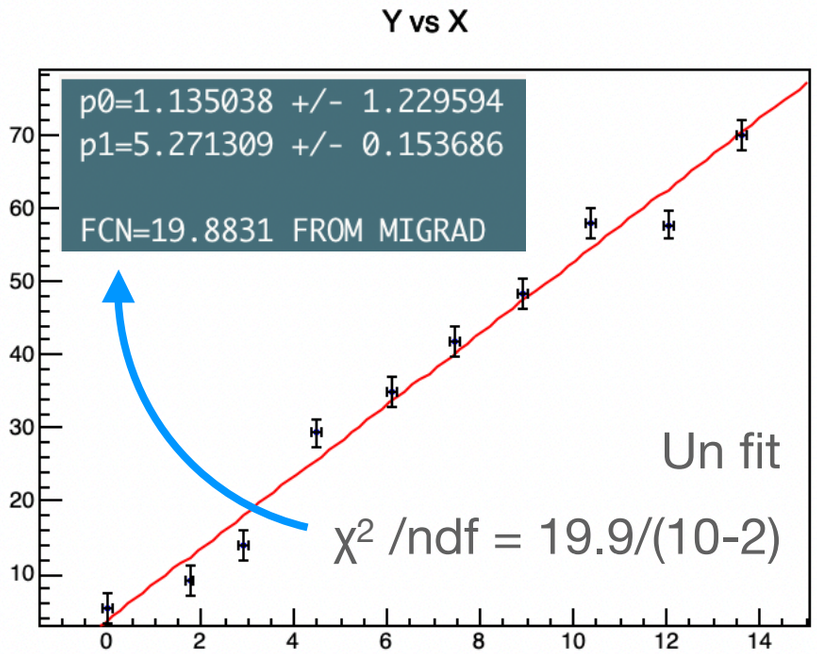
- x -> fluttua gaussianamente con sigma $\sigma_x = 0.1$
- y -> fluttua gaussianamente con sigma $\sigma_y = 2.0$
-

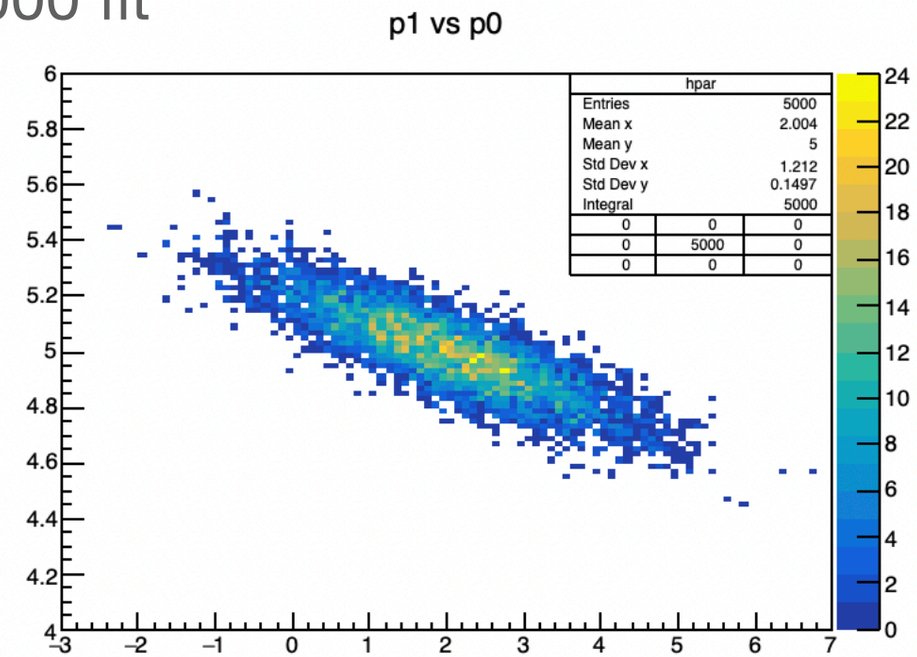
```
// definisco una funzione parametrica
TF1 * fun= new TF1("fun","[0]+[1]*x",0,20);
TGraphErrors *gr;
for (int iter=0; iter<5000; ++iter)
  {
    for (int i=0; i<NMISURE; i++) {
      // questi sono i valori "veri" del processo fisico^M
      xTrue[i]=i*stepX;
      yTrue[i]=5*xTrue[i]+2;
      // assume che gli errori di misura sulla y^M
      // siano distribuiti gaussianamente con sigma=0.1^M
      double eps=R.Gaus(0,2.);
      // questo e' il valore misurato di Y.
      y[i]=yTrue[i]+eps;
      eps=R.Gaus(0,0.1);
      // questo e' il valore misurato di X.
      x[i]=xTrue[i]+eps;
      // Errori sulle x nulli e sulle y sempre uguali a 1, 0.1, 0.01^M
      ex[i]=0.1;
      ey[i]=2.;
    }
    // costruisco il grafico^M
    gr= new TGraphErrors(NMISURE,x,y,ex,ey);
    // aspetto estetico.....^M
    gr->SetTitle("Y vs X");
    gr->SetMarkerColor(4);
    gr->SetMarkerStyle(21);
    gr->SetMarkerSize(0.2);
    // Disegno i punti "A" sta per disegna gli assi, "P"^M
    gr->Draw("AP");^M
    // La uso per fare un fit^M
    gr->Fit("fun");^M
    // estraggo i risultati della regressione^M
    TVirtualFitter * fitter = TVirtualFitter::GetFitter();
    // Estraggo e stampo il valore dei parametri.^M
    double p1=fitter->GetParameter(1);
    double p0=fitter->GetParameter(0);
    printf("\n********************** p1=%f p0=%f\n",p1,p0);
    // estraggo e stampo la matrice di covarianza^M
    double * covMatrix = fitter->GetCovarianceMatrix();
    printf("\n c11=%f   c12=%f\n",covMatrix[0],covMatrix[1]);
    printf(" c21=%f   c22=%f\n ",covMatrix[2],covMatrix[3]);
    printf("\n p0=%f +/- %f\n",p0,sqrt(covMatrix[0]));
    printf(" p1=%f +/- %f\n\n",p1,sqrt(covMatrix[3]));

    hpar->Fill(p0,p1);
}
```

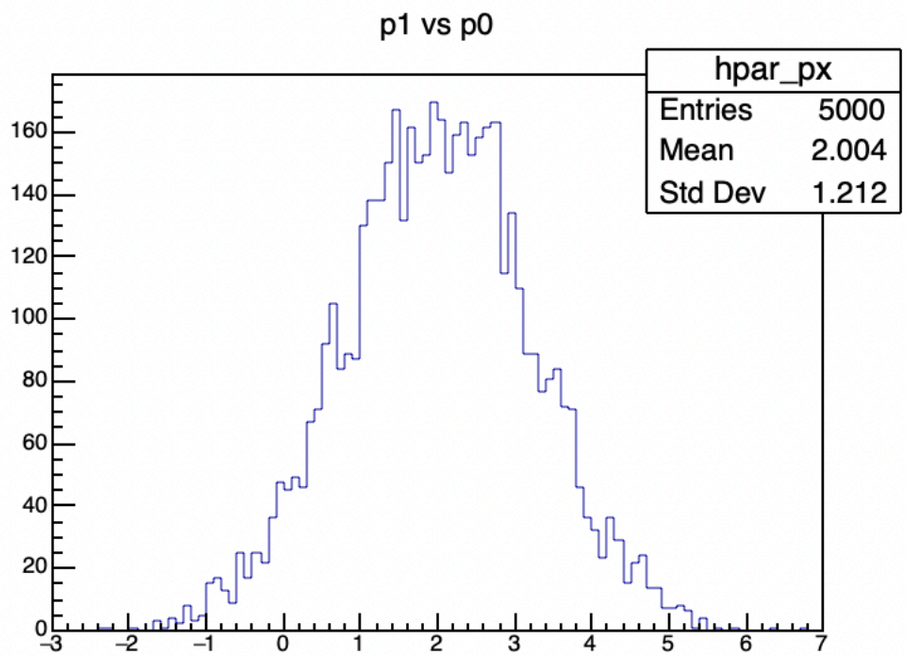
Regressione.C

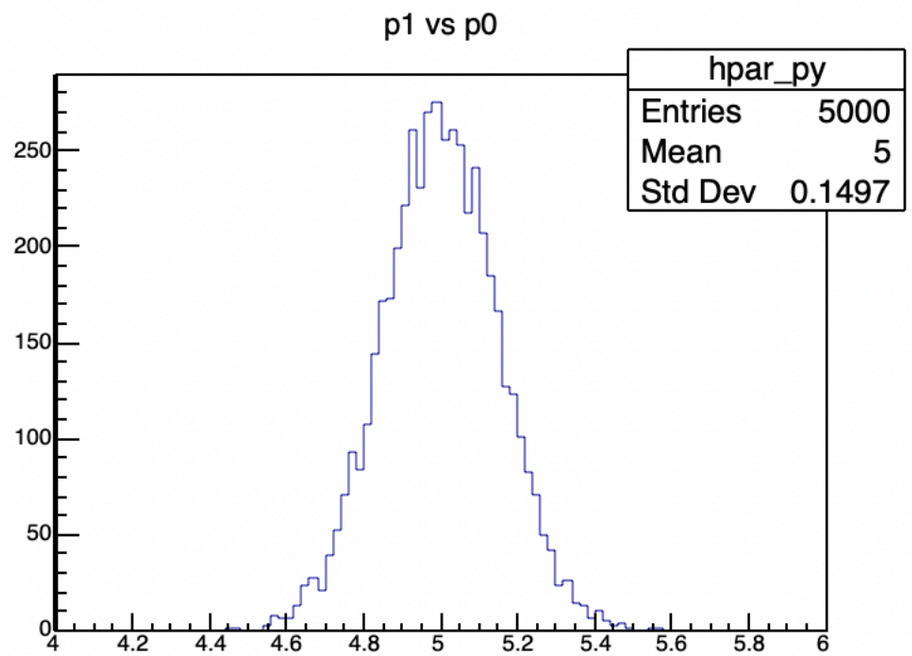
ESEMPIO 1

5000 set di misure e
5000 fit







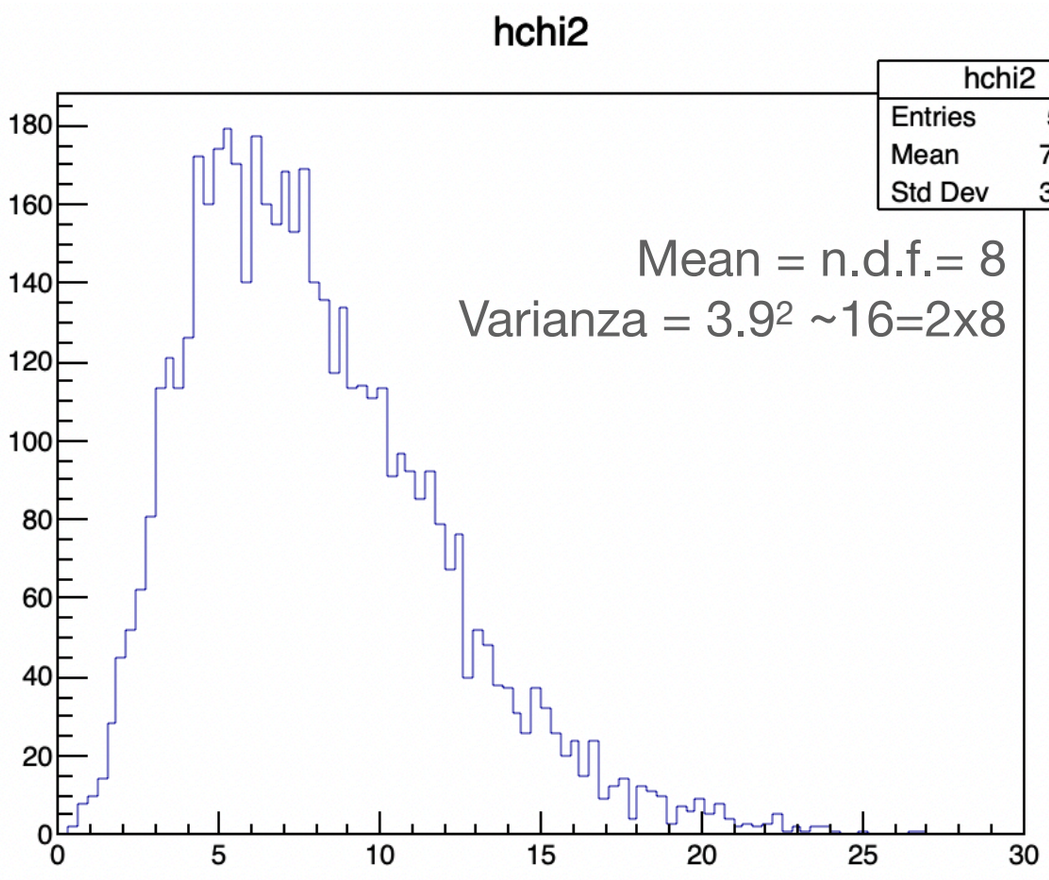


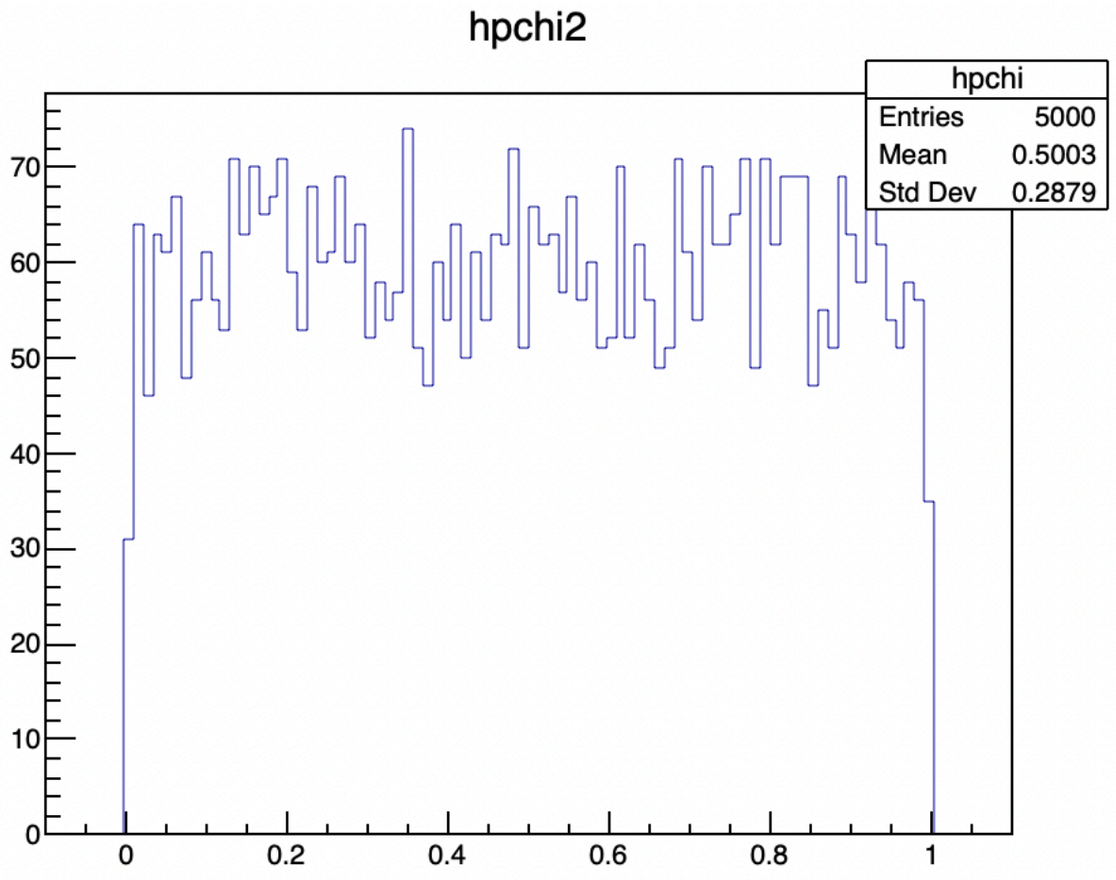
ESEMPIO 1

Regressione.C

5000 set di misure e
5000 fit

```
double chi2 = fun->GetChisquare();
double pchi2 = fun->GetProb();
```



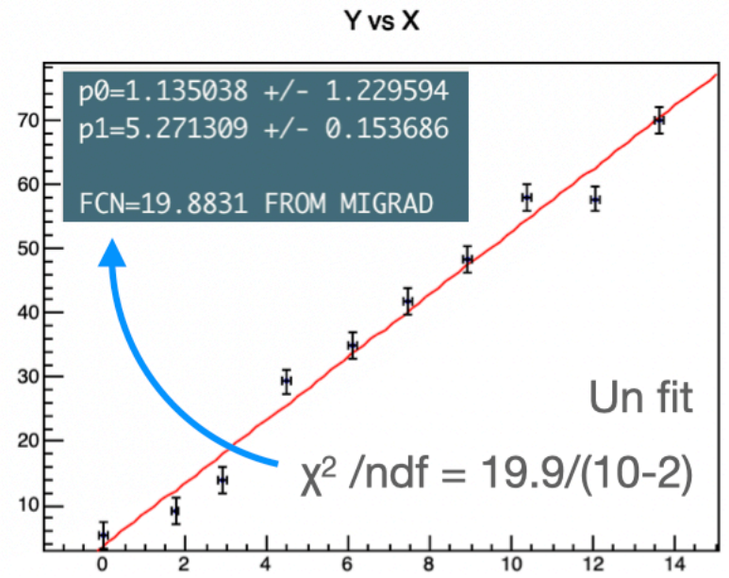


Se x è una variabile aleatoria distribuita secondo la pdf f(x), allora f(x) è distribuita uniformemente tra 0 e 1

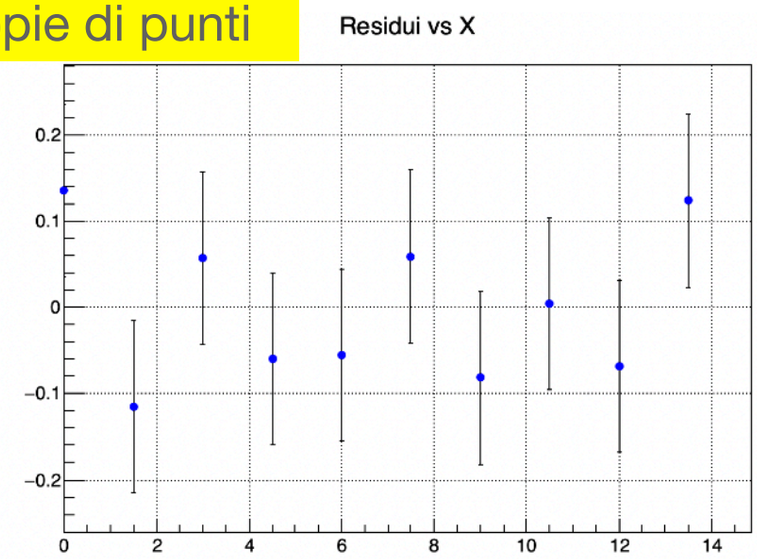
Dipartimento di Matematica e Fisica
"Ennio De Giorgi"
INFN

ESEMPIO 2

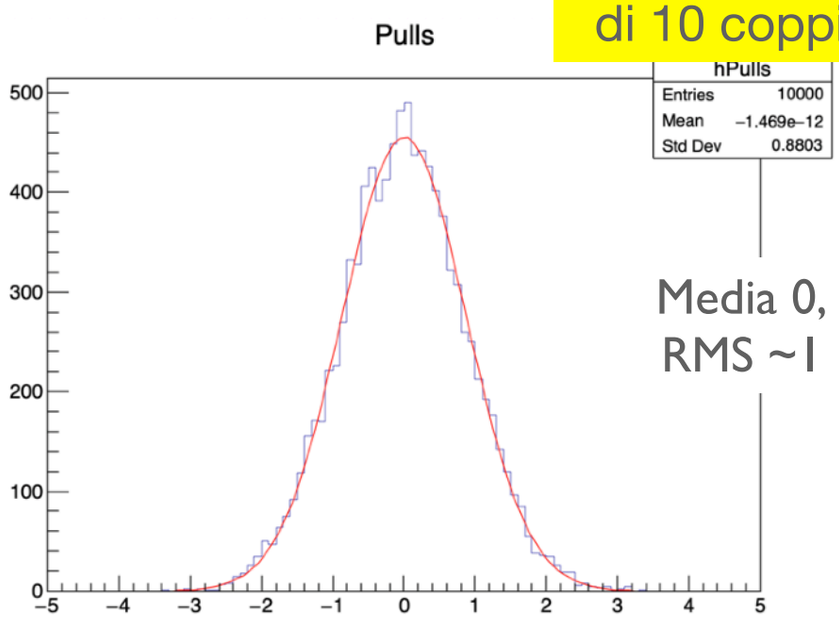
- Residui e Residui normalizzati
  - Per valutare la bonta' di ogni fit e' utile studiare (oltre al chi2) i residui
    - $R_i = y_i - f(x_i; \vec{\lambda}_0)$
    - Che devono essere distribuiti casualmente attorno a 0, essendo statisticamente compatibili con fluttuazioni gaussiane relative a un deviazione standard σ_i



Residui in un fit lineare a 10 coppie di punti



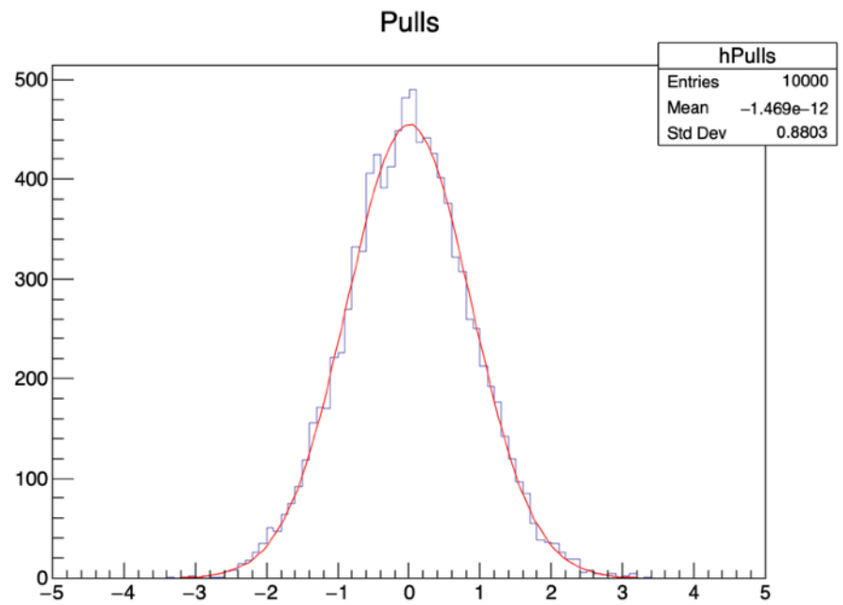
Residui normalizzato su fit lineari a 1000 set di 10 coppie di punti



ESEMPIO 2

Residui normalizzato su fit lineari a 1000 set di 10 coppie di punti

Media 0,
RMS ~1 PERCHE' NON PROPRIO 1

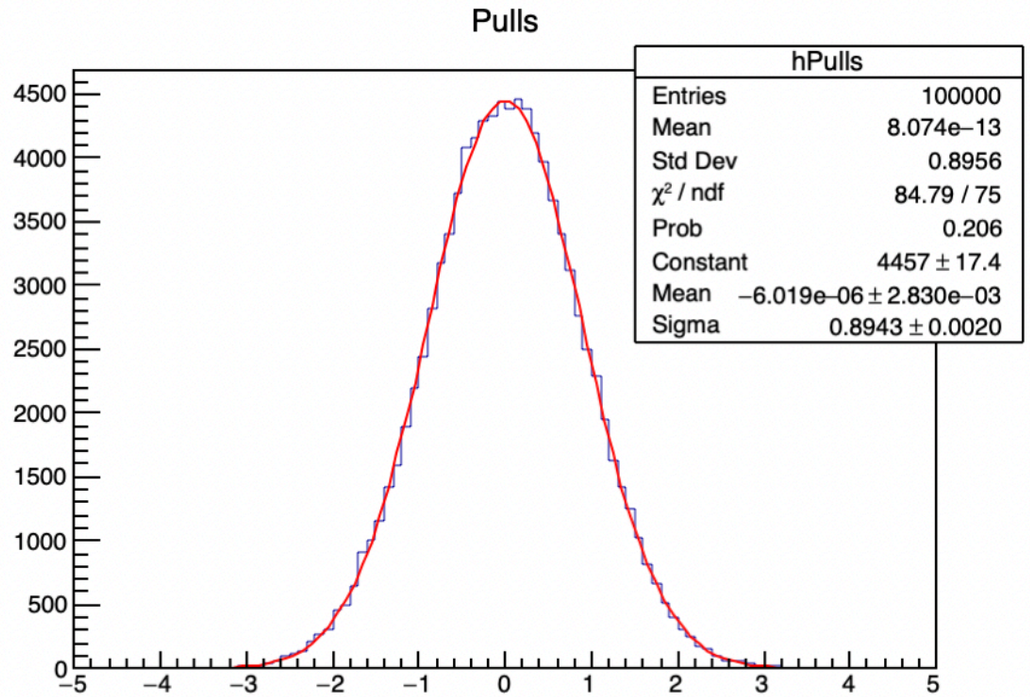


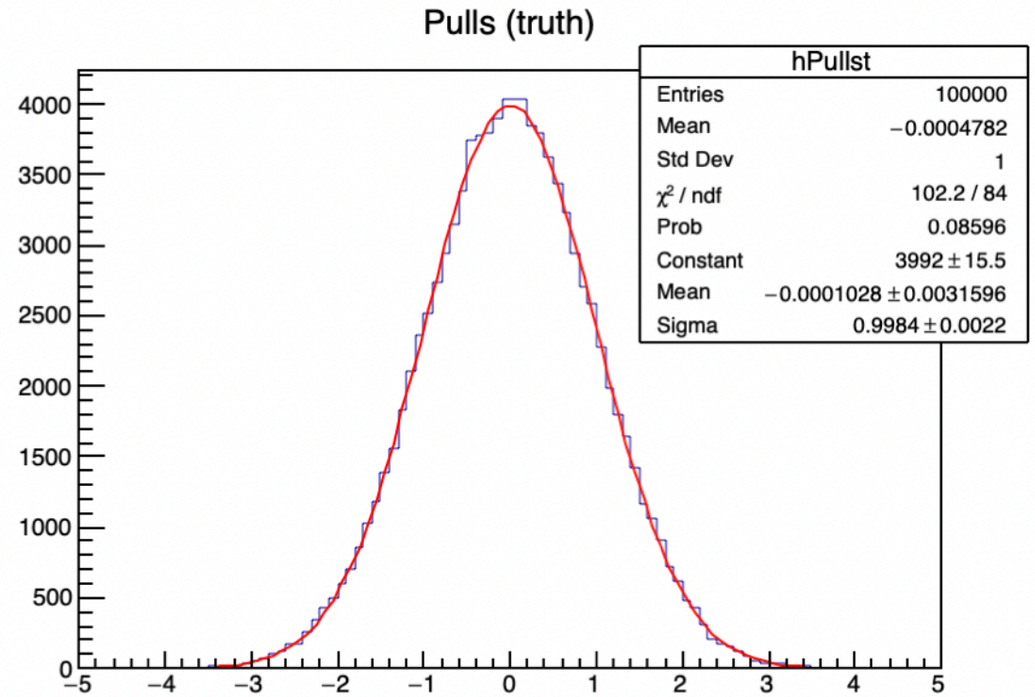
ESEMPIO 2

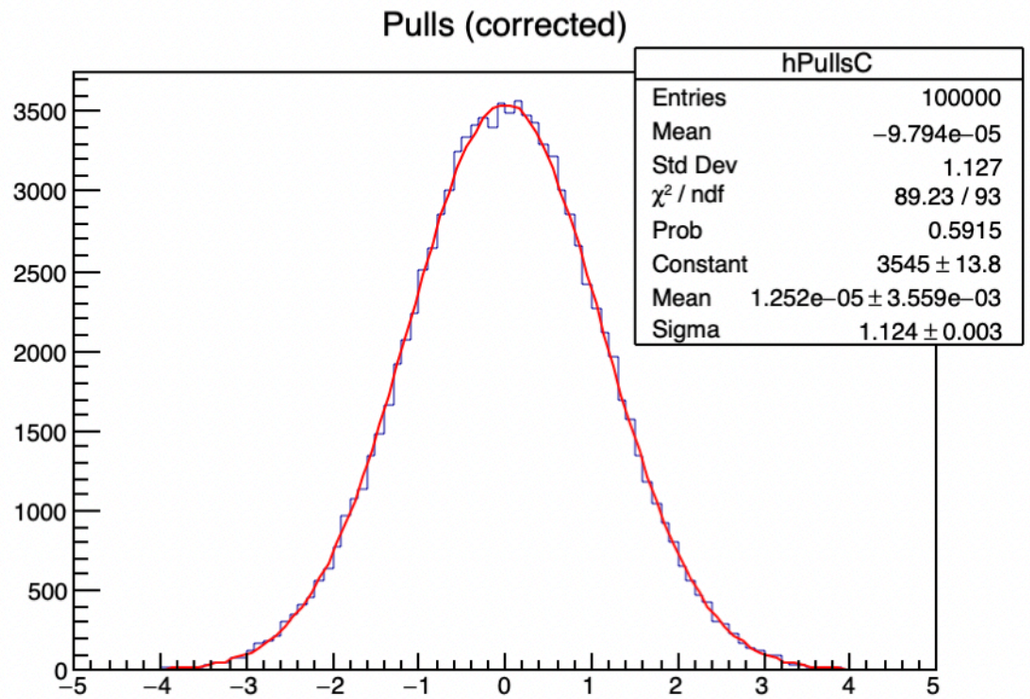
Residui I.C

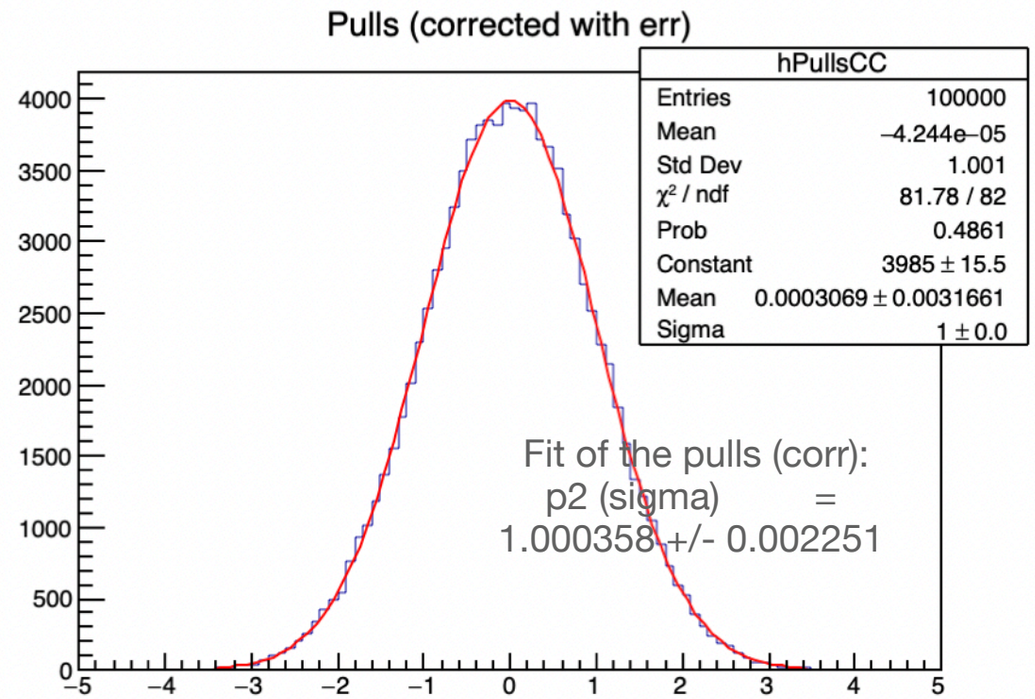
Ci sono Pull e Pull

10.000 esp con 10 punti ciascuno



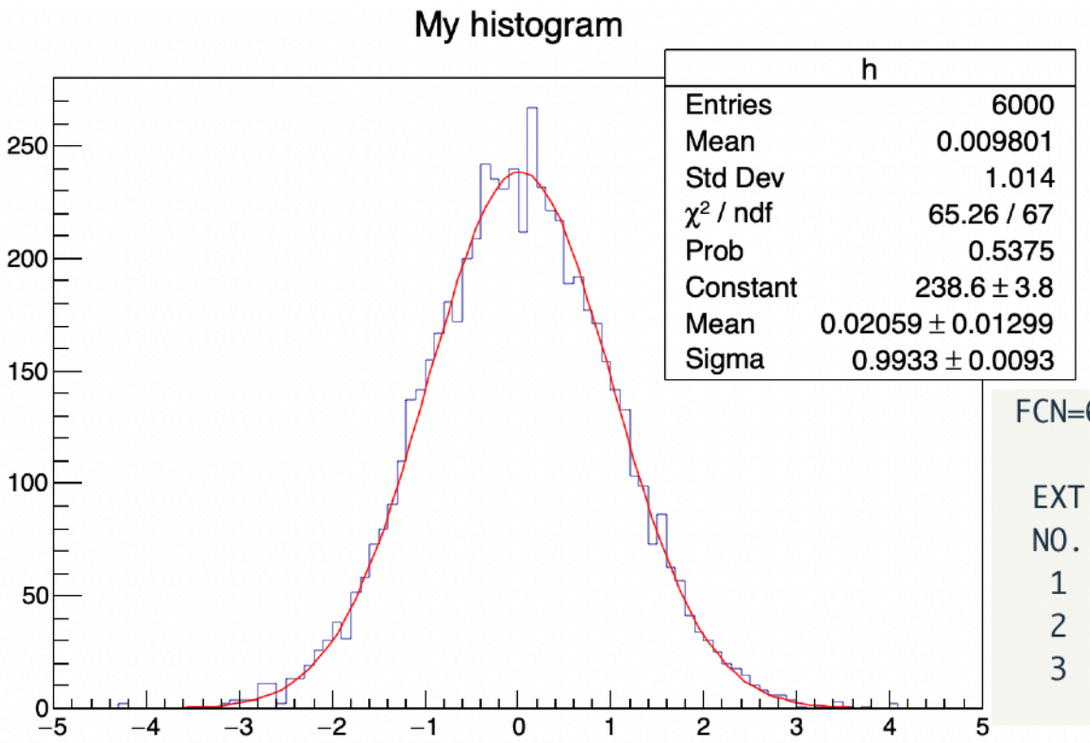






ESEMPIO 3

CurveLivello.C



Generato estraendo da una gaussiana normale

Fit con una gaussiana, tutti i e parametri da determinare contemporaneamente nel fit

```
FCN=65.2581 FROM MIGRAD    STATUS=CONVERGED      62 CALLS          63 TOTAL
                     EDM=9.11307e-10    STRATEGY= 1      ERROR MATRIX ACCURATE
 EXT PARAMETER                                   STEP         FIRST
 NO.   NAME      VALUE            ERROR          SIZE      DERIVATIVE
  1  Constant     2.38594e+02   3.80928e+00   1.23098e-02  -6.27631e-06
  2  Mean         2.05919e-02   1.29916e-02   5.15403e-05   2.31062e-03
  3  Sigma        9.93349e-01   9.33666e-03   1.00053e-05   2.46257e-03
```

```
c11=14.510645   c12=-0.001761   c13=-0.020707
c21=-0.001761   c22=0.000169   c23=0.000007
c31=-0.020707   c32=0.000007   c33=0.000087
```

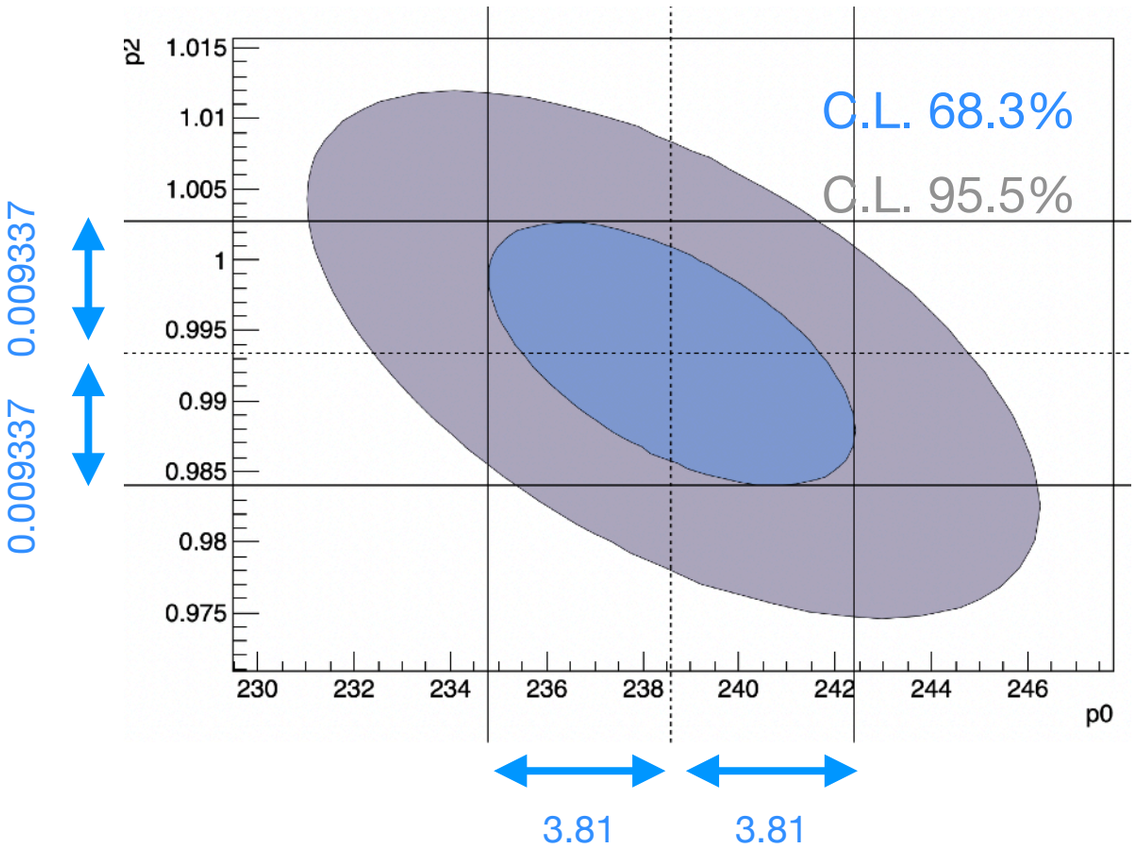
Matrice di covarianza

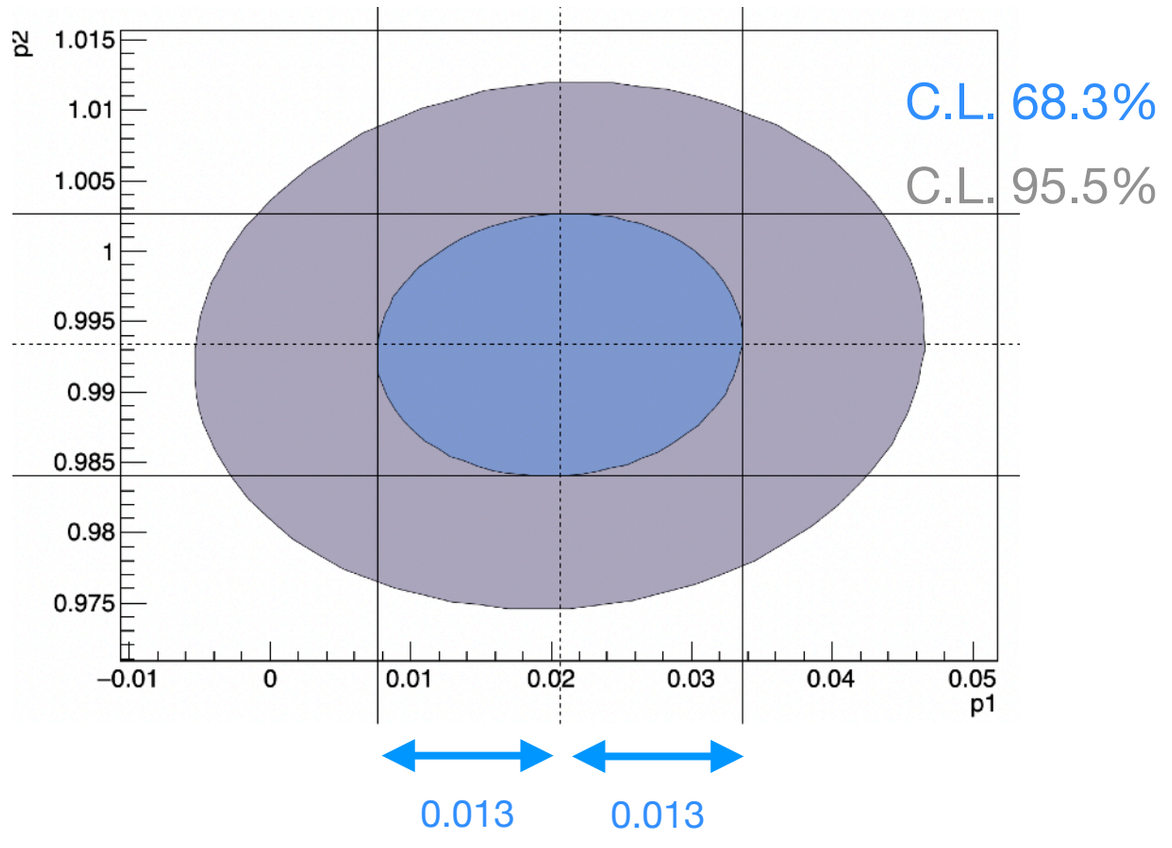
```
p2=0.993349 err 0.009337
p1=0.020592 err 0.012992
p0=238.593819 err 3.809284
```

Parametro con i loro errori

ESEMPIO 3

CurveLivello.C

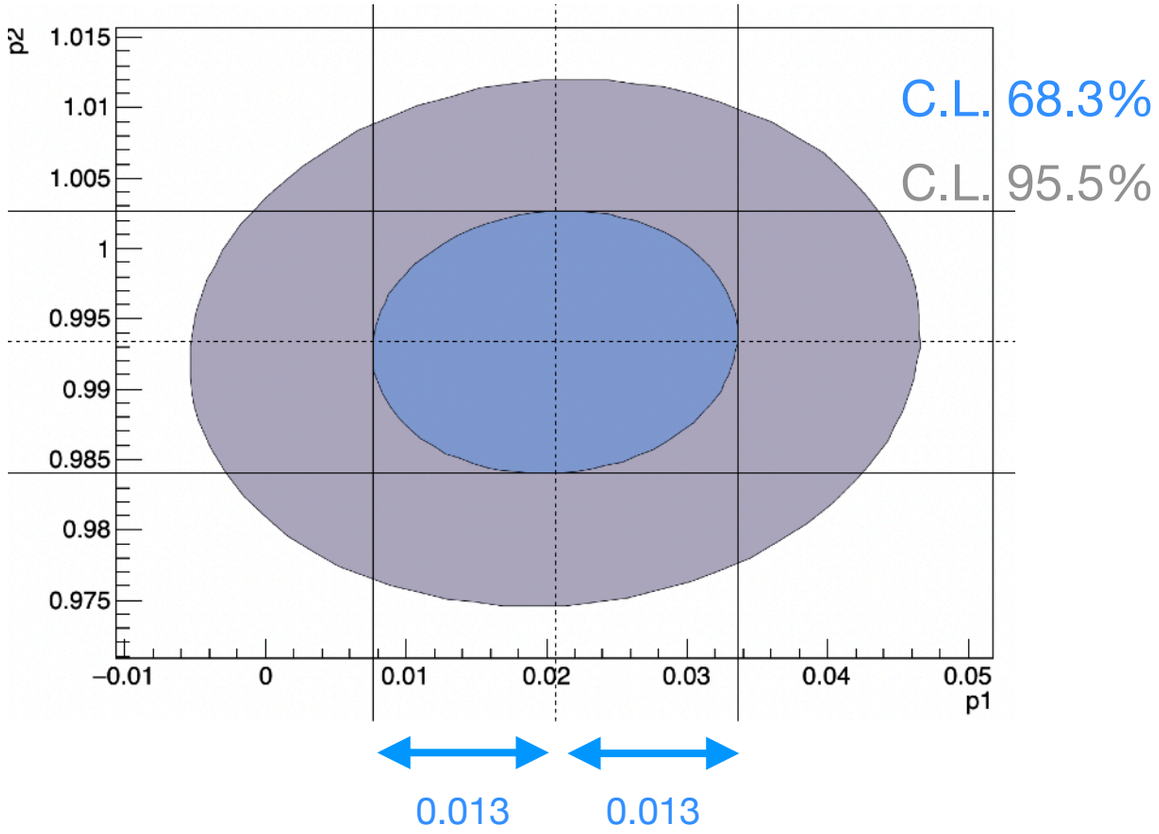





```
c11=14.510645   c12=-0.001761   c13=-0.020707
c21=-0.001761   c22=0.000169    c23=0.000007
c31=-0.020707   c32=0.000007    c33=0.000087

p2=0.993349 err 0.009337
p1=0.020592 err 0.012992
p0=238.593819 err 3.809284
```

CurveLivello.C

ESEMPIO 3




```
c11=14.510645   c12=-0.001761   c13=-0.020707
c21=-0.001761   c22=0.000169    c23=0.000007
c31=-0.020707   c32=0.000007    c33=0.000087

p2=0.993349 err 0.009337
p1=0.020592 err 0.012992
p0=238.593819 err 3.809284
```

```
// Chiedo a codice di graficarmi la regione di confidenza^M
// ad una sigma (68.3%) e a due sigma (95.5% ) entro la quale si possonomuovere
// i parametri 1 e 2 (media e varianza)

// Per ottenere curve di livello a n-sigma, l'argomento deve essere n^2.
gMinuit->SetErrorDef(4.); //la variazione del chi2 che definisce la regione dipendera' dal numero di parametri liberi
TGraph *gr11 = (TGraph*)gMinuit->Contour(80,1,2);
gr11->SetFillColor(40);
gr11->GetXaxis()->SetTitle("p1");
gr11->GetYaxis()->SetTitle("p2");
gr11->Draw("alf");
// Per ottenere curve di livello a n-sigma, l'argomento deve essere n^2.
gMinuit->SetErrorDef(1.); //la variazione del chi2 che definisce la regione dipendera' dal numero di parametri liberi
TGraph *gr12 = (TGraph*)gMinuit->Contour(80,1,2);
gr12->SetFillColor(38);
gr12->Draw("lf");
```