

Ulteriori Conoscenze di Informatica e Statistica

Carlo Meneghini

Dip. di fisica - via della Vasca Navale 84,
st. 83 (I piano) tel.: 06 55 17 72 17

meneghini@fis.uniroma3.it

Breve puntualizzazione sugli intervalli di confidenza

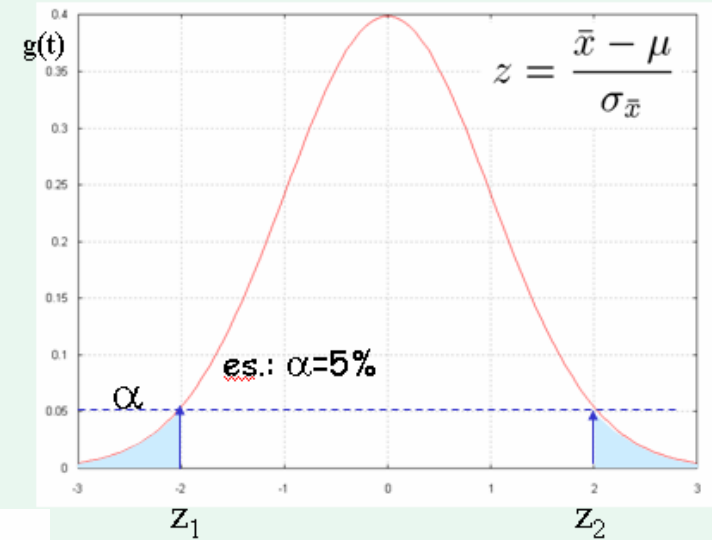
X : variabile aleatoria che segue una distribuzione normale $N(\mu, \sigma)$

$$z = \frac{x - \mu}{\sigma}$$

z : variabile aleatoria che segue una distribuzione normale standard $N(0,1)$

α : rischio di osservare un valore di z lontano da 0

$$\alpha = P(z < z_1) + P(z > z_2) = \int_{-\infty}^{z_1} g(z) dz + \int_{z_2}^{\infty} g(z) dz = 2 \int_{z_L}^{\infty} g(z) dz$$



$1-\alpha$: probabilità di osservare z vicino 0

$$P(z_1 < z < z_2) = 1 - \alpha$$

$$\left[z_1 < \frac{x - \mu}{\sigma} < z_2 \right] = [\mu - z_L \sigma < x < \mu + z_L \sigma]$$

Se osservo un valore X_o con varianza σ^2 questo rappresenta la migliore stima che ho, per ora, del valore medio. Quindi l'intervallo:

$$[X_o - z_L \sigma; X_o + z_L \sigma]$$

rappresenta l'*intervallo di confidenza*: la probabilità che una nuova misura fornisca un valore compreso nell'intervallo è $(1-\alpha)\%$

La funzione EXCEL CONFIDENZA(α, σ, N) calcola i limiti dell'intervallo di confidenza $z_L \sigma$ per una variabile aleatoria con distribuzione Normale.

- Singola misura X_0 σ : CONFIDENZA($\alpha, \sigma, 1$)
- Media di N misure X_0 σ : CONFIDENZA(α, σ, N)
- Media di N misure X_0 $\sigma_{\langle x \rangle}$: CONFIDENZA($\alpha, \sigma_{\langle x \rangle}, 1$)

$$\frac{1}{N} \sum_{i=1}^N x_i = \bar{x} = \langle x \rangle$$

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1} = \sigma^2$$

Se il valore osservato e la sua varianza sono stati determinati utilizzando un procedimento che combina diversi effetti bisogna tener conto dei gradi di libertà effettivi del procedimento e non si può utilizzare la funzione z per il calcolo dell'intervallo di confidenza

abile x è costruita utilizzando K variabili aleatorie (esempio) con errori σ_k la variabile aleatoria

$$t_\nu = \frac{x - \mu}{\sigma}$$

Segue una distribuzione t-Student con $\nu = K-1$. L'errore si calcola utilizzando la regole standard per la propagazione degli errori statistici

Se $X=a+b$ con errori σ_a σ_b , si ha $\sigma_x = (\sigma_a^2 + \sigma_b^2)^{1/2}$

La distribuzione t-Student approssima una distribuzione Gaussiana per ν grande

La funzione Excel **INV.T(a,n)** calcola gli estremi t_L dell'intervallo di confidenza.
L'intervallo di confidenza è:

$$[X_o - t_L \sigma; X_o + t_L \sigma]$$

t-Student	
ν	t_L
2	4.30
5	2.57
10	2.23
50	2.01
100	1.98
500	1.96
1000	1.96
z_L	
CONFIDENZA(0.05,1,1)	1.96

La variabile χ^2

$z_1, z_2, \dots, z_N$ sono variabili aleatorie $N(0,1)$

$$Z = \frac{X - \mu}{\sigma}$$

$$\chi_\nu^2 = \sum_i z_i^2$$

$$f(x, \nu) = \frac{x^{\nu/2-1}}{2^{\nu/2} \Gamma(\nu/2)} e^{-x/2}$$

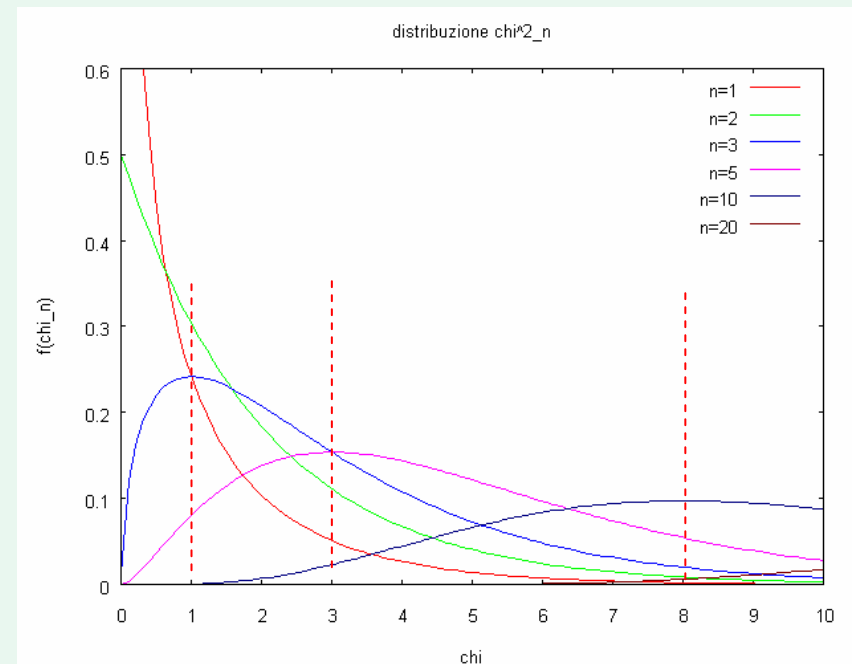
densità di
probabilità

$$\overline{\chi_\nu^2} = \nu$$

$$\sigma_{\chi_\nu^2}^2 = 2\nu$$

$$\chi_{\nu_1}^2 + \chi_{\nu_2}^2 = \chi_{\nu_1 + \nu_2}^2$$

massimo per $\chi = \nu - 2$

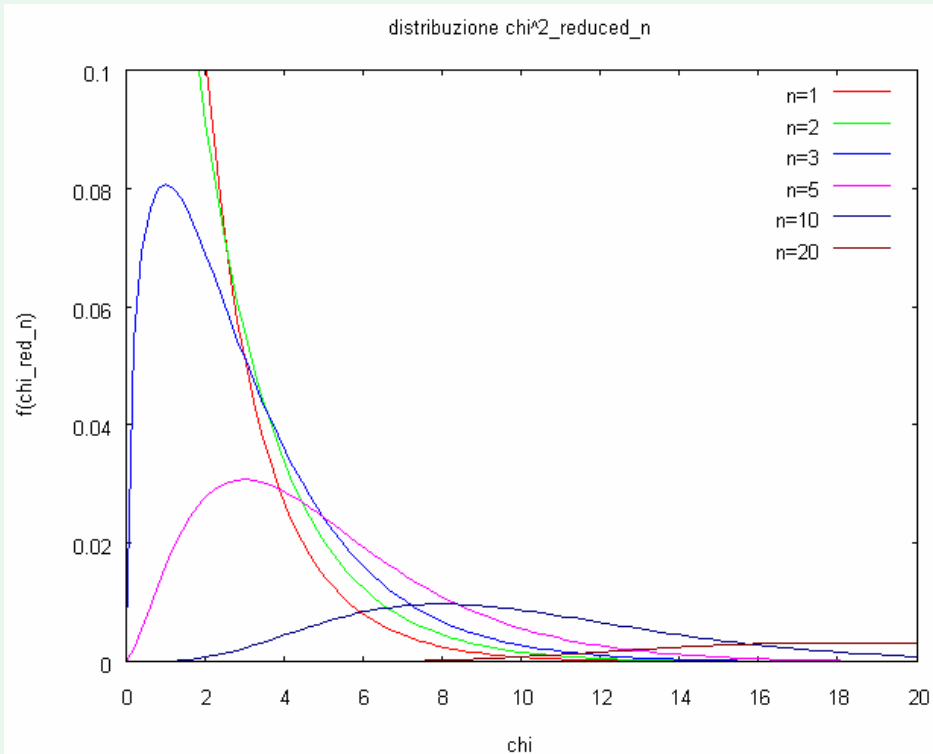


La variabile χ^2 ridotto

$$\tilde{\chi}_{\nu}^2 = \frac{\chi_{\nu}^2}{\nu}$$

$$\sigma_{\tilde{\chi}_{\nu}^2}^2 = \frac{2}{\nu}$$

$$\overline{\tilde{\chi}_{\nu}^2} = \nu$$



Il rapporto di due variabili χ^2 è
una funzione F

$$F(\nu_1, \nu_2) = \frac{\chi_{\nu_1}^2 / (\nu_1 - 1)}{\chi_{\nu_2}^2 / (\nu_2 - 1)} = \frac{\tilde{\chi}_{\nu_1}^2}{\tilde{\chi}_{\nu_2}^2}$$

X: variabile aleatoria che si distribuisce con legge normale attorno al suo valore medio μ con varianza σ^2 .

$$P(x_i) = f(x_i)dx = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{(x_i - \mu_v)^2}{2\sigma_v^2}} dx$$

La probabilità di osservare un dato set di valori x_i ($i=1, \dots, N$) effettuando N misure è:

$$P(x_1; x_2; \dots x_N) = \prod_{i=1}^N f(x_i)dx = \Lambda dx$$

L'insieme di valori più probabili è quello che massimizza Λ
criterio della massima verosimiglianza (maximum likelihood criterion).

$$\Lambda = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{(x_i - \mu_v)^2}{2\sigma_v^2}}$$

Es.: stima del valore medio della distribuzione.

trovare il valore μ che rende massima Λ

$$\log(\Lambda) = \sum_{i=1}^N \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right) = - \sum_{i=1}^N \log \left(\sqrt{2\pi\sigma^2} \right) - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$\left. \frac{\partial \Lambda}{\partial \mu} \right|_{\mu=\mu_s} = 0 \quad \text{equivale a :} \quad \left. \frac{\partial}{\partial \mu} \left(\sum_{i=1}^N \frac{(x_i - \mu)^2}{\sigma^2} \right) \right|_{\mu=\mu_s} = 0$$

$$\mu_s = \frac{1}{N} \sum_i x_i$$

Insieme di misure sperimentali: $X, Y^{\text{exp}} : (x_1; y_1), (x_2; y_2), \dots (x_N; y_N)$

Legge fisica: $Y^{\text{th}} = f^{\text{th}}(X; p_1; p_2; \dots p_k),$

$$Y_i^{\text{exp}} = f_i(p_1, p_2, \dots p_m) + \epsilon_i$$

$$z_i = \frac{Y_i^{\text{exp}} - f_i(p_1, p_2, \dots p_m)}{\sigma_i^2}$$

z_i segue una distribuzione normale standard: $N(0,1)$

$$\Lambda(p_1, p_2, \dots p_m) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(Y_i^{\text{exp}} - f_i)^2}{2\sigma_i^2}}$$

$$\text{Min} \left(\sum_{i=1}^N \frac{(Y_i^{\text{exp}} - f_i(p_1, p_2, \dots p_m))^2}{\sigma_i^2} \right)$$

Test di ipotesi

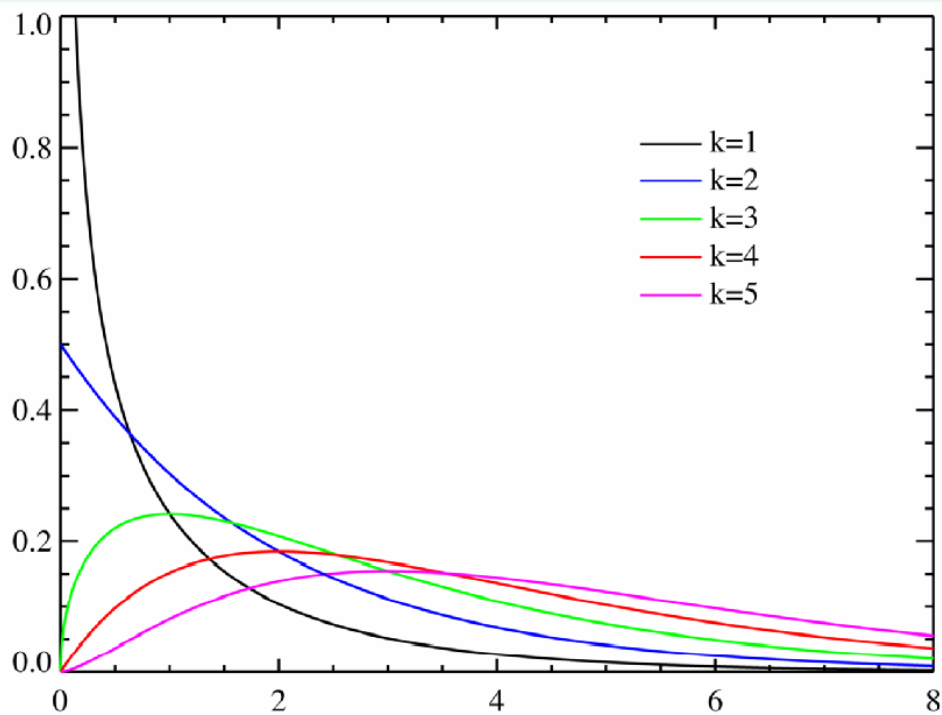
Possiamo utilizzare la funzione di massima verosimiglianza (e quindi il χ^2 associato) per effettuare un test di ipotesi: vogliamo sottoporre a test l'ipotesi che il modello scelto (relazione funzionale e/o set di parametri) sia compatibile con i dati sperimentali.

$$\chi^2_\nu = \sum_{i=1}^N \frac{(\epsilon_i)^2}{\sigma_i^2} \sim \nu$$

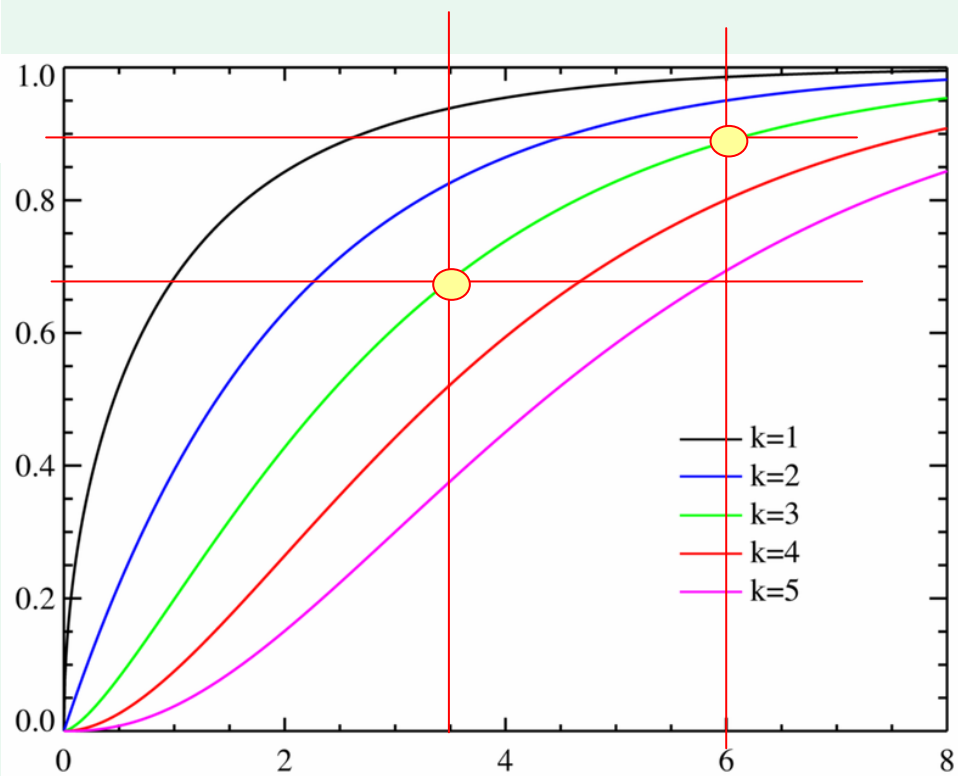
La funzione Excel INV.CHI(α ; ν) fornisce il valore limite di χ^2_ν associato alla probabilità α .

La funzione Excel IDISTRIB.CHI(χ ; ν) fornisce la probabilità associata al valore di χ^2_ν osservato (probabilità di osservare valori di χ^2 maggiori)

Attenzione: il test del χ^2 è un test di **reiezione delle ipotesi**, quindi lo scopo del test non è verificare se i dati siano compatibili con l'ipotesi ma quantificare il rischio di sbagliare affermando che i dati **non sono compatibili** con il modello. L'idea alla base del test è che, se la probabilità associata al valore di χ^2 calcolato in base ai dati sperimentali è piccola, questo equivale a dire che la probabilità di osservare questo valore per una distribuzione casuale dei dati attorno all'andamento teorico è piccola e quindi bisogna rifiutare il modello. Se, al contrario non posso rifiutare il modello non posso escludere che il modello sia errato ma non ho la sensibilità per verificarlo.



Distribuzione
integrata.



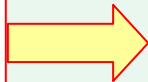
Fit (regressione) non lineare

Insieme di misure sperimentali: $X, Y^{\text{exp}} : (x_1; y_1), (x_2; y_2), \dots (x_N; y_N)$

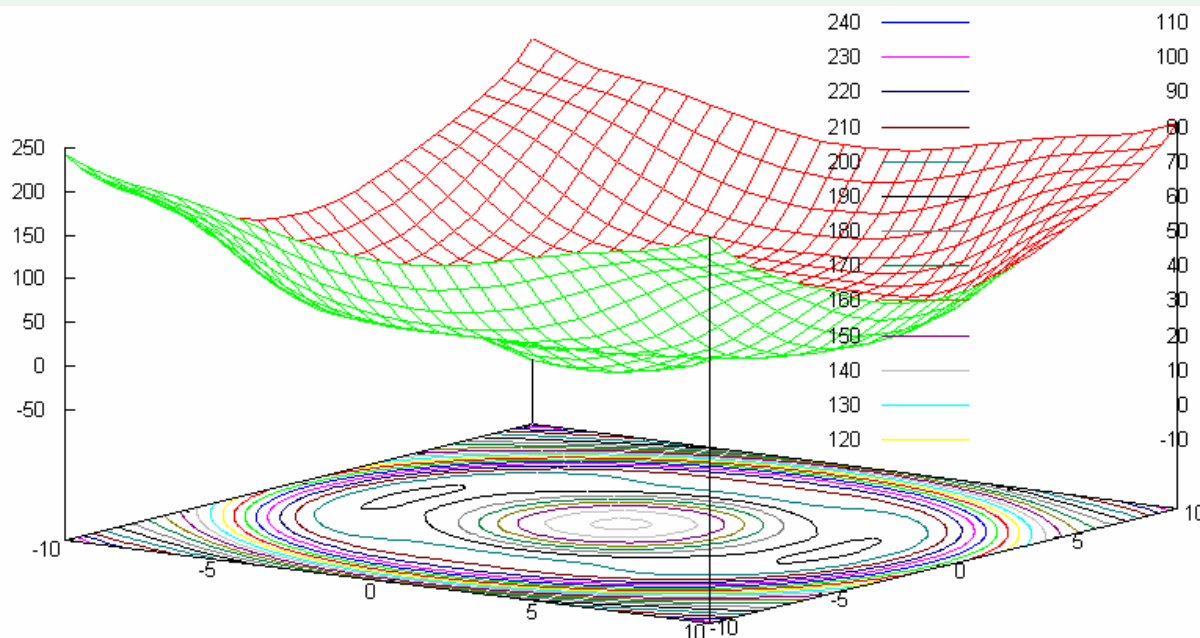
Legge fisica: $Y^{\text{th}} = f^{\text{th}}(X; p_1; p_2; \dots p_k),$

MAX

$$\Lambda = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} dx$$



$$\text{Min} \left(\sum_i \frac{(Y_i^{\text{exp}} - Y_i^{\text{th}})^2}{\sigma_i^2} \right)$$



$$\frac{\partial \chi^2}{\partial p_i} = 0$$

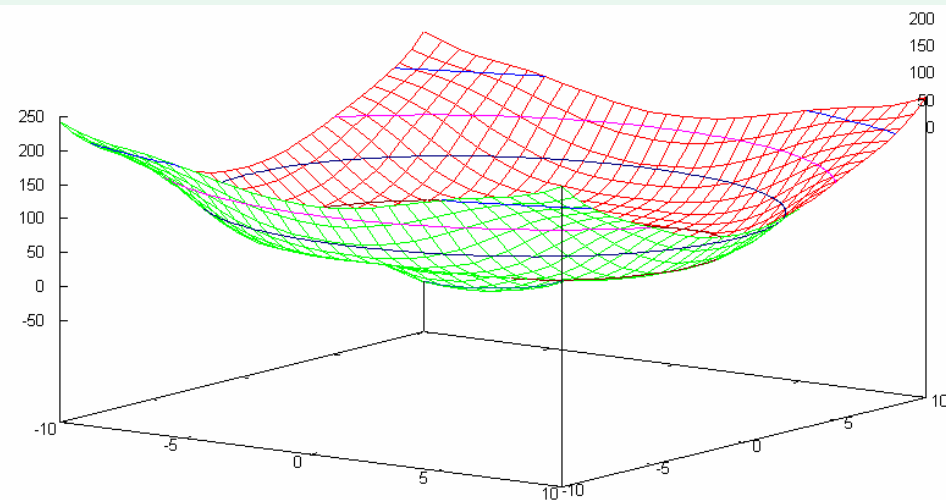
$$\frac{\partial^2 \chi^2}{\partial p_i^2} > 0$$

Intervalli di confidenza

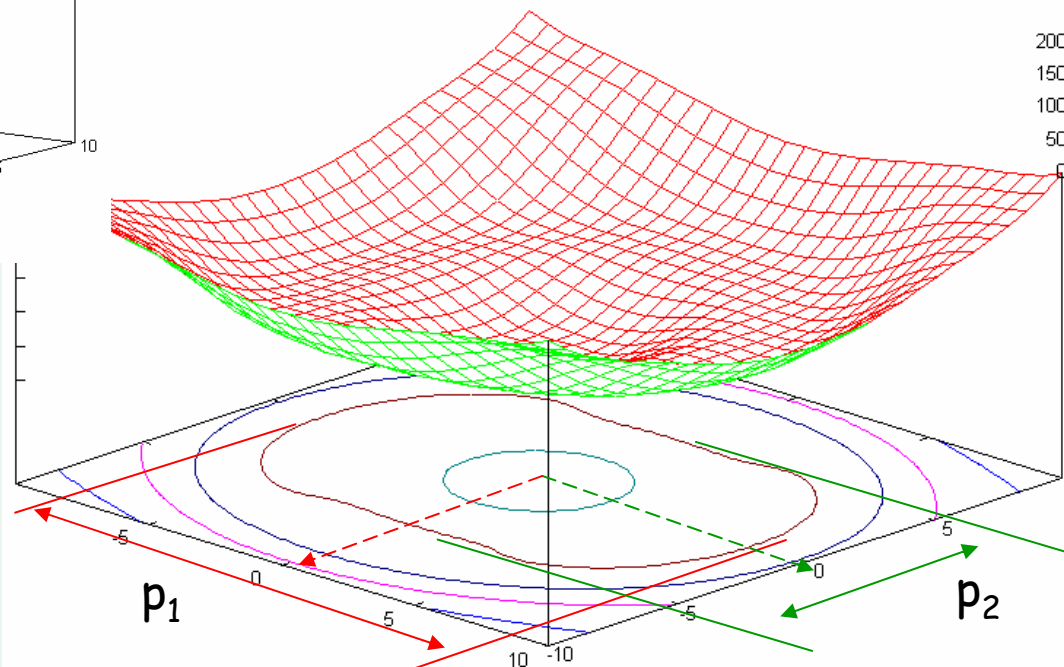
una volta trovati i valori dei parametri p_i che determinano un minimo della funzione χ^2 , quale è l'errore associato ad essi? Quale affidabilità dare ai parametri in termini di *confidenza*?

$$\chi^2(max) = \chi^2(min) + C$$

$$C = \sigma_{\chi^2}, 2\sigma_{\chi^2} \text{ etc... dove } \sigma_{\chi^2} = \sqrt{2\nu}$$



$$[p_i^{min} - \delta p_i^L; p_i^{min} + \delta p_i^R]$$



Quale modello è meglio?

$$\chi^2_{\nu_2} < \chi^2_{\nu_1}$$

$$\nu_2 > \nu_1$$

$$F(\nu_2 - \nu_1, \nu_2) = \frac{(\chi^2_{\nu_1} - \chi^2_{\nu_2})/(\nu_2 - \nu_1)}{\chi^2_{\nu_1}/\nu_2}$$

Si confronta la variabile F così ottenuta con la sua distribuzione statistica, si rifiuta il modello più complesso se la probabilità associata al valore di F ottenuto è piccola (minore di un limite che rappresenta il rischio di sbagliare rifiutando il modello più complesso).