25 MEDIE PESATE

Come combinare misure separate?

Esempio, 2 misure

Misura di A: $x = x_A \pm \sigma_A$

Misura di B: $x = x_B \pm \sigma_B$

Se si effettua la media aritmetica:

$$\frac{x_A + x_B}{2}$$

si da' eguale peso alle misure senza tener conto dell'incertezza, che in generale possono essere diverse.

Calcoliamo la misura più probabile.

Le misure siano distribuite normalmente, se X è il valore vero, le probabilità di ottenere x_A e x_B sono rispettivamente:

$$P(x_A) \propto \frac{1}{\sigma_A} e^{-\frac{(x_A - X)^2}{2\sigma_A{}^2}}$$

$$P(x_B) \propto \frac{1}{\sigma_B} e^{-\frac{(x_B - X)^2}{2\sigma_B{}^2}}$$

se le misure sono indipendenti si ha

$$P(x_A, x_B) = P(x_A)P(x_B) \propto \frac{1}{\sigma_A \sigma_B} e^{-\frac{\chi^2}{2}}$$

dove

$$\chi^2 = \frac{(x_A - X)^2}{\sigma_A^2} + \frac{(x_B - X)^2}{\sigma_B^2}$$

è la somma dei quadrati:

Per il *principio della massima verosimiglianza* la miglior stima di X si ha quando la probabilità è massima e quindi χ^2 è minimo:

$$\frac{\partial \chi^2}{\partial X} = -2\frac{x_A - X}{\sigma_A^2} - 2\frac{x_B - X}{\sigma_B^2} = 0$$

da cui, ricavando X, si ottiene:

$$x_{best} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_B}{\sigma_B^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}}$$

definendo i pesi:

$$w_A = \frac{1}{\sigma_A^{\,2}} \qquad\qquad w_B = \frac{1}{\sigma_B^{\,2}}$$

si ha:

$$x_{best} = \frac{w_A x_A + w_B x_B}{w_A + w_B}$$

Il valore stimato si avvicina alla misura che pesa maggiormente, cioè quella che ha una incertezza minore:

$$w_A > w_B \Leftrightarrow \sigma_A < \sigma_B$$

Se si hanno N misure:

$$x_1 \pm \sigma_1\,,\; x_2 \pm \sigma_2\,,\; x_3 \pm \sigma_3\,,\ldots\ldots\ldots,\; x_N \pm \sigma_N$$

$$x_{best} = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i} \qquad \text{con} \qquad w_i = \frac{1}{\sigma_i^2}$$

Note:

- $w_i = \frac{1}{\sigma_i^2}$ proporzionale all'inverso del quadrato
- se $\sigma_A = \sigma_B \Rightarrow w_A = w_B \Rightarrow x_{best} = \frac{x_A + x_B}{2}$
- propagando gli errori si ottiene: $\sigma_{best} = \frac{1}{\sqrt{\sum_i w_i}}$

26 METODO DEI MINIMI QUADRATI

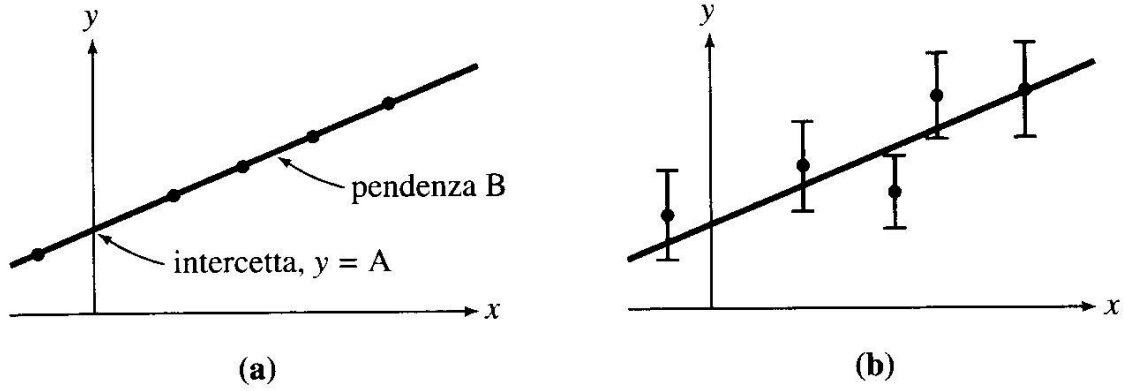
Regressione lineare.

Esempio legge fisica:

$$v=v_0+at$$

Si siano fatte misure delle quantità x e y, se disposte su un grafico i punti di coordinate x_i e y_i mostrano un andamento lineare, possiamo cercare la retta di miglior accordo:

$$y=A+Bx$$



Supponiamo che:

- l'incertezza sulle x sia trascurabile
- le incertezze sulle y siano tutte uguali[1]
- le misure y_i siano distribuite normalmente con $\sigma_i=\sigma_y$.

[1] Se gli errori non sono tutti uguali il metodo può essere comunque applicato sostituendo ai valori y_i le misure pesate sull'errore.

Siano A e B i parametri della retta allora per ogni x_i possiamo calcolare il valore di y_i corrispondente.

La probabilità di ottenere il valore osservato y_i è:

$$P_{A,B}(y_i) \propto \frac{1}{\sigma_y} e^{-\frac{(y_i - A - Bx_i)^2}{2\sigma_y^2}}$$

se abbiamo effettuato N coppie di misure indipendenti:

$$P_{A,B}(y_1,...., y_N) = P_{A,B}(y_1)....P_{A,B}(y_N) \propto \frac{1}{\sigma_y^N} e^{-\frac{\chi^2}{2}}$$

dove

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

La migliore stima per i parametri A e B si ottiene per i valori massimi di probabilità e quindi per i valor minimi di χ^2:

$$\frac{\partial \chi^2}{\partial A} = -\frac{2}{\sigma_y^2} \sum_i (y_i - A - Bx_i) = 0$$

$$\frac{\partial \chi^2}{\partial B} = -\frac{2}{\sigma_y{}^2} \sum_i x_i \left(y_i - A - B x_i \right) = 0$$

da cui si ottiene il sistema:

$$AN + B\sum_i x_i = \sum_i y_i$$
$$A\sum_i x_i + B\sum_i x_i{}^2 = \sum_i x_i y_i$$

le cui soluzioni sono:

$$A = \frac{\left(\sum_i x_i{}^2\right)\left(\sum_i y_i\right) - \left(\sum_i x_i\right)\left(\sum_i x_i y_i\right)}{\Delta}$$

$$\mathbf{B} = \frac{N\left(\sum_i x_i y_i\right) - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{\Delta}$$

dove

$$\Delta = N\left(\sum_i x_i{}^2\right) - \left(\sum_i x_i\right)^2$$

L'incertezza su y e':

$$\sigma_y^{\,2} = \frac{1}{N} \sum_i \left(y_i - A - B x_i \right)^2$$

o più correttamente:

$$\sigma_y^2 = \frac{1}{N-2}\sum_i (y_i - A - Bx_i)^2$$

Il fattore N-2 corregge la sottostima dovuta la fatto che, come mostrato prima, i valori di A e B rendono minima la sommatoria. Dobbiamo dividere per N-2 perché abbiamo già utilizzato i dati per stimare 2 parametri (A e B) e quindi il numero di gradi libertà si è ridotto di 2. In generale se si sono stimati ν parametri si hanno

N-ν gradi di libertà

differenza tra il numero delle misure e il numero di parametri stimati.

Se avessimo solo 2 punti misurati (N=2) la retta sarebbe univocamente determinata ma l'incertezza sarebbe massima, si deve quindi avere

$$\sigma_y = \frac{0}{0}$$

Nel caso di N grande la differenza è trascurabile.

Gli errori su A e B si ottengono propagando l'errore su y:

$$\sigma_A^2 = \sigma_y^2 \frac{\sum_i x_i^2}{\Delta}$$

$$\sigma_B^2 = \frac{N\sigma_y^2}{\Delta}$$

Regressione lineare pesata per errori variabili

Se gli errori sulle misure y non sono tutti uguali bisogna introdurre il peso di ogni misura:

$$w_{y_i} = \frac{1}{\sigma_{y_i}^2}$$

la variabile χ^2 che deve essere minima per avere la massima probabilità di ottenere le N misure di y, diventa:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_{y_i}^2} =$$

$$= \sum_{i=1}^{N} \left(\sqrt{w_{y_i}}\, y_i - \sqrt{w_{y_i}}\, A - \sqrt{w_{y_i}}\, Bx_i \right)^2$$

Derivando rispetto ad A e rispetto B e uguagliando le derivate a zero si ottiene il sistema di equazioni da cui si ricavano i coefficienti A e B:

$$A = \frac{\left(\sum_i w_{y_i} x_i^{\,2}\right)\left(\sum_i w_{y_i} y_i\right) - \left(\sum_i w_{y_i} x_i\right)\left(\sum_i w_{y_i} x_i y_i\right)}{\Delta}$$

$$\mathbf{B} = \frac{\left(\sum_i w_{y_i}\right)\left(\sum_i w_{y_i} x_i y_i\right) - \left(\sum_i w_{y_i} x_i\right)\left(\sum_i w_{y_i} y_i\right)}{\Delta}$$

dove

$$\Delta = \left(\sum_i w_{y_i}\right)\left(\sum_i w_{y_i} x_i^{\,2}\right) - \left(\sum_i w_{y_i} x_i\right)^2$$

Le incertezze in A e B sono in questo caso:

$$\sigma_A^{\,2} = \frac{\sum_i w_i x_i^{\,2}}{\Delta}$$

$$\sigma_B^{\,2} = \frac{\sum_i w_i}{\Delta}$$

Il metodo dei minimi quadrati può essere esteso ad un generico polinomio di grado *n*

$$y = A + Bx + Cx^2 + ... + Hx^n$$

nel caso *n=2* si ha la **regressione parabolica**:

$$y = A + Bx + Cx^2$$

Analogamente al caso precedente si ha la miglior stima per i parametri A, B e C quando è massima la probabilità di ottenere i valori osservati y_i . Questa si ottiene per i valor minimi di χ^2:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - A - Bx_i - Cx_i^2\right)^2}{\sigma_y^2}$$

derivando rispetto ad A, B e C e ponendo uguali a zero le derivate si ottengono l'equazioni normali:

$$AN + B\sum_i x_i + C\sum_i x_i^2 = \sum_i y_i$$

$$A\sum_i x_i + B\sum_i x_i^{\,2} + C\sum_i x_i^{\,3} = \sum_i x_i y_i$$
$$A\sum_i x_i^{\,2} + B\sum_i x_i^{\,3} + C\sum_i x_i^{\,4} = \sum_i x_i^{\,2} y_i$$

Risolvendo questo sistema si ottengono le stime per i parametri A, B e C.

In linea di principio a qualunque funzione y=f(x) può essere applicato il metodo, in pratica non sempre il sistema di equazioni normali è risolvibile in maniera semplice o è impossibile.

Regressione esponenziale

Consideriamo si abbia una relazione esponenziale tra le variabili x e y

$$y = Ae^{Bx}$$

dobbiamo determinare le costanti A e B.
La relazione esponenziale può essere *linearizzata*:

$$z = \ln y = \ln A + Bx$$

Si può verificare la relazione lineare tra i valori di x e quelli di z graficando in scala semi-logaritmica i dati x,y.

Ponendo $A'=\ln A$ si calcolano i coefficienti A' e B tramite una regressione lineare dei dati (x_i,z_i), essendo z_i il logaritmo di y_i. Infine si calcola $A=e^{A'}$.

Si noti che l'ipotesi fatta per la regressione lineare che gli errori su y_i fossero tutti uguali non è più valida, infatti:

$$\sigma_z = \frac{dz}{dy}\sigma_y = \frac{\sigma_y}{y}$$

a rigore bisognerebbe applicare il metodo dei minimi quadrati pesati, spesso però si applica il metodo normale essendo la variazione di σ_z piccola.

27 COVARIANZA E CORRELAZIONE

Supponiamo che per misurare la quantità q(x,y) si misurino N coppie di valori x_i, y_i da cui si possono ricavare N valori di:

$$q_i = q(x_i, y_i)$$

Assumiamo di avere piccole incertezze e che tutti i numeri x_i, y_i siano vicini rispettivamente a $\bar{x}$ e $\bar{y}$ allora si può approssimare:

$$q_i = q(x_i, y_i) \approx q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y})$$

dove le derivate sono calcolate nel punto $x = \bar{x}$, $y = \bar{y}$ e sono quindi tutte uguali.

Si ha:

$$\bar{q} = \frac{1}{N}\sum_{i=1}^{N} q_i =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left[q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y}) \right]$$

Essendo il secondo e il terzo termine nulli si ha:

$$\overline{q} = q\left(\overline{x}, \overline{y}\right)$$

L deviazione standard degli N valori q_i è

$$\sigma_q^2 = \frac{1}{N}\sum_{i=1}^{N}\left(q_i - \overline{q}\right)^2$$

e sostituendo le relazioni precedenti

$$\sigma_q^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial q}{\partial x}\left(x_i - \overline{x}\right) + \frac{\partial q}{\partial y}\left(y_i - \overline{y}\right)\right]^2 =$$

$$= \left(\frac{\partial q}{\partial x}\right)^2 \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2 + \left(\frac{\partial q}{\partial y}\right)^2 \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2 +$$

$$+ 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

Definendo ***covarianza*** di x e y la quantità :

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

(si noti che ha dimensioni di una varianza (σ^2))

si ottiene:

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\sigma_{xy}$$

Questa espressione fornisce la deviazione standard di q sia che le misure siano o non siano indipendenti e distribuite normalmente.

Se le grandezze sono indipendenti, dopo molte misure, la covarianza tende a zero.

La covarianza dipende dal segno dei termini $\left(x_i - \bar{x}\right)$ e $\left(y_i - \bar{y}\right)$. Essi possono essere entrambi sempre positivi (sovrastima), entrambi sempre negativi (sottostima) o sempre discordi, solo se non sono indipendenti. Se invece sono indipendenti il prodotto di questi termini potrà essere positivo o negativo e quindi la somma per un gran numero di valori sarà nulla.

Quindi per misure x e y indipendenti si ha:

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2$$

Se le misure x e y non sono indipendenti si ha

$$\sigma_{xy} \neq 0$$

le misure x e y sono correlate.

Si può dimostrare che vale la disuguaglianza di Schwarz:

$$\left|\sigma_{xy}\right| \leq \sigma_x \sigma_y$$

da cui si ha:

$$\sigma_q^2 \leq \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 + 2\left|\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right|\sigma_x \sigma_y =$$

$$= \left[\left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y\right]^2$$

cioè

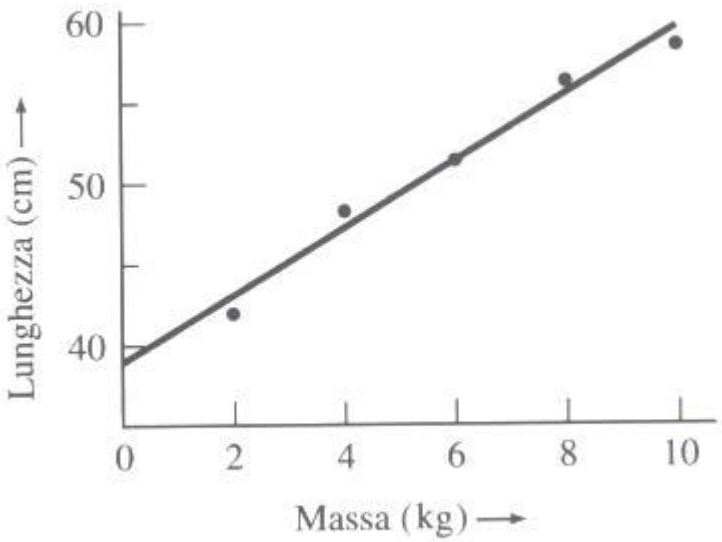
$$\sigma_q \leq \left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y$$

quindi

$\left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y$ costituisce un limite superiore per l'incertezza di q sia che gli errori su x e y siano o non siano indipendenti e normalmente distribuiti.

Coefficiente di correlazione

Consideriamo due variabili misurate x e y



Vogliamo verificare se esiste una relazione lineare tra esse:

$$y=A+Bx$$

Possiamo per esempio applicare il metodo dei minimi quadrati e poi conoscendo l'errore verificare se i punti giacciono sufficientemente vicini alla retta di regressione.

Se non conosciamo l'errore dobbiamo agire diversamente.

Notiamo che una relazione lineare implica che le variabili siano correlate (anche se in generale possono essere legate da una relazione non necessariamente lineare).

Definiamo il **coefficiente di correlazione**:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

dalla disuguaglianza di Schwarz si ha:

$$-1 \leq r \leq 1$$

infatti se vale la relazione lineare:

$$y_i = A + Bx_i$$

per ogni (x_i, y_i) allora si ha anche:

$$\bar{y} = A + B\bar{x}$$

e quindi, sottraendo membro a membro:

$$y_i - \bar{y} = B(x_i - \bar{x})$$

per qualunque *i*, e sostituendo in *r*:

$$r = \frac{B\sum_i (x_i - \bar{x})^2}{\sqrt{\left(B\sum_i (x_i - \bar{x})^2\right)^2}} = \frac{B}{|B|} = \pm 1$$

Se B>0 allora r=1, se B<0 allora r=-1, nel caso di correlazione lineare.

Nella pratica,

- se le variabili sono correlate linearmente $r \to 1$ per N grande
- se le variabili non sono correlate $r \to 0$ per N grande
- se $r \to -1$ la correlazione è negativa.

SIGNIFICATO QUANTITATIVO DI r

Sia

$$P_N(|r| \geq r_o)$$

la probabilità che N misure di 2 variabili non correlate diano un coefficiente di correlazione r più grande di un particolare r_o. Questa probabilità può essere ricavata, per esempio da tabelle:

Tabella La probabilità $P_N(|r| \geq r_0)$ che N misure di due variabili non correlate x ed y producano un coefficiente di correlazione con $|r| \geq r_0$. I valori dati sono probabilità percentuali e gli spazi vuoti indicano valori inferiori allo 0,05%

	r_0										
N	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
3	100	94	87	81	74	67	59	51	41	29	0
6	100	85	70	56	43	31	21	12	6	1	0
10	100	78	58	40	25	14	7	2	0,5		0
20	100	67	40	20	8	2	0,5	0,1			0
50	100	49	16	3	0,4						0

Nota: per $r_0=1$ si ha 0 perché r è una variabile continua e quindi la probabilità di ottenere esattamente un valore è nulla

- se P_N è grande le variabili non sono correlate
- se P_N è piccolo le variabili sono correlate

Al crescere di N la probabilità che le misure non correlate diano $|r| \geq r_o$ diminuisce, infatti per $N \to \infty$ si ha $r \to 0$ (quindi minore di qualunque r_o).

$P_N(|r| \geq r_o) \leq 5\%$ correlazione significativa

$P_N(|r| \geq r_o) \leq 1\%$ correlazione altamente significativa

Esempio:

N=3 $r_o = 0.7$ si ha $P_N(|r| \geq r_o) = 51\%$ non possiamo dire che le variabili siano correlate

N=3 $r_o = 0.9$ si ha $P_N(|r| \geq r_o) = 29\%$ non possiamo dire che le variabili siano correlate

N=6 $r_o = 0.7$ si ha $P_N(|r| \geq r_o) = 12\%$ non è sufficiente per dire che le variabili siano correlate.

N=20 $r_o = 0.7$ si ha $P_N(|r| \geq r_o) = 0.1\%$ la correlazione è altamente significativa.

ALCUNE PRECISAZIONI SUI CONCETTI DI COVARIANZA E CORRELAZIONE

Per variabili aleatorie indipendenti la covarianza e quindi anche il coefficiente di correlazione, sono nulli. Tuttavia non è vero il contrario: la covarianza e la correlazione possono essere nulle ma le variabili dipendenti:

- Indipendenza delle variabili implica correlazione nulla
- Correlazione nulla non implica necessariamente indipendenza delle variabili

Si può formulare il seguente Teorema:
Condizione necessaria ma non sufficiente affinché due variabili siano indipendenti è che esse non siano correlate

Il coefficiente di correlazione non può essere preso quindi come indicatore assoluto di dipendenza tra due variabili, esso può indicare se tra le variabili c'è dipendenza lineare:

- se $r = \pm 1$ esiste dipendenza lineare (positiva o negativa)

Consideriamo il seguente esempio.

Siano U e V due variabili con la stessa distribuzione di probabilità, siano X=U+V e Y=U-V, per esempio U sia il lancio di un dado e V il lancio di un secondo dado, allora:

$$\frac{1}{N}\sum_i x_i y_i = \frac{1}{N}\sum_i u_i^2 - \frac{1}{N}\sum_i v_i^2 = 0$$

e

$$\bar{y} = \frac{1}{N}\sum_i y_i = \frac{1}{N}\sum_i u_i - \frac{1}{N}\sum_i v_i = \bar{u} - \bar{v} = 0$$

quindi:

$$\sigma_{xy} = \frac{1}{N}\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \frac{1}{N}\sum_i x_i y_i - \overline{\bar{x}\bar{y}} = 0$$

La covarianza (e quindi anche il coefficiente di correlazione) è nulla ma le variabili (X e Y) non sono indipendenti.

Infine si noti che se la correlazione tra due variabili non è lineare il coefficiente di correlazione **r** perde di significato, quindi non può essere utilizzato per determinare il grado di legame tra le variabili stesse. In questi casi si deve far uso del *rapporto di correlazione empirica di Pearson.*

28 DISTRIBUZIONE BINOMIALE

Supponiamo di lanciare 3 dadi e volere calcolare la probabilità di ottenere $\nu = 0, 1, 2, 3$ assi.
La probabilità di ottenere una asso con un dado è 1/6, quindi la probabilità di ottenere 3 assi con 3 dadi (n=3, equivalente ad effettuare 3 lanci di un unico dado) è:

$$p(\upsilon=3,n=3) = \left(\frac{1}{6}\right)^3 \cong 0.5\%$$

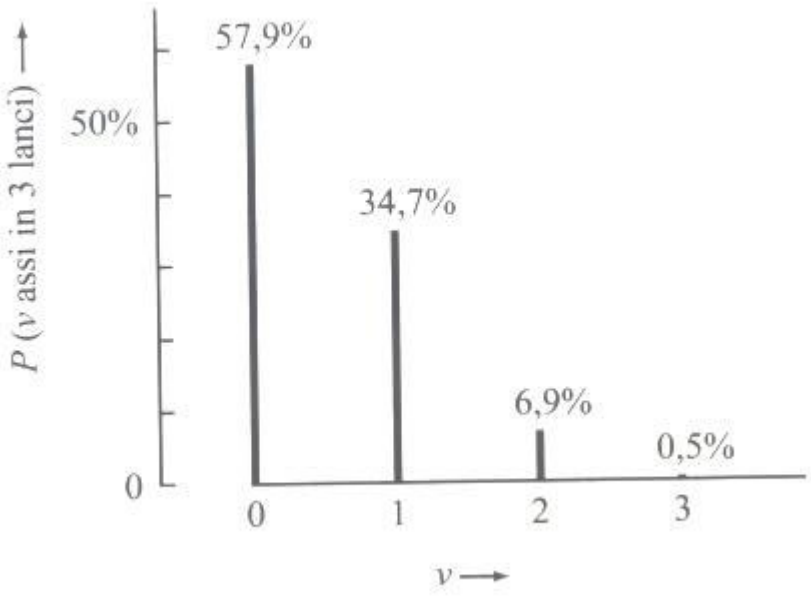
La probabilità di ottenere 2 assi ($\nu = 2$) in 3 lanci (evento A) è per esempio:

$$p(A,A,nonA) = \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right) \cong 2.3\%$$

essendo 5/6 la probabilità di non ottenere un asso.
Lo stesso risultato si può però ottenere anche con (A,non A, A) e (non A, A, A). Quindi si ha:

$$p(\upsilon=2,n=3) = 3\left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right) \cong 6.9\%$$

analogamente si può calcolare la probabilità di ottenere un solo asso in 3 lanci (34.7%) e nessun asso in 3 lanci (57.9%).



Si abbiano n prove indipendenti con:

- p probabilità di successo
- q=1-p probabilità di insuccesso

La probabilità di ottenere ν successi in n prove:

$$B_{n,p}(\upsilon) = \frac{n(n-1).....(n-\upsilon+1)}{1\times2\times3\times.....\times\upsilon} p^{\upsilon} q^{n-\upsilon}$$

dove

$$\frac{n(n-1).....(n-\upsilon+1)}{1\times2\times3\times.....\times\upsilon} = \frac{n!}{\upsilon!(n-\upsilon)!} = \binom{n}{\upsilon}$$

è il **coefficiente binomiale.**

Lo sviluppo binomiale è

$$(p+q)^n = \sum_{\upsilon=0}^{n} \binom{n}{\nu} p^{\upsilon} q^{n-\upsilon}$$

la distribuzione binomiale è:

$$B_{n,p}(\upsilon) = \binom{n}{\nu} p^{\upsilon} q^{n-\upsilon}$$

Esempio: calcoliamo la probabilità di trovare 2 assi in 3 lanci:

$$\upsilon=2, \qquad n=3,$$

$$p = \frac{1}{6}, \quad q = \frac{5}{6}$$

$$P = 3\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$$

- p^{υ} probabilità di ottenere tutti i successi in υ prove
- $q^{n-\upsilon}$ probabilità di ottenere insuccessi in tutte le rimanenti prove
- $\binom{n}{\nu}$ il numero di diversi ordini in cui si possono ottenere υ successi in n prove.

Proprietà della distribuzione binomiale

Dalla sviluppo binomiale essendo p+q=1 si ha la condizione di normalizzazione.

Valor medio:

$$\overline{\upsilon} = \sum_{\upsilon=1}^{n} \upsilon \binom{n}{\nu} p^{\upsilon} q^{n-\upsilon}$$

ed essendo

$$\upsilon \binom{n}{\nu} = \upsilon \frac{n!}{(n-\upsilon)!\upsilon!} = n \frac{(n-1)!}{(n-\upsilon)!(\upsilon-1)!} = n \binom{n-1}{\nu-1}$$

e quindi

$$\overline{\upsilon} = np \sum_{(\upsilon-1)=0}^{n} \binom{n-1}{\nu-1} p^{\upsilon-1} q^{n-\upsilon} = np$$

avendo posto m=n-1 e k=ν-1

Deviazione standard:

Si può dimostrare che per una qualunque distribuzione vale la relazione:

$$\sigma_x^2 = \overline{x^2} - \left(\overline{x}\right)^2$$

infatti

$$\sigma_x^2 = \overline{\left(x - \overline{x}\right)^2} = \overline{x^2 - 2x\overline{x} + \left(\overline{x}\right)^2} =$$
$$\overline{x^2} - 2\overline{x}\overline{x} + \left(\overline{x}\right)^2 = \overline{x^2} - \left(\overline{x}\right)^2$$

Quindi possiamo calcolare la varianza della distribuzione binomiale come segue:

$$\sigma_\upsilon^2 = \sum_{\upsilon=0}^{n} \upsilon^2 \binom{n}{\nu} p^\upsilon q^{n-\upsilon} - \left(np\right)^2$$

dove il primo termine a destra dell'uguale, ponendo

$\upsilon^2 = \upsilon(\upsilon - 1) + \upsilon$ diventa:

$$\sum_{\upsilon=2}^{n} \upsilon(\upsilon-1)\frac{n!}{\upsilon!(n-\upsilon)!}\,p^{\upsilon}\,q^{n-\upsilon} + \sum_{\upsilon=1}^{n} \upsilon\,\frac{n!}{\upsilon!(n-\upsilon)!}\,p^{\upsilon}\,q^{n-\upsilon} =$$

$$n(n-1)p^2 \sum_{(\upsilon-2)=0}^{n} \frac{(n-2)!}{(\upsilon-2)!(n-\upsilon)!}\,p^{\upsilon-2}\,q^{n-\upsilon} + np =$$

$$n(n-1)p^2 + np$$

avendo posto m=n-2 e k=ν-2, quindi

$$\sigma_{\upsilon}^2 = n(n-1)p^2 + np - (np)^2 = np(1-p)$$

$$\sigma_{\upsilon} = \sqrt{np(1-p)} = \sqrt{npq}$$

Se p=1/2 (esempio lancio di una moneta) il numero medio di successi è proprio n/2, inoltre la distribuzione è simmetrica:

$$B_{n,1/2}(\upsilon) = B_{n,1/2}(n-\upsilon)$$

Esempio:

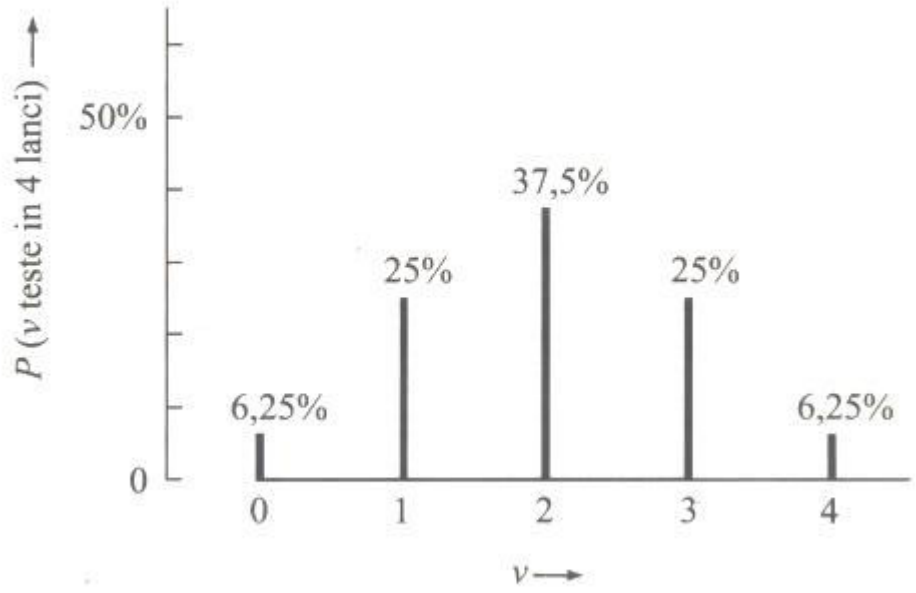
Si lanci 4 volte una moneta (n=4) possiamo facilmente calcolare le probabilità che escano ν

test $p(\upsilon, n=4) = \binom{4}{\nu}\left(\frac{1}{2}\right)^{\nu}\left(\frac{1}{2}\right)^{4-\nu} = \binom{4}{\nu}\left(\frac{1}{2}\right)^{4}$ e in

4 lanci, essendo p=q=1/2 :

per esempio:

$$p(\upsilon=0, n=4) = \left(\frac{1}{2}\right)^{4} = 0.0625$$



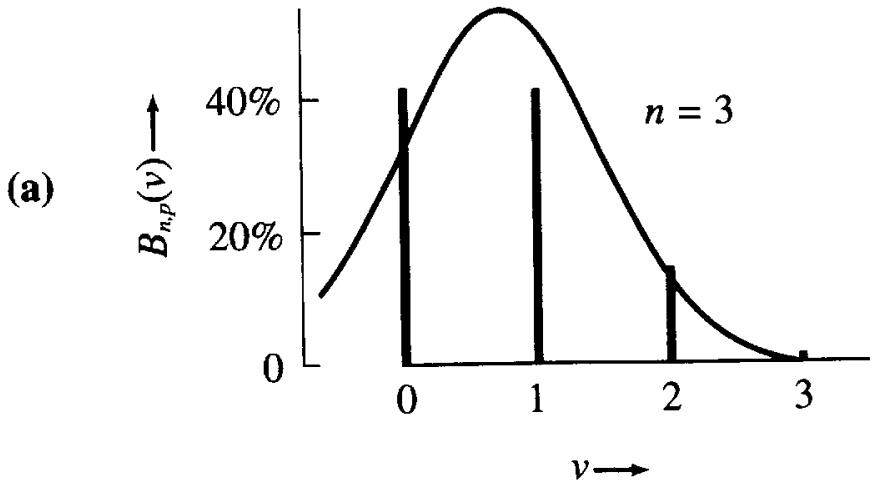
Si può dimostrare che **per n grande** la distribuzione binomiale tende alla distribuzione Gaussiana:

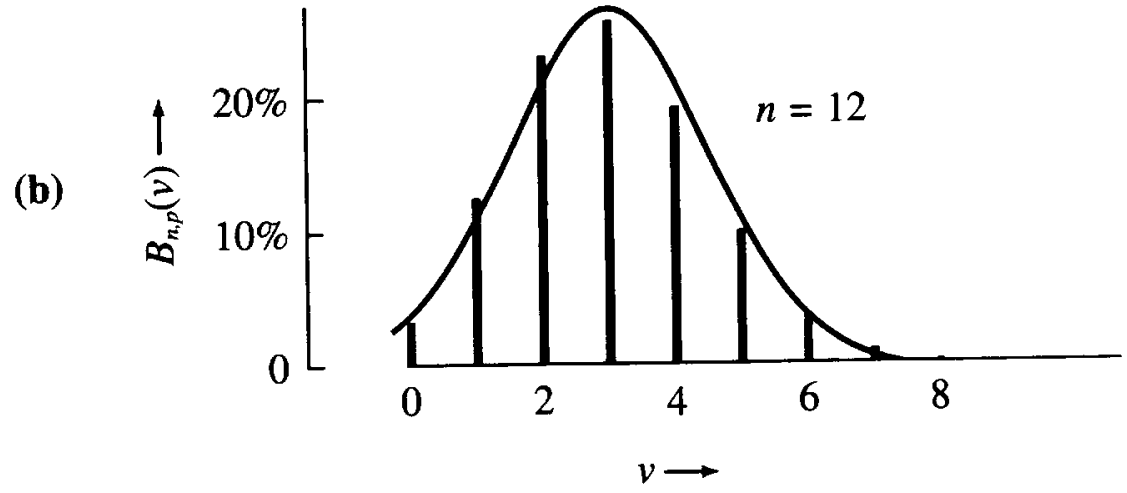
$$B_{n,p}(\upsilon) \approx G_{X,\sigma}(\upsilon)$$

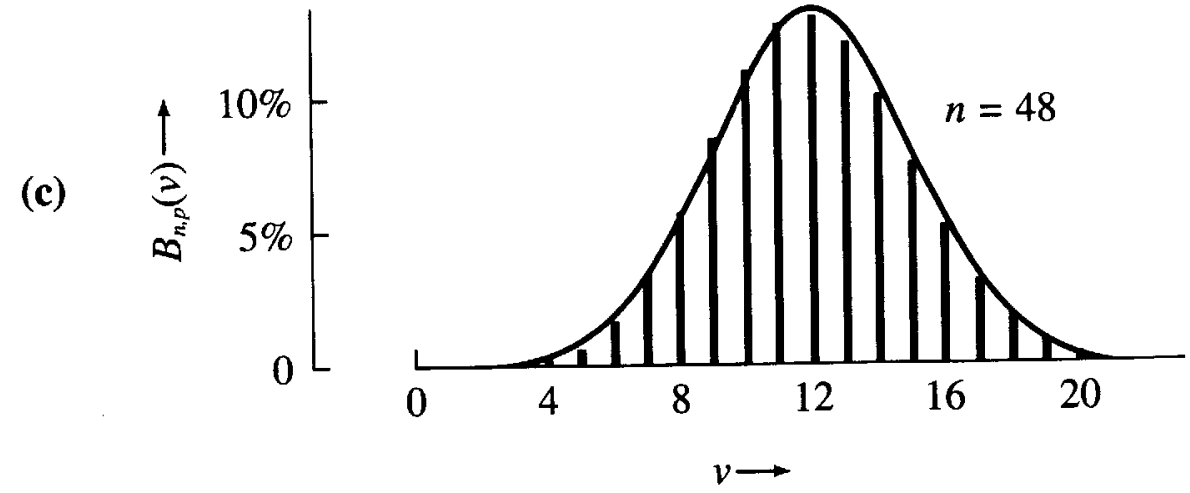
con

$$X = np \qquad\qquad \sigma = \sqrt{npq}$$

p=1/4


(a)
40%
20%
0
$B_{n,p}(v)$
$n = 3$
0
1
2
3
v


(b)
20%
10%
0
$B_{n,p}(v)$
$n = 12$
0
2
4
6
8
v


(c)
10%
5%
0
$B_{n,p}(v)$
$n = 48$
0
4
8
12
16
20
v

Possiamo ora giustificare l'affermazione per cui una misura soggetta a molti piccoli errori casuali è distribuita normalmente.

Se X è il valore vero assumiamo che le misure siano affette solo da errori casuali indipendenti, con n sorgenti di errore. Se questi errori hanno la stessa dimensione fissata ε allora per ogni sorgente si può avere con uguale probabilità (1/2) una sovrastima o una sottostima di X:

$x=X+\varepsilon$ o $x=X-\varepsilon$ ugualmente probabili

Se le sorgenti sono 2 si può avere:

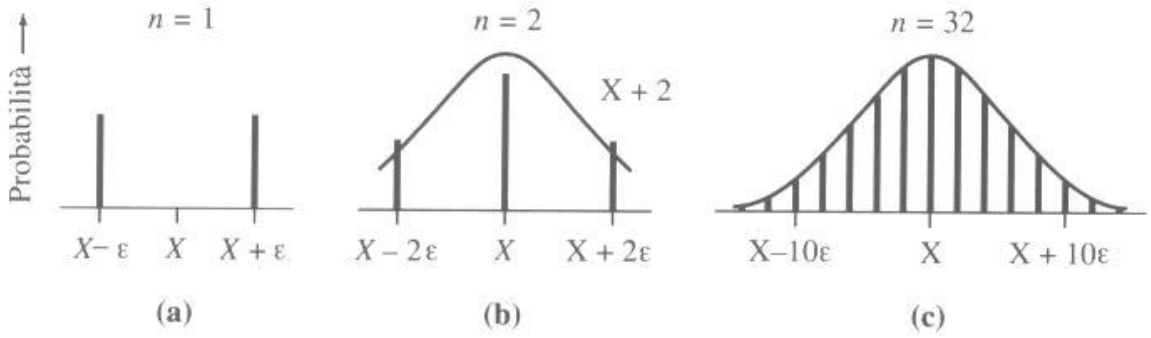
$x=X+2\varepsilon$, $x=X$ (gli errori si annullano) oppure $x=X-2\varepsilon$

In generale per n sorgenti, se υ danno errori positivi (n-υ) negativi si ha:

$$x = X+\upsilon\varepsilon - (n-\upsilon)\varepsilon = X+(2\upsilon-n)\varepsilon$$

La probabilità di ottenere questo risultato è dato dalla distribuzione binomiale

$$P(\nu \text{ errori positivi}) = B_{n,1/2}(\upsilon)$$



Propagando l'errore si ha:

$$\sigma_x = 2\varepsilon\sigma_\upsilon = \varepsilon\sqrt{n}$$

essendo

$$\sigma_\upsilon = \sqrt{np(1-p)} = \sqrt{\frac{n}{4}}$$

ponendo

$$n \to \infty \quad \text{e} \quad \varepsilon \to 0$$

in modo che $\sigma_x = \varepsilon\sqrt{n}$ rimanga finita, la distribuzione binomiale si approssima alla distribuzione gaussiana con centro X e larghezza σ_x. Inoltre poiché $\varepsilon \to 0$ la distribuzione discreta tende ad una distribuzione continua che è appunto la gaussiana.

Verifica di ipotesi

- Esempio: *verifica dell'efficacia di una sciolina*

Se effettuiamo 10 prove quante volte si deve verificare che gli sci trattati con la nuova sciolina siano più veloci, perché si possa dire che la sciolina è veramente efficace?

Ipotesi statistica:

Ipotesi nulla:la nuova sciolina non provoca alcun effetto

Se l'ipotesi nulla è corretta la probabilità di vincere per uno sci trattato rispetto ad uno non trattato è p=1/2

La probabilità che gli sci trattati ottengano ν successi è:

$$P(\nu \text{ vittorie in } 10 \text{ corse}) = B_{10,1/2}(\upsilon)$$

$$= \frac{10!}{\nu!(10-\nu)!}\left(\frac{1}{2}\right)^{10}$$

la probabilità che gli sci trattati vincano 10 corse è quindi 0.1%.

Se gli sci trattati vincessero tutte le 10 corse sarebbe molto improbabile che l'ipotesi nulla sia corretta.

Supponiamo che le vittorie siano 8 su 10 corse.

La probabilità di 8 o più vittorie è:

P(8 o più vittorie in 10 corse)=

= P(8 vittorie)+P(9 vittorie)+P(10 vittorie)≅ 5.5%

Si hanno 2 possibilità:

a) l'ipotesi nulla è corretta ma per caso è capitato un evento improbabile
b) l'ipotesi nulla è falsa: la sciolina è efficace

Si fissa un limite, p.e. il 5% (risultato significativo) o 1% (risultato altamente significativo) al di sotto del quale rifiutiamo l'ipotesi nulla (alternativa b).

Nel nostro caso 5.5% il risultato non è significativo e non è sufficiente per concludere che la sciolina funziona. Se avessimo ottenuto 10 vittorie (P=0.1%) il risultato sarebbe altamente significativo e potremmo concludere che la sciolina funziona.

PROCEDIMENTO GENERALE

- n prove indipendenti con due sole possibilità: successo o insuccesso
- ipotesi: si assume un valore per la probabilità di successo p
- il numero medio di successi in n prove è np
- se il numero di successi ν è vicino a np allora non c'è ragione di dubitare che l'ipotesi sia corretta.
- se il numero di successi ν è sensibilmente più grande di np, calcoliamo la probabilità, secondo l'ipotesi, di ottenere ν successi o più.
- Se questa probabilità e' minore del livello di significatività (5% o 1%), il valore osservato è improbabile e quindi l'ipotesi deve essere rigettata.
- se il numero di successi ν è sensibilmente più piccolo del valore medio np, dovremo calcolare la probabilità di ottenere ν successi o meno e procedere in maniera analoga.

Si noti che in alcuni casi la probabilità rilevante è la *probabilità a due code* di ottenere un risultato che differisca dalla media attesa di tanto quanto il valore realmente ottenuto o più.

29 DISTRIBUZIONE DI POISSON

Si voglia calcolare la probabilità di effettuare ν conteggi di eventi in un intervallo di tempo definito, se il numero di esperimenti n è grande e la probabilità p degli eventi è piccola (eventi rari), allora la distribuzione binomiale può essere approssimata dalla distribuzione di Poisson:

$$P_\mu(\upsilon) = e^{-\mu}\,\frac{\mu^\upsilon}{\upsilon!}$$

μ è un parametro positivo.

Caratteristiche della distribuzione:
Condizione di normalizzazione:

$$\sum_{\upsilon=0}^{\infty} P_\mu(\upsilon) = e^{-\mu}\sum_{\upsilon=0}^{\infty}\frac{\mu^\upsilon}{\upsilon!} = 1$$

essendo

$$e^{\mu} = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + ...$$

Valor medio

$$\overline{\upsilon} = \sum_{\upsilon=0}^{\infty} \upsilon P_{\mu}(\upsilon) = \sum_{\upsilon=0}^{\infty} \upsilon e^{-\mu} \frac{\mu^{\upsilon}}{\upsilon!} =$$

$$\mu e^{-\mu} \sum_{\upsilon-1=0}^{\infty} \frac{\mu^{\upsilon-1}}{(\upsilon-1)!} = \mu e^{-\mu} e^{\mu} = \mu$$

avendo posto e k=ν-1 ed essendo

$$e^{\mu} = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + ...$$

Deviazione standard

$$\sigma_\upsilon{}^2 = \overline{\upsilon^2} - \overline{\upsilon}^2 =$$

utilizzando l'uguaglianza $\upsilon^2 = \upsilon(\upsilon-1) + \upsilon$ si ha

$$\sum_{\upsilon=0}^{\infty} \upsilon^2 e^{-\mu} \frac{\mu^\upsilon}{\upsilon!} - \overline{\upsilon}^2 =$$

$$\sum_{\upsilon=2}^{\infty} \upsilon(\upsilon-1) e^{-\mu} \frac{\mu^\upsilon}{\upsilon!} + \sum_{\upsilon=0}^{\infty} \upsilon e^{-\mu} \frac{\mu^\upsilon}{\upsilon!} - \overline{\upsilon}^2 =$$

$$\mu^2 e^{-\mu} \sum_{\upsilon-2=0}^{\infty} \frac{\mu^{\upsilon-2}}{(\upsilon-2)!} + \mu - \mu^2 = \mu$$

essendo $\sum_{\upsilon-2=0}^{\infty} \frac{\mu^{\upsilon-2}}{(\upsilon-2)!} = e^\mu$ avendo posto e k = ν-2.

$$\sigma_\upsilon = \sqrt{\mu}$$

Negli esperimenti di conteggio la deviazione standard è uguale alla radice del valor medio.

Esempio, decadimenti radioattivi

Un campione di torio radioattivo emette particelle alfa ad un tasso di 1.5 al minuto

Quale è la probabilità di osservare ν particelle per ν= 0, 1, 2, 3, 4 e per ν ≥ 5 in 2 minuti?

Il valore atteso è il prodotto tra il tasso medio e il periodo (T=2 min.)

$$\mu = 1.5 \text{ x } 2 = 3$$

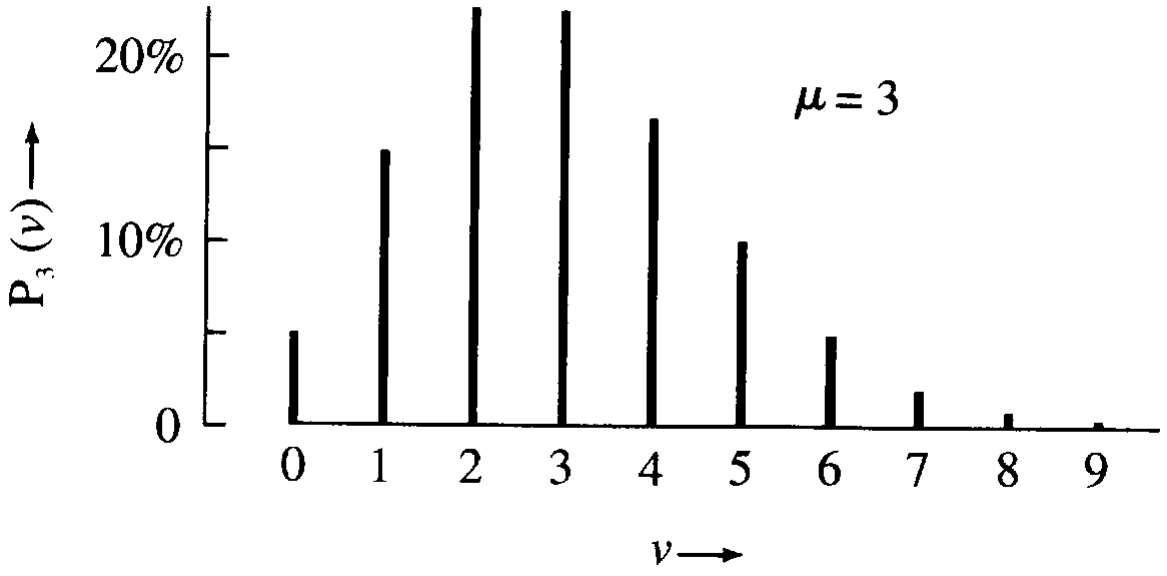
Si noti che non significa che ci aspettiamo di trovare 3 particelle in qualsiasi singola prova ma,

$$P(\upsilon \text{ particelle}) = P_3(\upsilon) = e^{-3}\frac{3^\upsilon}{\upsilon!}$$

da cui, per ν = 3

$$P_3(3) = e^{-3}\frac{3^3}{3!} = 0.22 = 22\%$$

quindi si troverà 3 solo nel 22% dei casi.



Si ha:

ν	0	1	2	3	4
$P_3(\upsilon)$	5%	15%	22%	22%	17%

$$P_3(\upsilon \geq 5) = 100\% - P_3(0) - P_3(1) - P_3(2) - P_3(3) - P_3(4) = 19\%$$

Utilizzando il principio della massima verosimiglianza si dimostra che se effettuiamo un esperimento di conteggio una volta trovando come risultato ν si ha:

$$\mu_{best} = \upsilon$$

e quindi, come risultato possiamo scrivere

$$\upsilon \pm \sqrt{\nu}$$

Se invece ripetiamo l'esperimento N volte si ha:

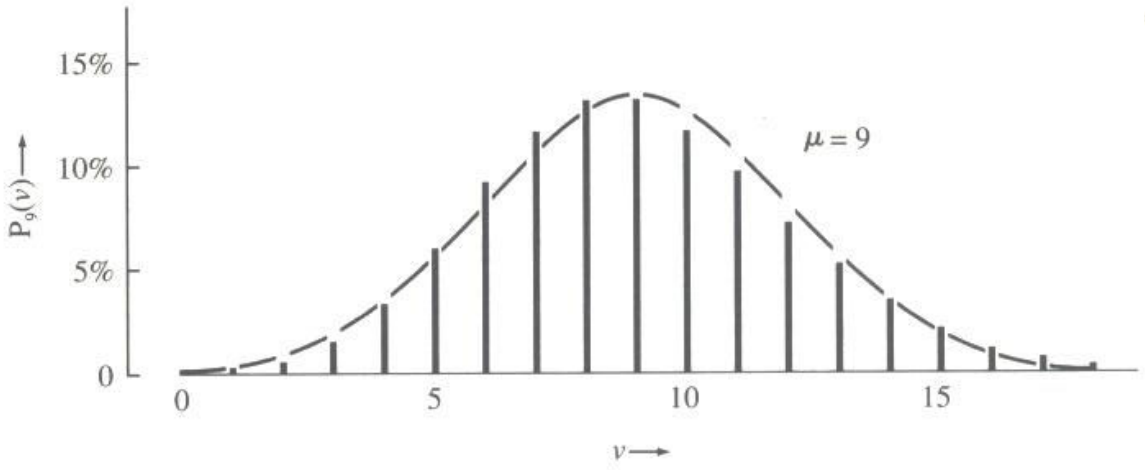
$$\mu_{best} = \frac{1}{N} \sum_{i=1}^{N} \upsilon_i$$

Si può mostrare che per $\mu \to \infty$ si ha:

$$P_\mu(\upsilon) \approx G_{X,\sigma}(\upsilon)$$

con

$$X = \mu \qquad \sigma = \sqrt{\mu}$$



30 TEST DEL CHI-QUADRATO (χ^2)

Supponiamo di voler confrontare la distribuzione di una serie di misure con una distribuzione teorica attesa.

Si abbiano N misure ed n intervalli con k=1, 2,….., n

- Sia O_k il numero di osservazioni che cadono nell'intervallo k-esimo
- Sia E_k il numero di misure nell'intervallo atteso secondo la distribuzione teorica

Definiamo la variabile χ^2:

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

Se $\chi^2 \leq n$ l'accordo è accettabile

Se $\chi^2 >> n$ il disaccordo è significativo

Infatti, se ripetessimo tante volte la serie di misure, allora il numero di misure in ogni intervallo equivarrebbe ad un esperimento di conteggio con distribuzione di probabilità di Poisson. Il valore medio e la deviazione standard saranno quindi:

$$\overline{O}_k = E_k \quad \text{e} \quad \sigma_{O_k} = \sqrt{E_k}$$

Se l'accordo è buono ci si aspetta che la differenza tra valori osservati e valori attesi sia paragonabile alla dimensione attesa delle fluttuazioni:

$$\left(O_k - E_k\right)^2 \approx E_k$$

e quindi

$$\frac{\left(O_k - E_k\right)^2}{E_k} \approx 1 \Rightarrow \chi^2 \leq n$$

Per trovare E_k si moltiplica la probabilità corrispondente all'intervallo k-esimo per il numero totale di misure N:

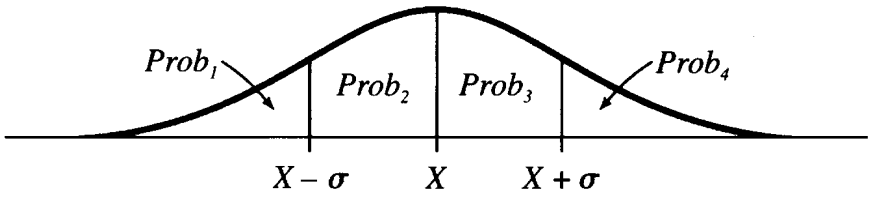
$$E_k = P_k \times N$$

Esempio distribuzione normale.

Si abbiano N=40 dati misurati da cui si stimano i parametri:

$$X = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - X)^2}$$

numero k nell'intervallo	*valori di x nell'intervallo*		*osservazioni O_k nell'intervallo*
1	$x < X - \sigma$	(o $x < 683{,}3$)	8
2	$X - \sigma < x < X$	(o $683{,}3 < x < 730{,}1$)	10
3	$X < x < X + \sigma$	(o $730{,}1 < x < 776{,}9$)	16
4	$X + \sigma < x$	(o $776{,}9 < x$)	6



$$P_2 + P_3 = 68\% \Rightarrow P_2 = P_3 = 0.34$$

$$P_1 + P_4 = (100-68)\% = 32\% \Rightarrow P_1 = P_4 = 0.16$$

intervallo k	*probabilità* P_k	*numero atteso* $E_k = NP_k$	*numero osservato* O_k
1	16%	6,4	8
2	34%	13,6	10
3	34%	13,6	16
4	16%	6,4	6

N=40

Si ha:

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} =$$

$$\frac{(1.6)^2}{6.4} + \frac{(-3.6)^2}{13.6} + \frac{(2.4)^2}{13.6} + \frac{(-0.4)^2}{6.4} = 1.80$$

CASO DI UNA VARIABILE CONTINUA

Sia f(x) la distribuzione attesa, allora

$$P(a < x < b) = \int_a^b f(x)dx$$

quindi per il k-esimo intervallo, il numero atteso di misure dopo N misure totali è

$$E_k = N \times P(a_k < x < a_{k+1}) = N \int_{a_k}^{a_{k+1}} f(x)dx$$

CASO DI UNA VARIABILE DISCRETA

Nel caso di una variabile discreta può non essere necessario definire gli intervalli possiamo infatti considerare i valori 'puntuali' di probabilità. Tuttavia in alcuni casi può essere utile raggruppare risultati diversi in un unico intervallo, come nell'esempio mostrato in tabella.

Tabella Occorrenza attesa di ν assi ($\nu = 0,1,...,5$) dopo aver lanciato cinque dadi 200 volte

risultato	*occorenze attese*	*numero dell'intervallo* k	*numero atteso* E_k
nessun asso	80,4	1	80,4
uno	80,4	2	80,4
due	32,2	3	32,2
tre	6,4	4	7,0
quattro	0,6		
cinque	0,03		

Si osservi che in generale gli intervalli devono essere scelti in modo tale da avere

$$E_k \geq 5$$

Inoltre il numero di intervalli (e di conseguenza il numero totale di misure) non deve essere troppo piccolo.

Altre forme di χ^2

In generale si può definire una variabile χ^2 come:

$$\chi^2 = \sum_{k=1}^{n} \left(\frac{\text{valore osservato} - \text{valore atteso}}{\text{deviazione standard}} \right)^2$$

Si può per esempio verificare la regressione del tipo y=f(x):
sia l'incertezza su x_k trascurabile e quella sulle y_k nota e uguale a σ_k, allora

$$\chi^2 = \sum_{k=1}^{n} \left(\frac{y_k - f(x_k)}{\sigma_k} \right)^2$$

Numero di gradi di libertà

Si definisce numero di gradi di libertà *d* il numero di dati misurati meno il numero di parametri da determinare dai dati:

$$d=n-c$$

Nel caso del χ^2 precedentemente studiato n corrisponde al numero di intervalli e c è il numero di parametri calcolati dai dati per ottenere i valori attesi (vincoli)

Se si usa la relazione $N = \sum_{k=1}^{n} O_k$ si ha c=1

Se si stimano i parametri della distribuzione gaussiana X e σ allora c=3

Deve sempre essere d ≥ 1

Si può dimostrare che il valor medio del chi-quadro tende al numero di gradi di libertà:

$$\overline{\chi^2}(atteso) = d$$

Quindi se il risultato del test indica che si ha:

$$\chi^2 >> \mathrm{d}$$

allora l'esito è negativo.

Si noti che poiché

$$\overline{\chi^2} = d = n - c$$

se $\boldsymbol{c}$ cresce , $\overline{\chi^2}$ diminuisce, in accordo con il fatto che i parametri sono stati calcolati con i dati e quindi l'accordo deve migliorare.

Si definisce **chi-quadro ridotto** la variabile:

$$\tilde{\chi}^2 = \frac{\chi^2}{d}$$

per quanto visto in precedenza si ha:

$$\overline{\tilde{\chi}^2}(atteso) = 1$$

Quindi se troviamo

$$\tilde{\chi}^2 \approx 1$$

non dobbiamo dubitare della distribuzione attesa, se invece

$$\tilde{\chi}^2 >> 1$$

allora è improbabile che la distribuzione attesa sia corretta.

Probabilità del χ^2

Dobbiamo definire quanto maggiore di uno deve essere il chi-quadro ridotto perché l'ipotesi sia scartata.

Esempio

$\tilde{\chi}^2 = 1.8$ è sufficientemente maggiore di uno per scartare l'ipotesi?

Supponiamo di aver calcolato $\tilde{\chi}_0^2$ (osservato), si può calcolare la probabilità di ottenere un valore maggiore o uguale di $\tilde{\chi}_0^2$, nell'ipotesi che le misure siano distribuite secondo la distribuzione attesa:

$$P(\tilde{\chi}^2 \geq \tilde{\chi}_0^2)$$

se questa probabilità è alta allora il valore $\tilde{\chi}_0^2$ è accettabile e non vi è motivo di rigettare la distribuzione attesa.

Il livello di accettazione (significatività) è scelto a priori, per esempio 5% (o 1%).

Se

$$P(\tilde{\chi}^2 \geq \tilde{\chi}_0^2) \leq 5\%$$

allora si deve rigettare l'ipotesi.

La probabilità P è funzione del numero di gradi di libertà e si può ricavare dalle tabelle.

	$\tilde{\chi}_o^2$												
d	0	0,25	0,5	0,75	1,0	1,25	1,5	1,75	2	3	4	5	6
1	100	62	48	39	32	26	22	19	16	8	5	3	1
2	100	78	61	47	37	29	22	17	14	5	2	0,7	0,2
3	100	86	68	52	39	29	21	15	11	3	0,7	0,2	
5	100	94	78	59	42	28	19	12	8	1	0,1		
10	100	99	89	68	44	25	13	6	3	0,1			
15	100	100	94	73	45	23	10	4	1				

Si osservi che:

- Il calcolo di queste probabilità tratta i numeri osservati O_k come variabili continue che sono distribuite gaussianamente attorno ai valori attesi E_k.
- O_k è un variabile discreta distribuita secondo la distribuzione di Poisson
- Questa distribuzione è ben approssimata da quella di Gauss se i numeri in gioco non sono troppo piccoli.
- Per questo il conteggio atteso per ogni intervallo E_k non deve essere troppo piccolo (almeno 5) e così pure il numero di intervalli.

ESEMPI SUL TEST DEL χ^2

Esempio 1

Distribuzione gaussiana

Tabella Misure delle altezze di 200 maschi adulti

intervalli k	*intervalli di altezze*	*numero* O_k *osservato*	*numero* E_k *atteso*
1	minore di $X - 1{,}5\sigma$	14	13,4
2	tra $X - 1{,}5\sigma$ e $X - \sigma$	29	18,3
3	tra $X - \sigma$ e $X - 0{,}5\sigma$	30	30,0
4	tra $X - 0{,}5\sigma$ e X	27	38,3
5	tra X e $X + 0{,}5\sigma$	28	38,3
6	tra $X + 0{,}5\sigma$ e $X + \sigma$	31	30,0
7	tra $X + \sigma$ e $X + 1{,}5\sigma$	28	18,3
8	più di $X + 1{,}5\sigma$	13	13,4

Per calcolare i valori attesi E_k si sono utilizzati 3 parametri X, σ, e il numero totale del campione N, quindi

d=8-3=5

$$\tilde{\chi}^2 = \frac{1}{5}\sum_{k=1}^{8}\frac{(O_k - E_k)^2}{E_k} = 3.5$$

Essendo maggiore di 1 si sospetta che le altezze non seguano la distribuzione gaussiana.

Dalle tabelle si ha:

$$P_5(\tilde{\chi}^2 \geq 3.5) \cong 0.5\%$$

quindi è molto improbabile che le altezze siano distribuite normalmente.

Possiamo rifiutare l'ipotesi di distribuzione normale al livello di significatività dell'1%.

Esempio 2

Se i dadi sono buoni si deve ottenere una distribuzione binomiale.

Tabella Distribuzione del numero di assi in 200 lanci di 5 dadi

intervallo k	*risultati per intervallo*	*numero* E_k *atteso*	*numero* O_k *osservato*
1	nessun asso	80,4	60
2	un asso	80,4	88
3	due assi	32,2	39
4	3, 4. o 5 assi	7.0	13

Si hanno 4 intervalli e 1 vincolo (il numero totale di dati, nessun parametro da determinare con i dati).

Il numero atteso E_k e' dato dal prodotto di N=200 dati totali per la probabilità data dalla distribuzione binomiale $B_{5,1/6}(\nu)$.

$$\tilde{\chi}^2 = \frac{1}{3}\sum_{k=1}^{4}\frac{(O_k - E_k)^2}{E_k} = 4.16$$

$$P_3(\tilde{\chi}^2 \geq 4.16) \cong 0.7\%$$

se i dadi sono buoni.

Quindi si conclude che i dadi sono quasi certamente **non** buoni.

Esempio 3

Distribuzione di Poisson per conteggi.

Numeri di particelle di raggi cosmici osservati in 100 intervalli distinti

Conteggi ν in 1 minuto	accadimenti	Numero intervallo k	O_k	E_k
Nessuno	7	1	7	7.5
Uno	17	2	17	19.4
Due	29	3	29	25.2
Tre	20	4	20	21.7
Quattro	16	5	16	14.1
Cinque	8 |			
Sei	1 |			
Sette	2 |	6	11	12.1
Otto o più	0 |			
totale	100			

L'analisi dei numeri della seconda colonna suggerisce immediatamente di raggruppare tutti i conteggi per $\nu \geq 5$.
Per calcolare i numeri attesi E_k, si è stimato il parametro μ della distribuzione di Poisson come $\overline{\upsilon}$ (si dimostra facilmente che questa è la miglior stima); inoltre si è usato il numero totale delle osservazioni N. Quindi i gradi di libertà sono d=6-2=4. Si ha:

$$\tilde{\chi}^2 = \frac{1}{4}\sum_{k=1}^{6}\frac{(O_k - E_k)^2}{E_k} = 0.35$$

$$P_4(\tilde{\chi}^2 \geq 0.35) \cong 85\%$$

Quindi non c'è motivo di dubitare della distribuzione di Poisson attesa.

Si noti che il valore atteso per il chi-quadro ridotto è 1 quindi il fatto che il risultato sia minore di uno è solo una fluttuazione dal valore medio e non dà una evidenza maggiore che la distribuzione sia quella di Poisson.

31 TEST DI STUDENT

Sia X una variabile casuale distribuita come il χ^2 con N gradi di libertà ed *u* una seconda variabile casuale, indipendente dalla prima e avente distribuzione normale standardizza (valor medio nullo e varianza 1).

Consideriamo la nuova variabile casuale *t* definita attraverso la:

$$t = \frac{u}{\sqrt{\frac{X}{N}}}$$

Si può dimostrare che la funzione densità di probabilità relativa alla variabile casuale *t* e' data dalla

$$f(t;N) = \frac{T_N}{\left(1+\frac{t^2}{N}\right)^{\frac{N+1}{2}}}$$

che si chiama **distribuzione di Student** ad N gradi di libertà.

Il coefficiente T_N è una costante che viene fissata dalla condizione di normalizzazione.

Se N viene fatto tendere all'infinito il denominatore della funzione tende a $e^{\frac{t^2}{2}}$ e dunque la distribuzione di Student tende alla distribuzione normale con media 0 e varianza 1.

Anche la forma della funzione di Student ricorda molto quella Normale, soltanto, rispetto a dati che seguano questa distribuzione, valori elevati dello scarto sono relativamente più probabili.

(Per N>35 la distribuzione di Student si può approssimare con la distribuzione Normale con media 0 e varianza 1).

La distribuzione di Student è simmetrica, quindi tutti i momenti di ordine dispari (compreso il valor medio) sono nulli; mentre la varianza della distribuzione è

$$VAR(t) = \frac{N}{N-2} \qquad \text{(se N>2)}$$

Indicando con $\bar{x}$ la media aritmetica di un campione di dimensione N, estratto a caso da una popolazione con distribuzione Normale, avente valore medio μ e varianza σ^2 e con *s* la stima della deviazione standard ottenuta dal campione stesso:

$$s^2 = \frac{\sum_i \left(x_i - \bar{x}\right)^2}{N-1}$$

e considerando che la variabile casuale

$$\left(N-1\right)\frac{s^2}{\sigma^2}$$

è distribuita come il χ^2 ad N-1 gradi di libertà e inoltre

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

segue la legge normale con media 0 e varianza 1, si ha che la variabile casuale

$$t = \frac{u}{\sqrt{\frac{\left(N-1\right)\frac{s^2}{\sigma^2}}{N-1}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

segue la distribuzione di Student ad N-1 gradi di libertà.

Quando la dimensione del campione non è sufficientemente grande da poter utilizzare l'approssimazione normale (piccoli campioni), è necessario usare le distribuzioni esatte delle variabili campionarie di cui ci si serve per effettuare il test.

Si voglia effettuare il confronto tra diversi campioni, più precisamente tra due medie campionarie ($\alpha = 1,2$), stimando la varianza della popolazione in basi ai dati stessi.

Indichiamo con n_α il numero di misure, con $\bar{y}_\alpha$ la media del campione e con S^2_α la varianza campionaria:

$$S^2_\alpha = \frac{1}{n_\alpha - 1} \sum_i \left(y_{\alpha_i} - \bar{y}_\alpha \right)^2$$

Si hanno due campioni:

Campione 1: $n_1 , \bar{y}_1 , S^2_1$

Campione 2: $n_2 , \bar{y}_2 , S^2_2$

Vogliamo verificare l'ipotesi nulla: $H_o : \mu_1 = \mu_2$

che i valori medi siano uguali, cioè i due campioni provengono dalla stessa popolazione. La differenza tra le medie campionarie

$\bar{y}_1 - \bar{y}_2$ segue la distribuzione di **Student** nella variabile:

$$t_\nu = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_{\bar{y}_1 - \bar{y}_2}}$$

con un numero di gradi di libertà $\nu = n_1 + n_2 - 2$

che per l'ipotesi nulla diventa:

$$t_\nu = \frac{\bar{y}_1 - \bar{y}_2}{S_{\bar{y}_1 - \bar{y}_2}}$$

La stima campionaria della varianza $\sigma^2_{\bar{y}_1 - \bar{y}_2}$, indicata con

$S^2_{\bar{y}_1 - \bar{y}_2}$, si ottiene tramite la seguente espressione:

$$S_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{(n_1 + n_2)}{n_1 n_2} \quad \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

Il test è bilaterale (a due code), prefissato il livello di significatività ad esempio al 5% se il **valore assoluto** della variabile t_ν di Student è maggiore del valore critico corrispondente al 5% con ν gradi di libertà (ricavabile dalle tabelle), si deve rifiutare l'ipotesi, altrimenti si accetta l'ipotesi nulla.

Esempio:
Si abbiano 2 campioni:

$$n_1 = 10 \quad , \quad \bar{y}_2 = 79.0 \quad , \quad S_1 = 0.7$$

$$n_2 = 10 \quad , \quad \bar{y}_2 = 81.0 \quad , \quad S_2 = 0.47$$

con $\nu = n_1 + n_2 - 2 = 18$ gradi di libertà.

Essendo $n_1 = n_2 = n$ si ha:

$$S_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{S_1^2 + S_2^2}{n}}$$

e quindi

$$t_\nu = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2 + S_2^2}{n}}} = -7.75$$

Il valore critico al livello del 5% con $\nu = 18$ è:

$$t_{crit} = 2.1$$

quindi l'ipotesi nulla che i campioni provengano dalla stessa popolazione deve essere rifiutata, i due campioni sono incompatibili.

TABELLE

Tabella La probabilità percentuale $P_d(\tilde{\chi}^2 \geq \tilde{\chi}^2_o)$ di ottenere un valore di $\tilde{\chi}^2 \geq \tilde{\chi}^2_o$ in un esperimento con d gradi di libertà, come funzione di d e $\tilde{\chi}^2_o$. (Gli spazi bianchi indicano probabilità minori di 0.05%.)

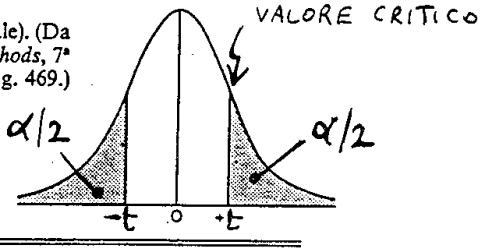
	$\tilde{\chi}_o^2$															
d	0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	8.0	10.0	
1	100	48	32	22	16	11	8.3	6.1	4.6	3.4	2.5	1.9	1.4	0.5	0.2	
2	100	61	37	22	14	8.2	5.0	3.0	1.8	1.1	0.7	0.4	0.2			
3	100	68	39	21	11	5.8	2.9	1.5	0.7	0.4	0.2	0.1				
4	100	74	41	20	9.2	4.0	1.7	0.7	0.3	0.1	0.1					
5	100	78	42	19	7.5	2.9	1.0	0.4	0.1							
	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
1	100	65	53	44	37	32	27	24	21	18	16	14	12	11	9.4	8.3
2	100	82	67	55	45	37	30	25	20	17	14	11	9.1	7.4	6.1	5.0
3	100	90	75	61	49	39	31	24	19	14	11	8.6	6.6	5.0	3.8	2.9
4	100	94	81	66	52	41	31	23	17	13	9.2	6.6	4.8	3.4	2.4	1.7
5	100	96	85	70	55	42	31	22	16	11	7.5	5.1	3.5	2.3	1.6	1.0
6	100	98	88	73	57	42	30	21	14	9.5	6.2	4.0	2.5	1.6	1.0	0.6
7	100	99	90	76	59	43	30	20	13	8.2	5.1	3.1	1.9	1.1	0.7	0.4
8	100	99	92	78	60	43	29	19	12	7.2	4.2	2.4	1.4	0.8	0.4	0.2
9	100	99	94	80	62	44	29	18	11	6.3	3.5	1.9	1.0	0.5	0.3	0.1
10	100	100	95	82	63	44	29	17	10	5.5	2.9	1.5	0.8	0.4	0.2	0.1
11	100	100	96	83	64	44	28	16	9.1	4.8	2.4	1.2	0.6	0.3	0.1	0.1
12	100	100	96	84	65	45	28	16	8.4	4.2	2.0	0.9	0.4	0.2	0.1	
13	100	100	97	86	66	45	27	15	7.7	3.7	1.7	0.7	0.3	0.1	0.1	
14	100	100	98	87	67	45	27	14	7.1	3.3	1.4	0.6	0.2	0.1		
15	100	100	98	88	68	45	26	14	6.5	2.9	1.2	0.5	0.2	0.1		
16	100	100	98	89	69	45	26	13	6.0	2.5	1.0	0.4	0.1			
17	100	100	99	90	70	45	25	12	5.5	2.2	0.8	0.3	0.1			
18	100	100	99	90	70	46	25	12	5.1	2.0	0.7	0.2	0.1			
19	100	100	99	91	71	46	25	11	4.7	1.7	0.6	0.2	0.1			
20	100	100	99	92	72	46	24	11	4.3	1.5	0.5	0.1				
22	100	100	99	93	73	46	23	10	3.7	1.2	0.4	0.1				
24	100	100	100	94	74	46	23	9.2	3.2	0.9	0.3	0.1				
26	100	100	100	95	75	46	22	8.5	2.7	0.7	0.2					
28	100	100	100	95	76	46	21	7.8	2.3	0.6	0.1					
30	100	100	100	96	77	47	21	7.2	2.0	0.5	0.1					

Tabella La probabilità percentuale, $P(\text{entro } t\sigma) = \int_{X-t\sigma}^{X+t\sigma} G_{X,\sigma}(x)dx$, come funzione di t.

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.80	1.60	2.39	3.19	3.99	4.78	5.58	6.38	7.17
0.1	7.97	8.76	9.55	10.34	11.13	11.92	12.71	13.50	14.28	15.07
0.2	15.85	16.63	17.41	18.19	18.97	19.74	20.51	21.28	22.05	22.82
0.3	23.58	24.34	25.10	25.86	26.61	27.37	28.12	28.86	29.61	30.35
0.4	31.08	31.82	32.55	33.28	34.01	34.73	35.45	36.16	36.88	37.59
0.5	38.29	38.99	39.69	40.39	41.08	41.77	42.45	43.13	43.81	44.48
0.6	45.15	45.81	46.47	47.13	47.78	48.43	49.07	49.71	50.35	50.98
0.7	51.61	52.23	52.85	53.46	54.07	54.67	55.27	55.87	56.46	57.05
0.8	57.63	58.21	58.78	59.35	59.91	60.47	61.02	61.57	62.11	62.65
0.9	63.19	63.72	64.24	64.76	65.28	65.79	66.29	66.80	67.29	67.78
1.0	68.27	68.75	69.23	69.70	70.17	70.63	71.09	71.54	71.99	72.43
1.1	72.87	73.30	73.73	74.15	74.57	74.99	75.40	75.80	76.20	76.60
1.2	76.99	77.37	77.75	78.13	78.50	78.87	79.23	79.59	79.95	80.29
1.3	80.64	80.98	81.32	81.65	81.98	82.30	82.62	82.93	83.24	83.55
1.4	83.85	84.15	84.44	84.73	85.01	85.29	85.57	85.84	86.11	86.38
1.5	86.64	86.90	87.15	87.40	87.64	87.89	88.12	88.36	88.59	88.82
1.6	89.04	89.26	89.48	89.69	89.90	90.11	90.31	90.51	90.70	90.90
1.7	91.09	91.27	91.46	91.64	91.81	91.99	92.16	92.33	92.49	92.65
1.8	92.81	92.97	93.12	93.28	93.42	93.57	93.71	93.85	93.99	94.12
1.9	94.26	94.39	94.51	94.64	94.76	94.88	95.00	95.12	95.23	95.34
2.0	95.45	95.56	95.66	95.76	95.86	95.96	96.06	96.15	96.25	96.34
2.1	96.43	96.51	96.60	96.68	96.76	96.84	96.92	97.00	97.07	97.15
2.2	97.22	97.29	97.36	97.43	97.49	97.56	97.62	97.68	97.74	97.80
2.3	97.86	97.91	97.97	98.02	98.07	98.12	98.17	98.22	98.27	98.32
2.4	98.36	98.40	98.45	98.49	98.53	98.57	98.61	98.65	98.69	98.72
2.5	98.76	98.79	98.83	98.86	98.89	98.92	98.95	98.98	99.01	99.04
2.6	99.07	99.09	99.12	99.15	99.17	99.20	99.22	99.24	99.26	99.29
2.7	99.31	99.33	99.35	99.37	99.39	99.40	99.42	99.44	99.46	99.47
2.8	99.49	99.50	99.52	99.53	99.55	99.56	99.58	99.59	99.60	99.61
2.9	99.63	99.64	99.65	99.66	99.67	99.68	99.69	99.70	99.71	99.72
3.0	99.73									
3.5	99.95									
4.0	99.994									
4.5	99.9993									
5.0	99.99994									

Tavola **Valori critici della distribuzione di *t*** (test bilaterale). (Da George W. Snedecor e William G. Cochran, *Statistical Methods*, 7ª ed., The Iowa State University Press, Ames-Iowa, USA, pag. 469.)

$$\frac{\alpha}{2} = \int_{-\infty}^{-x} S(t)\,dt$$



Gradi di libertà	α (SIGNIFICATIVITA')								
	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	0.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	.765	0.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	.741	.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
5	.727	.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	.718	.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	.711	.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	.706	.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	.703	.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	.700	.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	.697	.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	.695	.873	1.356	1.782	2.179	2.560	3.055	3.428	4.318
13	.694	.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	.692	.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	.691	.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	.690	.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	.689	.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	.688	.862	1.330	1.734	2.101	2.445	2.878	3.197	3.922
19	.688	.861	1.328	1.729	2.093	2.433	2.861	3.174	3.883
20	.687	.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	.686	.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	.686	.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	.685	.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	.685	.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	.684	.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	.684	.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	.684	.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	.683	.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	.683	.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	.683	.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
35	.682	.852	1.306	1.690	2.030	2.342	2.724	2.996	3.591
40	.681	.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
45	.680	.850	1.301	1.680	2.014	2.319	2.690	2.952	3.520
50	.680	.849	1.299	1.676	2.008	2.310	2.678	2.937	3.496
55	.679	.849	1.297	1.673	2.004	2.304	2.669	2.925	3.476
60	.679	.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
70	.678	.847	1.294	1.667	1.994	2.290	2.648	2.899	3.435
80	.678	.847	1.293	1.665	1.989	2.284	2.638	2.887	3.416
90	.678	.846	1.291	1.662	1.986	2.279	2.631	2.878	3.402
100	.677	.846	1.290	1.661	1.982	2.276	2.625	2.871	3.390
120	.677	.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373
∞	.6745	.8416	1.2816	1.6448	1.9600	2.2414	2.5758	2.8070	3.2905