Valutazione dell'indipendenza fra variabili qualitative: Il test del $\chi 2$

A cura di

Francesco Fabi

ce3s

CENTRO STUDI STATISTICI E SOCIALI

Tabelle di contingenza

Tabelle di contingenza

- Si utilizzano per confrontare proporzioni di caratteristiche qualitative in più popolazioni
- Si usano per classificare unità statistiche secondo due o più caratteri

Tabelle di contingenza
Esempio

Mancinismo vs. Genere

Mano dominante: sinistra vs. destra

Genere: Maschio vs. Femmina

- 2 categorie per ogni variabile, si ottiene una tabella 2X2

- Supponiamo di esaminare un campione di 300 unità

Tabelle di contingenza
Esempio

Le unità sono organizzate in una Tabella 2X2:

Numerosità campionaria

n = 300:

120 Femmine, 12 mancine

180 Maschi, 24 mancini

Genere	Mano dominante		
	Sinistra	Destra	
Femmina	12		120
Maschio	24		180
	36		300

Test χ^2 per la differenza tra due proporzioni

H_0: $\pi_1 = \pi_2$ (la proporzione di femmine mancine è uguale alla proporzione di maschi mancini)

H_1: $\pi_1 \neq \pi_2$ (Le due proporzioni non sono uguali– la mano dominante non è indipendente dal genere)

- Se H_0 è vera, la proporzione di femmine mancine deve essere uguale a quella di maschi mancini nella popolazione
- Le due proporzioni devono allora essere uguali alla proporzione della popolazione totale

Esempio: calcolo della frequenza media

La frequenza media:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{X}{n}$$

120 Femmine, 12 mancine

180 Maschi, 24 mancini

$$\bar{p} = \frac{12 + 24}{120 + 180} = \frac{36}{300} = 0.12$$

La proporzione globale di mancini è 12%

Calcolo delle frequenze attese

- Per ottenere la frequenza attesa delle femmine mancine si moltiplica la frequenza media per il totale di femmine
- Per ottenere la frequenza attesa dei maschi mancini si moltiplica la frequenza media per il totale di maschi

Se le due proporzioni sono uguali

P(mancinismo | Femmina) = P(mancinismo | Maschio) = 0,12

Il valore atteso è **(0,12)(120) = 14,4 per le femmine mancine**
(0,12)(180) = 21,6 per I maschi mancini

La statistica test χ^2

La statistica test χ^2 è:

$$\chi^2 = \sum_{\text{su tutte le celle}} \frac{(f_o - f_e)^2}{f_e}$$

- dove:

  f_o = frequenza osservata in una cella

  f_e = frequenza attesa nella stessa cella se H_0 è vera

χ^2 per il caso 2 x 2 ha 1 grado di libertà

(si assume che ogni cella abbia frequenza attesa almeno =5

La statistica χ^2

Genere	Mano dominante		
	Sinistra	Destra	
Femmina	Osservata = 12 Attesa = 14.4	Osservata = 108 Attesa = 105.6	120
Maschio	Osservata = 24 Attesa = 21.6	Osservata = 156 Attesa = 158.4	180
	36	264	300

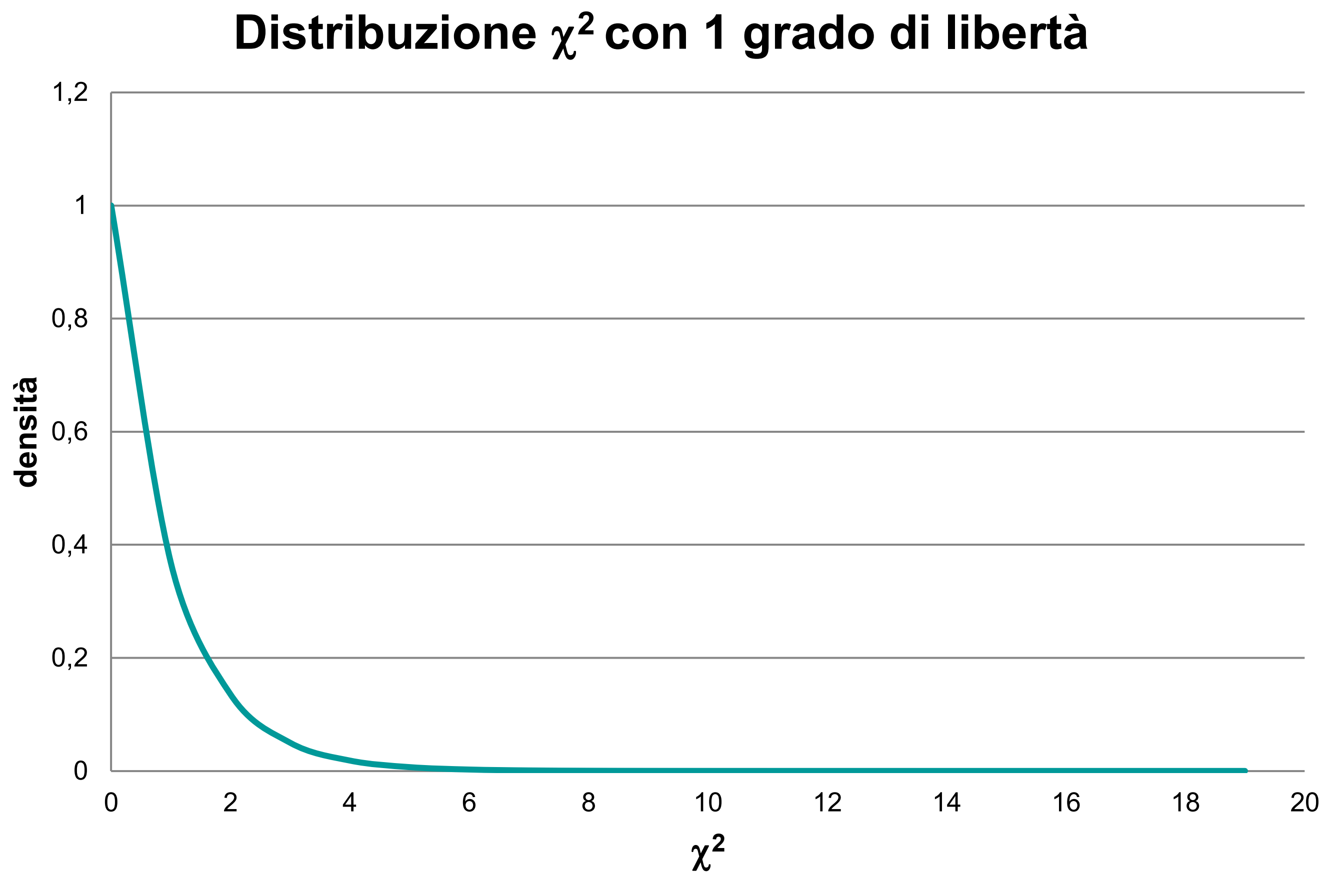
La statistica test è:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(12-14,4)^2}{14,4} + \frac{(108-105,6)^2}{105,6} + \frac{(24-21,6)^2}{21,6} + \frac{(156-158,4)^2}{158,4} = 0,7576$$

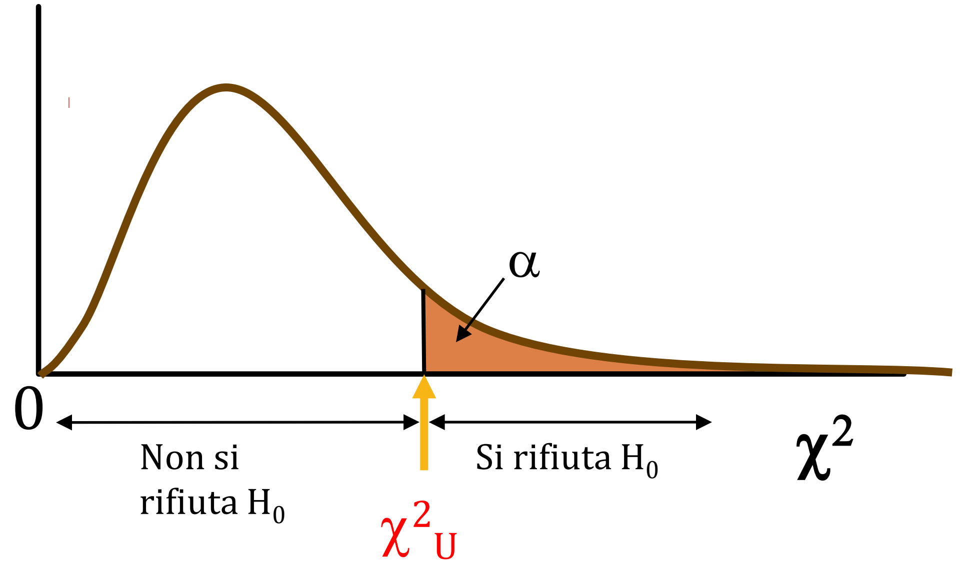
Distribuzione della statistica test

La statistica test χ^2 segue approssimativamente la distribuzione χ^2 con un grado di libertà



Regola di decisione generale

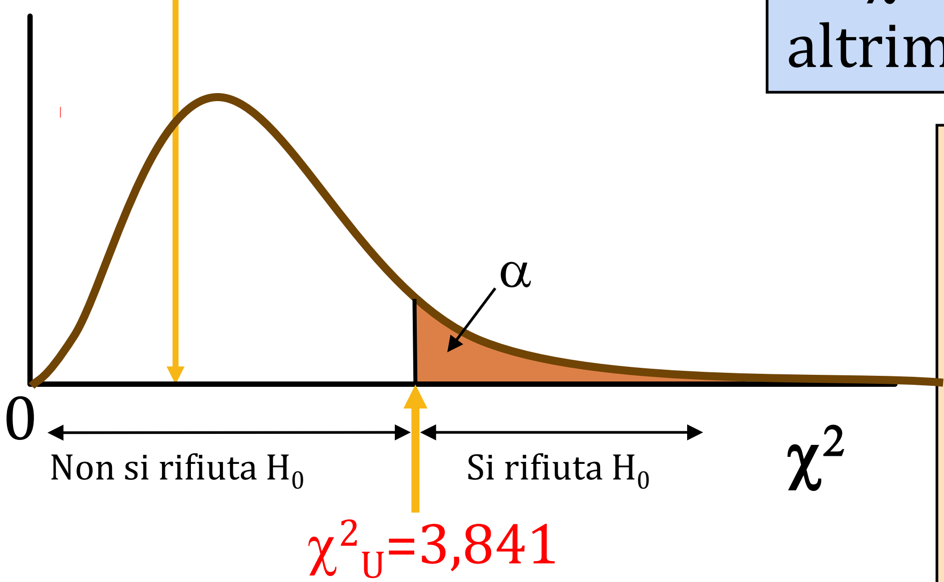
Regola di decisione:
Se $\chi^2 > \chi^2_U$, si rifiuta H_0, altrimenti non si rifiuta H_0



Regola di decisione

La statistica test è $\chi^2 = \boxed{0{,}7576}$, χ^2_U con 1 gl = 3,841

Regola di decisione:
Se $\chi^2 > 3{,}841$, si rifiuta H_0,
altrimenti, non si rifiuta H_0



$\chi^2 = 0{,}7576 < \chi^2_U = 3{,}841$,
Quindi non si rifiuta H_0 e si conclude che non c'è sufficiente evidenza di una differenza significativa tra le due proporzioni al livello $\alpha = 0{,}05$

Il test χ^2 di indipendenza

- estende il concetto di tabella di contingenza a r righe e c colonne

H_0: le due variabili sono indipendenti

H_1: le due variabili non sono indipendenti

Il test χ^2 di indipendenza

La statistica test χ^2 è:

$$\chi^2 = \sum_{\text{su tutte le celle}} \frac{(f_o - f_e)^2}{f_e}$$

- dove:

  f_o = frequenza osservata in una cella della tabella rxc

  f_e = frequenza attesa nella stessa cella se H_0 è vera

χ^2 per tabelle r x c ha (r-1)(c-1) gradi di libertà

(Si assume che ogni cella abbia frequenza attesa almeno =1)

Frequenze attese

- Frequenza attese:

$$f_e = \frac{\text{totale di riga} \times \text{totale di colonna}}{\text{n}}$$

dove:

totale di riga = somma di tutte le frequenze nella riga

totale di colonna = somma di tutte le frequenze nella colonna

n = numerosità campionaria totale

Regola di decisione

- La regola di decisione è

> Regola di decisione:
> Se $\chi^2 > \chi^2_U$, si rifiuta H_0,
> altrimenti non si rifiuta H_0

Dove χ^2_U è il valore critico della distribuzione χ^2 con $(r - 1)(c - 1)$ gradi di libertà

Esempio

- Nella tabella sono riportati i dati rilevati da un cardiologo sulla presenza di un certo anticorpo nel sangue e la gravità di attacco cardiaco

Attacco cardiaco	Anticorpo		Totale
	+	-	
+++	85	40	125
++	125	95	220
+	150	145	295
Totale	360	280	640

Esempio

- Tabella attesa sotto l'ipotesi H_0

Attacco cardiaco	Anticorpo		Totale
	+	-	
+++	70,31	54,69	125
++	123,75	93,25	220
+	165,94	129,06	295
Totale	360	280	640

Esempio: la statistica test

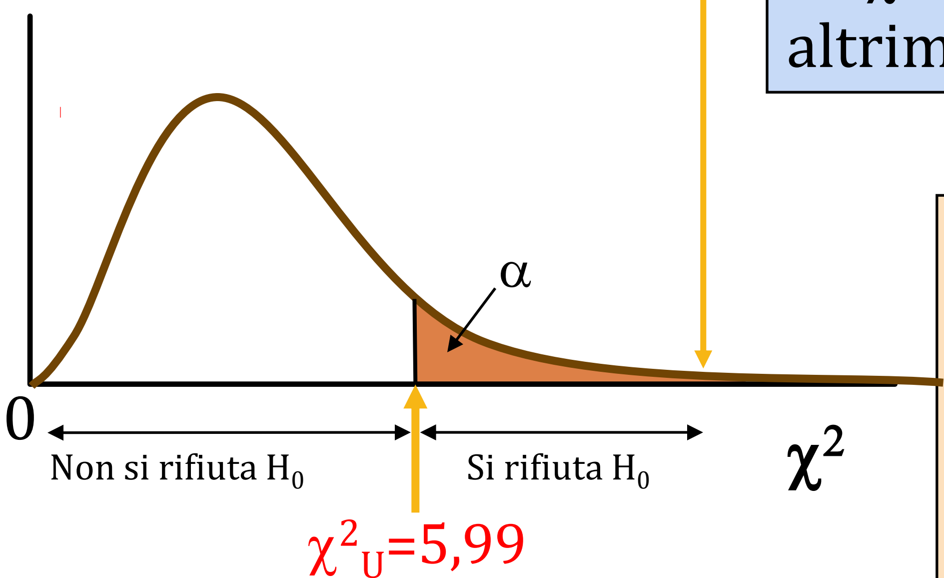
- Il valore della statistica test è χ_2=10,54 e $p<0.01$

$\chi^2_U = 5{,}99$ per $\alpha = 0{,}05$ (gradi di libertà (3-1)x(2-1)=2)

Decisione

La statistica test è $\chi^2 = 10{,}54$, χ^2_U con 2 gl $= 5{,}99$

Regola di decisione:
Se $\chi^2 > 5{,}99$, si rifiuta H_0,
altrimenti, non si rifiuta H_0



In questo caso,
$\chi^2 = 10{,}54 > \chi^2_U = 5{,}99$,
quindi si rifiuta H_0

Conclusione: c'è evidenza che l'anticorpo influenzi la gravità dell'attacco al livello $\alpha = 0{,}05$