Metodi statistici per la ricerca sociale

Capitolo 8. L'analisi dell'Associazione tra Variabili Categoriali

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM–88 SOCIOLOGIA E RICERCA SOCIALE

Analisi dell'associazione tra due caratteri

- Lo studio di un fenomeno collettivo richiede in genere la rilevazione, su ogni singola unità statistica, di molteplici caratteri che devono essere analizzati simultaneamente
- Obiettivo: Analizzare la relazione che intercorre tra le variabili che caratterizzano un certo fenomeno (Studio dell'associazione tra variabili)
- Si dice che esiste associazione tra due variabili se certi valori di una variabile tendono ad essere più frequenti in corrispondenza di certi valori dell'altra variabile
- Un'analisi di associazione tra due variabili è un'**analisi bivariata** perché coinvolge due variabili
- La natura delle variabili che si rilevano consente o suggerisce alcune analisi, escludendone altre

Caratteri Qualitativi (Nominali/Ordinali) *versus* Caratteri Quantitativi (Discreti/Continui)

Indagine sulla soddisfazione dei clienti

Cliente	Zona di Residenza	Classe di età	Livello di istruzione	Livello di soddisfazione
1	Centro	18-24	Obbligo	Medio
2	Nord	45-65	Laurea	Molto Basso
3	Sud-isole	45-65	Laurea	Basso
4	Nord	Oltre 65	Diploma	Basso
5	Centro	25-44	Laurea	Molto basso
6	Sud-Isole	Oltre 65	Obbligo	Alto

Obiettivo: Studio dell'associazione tra 'Livello di soddisfazione' e 'Livello di istruzione'

Distribuzioni doppie di frequenze (Tabelle di Contingenza)

- La distribuzione unitaria doppia riferita a due caratteri può essere sintetizzata attraverso una **distribuzione doppia di frequenze** (tabella di frequenze a doppia entrata)
- Dati due caratteri, X e Y, aventi rispettivamente r e c modalità, si definisce distribuzione doppia di frequenze (assolute) l'insieme delle frequenze (assolute) congiunte n_{ij} per $i = 1, \ldots, r$ e $j = 1, \ldots, c$

  n_{ij} = numero di unità che presentano congiuntamente la modalità i-esima del carattere X e la modalità j-esima del carattere Y

Distribuzione doppia di frequenze (Tabella di contingenza $r \times c$)

			Y			
X	y_1	$\cdots$	y_j	$\cdots$	y_c	Totale
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rc}	$n_{r.}$
Totale	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

Costruire la distribuzione doppia di frequenze assolute

Distribuzione unitaria doppia

Soggetto	Livello di istruzione	Livello di Soddisfazione soddisfazione
1	Obbligo	Medio
2	Laurea	Molto Basso
3	Laurea	Basso
4	Diploma	Basso
5	Laurea	Molto basso
6	Obbligo	Alto

Distribuzione doppia di frequenze assolute

Livello di istruzione	Livello di soddisfazione				
	Molto Basso	Basso	Medio	Alto	Totale
Obbligo	0	0	1	1	2
Diploma	0	1	0	0	1
Laurea	2	1	0	0	3
Totale	2	2	1	1	6

Distribuzione doppia di frequenze (Tabella di contingenza 3×4)

Livello di istruzione	Livello di soddisfazione				
	Molto Basso	Basso	Medio	Alto	Totale
Obbligo	40	50	20	40	150
Diploma	30	48	15	7	100
Laurea	26	14	7	3	50
Totale	96	112	42	50	300

Distribuzioni di frequenze marginali

Distribuzione di frequenza marginale della variabile X (ultima colonna):

X	$n_{i.}$
x_1	$n_{1.}$
$\vdots$	$\vdots$
x_i	$n_{i.}$
$\vdots$	$\vdots$
x_r	$n_{r.}$
Totale	n

Esempio: Distribuzione di frequenza marginale della variabile "Livello di istruzione"

Livello di istruzione	$n_{i.}$
Obbligo	150
Diploma	100
Laurea	50
Totale	300

Distribuzioni di frequenze marginali

Distribuzione di frequenza marginale di Y (ultima riga):

Y	y_1	$\cdots$	y_j	$\cdots$	y_c	Totale
Frequenza	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

Esempio: Distribuzione di frequenza marginale della variabile "Livello di soddisfazione"

Livello di soddisfazione	Molto Basso	Basso	Medio	Alto	Totale
$n_{.j}$	96	112	42	50	300

Frequenze relative e percentuali

Frequenze relative e percentuali congiunte

$$f_{ij} = \frac{n_{ij}}{n} \qquad \text{e} \qquad p_{ij} = \frac{n_{ij}}{n} \cdot 100 = f_{ij} \cdot 100$$

- Si noti che

$$\sum_{j=1}^{c} \sum_{i=1}^{r} f_{ij} = 1 \qquad \text{e} \qquad \sum_{j=1}^{c} \sum_{i=1}^{r} p_{ij} = 100$$

Distribuzione doppia di frequenze relative (percentuali)

Livello di istruzione	Livello di soddisfazione				Totale
	Molto Basso	Basso	Medio	Alto	
Obbligo	0.133 (13.3%)	0.167 (16.7%)	0.067 (6.7%)	0.133 (13.3%)	
Diploma	0.100 (10.0%)	0.160 (16.0%)	0.050 (5.0%)	0.023 (2.3%)	
Laurea	0.087 (8.7%)	0.047 (4.7%)	0.023 (2.3%)	0.010 (1.0%)	
Totale					1 (100%)

Frequenze relative e percentuali

Frequenze relative e percentuali marginali

- Frequenze relative e percentuali marginali di riga

$$f_{i.} = \frac{n_{i.}}{n} = f_{i1} + \cdots + f_{ic} \qquad \text{e} \qquad p_{i.} = \frac{n_{i.}}{n} \cdot 100 = f_{i.} \cdot 100$$

$$\sum_{i=1}^{r} f_{i.} = 1 \qquad \text{e} \qquad \sum_{i=1}^{r} p_{i.} = 100$$

- Frequenze relative e percentuali marginali di colonna

$$f_{.j} = \frac{n_{.j}}{n} = f_{1j} + \cdots + f_{rj} \qquad \text{e} \qquad p_{.j} = \frac{n_{.j}}{n} \cdot 100 = f_{.j} \cdot 100$$

$$\sum_{j=1}^{c} f_{.j} = 1 \qquad \text{e} \qquad \sum_{j=1}^{c} p_{.j} = 100$$

Distribuzione doppia di frequenze relative (percentuali)

Livello di istruzione	Livello di soddisfazione				
	Molto Basso	Basso	Medio	Alto	Totale
Obbligo	0.133 (13.3%)	0.167 (16.7%)	0.067 (6.7%)	0.133 (13.3%)	0.500 (50.0%)
Diploma	0.100 (10.0%)	0.160 (16.0%)	0.050 (5.0%)	0.023 (2.3%)	0.333 (33.3%)
Laurea	0.087 (8.7%)	0.047 (4.7%)	0.023 (2.3%)	0.010 (1.0%)	0.167 (16.7%)
Totale	0.320 (32.0%)	0.373 (37.3%)	0.140 (14.0%)	0.167 (16.7%)	1.000 (100%)

- Quando si riportano le distribuzioni di frequenza relative (percentuali) invece che le distribuzioni di frequenza assolute, è utile includere i totali in modo che il lettore se lo desidera possa ricavare le distribuzioni di frequenza assolute.

Frequenze condizionate

Frequenze relative e percentuali condizionate

- Frequenze relative e percentuali condizionate di riga (condizionate al carattere posto in testa alle righe):

$$f_{j|i} = \frac{n_{ij}}{n_{i.}} \qquad \text{e} \qquad p_{j|i} = \cdot 100 = \frac{n_{ij}}{n_{i.}} \cdot 100 = f_{j|i} \cdot 100$$

$$\sum_{j=1}^{c} f_{j|i} = 1 \qquad \text{e} \qquad \sum_{j=1}^{c} p_{j|i} = 100$$

- Frequenze relative e percentuali condizionate di colonna (condizionate al carattere posto in testa alle colonne):

$$f_{i|j} = \frac{n_{ij}}{n_{.j}} \qquad \text{e} \qquad p_{i|j} = \frac{n_{ij}}{n_{.j}} \cdot 100 = f_{i|j} \cdot 100$$

$$\sum_{i=1}^{r} f_{i|j} = 1 \qquad \text{e} \qquad \sum_{i=1}^{r} p_{i|j} = 100$$

Distribuzioni di frequenza condizionate della variabile "Livello di soddisfazione" data la variabile "Livello di istruzione"

Livello di Istruzione: Obbligo

	Molto Basso	Basso	Medio	Alto	Totale
$n_{j\|i=1}$	40	50	20	40	150
$f_{j\|i=1}$	0.267	0.333	0.133	0.267	1
$p_{j\|i=1}$	26.7%	33.3%	13.3%	26.7%	100%

Livello di Istruzione: Diploma

	Molto Basso	Basso	Medio	Alto	Totale
$n_{j\|i=2}$	30	48	15	7	100
$f_{j\|i=2}$	0.300	0.480	0.150	0.070	1
$p_{j\|i=2}$	30.0%	48.0%	15.0%	7.0%	100%

Livello di Istruzione: Laurea

	Molto Basso	Basso	Medio	Alto	Totale
$n_{j\|i=3}$	26	14	7	3	50
$f_{j\|i=3}$	0.520	0.280	0.140	0.060	1
$p_{j\|i=3}$	52.0%	28.0%	14.0%	6.0%	100%

Distribuzioni di frequenza condizionate della variabile "Livello di istruzione" data la variabile "Livello di soddisfazione"

Molto Basso

Livello di istruzione	$n_{i\|j=1}$	$f_{i\|j=1}$	$p_{i\|j=1}$
Obbligo	40	0.417	41.7%
Diploma	30	0.313	31.3%
Laurea	26	0.271	27.1%
Totale	96	1.000	100.0%

Basso

Livello di istruzione	$n_{i\|j=2}$	$f_{i\|j=2}$	$p_{i\|j=2}$
Obbligo	50	0.446	44.6%
Diplota	48	0.429	42.9%
Laurea	14	0.125	12.5%
Totale	112	1.000	100.0%

Medio

Livello di istruzione	$n_{i\|j=3}$	$f_{i\|j=3}$	$p_{i\|j=3}$
Obbligo	20	0.476	47.6%
Diploma	15	0.357	35.7%
Laurea	7	0.167	16.7%
Totale	42	1.000	100.0%

Alto

Livello di istruzione	$n_{i\|j=4}$	$f_{i\|j=4}$	$p_{i\|j=4}$
Obbligo	40	0.800	80.0%
Diploma	7	0.140	14.0%
Laurea	3	0.060	6.0%
Totale	50	1.000	100.0%

Indipendenza statistica

Tra due caratteri esiste indipendenza statistica quando la conoscenza della modalità di uno dei due caratteri non migliora la "previsione" della modalità dell'altro

- L'indipendenza statistica è un concetto simmetrico
  - ✓ Dati due caratteri X e Y, se X è indipendente da Y allora anche Y è indipendente da X e viceversa
  - ✓ Si dice che i due caratteri X e Y sono (statisticamente) indipendenti
- La tabella doppia di frequenze è uno strumento utile per studiare le relazioni che esistono tra due caratteri (in particolare tra due caratteri qualitativi)
  - ✓ Due variabili sono statisticamente indipendenti se nella popolazione le distribuzioni condizionate di una variabile rispetto a ciascuna categoria dell'altra sono identiche
  - ✓ Due variabili sono statisticamente dipendenti se nella popolazione le distribuzioni condizionate di una variabile rispetto a ciascuna categoria dell'altra NON sono identiche

Studio dell'associazione tra due caratteri

Siano X e Y due caratteri aventi rispettivamente r e c modalità:

$$x_1, \ldots, x_i, \ldots, x_r \qquad y_1, \ldots, y_j, \ldots, y_c$$

I due caratteri X e Y si dicono indipendenti se (nella popolazione) le distribuzioni relative condizionate di un carattere rispetto all'altro sono tutte uguali tra loro, ossia se, nella popolazione,

- le distribuzioni di frequenza condizionate relative del carattere X rispetto alle modalità del carattere Y, sono tutte uguali tra loro e uguali alla distribuzione marginale relativa del carattere X

$$f_{i|1} = f_{i|2} = \cdots = f_{i|j} = \cdots = f_{i|c} = f_{i.} \qquad i = 1, \ldots, r$$

- le distribuzioni di frequenza condizionate relative del carattere Y rispetto alle modalità del carattere X sono tutte uguali tra loro e uguali alla distribuzione marginale relativa del carattere Y

$$f_{j|1} = f_{j|2} = \cdots = f_{j|i} = \cdots = f_{j|r} = f_{.j} \qquad j = 1, \ldots, c$$

Associazione tra zona di residenza e status occupazionale

Zona di residenza (X)	Status Occupazionale (Y) Occupato (O)	Disoccupato (D)	Fuori dalla forza lavoro (FFL)	Totale
Nord	6	18	2	26
Centro	3	9	1	13
Sud	9	27	3	39
Isole	15	45	5	65
Totale	33	99	11	143

Distribuzioni relative condizionate dello status occupazionale rispetto alla zona di residenza

Zona di residenza	Status occupazionale O	D	FFL	Totale
Nord	0.231	0.692	0.077	1.000
Centro	0.231	0.692	0.077	1.000
Sud	0.231	0.692	0.077	1.000
Isole	0.231	0.692	0.077	1.000
Totale	0.231	0.692	0.077	1.000

Distribuzioni relative condizionate della zona di residenza rispetto allo status occupazionale

Zona di residenza	Status occupazionale O	D	FFL	Totale
Nord	0.182	0.182	0.182	0.182
Centro	0.091	0.091	0.091	0.091
Sud	0.273	0.273	0.273	0.273
Isole	0.455	0.455	0.455	0.455
Totale	1.000	1.000	1.000	1.000

"Zona di residenza" (X) e "Status occupazionale" (Y) sono statisticamente indipendenti

Frequenze assolute congiunte per caratteri indipendenti

Se due caratteri sono indipendenti, allora

$$f_{i|j} \equiv \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} \equiv f_{i.} \qquad \text{e} \qquad f_{j|i} \equiv \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \equiv f_{.j}$$

$$\Downarrow$$

Frequenza assoluta congiunta =

$$\frac{\text{Totale marginale di riga} \cdot \text{Totale marginale di colonna}}{\text{Numero delle osservazioni}}$$

$$\Updownarrow$$

$$n_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

Associazione tra zona di residenza e status occupazionale

Zona di residenza	Status Occupazionale: Occupato	Disoccupato	Fuori dalla forza lavoro	Totale
Nord	$6 = \frac{33 \cdot 26}{143}$	$18 = \frac{99 \cdot 26}{143}$	$2 = \frac{11 \cdot 26}{143}$	26
Centro	$3 = \frac{33 \cdot 13}{143}$	$9 = \frac{99 \cdot 13}{143}$	$1 = \frac{11 \cdot 13}{143}$	13
Sud	$9 = \frac{33 \cdot 39}{143}$	$27 = \frac{99 \cdot 39}{143}$	$3 = \frac{11 \cdot 39}{143}$	39
Isole	$15 = \frac{33 \cdot 65}{143}$	$45 = \frac{99 \cdot 65}{143}$	$5 = \frac{11 \cdot 65}{143}$	65
Totale	33	99	11	143

Frequenze teoriche di indipendenza

Le frequenze teoriche di una tabella doppia ottenute nell'ipotesi di indipendenza usando l'espressione

$$\frac{n_{i.} \cdot n_{.j}}{n}$$

sono dette **frequenze teoriche di indipendenza** (**frequenze teoriche**) e si indicano con $\hat{n}_{ij}$ (oppure f_e)

Frequenze osservate e frequenze teoriche

Distribuzione di frequenza doppia

Livello di istruzione	Livello di soddisfazione				
	Molto Basso	Basso	Medio	Alto	Totale
Obbligo	40	50	20	40	150
Diploma	30	48	15	7	100
Laurea	26	14	7	3	50
Totale	96	112	42	50	300

Tabella di indipendenza

Livello di istruzione	Livello di soddisfazione				
	Molto Basso	Basso	Medio	Alto	Totale
Obbligo	48	56	21	25	150
Diploma	32	37.3	14	16.7	100
Laurea	16	18.7	7	8.3	50
Totale	96	112	42	50	300

Test χ^2 di indipendenza

(Karl Pearson, 1900)

- Indipendenza = Concetto relativo alla popolazioine
- Test χ^2 di indipendenza

H_0 : Le variabili sono statisticamente indipendenti:

$$f_{i|1} = \cdots = f_{i|j} = \cdots f_{i|c} = f_{i.} \quad \text{per ogni } i$$

$$\Longleftrightarrow$$

$$f_{j|1} = \cdots = f_{j|i} = \cdots f_{j|r} = f_{.j} \quad \text{per ogni } j$$

H_a : Le variabili NON sono statisticamente indipendenti

- Statistica test

Contingenze (Residui)

- Differenze tra frequenze osservate e frequenze teoriche (Contingenze)

$$\text{Frequenza osservata} - \text{Frequenza teorica} = n_{ij} - \hat{n}_{ij} = n_{ij} - \frac{n_{i.} \cdot n_{.j}}{n}$$

- La somma delle differenze tra frequenze osservate e frequenze teorica è sempre nulla

$$\sum_i \left(n_{ij} - \hat{n}_{ij} \right) = \sum_j \left(n_{ij} - \hat{n}_{ij} \right) = \sum_i \sum_j \left(n_{ij} - \hat{n}_{ij} \right) = 0$$

Esempio: Il nettare di agave pastorizzato per la cura della tosse

Distribuzioni di frequenze congiunte e marginali

	La tosse disturba il sonno				
Agave	Per niente	Poco	Abbastanza	Molto	Totale
Si	280 (32.2%)	140 (16.1%)	20 (2.3%)	40 (4.6%)	480 (55.2%)
No	120 (13.8%)	220 (25.3%)	10 (1.1%)	40 (4.6%)	390 (44.8%)
Totale	400 (46.0%)	360 (41.4%)	30 (3.4%)	80 (9.2%)	870 (100%)

Esempio

- Distribuzioni di frequenze relative condizionate (%) del livello di disturbo del sonno dovuto alla tosse dato il gruppo

	La tosse disturba il sonno				
Agave	Per niente	Poco	Abbastanza	Molto	Totale
Si	58.3	29.2	4.2	8.3	100.0
No	30.8	56.4	2.6	10.3	100.0

- Distribuzioni di frequenze relative condizionate (%) del gruppo dato il livello di disturbo del sonno dovuto alla tosse

	La tosse disturba il sonno			
Agave	Per niente	Poco	Abbastanza	Molto
Si	70.0	38.9	66.7	50.0
No	30.0	61.1	33.3	50.0
Totale	100.0	100.0	100.0	100.0

Frequenze osservate (frequenze teoriche)

- Distribuzione di frequenza doppia (Frequenze teoriche)

	La tosse disturba il sonno				
Agave	Per niente	Poco	Abbastanza	Molto	Totale
Si	280	140	20	4	480
	(220.7)	(198.6)	(16.6)	(44.1)	
No	120	220	10	4	390
	(179.3)	(161.4)	(13.4)	(35.9)	
Totale	400	360	30	80	870

- Frequenze osservate – Frequenze teoriche

	La tosse disturba il sonno				
Agave	Per niente	Poco	Abbastanza	Molto	Totale
Si	59.3	-58.6	3.4	-4.1	0.00
No	-59.3	220-161.4 = 58.6	-3.4	4.1	0.00
Totale	0.00	0.00	0.00	0.00	0.00

La statistica test χ^2 di Pearson

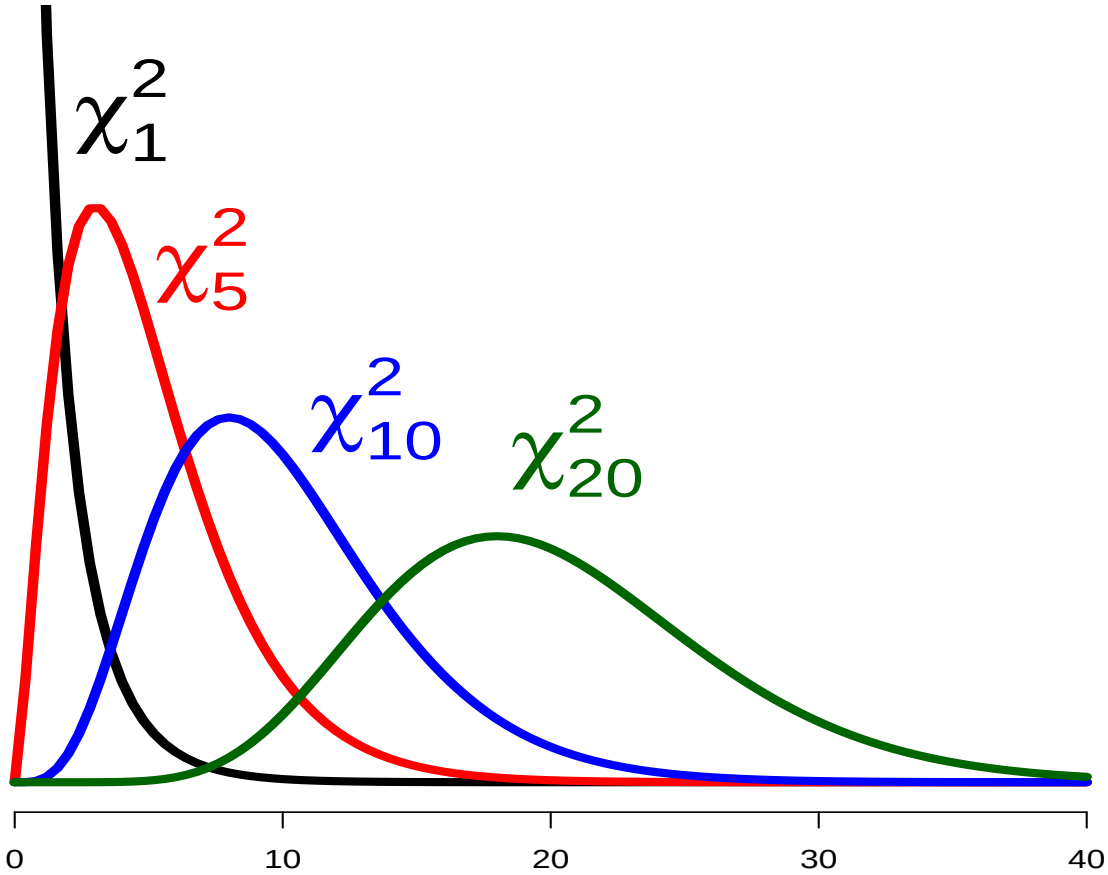
$$\chi^2 = \sum \frac{(\text{frequenze osservate - frequenze attese})^2}{\text{Frequenze attese}} = \sum \frac{(f_0 - f_e)^2}{f_e}$$

$$\Updownarrow$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

- $\chi^2 \geq 0$
- Sotto H_0, se il campione è sufficientemente grande la distribuzione campionaria della statistica test χ^2 si può approssimare con una distribuzione $\chi^2_{(r-1)\cdot(c-1)}$
- Significato dei gdl: Fissati i marginali e $(r-1)\cdot(c-1)$ frequenze sono automaticamente determinate le restanti frequenze

Funzione di densità di probabilità della v.a. χ^2_{gld}



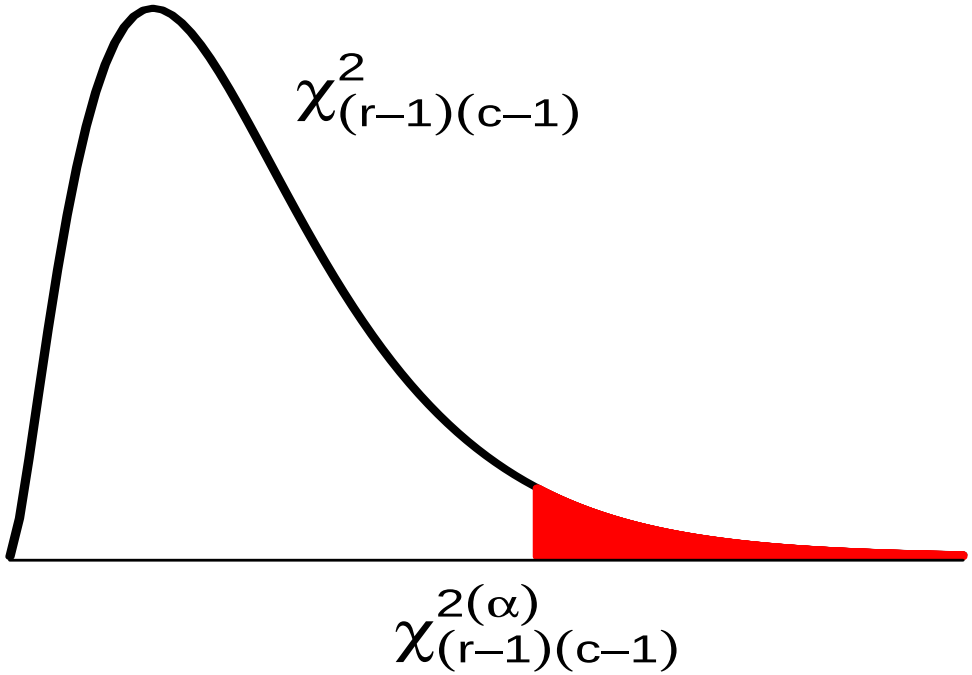
- La variabile aleatoria χ^2 assume solo valori positivi
- La funzione di densità è asimettrica con asimmetria positiva
- Media e deviazione standard dipendono dai gdl

$$\text{Media} = gld \qquad \text{Deviazione standard} = \sqrt{2 \cdot gdl}$$

Test χ^2 di indipendenza

- Fissare il livello di significatività del test, α, ossia la probabilità di commettere l'errore di prima specie
- Valore critico $\chi^2_{(r-1)\cdot(c-1)}(\alpha) : Pr\left(\chi^2_{(r-1)\cdot(c-1)} > \chi^2_{(r-1)\cdot(c-1)}(\alpha)\right) = \alpha$
- Regione critica

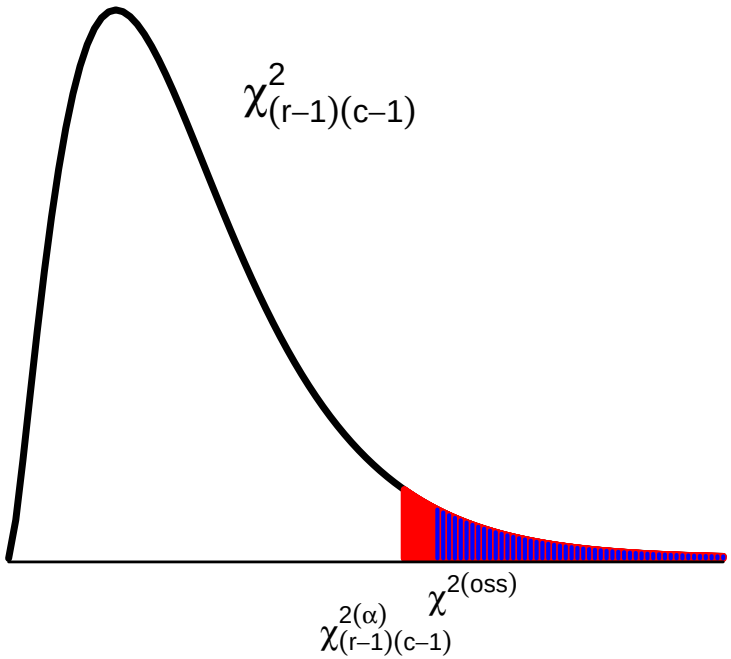
$$RC_\alpha : \chi^2 > \chi^2_{(r-1)\cdot(c-1)}(\alpha)$$

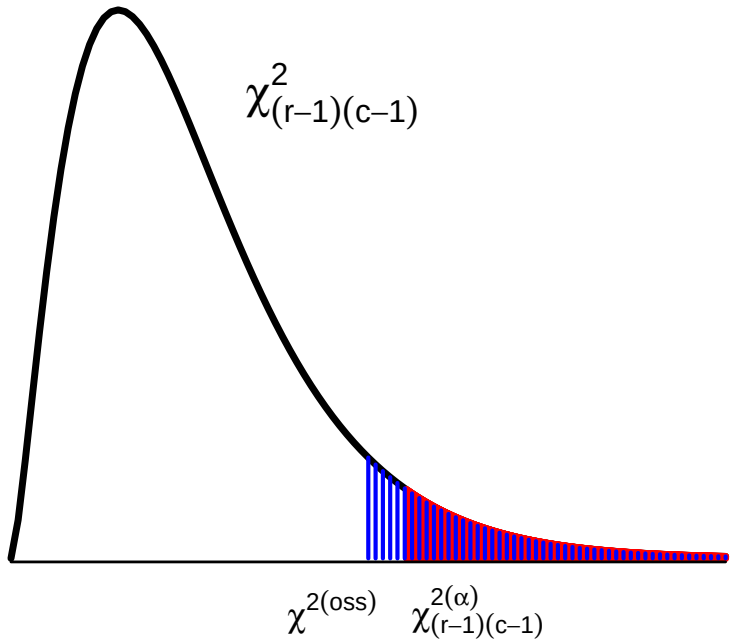


Test χ^2 di indipendenza: P–valore

p–valore = probabilità (sotto H_0) di osservare un valore della statistica test 'più estremo' del valore osservato

$$p - value = Pr\left(\chi^2_{(r-1)\cdot(c-1)} > \chi^{2,oss}; H_0\right)$$





Test di indipendenza: Riepilogo

1. Definire il sistema di ipotesi

$$\begin{aligned} H_0: &\quad \text{Le variabili sono statisticamente indipendenti} \\ H_a: &\quad \text{Le variabili Non sono statisticamente indipendenti} \end{aligned}$$

2. Fissare il livello di significatività del test, α
3. Statistica test

$$\chi^2 = \sum \frac{(\text{frequenze osservate - frequenze attese})^2}{\text{Frequenze attese}} = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\Updownarrow$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Sotto H_0, la statistica test χ^2 ha una distribuzione asintotica $\chi^2_{(r-1)\cdot(c-1)}$

4. Calcolare il valore osservato della statistica test: $\chi^{2,oss}$
5. Valore critico $\chi^2_{(r-1)\cdot(c-1)}(\alpha) \longrightarrow$ Regione critica $RC_\alpha : \chi^2 \geq \chi^2_{(r-1)\cdot(c-1)}(\alpha)$
6. Utilizzare il p-value

Esempio: Il nettare di agave pastorizzato per la cura della tosse

Distribuzioni di frequenze congiunte e marginali

	La tosse disturba il sonno				
Agave	Per niente	Poco	Abbastanza	Molto	Totale
Si	280 (32.2%)	140 (16.1%)	20 (2.3%)	40 (4.6%)	480 (55.2%)
No	120 (13.8%)	220 (25.3%)	10 (1.1%)	40 (4.6%)	390 (44.8%)
Totale	400 (46.0%)	360 (41.4%)	30 (3.4%)	80 (9.2%)	870 (100%)

- Ipotesi

  H_0 : Livello di disturbo del sonno dovuto alla tosse e uso del nettare di Agave sono statisticamente indipendenti

  H_a : Livello di disturbo del sonno dovuto alla tosse e uso del nettare di Agave Non sono statisticamente indipendenti

- Livello di significatività del test, $\alpha = 0.05$
- Statistica test: Statistica test χ^2 di Pearson

Statistica test χ^2 di Pearson: Esempio

Frequenze osservate - frequenze teoriche
(Frequenze osservate - frequenze teoriche)^2

	La tosse disturba il sonno			
Agave	Per niente	Poco	Abbastanza	Molto
Si	280 – 220.7 = 59.3	140 – 198.6 = –58.6	20 – 16.6 = 3.4	40 – 44.1 = –4.1
	(3517.7)	(3436.4)	(11.9)	(17.1)
No	120 – 179.3 = –59.3	220 – 161.4 = 58.6	10 – 13.4 = –3.4	40 – 35.9 = 4.1
	(3517.7)	(3436.4)	(11.9)	(17.1)

$$\begin{aligned}\chi^{2,oss} &= \frac{3517.7}{220.7} + \frac{3436.4}{198.6} + \frac{11.9}{16.6} + \frac{17.1}{44.1} + \frac{3517.7}{179.3} + \frac{3436.4}{161.4} + \frac{11.9}{13.4} + \frac{17.1}{35.9} \\ &= 15.940 + 17.301 + 0.718 + 0.388 + 19.618 + 21.294 + 0.884 + 0.477 \\ &= 76.62\end{aligned}$$

- Valore critico $\chi^2_{(2-1)\cdot(4-1)}(0.05) = 7.81 \longrightarrow$ Regione critica $R_\alpha : \chi^2 \geq 7.81$
- Utilizzare il *p*-value

$$p - value = Pr\left(\chi^2_{(2-1)\cdot(4-1)} > 76.62; H_0\right) = 0.0000$$

- $\chi^{2,oss} = 76.62 > 7.81 \Longrightarrow$ I dati mostrano evidenza contro l'ipotesi nulla di indipendenza al livello di significatività de 5%

Osservazioni

- L'approssimazione della distribuzione campionaria della statistica test χ^2 è adeguata per campioni di dimensioni elevata ($f_e = \hat{n}_{ij} \geq 5$ per ogni i, j)
- La statistica χ^2 considera le variabili come variabili misurate su scala nominale (riordinando le categorie si ottengono gli stessi risultati).

Test χ^2 di indipendenza per tabelle 2×2

	Y		
X	0	1	*Totale*
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
Totale	$n_{.1}$	$n_{.2}$	n

Distribuzioni di frequenza relative condizionate di Y dato X

	Y		
X	0	1	*Totale*
1	$f_{1\|1} = \frac{n_{11}}{n_{1.}}$	$f_{2\|1} = \frac{n_{12}}{n_{1.}}$	1
2	$f_{1\|2} = \frac{n_{21}}{n_{2.}}$	$f_{2\|2} = \frac{n_{22}}{n_{2.}}$	1

	Y		
X	0	1	*Totale*
1	$1 - \pi_1$	π_1	1
2	$1 - \pi_2$	π_2	1

$$\pi_1 = \frac{\sum_{i=1}^{n_1} Y_{i1}}{n_1} = \frac{n_{12}}{n_{1.}} \qquad \pi_2 = \frac{\sum_{i=1}^{n_2} Y_{i2}}{n_2} = \frac{n_{22}}{n_{2.}}$$

Test χ^2 di indipendenza per tabelle 2×2

$$\begin{array}{l} H_0 : X \text{ e } Y \text{ indipendenti} \\ H_a : X \text{ e } Y \text{ dipendenti} \end{array} \iff \begin{array}{l} H_0 : \pi_1 = \pi_2 \\ H_a : \pi_1 \neq \pi_2 \end{array}$$

- Test χ^2 di indipendenza $\longrightarrow$ Statistica χ^2

$$\chi^2 = \frac{(n_{11} - \hat{n}_{11})^2}{\hat{n}_{11}} + \frac{(n_{12} - \hat{n}_{12})^2}{\hat{n}_{12}} + \frac{(n_{21} - \hat{n}_{21})^2}{\hat{n}_{21}} + \frac{(n_{22} - \hat{n}_{22})^2}{\hat{n}_{22}}$$

  Per campioni di dimensione elevata, sotto H_0, $\chi^2 \approx \chi^2_1$

- Statistica test per il confronto tra proporzioni

$$Z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{\sqrt{\hat{\pi}(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  dove $\hat{\pi}_1 = \frac{n_{12}}{n_{1.}}$, $\hat{\pi}_2 = \frac{n_{22}}{n_{2.}}$ e $\hat{\pi} = \frac{n_{.2}}{n}$

  Per campioni di dimensione elevata, sotto H_0, $Z \approx N(0, 1)$

- $\chi^2 = Z^2$

Test χ^2 di indipendenza per tabelle 2×2

- Livello di significatività: α
- Regione di rifiuto – Test χ^2 di indipendenza

$$RC_\alpha = \chi^2 \geq \chi^2_1(\alpha)$$

- Regione di rifiuto – Test per il confronto tra proporzioni

$$RC_\alpha = z \leq -z_{\alpha/2} \text{ oppure } z \geq z_{\alpha/2}$$

- Quindi ...

$$RC_\alpha = \chi^2 \geq \chi^2_1(\alpha) = z^2 \geq z^2_{\alpha/2}$$

- p–valore:

$$p - value = 2 \cdot [1 - Pr(Z \leq |z^{oss}|; H_0)] = Pr\left(\chi^2_{(r-1)\cdot(c-1)} \geq \chi^{2,oss}; H_0\right)$$

Test χ^2 di indipendenza per tabelle 2×2: Esempio

- Obiettivo: Confrontare il tasso di occupazione di giovani svantaggiati che hanno seguito un corso di formazione professionale (π_2) con il tasso di occupazione di giovani svantaggiati che non hanno seguito un corso di formazione professionale (π_1)
- Osservazioni campionarie

Corso di formazione	Numerosità campionaria	Numero occupati	Proporzione campionaria
No	$n_1 = 80$	60	$\hat{\pi}_1 = \frac{60}{80} = 0.75$
Si	$n_2 = 120$	93	$\hat{\pi}_2 = \frac{93}{120} = 0.775$

Test χ^2 di indipendenza per tabelle 2×2: Esempio

Corso di formazione	*Status occupazionale* 0	1	*Totale*
No	20	60	80
Si	27	93	120
Totale	47	153	200

$$\hat{\pi}_1 = \frac{\sum_{i=1}^{n_1} y_{i1}}{n_1} = \frac{n_{12}}{n_{1.}} = \frac{60}{80} = 0.75 \qquad \hat{\pi}_2 = \frac{\sum_{i=1}^{n_2} y_{i2}}{n_2} = \frac{n_{22}}{n_{2.}} = \frac{93}{120} = 0.775$$

$$\text{Stima pooled: } \hat{\pi} = \frac{80 \cdot 0.75 + 120 \cdot 0.775}{80 + 120} = \frac{60 + 93}{80 + 120} = \frac{153}{200} = 0.765$$

Test χ^2 di indipendenza per tabelle 2×2: Esempio

- Valore osservato della statistica test

$$
\begin{aligned}
z &= \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.775 - 0.75}{\sqrt{0.765 \cdot (1 - 0.765)\left(\frac{1}{80} + \frac{1}{120}\right)}} \\
&= \frac{0.025}{\sqrt{0.1798(0.0125 + 0.0083)}} = \frac{0.025}{\sqrt{0.0037}} = \frac{0.025}{0.0612} = 0.4085
\end{aligned}
$$

- Valore osservato della statistica χ^2

$$
\begin{aligned}
\chi^2 &= \frac{(20 - 18.8)^2}{18.8} + \frac{(60 - 61.2)^2}{61.2} + \frac{(27 - 28.2)^2}{28.2} + \frac{(93 - 91.8)^2}{91.8} \\
&= \frac{1.44}{18.8} + \frac{1.44}{61.2} + \frac{1.44}{28.2} + \frac{1.44}{91.8} = 0.077 + 0.024 + 0.051 + 0.016 = 0.1669
\end{aligned}
$$

- $z^2 = (0.4085)^2 = 0.1669 = \chi^2$

Test χ^2 di indipendenza per tabelle 2×2

- Livello di significatività: $\alpha = 0.03 \Rightarrow \alpha/2 = 0.015$

$$z_{0.015} = 2.17 \qquad \chi^2_1(0.03) = 4.71 = (2.17)^2 = (z_{0.015})^2$$

- Regione rifiuto:

$$RC_{0.03} = Z \leq -2.17 \text{ oppure } Z \geq 2.17 \Longleftrightarrow Z^2 \geq 4.71 \Longleftrightarrow \chi^2 \geq 4.71$$

- Decisione: I valori osservati delle statistiche test NON appartengono alla regione di rifiuto: I dati NON mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 3%
- $P-$valore

$$p = 2 \cdot [1 - Pr\,(Z \leq |0.4085|; H_0)] = 0.6829 \text{ e } p = Pr\left(\chi^2_1 \geq 0.1669; H_0\right) = 0.6829$$

Misure di associazione in tabelle di contingenza

- La statistica χ^2 non misura l'associazione
- Il test χ^2 di indipendenza da informazioni sulla presenza versus l'assenza di associazione
- Il test χ^2 di indipendenza non da informazioni sulla natura e la forza dell'associazione
- Natura dell'associazione: Analisi dei Residui
  - ✓ Quanto si allontanano le frequenze osservate dalle frequenze teoriche di indipendenza
- Forza dell'associazione: Misure di associazione

Associazione

- Se due caratteri non sono statisticamente indipendenti, allora tra essi esiste una qualche **associazione**: Relazione di **dipendenza** o **interdipendenza** tra i caratteri
- In presenza di associazione alcune modalità di un carattere si presentano più frequentemente in corrispondenza di alcune modalità dell'altro carattere
- Una misura di associazione sintetizza la forza della dipendenza tra due variabili
- Misura di associazione nella popolazione = Parametro
- Misura di associazione campionaria = Statistica

Assenza di associazione: Indipendenza – Esempio

- Associazione tra zona di residenza e status occupazionale

Zona di residenza (X)	Status Occupazionale (Y) Occupato (O)	Disoccupato (D)	Fuori dalla forza lavoro (FFL)	Totale
Nord	6	18	2	26
Centro	3	9	1	13
Sud	9	27	3	39
Isole	15	45	5	65
Totale	33	99	11	143

- Distribuzioni di frequenza relativa condizionata dello status occupazionale data la zona di residenza

Zona di residenza (X)	Status Occupazionale (Y) Occupato (O)	Disoccupato (D)	Fuori dalla forza lavoro (FFL)	Totale
Nord	0.231	0.692	0.077	1.000
Centro	0.231	0.692	0.077	1.000
Sud	0.231	0.692	0.077	1.000
Isole	0.231	0.692	0.077	1.000
Totale	0.231	0.692	0.077	1.000

- *"Zona di residenza" e "Status occupazionale" sono statisticamente indipendenti*

Associazione perfetta: Esempio I

- Relazione tra "Classe retributiva" e "Posizione lavorativa" in una grande azienda

Posizione Lavorativa	Classe retributiva Bassa	Media	Alta	Totale
Dirigente	0	0	10	10
Funzionario	0	18	0	18
Impiegato	0	27	0	27
Operaio	105	0	0	105
Totale	105	45	10	160

- *La "classe retributiva" dipende perfettamente dalla "Posizione lavorativa"*

Associazione perfetta: Esempio II

- Relazione tra "Interesse nella lettura" e "Interesse nel cinema"

Interesse nella lettura	Interesse nel cinema			
	Basso	Medio	Alto	Totale
Basso	0	21	0	21
Medio	18	0	0	18
Alto	0	0	53	53
Totale	18	21	53	92

- *"Interesse nella lettura" e "Interesse nel cinema" sono perfettamente interdipendenti*

Misure di associazione

- Ogni situazione intermedia tra l'indipendenza e l'associazione perfetta esprime una certo grado di dipendenza o interdipendenza tra i caratteri
- Il livello di associazione sarà tanto maggiore quanto più la tabella osservata si discosta da quella di indipendenza a favore di una situazione di perfetta associazione
- Nello studio dell'interdipendenza si utilizzano indici basati su un approccio **simmetrico** rispetto al modo di trattare i due caratteri: Tali indici sono in genere calcolati utilizzando la distribuzione doppia di frequenze
- Nello studio della dipendenza si utilizzano indici basati su un approccio **asimmetrico** rispetto al modo di trattare i due caratteri: Tali indici sono in genere calcolati utilizzando la distribuzione condizionata di frequenze del carattere risposta al carattere esplicativo

Misure di associazione in tabelle 2×2

- Differenza tra proporzioni: $\pi_2 - \pi_1$
- Rapporto tra proporzioni (Rischio relativo): $RR = \dfrac{\pi_2}{\pi_1}$
- Rapporto delle quote (odds-ratio)

$$OR = \frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)}$$

Misure di associazione in tabelle 2×2 – Esempi

- Esempio I: Relazione tra frequenza di un corso di formazione e Status occupazionale

Corso di formazione	*Status occupazionale* 0	1	*Totale*
No	20	60	80
Si	27	93	120
Totale	47	153	200

$$\hat{\pi}_1 = \frac{\sum_{i=1}^{n_1} y_{i1}}{n_1} = \frac{n_{12}}{n_{1.}} = \frac{60}{80} = 0.75$$

$$\hat{\pi}_2 = \frac{\sum_{i=1}^{n_2} y_{i2}}{n_2} = \frac{n_{12}}{n_{2.}} = \frac{93}{120} = 0.775$$

- Esempio II: Relazione tra Infarto e Aspirina

Gruppo	*Infarto* 0. *No*	1. *Si*	*Totale*
Placebo	10829	221	11050
Aspirina	13020	105	13125
Totale	23849	326	24175

$$\hat{\pi}_1 = \frac{\sum_{i=1}^{n_1} y_{i1}}{n_1} = \frac{n_{12}}{n_{1.}} = \frac{221}{11050} = 0.02$$

$$\hat{\pi}_2 = \frac{\sum_{i=1}^{n_2} y_{i2}}{n_2} = \frac{n_{12}}{n_{2.}} = \frac{105}{13125} = 0.008$$

- Esempio III: Relazione tra lo Status di fumatore e cancro ai polmoni

Fumatore	*Cancro* 0. *Si*	1. *No*	*Totale*
No	1300	1200	2500
Si	1596	1404	3000
Totale	2896	2604	5500

$$\hat{\pi}_1 = \frac{\sum_{i=1}^{n_1} y_{i1}}{n_1} = \frac{n_{12}}{n_{1.}} = \frac{1300}{2500} = 0.52$$

$$\hat{\pi}_2 = \frac{\sum_{i=1}^{n_2} y_{i2}}{n_2} = \frac{n_{12}}{n_{2.}} = \frac{1596}{3000} = 0.532$$

Misure di associazione in tabelle 2×2: Differenza tra proporzioni

- $-1 \leq \pi_2 - \pi_1 \leq 1$
- $\pi_1 = \pi_2 \iff \pi_2 - \pi_1 = 0 \iff X$ e Y sono indipendenti
- Esempi

$$\begin{aligned}
\hat{\pi}_1 = 0.75 \quad e \quad \hat{\pi}_2 = 0.775 &\implies \hat{\pi}_2 - \hat{\pi}_1 = 0.775 - 0.75 = 0.025 \\
\hat{\pi}_1 = 0.02 \quad e \quad \hat{\pi}_2 = 0.008 &\implies \hat{\pi}_2 - \hat{\pi}_1 = 0.008 - 0.02 = -0.012 \\
\hat{\pi}_1 = 0.52 \quad e \quad \hat{\pi}_2 = 0.532 &\implies \hat{\pi}_2 - \hat{\pi}_1 = 0.532 - 0.52 = 0.012
\end{aligned}$$

- Test e intervalli di confidenza per la differenza tra proporzioni

Misure di associazione in tabelle 2×2: Rischio Relativo

- $0 \leq \dfrac{\pi_2}{\pi_1} < +\infty$
- $\pi_1 = \pi_2 \iff \dfrac{\pi_2}{\pi_1} = 1 \iff X$ e Y sono indipendenti
- Una stessa differenza di probabilità dà luogo a rischi relativi molto più elevati quando le probabilità sono vicine a 0, piuttosto che quando sono vicine a 0.5
  - ✓ Esempi

$$\hat{\pi}_1 = 0.75 \quad e \quad \hat{\pi}_2 = 0.775 \quad \Longrightarrow \quad \hat{RR} = \frac{\hat{\pi}_2}{\hat{\pi}_1} = \frac{0.775}{0.75} = 1.033$$

$$\hat{\pi}_1 = 0.02 \quad e \quad \hat{\pi}_2 = 0.008 \quad \Longrightarrow \quad \hat{RR} = \frac{\hat{\pi}_2}{\hat{\pi}_1} = \frac{0.008}{0.02} = 2.5$$

$$(\hat{\pi}_2 - \hat{\pi}_1 = -0.012)$$

$$\hat{\pi}_1 = 0.52 \quad e \quad \hat{\pi}_2 = 0.532 \quad \Longrightarrow \quad \hat{RR} = \frac{\hat{\pi}_2}{\hat{\pi}_1} = \frac{0.532}{0.52} = 1.023$$

$$(\hat{\pi}_2 - \hat{\pi}_1 = 0.012)$$

- Scambiando i due gruppi si ottengono dei valori reciproci del rischio relativo
  - ✓ Esempio: $\pi_1 = 0.2$ e $\pi_2 = 0.4$

$$RR = \frac{\pi_2}{\pi_1} = \frac{0.4}{0.2} = 2 \qquad \textit{versus} \qquad RR = \frac{\pi_1}{\pi_2} = \frac{0.2}{0.4} = 0.5 = \frac{1}{2}$$

Odds

$$odds = \frac{\text{Probabilità di successo}}{\text{Probabilità di insuccesso}} = \frac{\text{Probabilità di successo}}{1 - \text{Probabilità di successo}} = \frac{\pi}{1-\pi}$$

- $0 \leq odds < +\infty$
- Esempio I

  $$\pi = 0.75 \Longrightarrow odds = \frac{\pi}{1-\pi} = \frac{0.75}{1-0.75} = \frac{0.75}{0.25} = 3$$

  Un successo è circa 3 volte più probabile di un insuccesso: 3 successi ogni insuccesso (1 insuccesso ogni 3 successi)
- Esempio II

  $$\pi = 0.2 \Longrightarrow odds = \frac{\pi}{1-\pi} = \frac{0.2}{1-0.2} = \frac{0.2}{0.8} = \frac{1}{4} = 0.25$$

  Un insuccesso è circa 4 volte più probabile di un successo: 1 successo ogni 4 insuccessi
- Esempio III

  $$\pi = 0.5 \Longrightarrow odds = \frac{\pi}{1-\pi} = \frac{0.5}{1-0.5} = \frac{0.5}{0.5} = 1$$

  La probalità di successo è uguale alla probabilità di insuccesso

Odds e probabilità di successo

$$\text{Probabilità di successo} = \frac{\text{Odds}}{\text{odds} + 1}$$

- Esempio I

$$odds = 3 \Longrightarrow \pi = \frac{odds}{odds + 1} = \frac{3}{3 + 1} = \frac{3}{4} = 0.75$$

- Esempio II

$$odds = 0.25 \Longrightarrow \pi = \frac{odds}{odds + 1} = \frac{0.25}{0.25 + 1} = \frac{0.25}{1.25} = 0.2$$

- Esempio III

$$odds = 1 \Longrightarrow \pi = \frac{odds}{odds + 1} = \frac{1}{1 + 1} = \frac{1}{2} = 0.5$$

Odds in tabelle 2×2: Esempio I

Relazione tra frequenza di un corso di formazione e status occupazionale

Corso di formazione	*Status occupazionale* 0	1	*Totale*
No	20	60	80
Si	27	93	120
Totale	47	153	200

$$\hat{\pi}_1 = \frac{\sum_{i=1}^{n_1} y_{i1}}{n_1} = \frac{n_{12}}{n_{1.}} = \frac{60}{80} = 0.75$$

$$\hat{\pi}_2 = \frac{\sum_{i=1}^{n_2} y_{i2}}{n_2} = \frac{n_{12}}{n_{2.}} = \frac{93}{120} = 0.775$$

- Odds per coloro che non hanno seguito il corso di formazione

$$\widehat{odds}_1 = \frac{\hat{\pi}_1}{1 - \hat{\pi}_1} = \frac{0.75}{1 - 0.75} = \frac{0.75}{0.25} = 3$$

- Odds per coloro che hanno seguito il corso di formazione

$$\widehat{odds}_2 = \frac{\hat{\pi}_2}{1 - \hat{\pi}_2} = \frac{0.775}{1 - 0.775} = \frac{0.775}{0.225} = 3.44$$

Misure di associazione in tabelle 2×2: Rapporto degli odds

$$OR = \frac{Odds_2}{Odds_1} = \frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)}$$

- $0 \leq OR < +\infty$
- $\pi_1 = \pi_2 \iff OR = 1 \iff X$ e Y sono indipendenti
- Esempi

$$\hat{\pi}_1 = 0.75 \quad e \quad \hat{\pi}_2 = 0.775 \implies \widehat{OR} = \frac{0.775/(1-0.775)}{0.75/(1-0.75)} = \frac{3.44}{3} = 1.15$$

$$\hat{\pi}_1 = 0.02 \quad e \quad \hat{\pi}_2 = 0.008 \implies \widehat{OR} = \frac{0.008/(1-0.008)}{0.02/(1-0.02)} = \frac{0.0081}{0.0204} = 0.40$$

$$\hat{\pi}_1 = 0.52 \quad e \quad \hat{\pi}_2 = 0.532 \implies \widehat{OR} = \frac{0.532/(1-0.532)}{0.52/(1-0.52)} = \frac{1.14}{1.08} = 1.05$$

Proprietà dell'odds ratio

- $0 \leq OR < +\infty$
- $\pi_1 = \pi_2 \Longleftrightarrow OR = 1$
- $OR > 1 \Longrightarrow$ gli odds di successo nella popolazione 2 sono *maggiori* degli odds di successo nella popolazione 1
  - ✓ $OR > 1 \Longrightarrow$ gli odds di successo nella popolazione 1 sono *minori* degli odds di successo nella popolazione 2
- $OR < 1 \Longrightarrow$ gli odds di successo nella popolazione 1 sono *minori* degli odds di successo nella popolazione 2
  - ✓ $OR < 1 \Longrightarrow$ gli odds di successo nella popolazione 1 sono *maggiori* degli odds di successo nella popolazione 2
- Maggiore è la distanza del valore dell'odds ratio da 1 in una direzione, più forte è l'associazione
- Due valori dell'odds ratio che rappresentano la stessa associazione ma in direzioni opposte sono uno il reciproco dell'altro

$$OR = 4 \qquad \textit{versus} \qquad OR = \frac{1}{4} = 0.25$$

Proprietà dell'odds ratio

- L'odds ratio è una misura di associazione simmetrica: il valore dell'odds ratio è lo stesso indipendentemente da quale variabile è scelta come variabile risposta

Borsa di Studio	Abbandono degli studi Si	No	*Totale*
Si	150	850	1000
No	200	800	1000
Totale	350	1650	2000

Borsa di Studio	Abbandono degli studi Si	No	*Totale*
Si	0.15	0.85	1
No	0.20	0.80	1
Totale	0.175	0.825	1

$$OR = \frac{0.15/0.85}{0.20/0.80} = \frac{0.18}{0.25} = 0.706$$

Borsa di studio	Abbandono degli studi Si	No	*Totale*
Si	0.43	0.52	0.5
No	0.57	0.48	0.5
Totale	1	1	1

$$OR = \frac{0.43/0.57}{0.52/0.48} = \frac{0.75}{1.0625} = 0.706$$

Proprietà dell'odds ratio

- Odds ratio = Prodotto incrociato

$$OR = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

✓ Esempio

Borsa di studio	Abbandono degli studi: Si	Abbandono degli studi: No	*Totale*
Si	150	850	1000
No	200	800	1000
Totale	350	1650	2000

$$OR = \frac{150 \cdot 800}{200 \cdot 850} = \frac{120000}{170000} = 0.706$$

Proprietà dell'odds ratio

- Invarianza del rapporto degli odds. Se si scambiano i gruppi l'odds ratio diventa il reciproco

Borsa di studio	Abbandono degli studi: Si	No	*Totale*
Si	150	850	1000
No	200	800	1000
Totale	350	1650	2000

$\Longleftrightarrow$

Borsa di studio	Abbandono degli studi: Si	No	*Totale*
No	200	800	1000
Si	150	850	1000
Totale	350	1650	2000

$$OR = \frac{150 \cdot 800}{200 \cdot 850} = \frac{120000}{170000} = 0.706 \qquad OR = \frac{200 \cdot 850}{150 \cdot 800} = \frac{170000}{120000} = 1.42 = \frac{1}{0.706}$$

Proprietà dell'odds ratio

- Invarianza del rapporto degli odds: Se si cambiano sia i gruppi che le categorie della risposta l'odds ratio non varia

Borsa di studio	Abbandono degli studi Si	No	*Totale*
Si	150	850	1000
No	200	800	1000
Totale	350	1650	2000

$\Longleftrightarrow$

Borsa di studio	Abbandono degli studi No	Si	*Totale*
No	800	200	1000
Si	850	150	1000
Totale	1650	350	2000

$$OR = \frac{150 \cdot 800}{200 \cdot 850} = \frac{120000}{170000} = 0.706 \qquad OR = \frac{800 \cdot 150}{850 \cdot 200} = \frac{120000}{170000} = 0.706$$

Confronto tra differenza tra le proporzioni, rischio relativo, odds-ratio

- Il rischio relativo e l'odds-ratio sono misure relative mentre le differenza delle probabilità che è una misura espressa in termini assoluti
- Il rischio relativo è il più facile da interpretare dell'rischio relativo
  - ✓ Se $RR = 4$ allora la probabilità di successo in un gruppo è 4 volte la probabilità nell'altro. Se $OR = 4$ l'affermazione precedente non è vera, ma è vero gli odds in un gruppo sono 4 volte più grandi degli odds nell'altro.
- L'interpretazione del rischio relativo cambia se si cambia la scelta della classe dei successi, mentre l'interpretazione del l'odds-ratio non cambia se si cambia la scelta della classe dei successi
  - ✓ Esempio: $\pi_1 = 0.2$ e $\pi_2 = 0.4$

$$RR = \frac{\pi_2}{\pi_1} = \frac{0.4}{0.2} = 2 \qquad vs \qquad RR = \frac{1-\pi_2}{1-\pi_1} = \frac{0.6}{0.8} = 0.75$$

Mentre

$$OR = \frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} = \frac{0.67}{0.25} = 2.67 \quad vs \quad OR = \frac{(1-\pi_2)/\pi_2}{(1-\pi_1)/\pi_1} = \frac{1.5}{4} = 0.375 = \frac{1}{2.67}$$

Confronto tra differenza tra le proporzioni, rischio relativo, odds-ratio

- **Uguaglianza approssimata tra rischio relativo e odds-ratio**: Se le due probabilità π_1 e π_2 sono vicine a zero l'odds-ratio è approssimativamente uguale al rischio relativo

  ✓ Esempio: $\pi_1 = 0.002$ e $\pi_2 = 0.004$

$$RR = \frac{\pi_2}{\pi_1} = \frac{0.004}{0.002} = 2$$

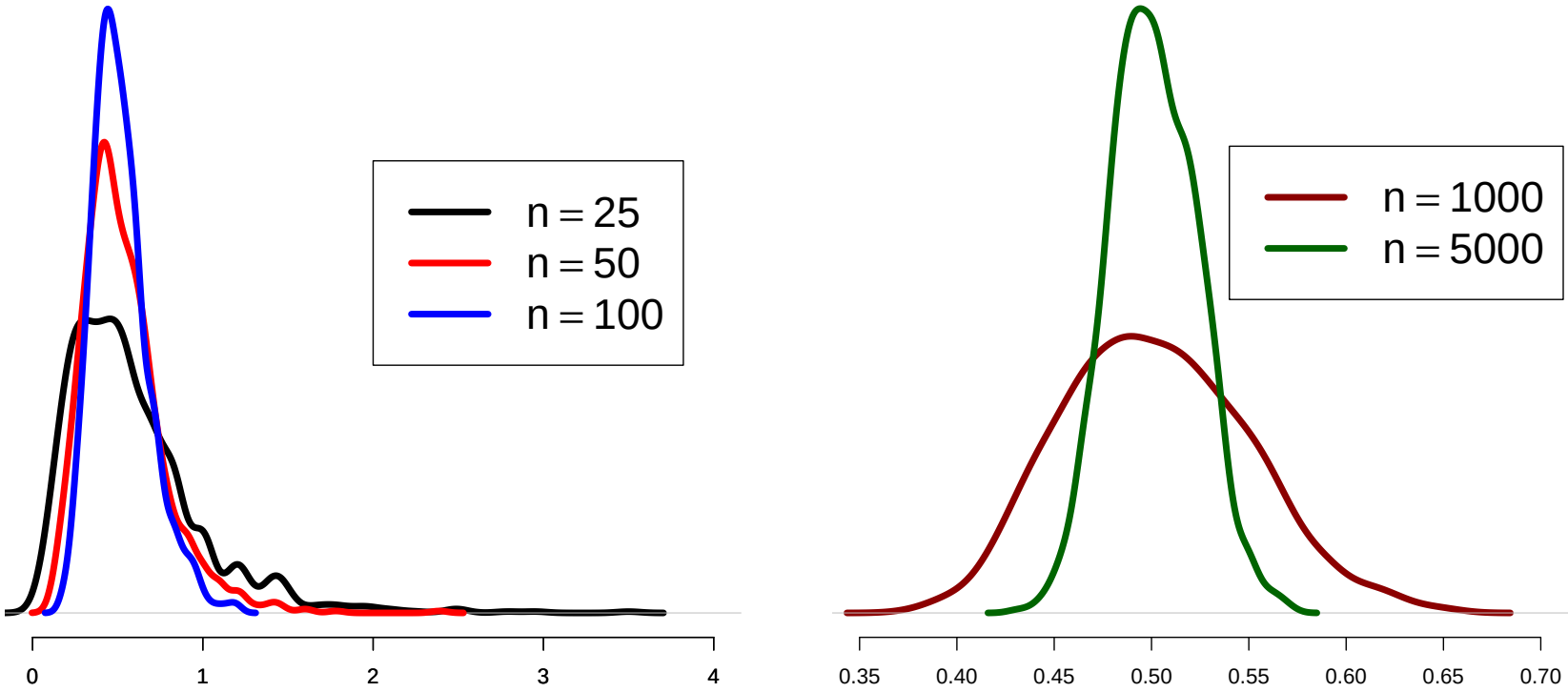
e

$$OR = \frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} = \frac{0.004016}{0.002004} = 2.004$$

Quindi $OR \approx RR$

Distribuzione campionaria dello stimatore dell'odds ratio

$Y_1 \sim Ber(\pi_1 = 0.6)$ $Y_2 \sim Ber(\pi_2 = 0.75)$ $n_1 = n_2 = n$



- La distribuzione campionaria dello stimatore dell'odds ratio è fortemente asimmatrica
- Per dimensioni campionarie molto grandi la distribuzione campionaria dello stimatore dell'odds ratio è approssimativamente normale

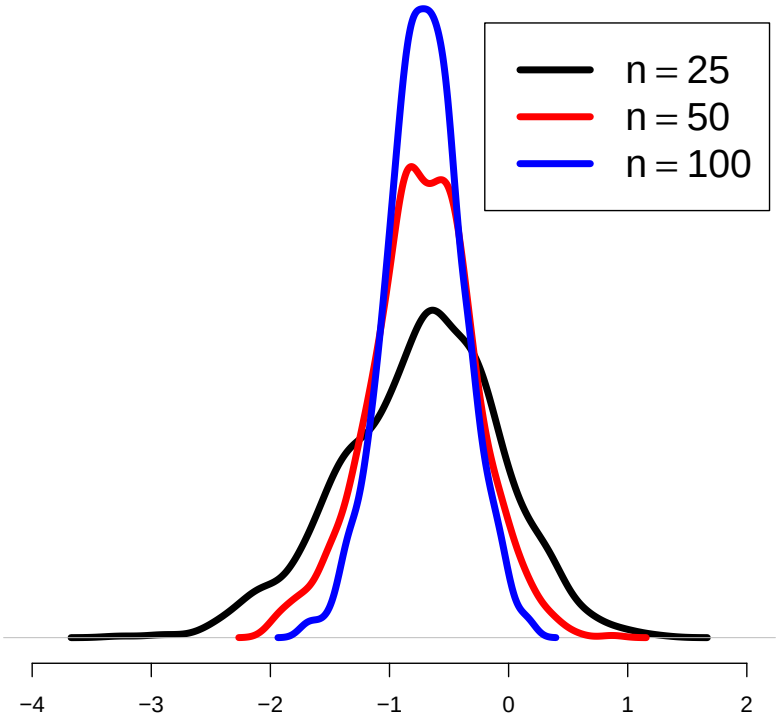
Distribuzione campionaria dello stimatore del logaritmo dell'odds ratio

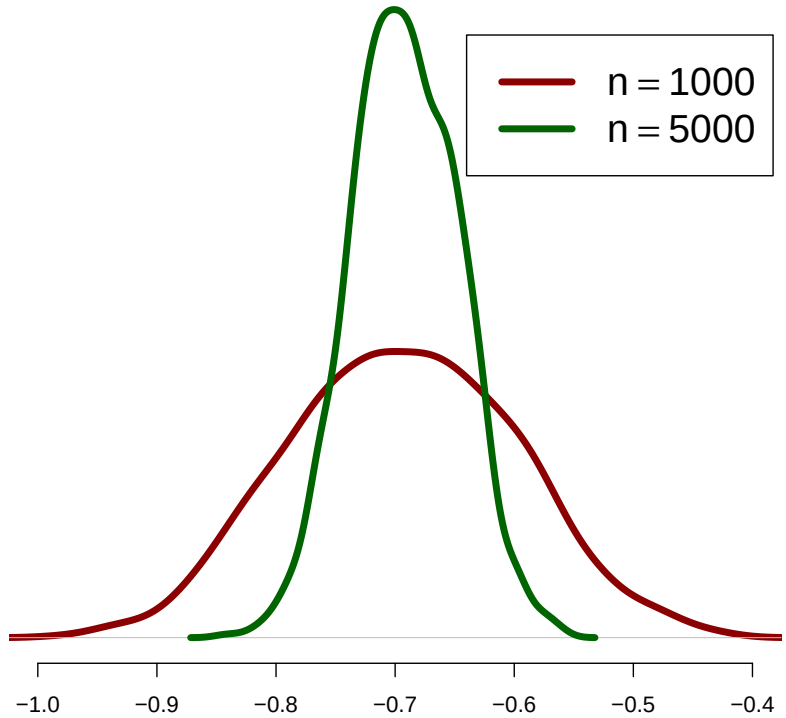
- Logaritmo del rapporto delle quote

$$\log(OR) = \log\left(\frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)}\right) = \log\left(\frac{\pi_2}{1-\pi_2}\right) - \log\left(\frac{\pi_1}{1-\pi_1}\right)$$

- Distribuzione campionaria dello stimatore del logaritmo dell'odds ratio

$$Y_1 \sim Ber(\pi_1 = 0.6) \qquad Y_2 \sim Ber(\pi_2 = 0.75) \qquad n_1 = n_2 = n$$





Distribuzione campionaria dello stimatore del logaritmo dell'odds ratio

	Y	
X	0	1
1	n_{11}	n_{12}
2	n_{21}	n_{22}

Per dimensioni campionarie sufficientemente grandi

$$\log(\hat{OR}) \approx N\left(\log(OR), \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)\right)$$

Intervallo di confidenza per l'odds ratio
Campioni di grandi dimensioni

- Intervallo di confidenza al livello di confidenza $1-\alpha$ per il logaritmo dell'odds ratio

$$IC_{1-\alpha}(log(OR)) = \left[\log(\widehat{OR}) - z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}; \right.$$
$$\left. \log(\widehat{OR}) + z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} \right]$$

- Intervallo di confidenza al livello di confidenza $1-\alpha$ per l'odds ratio

$$IC_{1-\alpha}(OR) = \left[\exp\left\{ \log(\widehat{OR}) - z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} \right\}; \right.$$
$$\left. \exp\left\{ \log(\widehat{OR}) + z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} \right\} \right]$$

Intervallo di confidenza per l'odds ratio: Esempio I

Relazione tra frequenza di un corso di formazione e status occupazionale

Corso di formazione	*Status occup.* 0	1	*Totale*
No	20	60	80
Si	27	93	120
Totale	47	153	200

$$\hat{\pi}_1 = \frac{60}{80} = 0.75 \qquad \widehat{odds}_1 = \frac{0.75}{1-0.75} = 3$$

$$\hat{\pi}_2 = \frac{93}{120} = 0.775 \qquad \widehat{odds}_2 = \frac{0.775}{1-0.775} = 3.44$$

- Odds -ratio e logaritmo dell'odds ratio

$$\widehat{OR} = \frac{3.44}{3} = 1.15 \qquad \log(\widehat{OR}) = \log(1.15) = 0.14$$

- Errore standard

$$\begin{aligned} \hat{es}.(\log(\widehat{OR})) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{20} + \frac{1}{60} + \frac{1}{27} + \frac{1}{93}} \\ &= \sqrt{0.05 + 0.017 + 0.037 + 0.0108} = \sqrt{0.1145} = 0.338 \end{aligned}$$

Intervallo di confidenza per l'odds ratio: Esempio I

Relazione tra frequenza di un corso di formazione e status occupazionale

- Intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$ per il logaritmo dell'odds ratio

$$\begin{aligned} IC_{0.95}(\log(OR)) &= [\log(1.15) - 1.96 \cdot 0.338; \log(1.15) + 1.96 \cdot 0.338] \\ &= [0.14 - 0.663; 0.14 + 0.663] = [-0.525; 0.801] \end{aligned}$$

- Intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$ per l'odds ratio

$$IC_{0.95}(OR) = [\exp\{-0.525\}, \exp\{0.801\}] = [0.592; 2.228]$$