Test di adattamento del chi-quadro

Il test di adattamento del chi-quadro confronta l’accordo (=adattamento) tra le frequenze osservate e le frequenze attese di dati organizzati in un certo numero di categorie qualitative

Test di adattamento del chi-quadro: esempio 1

Distribuzione dei gruppi sanguigni nella popolazione americana

Gruppo sanguigno	proporzione (%)
A	41
B	9
AB	4
0	46

Supponiamo di sospettare che la distribuzione dei gruppi sanguigni sia diversa tra le persone affette da tumore allo stomaco

Esempio 1

Campione cas. di 200 individui con tumore allo stomaco.

Gruppo sanguigno	frequenza osservata
A	92
B	20
AB	4
0	84

Ipotesi nulla: la distribuzione dei gruppi sanguigni nei pazienti affetti da tumore è la stessa della popolazione.

Possiamo rifiutarla?

Esempio 1

- H_0 $p_1 = 0{,}41$ $p_2 = 0{,}09$ $p_3 = 0{,}04$ $p_4 = 0{,}46$
- Frequenze osservate $N_1=92$ $N_2=20$ $N_3=4$ $N_4=84$
- Frequenze attese (v.a. binomiale di par. 200, p_i)

  $E_1 = 200*p_1 = 82$ $E_2 = 200*p_2 = 18$ $E_3 = 200*p_3 = 8$
  $E_4 = 200*p_4 = 92$

Esempio 1

- H_0 $p_1 = 0,41$ $p_2 = 0,09$ $p_3 = 0,04$ $p_4 = 0,46$
- Frequenze osservate $N_1=92$ $N_2=20$ $N_3=4$ $N_4=84$
- Frequenze attese (v.a. binomiale di par. 200, p_i)

  $E_1= 200*p_1= 82$ $E_2= 200*p_2= 18$ $E_3= 200*p_3= 8$
  $E_4= 200*p_4= 92$

Statistica del test

$$ST= \frac{(N_1 - E_1)^2}{E_1} + \frac{(N_2 - E_2)^2}{E_2} + \frac{(N_3 - E_3)^2}{E_3} + \frac{(N_4 - E_4)^2}{E_4} = 4,1374$$

Test di adattamento del chi-quadro

Cosa vuol dire "abbastanza grande" ?

(Pearson) per grandi valori di N (taglia del campione), ST ha una distribuzione approssimativamente chi-quadro con 3 gradi di libertà

(N_i non sono indipendenti, perché $N_1 + N_2 + N_3 + N_4 = N$;

E_i non sono indipendenti, perché $E_1 + E_2 + E_3 + E_4 = N$)

Test di adattamento del chi-quadro

Cosa vuol dire "abbastanza grande" ?
(Pearson) per grandi valori di N (taglia del campione), ST ha una distribuzione approssimativamente chi-quadro con 3 gradi di libertà
(N_i non sono indipendenti, perché $N_1 + N_2 + N_3 + N_4 = N$;
E_i non sono indipendenti, perché $E_1 + E_2 + E_3 + E_4 = N$)

$P(\chi^2 > 4{,}1374) = 0{,}24$ (non si può rifiutare H_0)

chidist(number, degrees_freedom)

Test di adattamento del chi-quadro

Campione di dimensione N, k categorie qualitative.

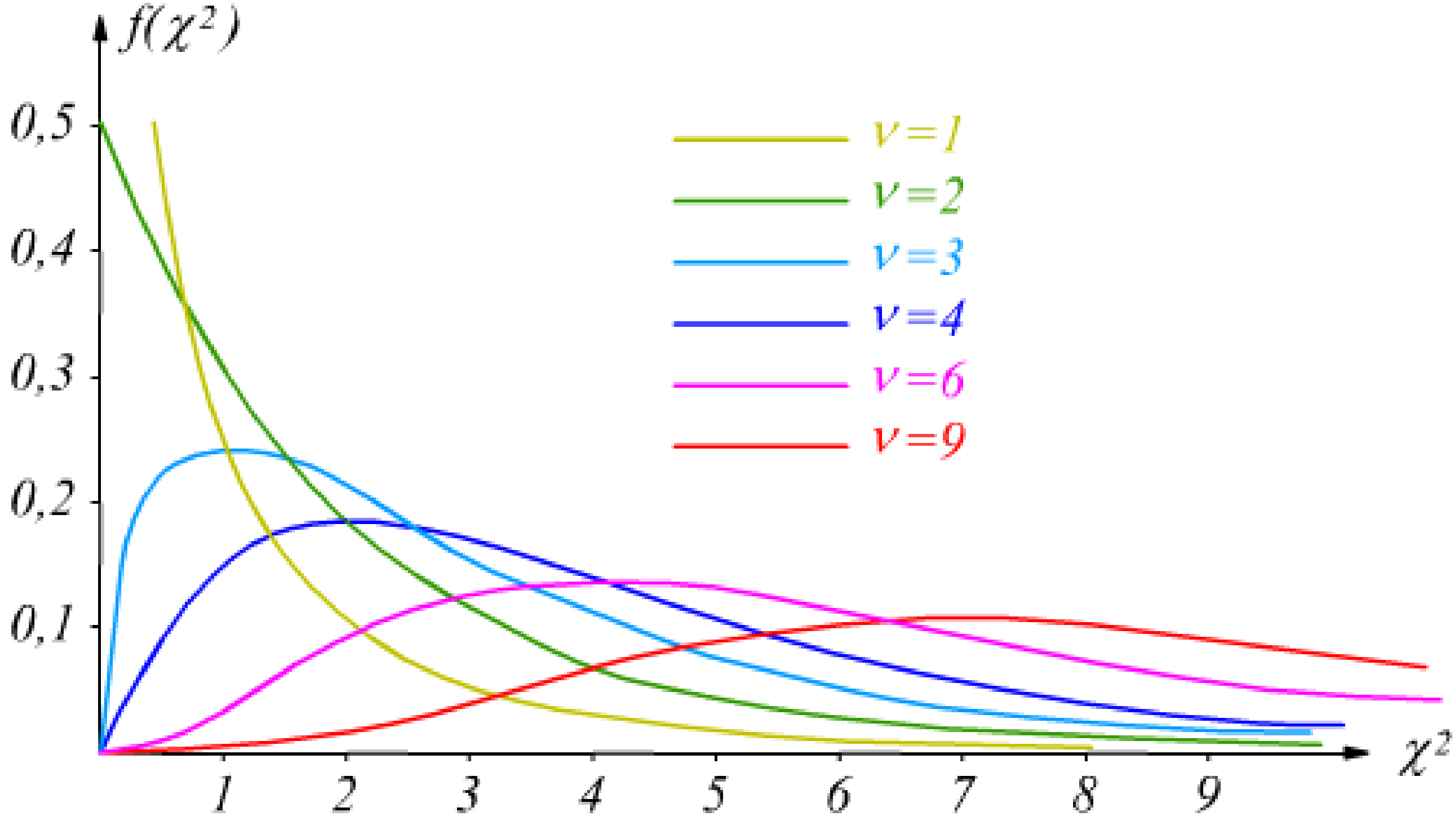
Frequenze osservate: $N_1, N_2, \ldots, N_k$.

$H_0 : p_1, p_2, \ldots, p_k$;   Frequenze attese: $E_i = p_i \, N$ i=1,…,k

Statistica del test   ST=$\sum_{i=1}^{k} \frac{(E_i - N_i)^2}{E_i}$ = a

P(χ^2> a)   (v.a. χ^2 con (k-1) gradi di libertà)

Densità della probabilità della variabile chi-quadro al variare dei gradi di libertà



Densità di probabilità della variabile χ^2

La funzione chidist(Number, degrees_fredom)
calcola l'area della coda della distribuzione a destra del valore "Number", ovvero la probabilità che χ^2 sia maggiore del valore assegnato.

Altrimenti, si usa la tabella che riporta i valori della variabile X^2 in corrispondenza dei più usati valori della significatività.

Tabella

Degrees of Freedom (df)	Probability (p)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Tabella

Degrees of Freedom (df)	Probability (p)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Tabella

Degrees of Freedom (df)	Probability (p)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Esempio 2

Per determinare se gli incidenti sul lavoro si verificano più frequentemente in certi giorni della settimana, vengono raccoli dati sugli incidenti in una fabbrica negli ultimi 12 mesi:

Lunedì 62
Martedì 47
Mercoledì 44 n totale di incidenti: 250
Giovedì 45
Venerdì 52

Verificare, a un livello del 5%, l'ipotesi che un incidente abbia la stessa probabilità di verificarsi in qualunque giorno della settimana.

Esempio 2

- H_0 : $p_i = 1/5$ $i=1, 2, 3, 4, 5$
- Frequenze osservate
- $N_1 = 62$, $N_2 = 47$, $N_3 = 44$, $N_4 = 45$, $N_5 = 52$
- Frequenze teoriche
- $E_i = N\,p_i = 250 * 1/5 = 50$
- $ST = \sum_i (E_i - N_i)^2 / E_i = 4{,}36$
- χ^2 a 4 gradi di libertà. $P(\chi^2 > 4{,}36)$

Esempio 2

Degrees of Freedom (df)	Probability (p)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		