

# Test chi-quadro

Finora abbiamo visto test d'ipotesi per testare ipotesi differenti, ma tutte concernenti il valore atteso di una o due popolazioni.

I precedenti test sono detti *parametrici* perché le ipotesi riguardano i valori dei parametri.

In questo capitolo vediamo come testare

- 1 l'ipotesi che la popolazione segua una legge fissata;
- 2 l'ipotesi che due variabili siano indipendenti

si tratta del test del chi-quadro di adattamento e di quello di indipendenza, due test *non parametrici*.

## Esempio: il dado

Partiamo con un esempio che ci aiuta a fissare le idee: ho un dado e mi chiedo se lanciandolo tutte le facce sono equiprobabili: chiamo  $p_1, p_2, p_3, p_4, p_5, p_6$  le probabilità che escano la faccia 1, 2, 3, 4, 5 e 6 rispettivamente.

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$$

Come al solito per decidere faremo  $n$  lanci. Occorre una statistica (= funzione degli  $n$  risultati) per decidere su  $H_0$ .

Idea: se  $H_0$  è vera la frequenza assoluta osservata di ogni faccia verrà “vicina” a  $n/6$ .

# Statistica per il dado

La statistica in questo caso particolare (fra poco vedremo la formula generale) è

$$Q = \sum_{i=1}^6 \frac{(F_a(i) - n/6)^2}{n/6},$$

dove  $F_a(i)$  è il numero di volte che abbiamo osservato la faccia  $i$  (= frequenza assoluta osservata di  $i$ ).

Se  $Q$  è abbastanza grande rifiuteremo  $H_0$ .

Resta da capire cosa significhi “abbastanza grande”.

## In generale

- 1 Dividiamo l'insieme dei possibili valori che le singole osservazioni possono assumere in  $k$  classi:  
 $C_1, C_2, \dots, C_k$ .
- 2 Chiamiamo  $p_i$  la probabilità che una osservazione appartenga alla classe  $C_i$ .
- 3 Decidiamo di fare  $n$  osservazioni.
- 4 Ognuna delle classi ha una frequenza assoluta teorica  $np_i$  e una frequenza assoluta osservata  $F_a(i)$ .
- 5 La statistica di riferimento è

$$Q = \sum_{i=1}^k \frac{(np_i - F_a(i))^2}{np_i}$$

## In generale

La statistica, a seconda dell'occorrenza, ha altre espressioni equivalenti:

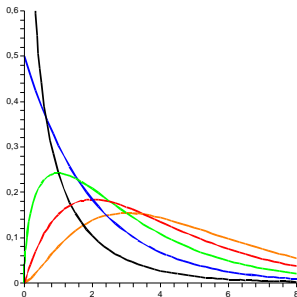
$$\begin{aligned} Q &:= \sum_{i=1}^{N_c} \frac{(np_i - F_a(i))^2}{np_i} = n \sum_{i=1}^{N_c} \frac{(p_i - F_r(i))^2}{p_i} \\ &= \sum_{i=1}^{N_c} \frac{F_a(i)^2}{np_i} - n = n \left( \sum_{i=1}^{N_c} \frac{F_r(i)^2}{p_i} - 1 \right) \end{aligned}$$

dove  $F_r(i) := F_a(i)/n$  è la *frequenza relativa* della classe  $i$ -esima.

## Legge chi-quadro

Ci serve ora un nuovo tipo di v.a. continua: la  $\chi^2(n)$ , che si legge “**chi-quadro a  $n$  gradi di libertà**”.

Per ogni  $n$ , la v.a.  $\chi^2(n)$  può assumere solo valori  $\geq 0$  e densità che ha forma diversa a seconda del valore di  $n$ .



**Nero** =  $\chi^2(1)$

**Blu** =  $\chi^2(2)$

**Verde** =  $\chi^2(3)$

**Rosso** =  $\chi^2(4)$

**Arancio** =  $\chi^2(5)$

# Teorema del chi-quadro

## TEOREMA PER TEST CHI-QUADRO

Se la legge delle  $X_1, \dots, X_n$  (che sono iid) è tale che  $\mathbb{P}(X_i \in C_i) = p_i$  per ogni  $i = 1, \dots, k$ , allora

$$Q = \sum_{i=1}^k \frac{(np_i - F_a(i))^2}{np_i}$$

è una v.a. la cui legge tende alla  $\chi^2(k-1)$  per  $n \rightarrow \infty$ .

Se le probabilità  $p_i$ , invece di essere date a priori, sono calcolate dopo aver stimato  $r$  parametri incogniti dai dati del campione, allora  $Q \rightarrow \chi^2(k-1-r)$ .

### Regola pratica per approssimare

Le approssimazioni  $Q \approx \chi^2(k-1)$  e  $Q \approx \chi^2(k-1-r)$  valgono se le probabilità  $p_i$  soddisfano  $np_i \geq 5$  per ogni  $i$ .

Test adattamento di livello  $\alpha$ 

Dato un campione casuale  $X_1, \dots, X_n$ , se le probabilità  $p_i$  sono determinate *senza stimare parametri* e se  $np_i \geq 5$  per ogni  $i \in \{1, \dots, k\}$ , allora il test di adattamento è

$H_0$	$\mathbb{P}(X_1 \in C_i) = p_i$ per ogni $i \in \{1, \dots, k\}$
$H_1$	$\exists i$ tale che $\mathbb{P}(X_1 \in C_i) \neq p_i$
Rifiutiamo $H_0$ se	$q > \chi^2_{1-\alpha}(k-1)$
$p$ -value: $\bar{\alpha}$ tale che	$q = \chi^2_{1-\bar{\alpha}}(k-1)$

dove  $\chi^2_{1-\alpha}(k-1)$  è il quantile  $1 - \alpha$  della legge  $\chi^2(k-1)$ .



Test adattamento di livello  $\alpha$ 

Dato un campione casuale  $X_1, \dots, X_n$ , se le probabilità  $p_i$  sono determinate *stimando  $r$  parametri* e se  $np_i \geq 5$  per ogni  $i \in \{1, \dots, k\}$ , allora il test di adattamento è

$H_0$	$\mathbb{P}(X_1 \in C_i) = p_i$ per ogni $i \in \{1, \dots, k\}$
$H_1$	$\exists i$ tale che $\mathbb{P}(X_1 \in C_i) \neq p_i$
Rifiutiamo $H_0$ se	$q > \chi^2_{1-\alpha}(k - r - 1)$
$p$ -value: $\bar{\alpha}$ tale che	$q = \chi^2_{1-\bar{\alpha}}(k - r - 1)$

dove  $\chi^2_{1-\alpha}(k - r - 1)$  è il quantile  $1 - \alpha$  della legge  $\chi^2(k - r - 1)$ .

## Tavole del chi-quadro

Tabella dei quantili  $\chi^2_\alpha(n)$  della legge chi-quadro

$$\alpha = P(Y \leq \chi^2_\alpha(n)) \quad \text{con } Y \sim \chi^2(n)$$

	0.01	0.025	0.05	0.95	0.975	0.99
1	0.00016	0.00098	0.00393	3.84146	5.02389	6.63490
2	0.02010	0.03064	0.10259	5.99146	7.37776	9.21034
3	0.11483	0.21580	0.35185	7.81473	9.34840	11.34487
4	0.29711	0.48442	0.71072	9.48773	11.14329	13.27670
5	0.55430	0.83121	1.14548	11.07050	12.83250	15.08627
6	0.87209	1.23734	1.63538	12.59159	14.44938	16.81189
7	1.23904	1.68987	2.16735	14.06714	16.01276	18.47531
8	1.64650	2.17973	2.73264	15.50731	17.53455	20.09024
9	2.08790	2.70039	3.32511	16.91898	19.02277	21.66599
10	2.55821	3.24697	3.94030	18.30704	20.48318	23.20925
11	3.05348	3.81575	4.57481	19.67514	21.92005	24.72497
12	3.57057	4.40379	5.22603	21.02607	23.33666	26.21697
13	4.10692	5.00875	5.89186	22.36203	24.73560	27.68825
14	4.66043	5.62873	6.57063	23.68479	26.11895	29.14124
15	5.22935	6.26214	7.26094	24.99579	27.48839	30.57791
16	5.81221	6.90766	7.96165	26.29623	28.84535	31.99993
17	6.40776	7.56419	8.67176	27.58711	30.19101	33.40866
18	7.01491	8.23075	9.39046	28.86930	31.52638	34.80531
19	7.63273	8.90652	10.11701	30.14353	32.85233	36.19087
20	8.26040	9.59078	10.85081	31.41043	34.16961	37.56623
21	8.89720	10.28290	11.59131	32.67057	35.47888	38.93217
22	9.54249	10.98232	12.33801	33.92444	36.78071	40.28936
23	10.19572	11.68855	13.09051	35.17246	38.07563	41.63840
24	10.85636	12.40115	13.84843	36.41503	39.36408	42.97982
25	11.52398	13.11972	14.61141	37.65248	40.64647	44.31410
26	12.19815	13.84390	15.37916	38.88514	41.92317	45.64168
27	12.87850	14.57338	16.15140	40.11327	43.19451	46.96294
28	13.56471	15.30786	16.92788	41.33714	44.46079	48.27824
29	14.25645	16.04707	17.70837	42.55697	45.72229	49.58788
30	14.95346	16.79077	18.49266	43.77297	46.97924	50.89218

Per  $n > 30$  si usa l'approssimazione normale:  $\chi^2_\alpha(n) \simeq z_\alpha \sqrt{2n} + n$

Come per la  $\mathcal{N}(0, 1)$  e per le Student  $t(n)$  anche per i quantili  $\chi^2(n)$  ci sono le tavole.

## Esercizio

Abbiamo osservato 2000 lanci di un dado, ecco il numero di volte che ciascuna faccia è stata osservata:

$i$	1	2	3	4	5	6
$F_a(i)$	314	322	316	344	316	388

Si può affermare che il dado non è equilibrato?

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$$

Il test  $\chi^2$  è applicabile poiché  $np_i = 2000/6 \approx 333.3 \geq 5$  per ogni  $i$ .

## La statistica

Calcoliamo

$$\begin{aligned}
 q &= \sum_{i=1}^6 \frac{(np_i - F_a(i))^2}{np_i} = \frac{(333.3 - 314)^2}{333.3} + \frac{(333.3 - 322)^2}{333.3} \\
 &\quad + \frac{(333.3 - 316)^2}{333.3} + \frac{(333.3 - 344)^2}{333.3} \\
 &\quad + \frac{(333.3 - 316)^2}{333.3} + \frac{(333.3 - 388)^2}{333.3} = 13.6
 \end{aligned}$$

Rifiutiamo  $H_0$ , con un livello  $\alpha$ , se questo numero è  $\geq \chi^2_{1-\alpha}(6-1)$ . Prendendo  $\alpha = 0.025$ , dato che  $\chi^2_{0.975}(5) = 12.82$ , rifiutiamo  $H_0$  e affermiamo che il dado non è equilibrato.

Se prendessimo  $\alpha = 0.01$ , dato che  $\chi^2_{0.99}(5) = 15.09$ , accetteremmo  $H_0$  e affermeremmo che non c'è sufficiente evidenza che il dado non sia equilibrato. Il  $p$ -value è compreso fra 0.025 e 0.01.

Esercizio:  $\mathcal{N}(0, 1)$ 

(Dall'eserciziario di Baldi-Giuliano-Ladelli, McGraw-Hill).

Un software statistico afferma di essere in grado di generare numeri a caso, in modo che la legge sia una  $\mathcal{N}(0, 1)$ .

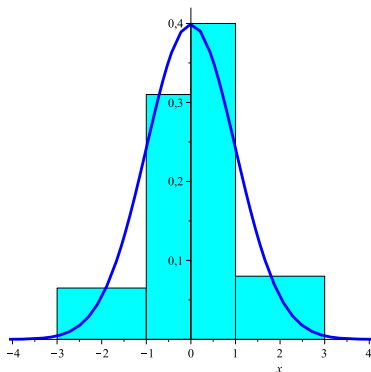
Vogliamo testare questa affermazione e osserviamo 100 numeri, suddividendoli in 4 classi.

Classe	$(-\infty, -1]$	$(-1, 0]$	$(0, 1]$	$(1, +\infty)$
$F_a(i)$	13	31	40	16

Possiamo dire che il software non è affidabile?

# L'adattamento

Si tratta di capire quanto le frequenze osservate siano vicine o meno alle previsioni teoriche della legge  $\mathcal{N}(0, 1)$ .



L'istogramma è costruito in modo che le aree dei rettangoli siano uguali alla frequenza relativa della base.

Dobbiamo calcolare le probabilità teoriche che una osservazione cada nelle classi:

$$\mathbb{P}(\mathcal{N}(0, 1) \leq -1) = \Phi(-1) = 0.16;$$

$$\mathbb{P}(-1 < \mathcal{N}(0, 1) \leq 0) = \Phi(0) - \Phi(-1) = 0.34; \text{ e per simmetria si ricavano le altre due.}$$

Classe	$(-\infty, -1]$	$(-1, 0]$	$(0, 1]$	$(1, +\infty)$
$F_a(i)$	13	31	40	16
Probabilità	0.16	0.34	0.34	0.16
$np_i$	16	34	34	16

Il test  $\chi^2$  è applicabile poiché  $np_i = 100p_i \geq 5$  per ogni  $i$ .

Calcoliamo

$$\begin{aligned} q &= \sum_{i=1}^k \frac{(np_i - F_a(i))^2}{np_i} \\ &= \frac{(13 - 16)^2}{16} + \frac{(31 - 34)^2}{34} \\ &\quad + \frac{(40 - 34)^2}{34} + \frac{(16 - 16)^2}{16} = 1.89 \end{aligned}$$

Rifiutiamo  $H_0$ , con un livello  $\alpha$ , se questo numero è  $\geq \chi^2_{1-\alpha}(4-1)$ . Con  $\alpha = 0.05$ , dato che  $\chi^2_{0.95}(3) = 7.815$ , accettiamo  $H_0$  (e il  $p$ -value è superiore a 0.05). Non c'è sufficiente evidenza che il software non sia affidabile.



Esercizio:  $\mathcal{N}(\mu, \sigma^2)$ 

La pressione massima misurata in 100 persone ha portato i seguenti dati (arrotondiamo all'intero):

Valori	Num.osservazioni	Valori	Num.osservazioni
113	1	115	3
116	6	117	5
118	11	119	18
120	9	121	12
122	13	123	7
124	8	125	3
126	2	127	1
128	1		

Il modello normale è valido per descrivere questi dati?

Stima per  $\mu$  e  $\sigma^2$ 

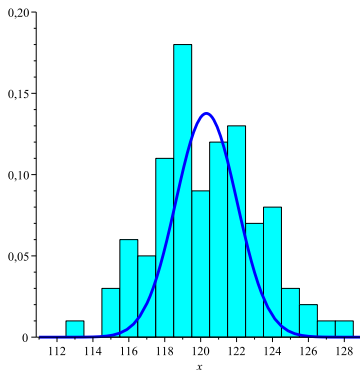
Valore atteso e varianza non sono dati, perciò li stimiamo con la media e la varianza campionarie. Utilizziamo la somma dei dati e la somma dei quadrati (forniti da un qualsiasi software matematico):

$$\sum_{i=1}^{100} x_i = 12032; \quad \sum_{i=1}^{100} x_i^2 = 1448522.$$

$$\bar{x}_n = 120.32; \quad s_n^2 = \frac{1448522}{99} - \frac{100}{99}(120.32)^2 = 8.40.$$

# Adattamento alla normale

Confrontiamo l'istogramma delle frequenze con la curva normale con valore atteso e varianza stimati.



## Divisione in classi

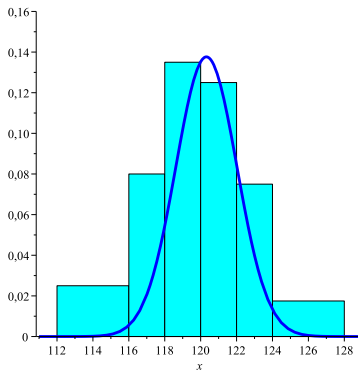
Per eseguire il test, occorre dividere i valori in classi in modo che in ciascuna di esse la  $\mathcal{N}(120.32, 8.40)$  abbia frequenza assoluta teorica  $\geq 5$ .

Si fanno tentativi, ad esempio questa divisione funziona (la frequenza teorica è  $n$  per la probabilità che una  $\mathcal{N}(120.32, 8.40)$  stia in quella classe):

Classi	Freq.ass.osservata	Freq.ass.teorica
$(-\infty, 116]$	10	6.81
$(116, 118]$	16	14.38
$(118, 120]$	27	24.43
$(120, 122]$	25	26.29
$(122, 124]$	15	17.89
$(124, +\infty)$	7	10.20

# Istogramma con le classi

Confrontiamo ora l'istogramma delle frequenze, che ha i valori suddivisi nelle classi, con la curva normale.



Calcoliamo

$$\begin{aligned} q &= \sum_{i=1}^k \frac{(np_i - F_a(i))^2}{np_i} \\ &= \frac{(6.811 - 10)^2}{6.81} + \frac{(14.38 - 16)^2}{14.375} + \frac{(24.43 - 27)^2}{24.43} \\ &\quad + \frac{(26.29 - 25)^2}{26.29} + \frac{(17.89 - 15)^2}{17.89} + \frac{(10.2 - 7)^2}{10.2} = 3.48 \end{aligned}$$

Rifiutiamo  $H_0$ , con un livello  $\alpha$ , se questo numero è  $\geq \chi^2_{1-\alpha}(6 - 2 - 1)$ . Con  $\alpha = 0.05$ , dato che  $\chi^2_{0.95}(3) = 7.815$ , accettiamo  $H_0$  (e il  $p$ -value è superiore a 0.05). Non c'è sufficiente evidenza che il modello normale non descriva bene la pressione arteriosa.

# Test indipendenza

Il test del chi-quadro può essere anche utilizzato quando si ha un campione di osservazioni accoppiate

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

per decidere se accettare o meno l'ipotesi

$H_0$ : le misurazioni  $X$  e le  $Y$  sono indipendenti.

## Costruiamo il test

- 1 Dividiamo l'insieme dei possibili valori che le singole  $X$  possono assumere in  $k$  classi:  $A_1, A_2, \dots, A_k$ . Allo stesso modo dividiamo l'insieme dei possibili valori che le singole  $Y$  possono assumere in  $j$  classi:  $B_1, B_2, \dots, B_j$ .

Otteniamo una tabella di questo tipo:

$\underbrace{Y \setminus X}$	$A_1$	$\dots$	$A_k$
$B_1$			
$\dots$			
$B_j$			



## Completiamo la tabella

- ② Facciamo  $n$  osservazioni.
- ③ Nella casella  $(A_i, B_m)$  mettiamo il numero di osservazioni accoppiate in cui la coordinata  $X$  sta in  $A_i$  e la  $Y$  in  $B_m$ . Dunque è la frequenza assoluta osservata della casella  $(A_i, B_m)$ : la indichiamo con  $F_a(i, m)$ .
- ④ Sommando sulle colonne otteniamo il numero di volte che abbiamo trovato  $X$  in  $A_i$ , la frequenza assoluta osservata della  $A_i$ :  $F_X(i)$ .
- ⑤ Sommando sulle righe otteniamo il numero di volte che abbiamo trovato  $Y$  in  $B_m$ , la frequenza assoluta osservata della  $B_m$ :  $F_Y(m)$ .

## Tabella delle frequenze

$\underbrace{Y}_{\setminus X}$	$A_1$	$\dots$	$A_k$	$F_Y$
$B_1$	$F_a(1, 1)$	$\dots$	$F_a(k, 1)$	$F_Y(1)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$B_j$	$F_a(1, j)$	$\dots$	$F_a(k, j)$	$F_Y(j)$
$F_X$	$F_X(1)$	$\dots$	$F_X(k)$	$n$

## Esempio

Campione con  $X$  e  $Y$  che assumono solo valori interi:

$(1,2), (1,1), (1,1), (2,3), (2,2),$   
 $(2,1), (2,1), (1,3), (2,3), (3,2)$

$Y \setminus X$	1	2	3	$F_Y$
1	2	2	0	4
2	1	1	1	3
3	1	2	0	3
$F_X$	4	5	1	10

## Stimiamo le probabilità

Stimiamo le probabilità per  $X$  e  $Y$ :

$$\mathbb{P}(X \in A_i) \approx \frac{F_X(i)}{n} = \hat{p}_X(i),$$

$$\mathbb{P}(Y \in B_m) \approx \frac{F_Y(m)}{n} = \hat{p}_Y(m).$$

Idea: se  $X$  e  $Y$  sono indipendenti vale

$$\mathbb{P}(X \in A_i, Y \in B_m) \approx \hat{p}_X(i)\hat{p}_Y(m)$$

e inoltre la casella  $(A_i, B_m)$  ha una frequenza assoluta teorica pari a

$$n\hat{p}_X(i)\hat{p}_Y(m) = \frac{F_X(i)F_Y(m)}{n}.$$

# Frequenze teoriche e osservate

Abbiamo una tabella di frequenze osservate

$\underbrace{Y} \setminus X$	$A_1$	$\dots$	$A_k$
$B_1$	$F_a(1, 1)$	$\dots$	$F_a(k, 1)$
$\dots$	$\dots$	$\dots$	$\dots$
$B_j$	$F_a(1, j)$	$\dots$	$F_a(k, j)$

e una di frequenze teoriche

$\underbrace{Y} \setminus X$	$A_1$	$\dots$	$A_k$
$B_1$	$\frac{F_X(1)F_Y(1)}{n}$	$\dots$	$\frac{F_X(k)F_Y(1)}{n}$
$\dots$	$\dots$	$\dots$	$\dots$
$B_j$	$\frac{F_X(1)F_Y(j)}{n}$	$\dots$	$\frac{F_X(k)F_Y(j)}{n}$

$\implies$  ci riconduciamo a un test di adattamento.

## Quante classi e quanti parametri stimati?

Le classi sono  $j \cdot k$ .

Abbiamo stimato  $k - 1$  probabilità per  $X$ :

$\hat{p}_X(1), \dots, \hat{p}_X(k - 1)$ . Infatti  $\hat{p}_X(k)$  viene ricavato dal fatto che deve essere  $\sum_{i=1}^k \hat{p}_X(i) = 1$ .

Allo stesso modo abbiamo stimato  $j - 1$  probabilità per  $Y$ :

$\hat{p}_Y(1), \dots, \hat{p}_Y(j - 1)$ .

I gradi di libertà della  $\chi^2$  saranno allora

$$j \cdot k - (j - 1) - (k - 1) - 1 = (j - 1)(k - 1).$$

# Test d'indipendenza

La statistica di riferimento è

$$Q = \sum_{i=1}^k \sum_{m=1}^j \frac{(F_X(i)F_Y(m)/n - F_a(i, m))^2}{F_X(i)F_Y(m)/n}$$

e il test

$H_0$	$X$ e $Y$ sono indipendenti
$H_1$	$X$ e $Y$ NON sono indipendenti
Rifiutiamo $H_0$ se	$q > \chi^2_{1-\alpha}((j-1) \cdot (k-1))$
$p$ -value: $\bar{\alpha}$ tale che	$q = \chi^2_{1-\bar{\alpha}}((j-1) \cdot (k-1))$

dove  $\chi^2_{1-\alpha}(h)$  è il quantile  $1 - \alpha$  della legge  $\chi^2(h)$ .

## Esercizio

(Dal libro *Appunti di Metodi matematici e statistici*, autore P.Baldi, editore CLUEB).

La *Cicindela fulgida* è una specie di coleottero. Si vuole capire se la sua colorazione (rosso brillante oppure non rosso) dipende dalla stagione oppure no. Si studiano  $n=671$  esemplari con risultato la seguente tabella:

$\underbrace{\text{periodo}} \setminus \text{colore}$	rosso	non rosso	$F_{\text{periodo}}$
inizio primavera	29	11	40
tarda primavera	273	191	464
inizio estate	8	31	39
tarda estate	64	64	128
$F_{\text{colore}}$	374	297	671



# Le frequenze teoriche

Le frequenze osservate

$\underbrace{\text{periodo}} \setminus \text{colore}$	rosso	non rosso
inizio primavera	29	11
tarda primavera	273	191
inizio estate	8	31
tarda estate	64	64

e quelle teoriche

$\underbrace{\text{periodo}} \setminus \text{colore}$	rosso	non rosso
inizio primavera	$\frac{40 \cdot 374}{671} = 22.3$	$\frac{40 \cdot 297}{671} = 17.7$
tarda primavera	$\frac{464 \cdot 374}{671} = 258.6$	$\frac{464 \cdot 297}{671} = 205.4$
inizio estate	$\frac{39 \cdot 374}{671} = 21.7$	$\frac{39 \cdot 297}{671} = 17.3$
tarda estate	$\frac{128 \cdot 374}{671} = 71.3$	$\frac{128 \cdot 297}{671} = 56.7$

Ci sono 8 classi: calcoliamo

$$q = \sum_{i=1}^8 \frac{(\text{freq.teoriche} - \text{freq.osservate})^2}{\text{freq.teoriche}} \\ = \dots = 27.55$$

Rifiutiamo  $H_0$ , con un livello  $\alpha$ , se questo numero è  $\geq \chi^2_{1-\alpha}((2-1)(4-1)) = \chi^2_{1-\alpha}(3)$ .

Con  $\alpha = 0.01$ , dato che  $\chi^2_{0.99}(3) = 11.34$ , rifiutiamo  $H_0$  (e il  $p$ -value è inferiore a 0.01): c'è sufficiente evidenza che il colore è correlato alla stagione.