Questo test serve a decidere se un certo campione segue una data legge, oppure no. Il test nasce per leggi discrete, ma come vedremo si può adattare anche a leggi continue.

Sia $X_1 \dots X_N$ un campione di v.a. indipendenti a valori in un insieme finito $E = \{k_1, \dots, k_m\}$. Poniamo, per $\underline{\theta} = (\theta_1, \dots, \theta_m)$ con $\theta_j > 0$, $\sum_{j=1}^{m} \theta_j = 1$,

$$P^{\underline{\theta}}(X_i = k_j) = \theta_j \quad , j = 1 \dots m,$$

ove l'indice $i \in \{1, \dots, N\}$ è arbitrario perché assumiamo che le X_i seguano tutte la stessa legge sconosciuta. I numeri θ_j sono quindi la densità discreta delle v.a. X_i, secondo la legge di parametro $\underline{\theta}$.

Vogliamo stabilire se il campione segue, o no, la legge di parametro $\underline{\theta}_0 = (p_1, \dots, p_m)$, ove supponiamo naturalmente $p_j > 0$ e $\sum_{j=1}^{m} p_j = 1$. Dunque

H_I: "il campione segue la legge determinata da $\underline{\theta}_0$",

H_A: "il campione segue una legge differente".

Per $j = 1, \dots, m$ definiamo l'effettivo empirico di k_j:

$$O_j(\omega) = \#\{i \in \{1, \dots, N\} : X_i(\omega) = k_j\}$$

(O_j è il numero di osservazioni che danno come risultato k_j) e l'effettivo teorico di k_j, $E_j = Np_j$ (E_j è il numero di volte

in cui, in teoria, dovremmo aspettarci di avere come risultato k_j, se la legge del campione è davvero quella determinata da $\underline{\theta_0}$). Infine definiamo la statistica di Pearson

$$T = \sum_{j=1}^{N} \frac{(O_j - E_j)^2}{E_j}.$$

Per questa statistica vale il teorema di Pearson, secondo il quale T ha approssimativamente legge $\chi^2(m-1)$, se N è abbastanza grande. Non dimostreremo questo fatto. Si considera valida l'approssimazione se si ha $Np_j \geq 5$ per ogni $j=1 \dots m$.

Di conseguenza, l'evento $\{T > \chi^2_{1-\alpha}(m-1)\}$ è una regione critica di livello α; ossia, se si calcola sulle osservazioni la statistica T e si trova un valore maggiore di $\chi^2_{1-\alpha}(m-1)$ l'ipotesi verrà rigettata.

Esempio Lanciamo un dado 2400 volte e troviamo questi risultati: esce 1 450 volte, 2 421 volte, 3 395 volte, 4 358 volte, 5 387 volte e 6 389 volte. Possiamo ipotizzare, al livello $\alpha = 0.05$, che il dado sia equilibrato?

Sia H_I: "il dado è equilibrato, cioè i numeri usciti seguono la legge uniforme",

H_A: "il dado non è equilibrato, cioè i numeri usciti seguono una diversa legge".

Allora poniamo porre $p_j = \frac{1}{6}$, $j=1\dots6$, $N=2400$, e stilare la seguente tabella:

k_j	O_j	E_j	$O_j - E_j$	$\frac{(O_j-E_j)^2}{E_j}$
1	450	400	50	6.15
2	421	400	21	1.10
3	395	400	-5	0.06
4	358	400	-42	4.41
5	387	400	-13	0.42
6	389	400	-11	0.30

Ne segue $T = 12.54$, mentre $\chi^2_{0.05} = 11.07$. Dunque, l'ipotesi, al livello 0.5, va rigettata.

Adottiamo il test a leggi continue. Sia $(X_1 \dots X_N)$ un campione di v.a. indipendenti, con funzione di ripartizione, secondo il parametro ϑ,

$$F_\vartheta(t) = P^\vartheta(X_i \le t), \quad t \in \mathbb{R},$$

per ogni $i=1\dots N$. Fissato ϑ_0, vogliamo decidere fra

H_I: "il campione ha F_{ϑ_0} come funzione di ripartizione",

e H_A: "il campione ha un'altra funzione di ripartizione".

Fissiamo $m-1$ numeri x_j con $-\infty < x_1 < x_2 < \dots < x_{m-1} < \infty$, e poniamo

$$I_1 =]-\infty, x_1], \; I_2 =]x_1, x_2], \dots, I_{m-1} =]x_{m-2}, x_{m-1}], \; I_m =]x_{m-1}, \infty[,$$

e definiamo infine, per $k=1,\dots,m$,

$$Y_j = k \iff X_j \in I_k .$$

Le Y_j sono v.a. discrete a valori in $\{1,\dots,m\}$, e si ha, se l'ipotesi è vera,

$$P^{\theta_0}(Y_j=k) = P^{\theta_0}(X_j \in I_k) = F_{\theta_0}(x_k) - F_{\theta_0}(x_{k-1}) \quad \text{se } 1<k<m,$$
$$P^{\theta_0}(Y_j=1) = P^{\theta_0}(X_j \in I_1) = F_{\theta_0}(x_1),$$
$$P^{\theta_0}(Y_j=m) = P^{\theta_0}(X_j \in I_m) = 1 - F_{\theta_0}(x_{m-1}).$$

Se si pone $p_k = P^{\theta_0}(Y_j=k)$, possiamo applicare il test del chi-quadro a $(Y_1,\dots,Y_m)$, con

$H_I' =$ "la legge di $Y_1,\dots,Y_m$ è determinata da $\underline{\theta_0}=\{p_1,\dots,p_m\}$"

$H_A' =$ "la legge di $Y_1,\dots,Y_m$ è un'altra".

e si accetterà o rigetterà l'ipotesi H_I se e solo se si accetta o si rigetta l'ipotesi H_I'.

Come si devono scegliere i numeri $x_1,\dots,x_{m-1}$? Devono essere <u>abbastanza</u> vicini, per poter distinguere fra leggi non troppo diverse fra loro, ma <u>non troppo</u> vicini, perché ciò potrebbe implicare $Np_j < 5$ per qualche j.

<u>Esempio</u> Supponiamo di avere 66 numeri; vogliamo decidere se la loro scelta segue, a livello 0.05, la legge $N(0,1)$ oppure no.

Anzitutto, bisogna scegliere gli intervalli I_j, $j=1,\dots,m$, ed i numeri p_j corrispondenti. Per esempio, fissiamo

$$p_j = \frac{1}{m}, \quad j=1,\dots,m,$$

e gli x_j, estremi degli intervalli I_j, in modo che

$$\Phi(x_1)=\frac{1}{m}, \quad \Phi(x_k)-\Phi(x_{k-1})=\frac{1}{m} \ (k=2,\dots,m-1), \quad 1-\Phi(x_{m-1})=\frac{1}{m}$$

vale a dire,

$$x_j = \phi_{j/m}, \quad j=1\dots m-1.$$

Possiamo scegliere $m=10$, cosicché $Np_j = \frac{66}{10} = 6.6 \geq 5$.

Supponiamo che i 66 numeri siano suddivisi, nei vari intervalli, così:

9 fra $-\infty$ e $\phi_{1/10} = -1.28$,
8 fra $\phi_{1/10}$ e $\phi_{2/10} = -0.84$,
11 fra $\phi_{2/10}$ e $\phi_{3/10} = -0.52$,
6 fra $\phi_{3/10}$ e $\phi_{4/10} = -0.25$,
2 fra $\phi_{4/10}$ e $\phi_{5/10} = 0$,
7 fra $\phi_{5/10}$ e $\phi_{6/10} = 0.25$,
8 fra $\phi_{6/10}$ e $\phi_{7/10} = 0.52$
5 fra $\phi_{7/10}$ e $\phi_{8/10} = 0.84$
5 fra $\phi_{8/10}$ e $\phi_{9/10} = 1.28$
5 fra $\phi_{9/10}$ e $+\infty$.

Si ha allora la seguente tabella:

k	O_k	E_k	$O_k - E_k$	$\frac{(O_k - E_k)^2}{E_k}$
1	9	6.6	2.4	0.87
2	8	6.6	1.4	0.30
3	11	6.6	4.4	2.93
4	6	6.6	−0.6	0.05
5	2	6.6	−4.6	3.21
6	7	6.6	0.4	0.02
7	8	6.6	1.4	0.30
8	5	6.6	−1.6	0.39
9	5	6.6	−1.6	0.39
10	5	6.6	−1.6	0.39

Ne segue $T = 8.85$, mentre $\chi^2_{0.95}(9) = 16.92$. Ne segue che, a livello 0.05, non possiamo respingere l'ipotesi.