

Statistica

Analisi dell'associazione tra due caratteri

Maura Mezzetti

maura.mezzetti@uniroma2.it

Associazione

- Condurre un'analisi bivariata significa tener conto contemporaneamente di due caratteri per ciascuna unità studiata.
- Esiste associazione tra due variabili se è più probabile che un determinato valore della prima variabile si verifichi in corrispondenza di determinati valori della seconda variabile
- Esiste associazione tra due caratteri X e Y se la modalità x_j di uno dei due caratteri si osserva più frequentemente insieme alla modalità y_j del carattere Y

Analisi bivariata

- Studio **congiunto** di due caratteri
- Analisi delle **relazioni tra due caratteri** osservati sulle stesse unità statistiche
- Studio del comportamento di **due caratteri considerati congiuntamente**
- Misura del **grado di associazione** tra due caratteri qualitativi, quantitativi e misti
- *il tipo di associazione dipende dalla natura dei caratteri*

Nome della relazione

Tipi di variabili

Associazione

Variabili qualitative

Correlazione/concordanza

Variabili quantitative

Differenza in media

Variabile qualitativa e
variabile quantitativa

Nome della relazione	Indice
Associazione	χ^2 - chiquadro Pearson V – Cramer
Correlazione/concordanza	ρ
Differenza in media	

Nome della relazione

Esempi

Associazione

- Maschi e femmine hanno diverse abitudini al fumo?
- Disoccupazione maschi vs disoccupazione femmine
- Scelta facoltà e scuola di provenienza
- Stato civile e titolo di studio

Correlazione/concordanza

- Reddito vs Consumi
- Reddito familiare vs Voto di laurea
- Prezzo casa e distanza da metropolitana

Differenza in media

- Reddito laureati vs reddito diplomati
- Prezzo casa e presenza metropolitana

Laureati in Corso e Genere:

	Femmine	Maschi	Totale
Laureati in corso	487	505	992
Laureati fuori corso	377	504	881
Totale	864	1009	

C'è una Differenza nella Velocità di Studio?

Laureati in Corso e Genere:

	Femmine	Maschi	Totale
Laureati in corso	487	505	992
Laureati fuori corso	377	504	881
Totale	864	1009	

Tra le femmine, qual è la percentuale delle lauree in corso?

$$487/864=56.37\%$$

Tra i maschi, qual è la percentuale delle lauree in corso?

$$505/1009=50.05\%$$

Laureati in Corso e Genere:

	Femmine	Maschi	Totale
Laureati in corso	487	505	992
Laureati fuori corso	377	504	881
Totale	864	1009	

Tra i laureati in corso, qual è la percentuale di femmine?

$$487/992=49.09\%$$

Tra i laureati FUORI corso, qual è la percentuale di femmine?

$$377/881=42.79\%$$

Analisi Bivariata

L'informazione statistica di base prende la forma di una distribuzione unitaria bivariata (o doppia):

Unità	Modalità di X	Modalità di Y
u_1	x_1	y_1
u_2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
u_n	x_n	y_n

Analisi Bivariata

La distribuzione può essere sintetizzata tramite *una tabella di contingenza* (o di frequenza) a doppia entrata, con I righe e J colonne, che contiene la classificazione incrociata di un numero n di unità costituenti il collettivo secondo le modalità uniche e mutualmente esclusive di due caratteri, X e Y , rispettivamente $x_1, \dots, x_i, \dots, x_J$ e $y_1, \dots, y_j, \dots, y_J$.

Tabella di frequenza doppia

		Valore di Y					totale
		y_1	y_2		y_i	y_J	
Valore di X	x_1	n_{11}	n_{12}		n_{1i}	n_{1J}	$n_{1\cdot}$
	x_2	n_{21}	n_{22}				$n_{2\cdot}$
	...					...	
	x_l	n_{l1}	n_{l1}			n_{lJ}	
totale		$n_{\cdot 1}$	$n_{\cdot 2}$			n_{*J}	n

Tabella di contingenza

- Una tabella di contingenza è una forma di sintesi usata per visualizzare simultaneamente due variabili categoriche. Le sue righe indicano le categorie di una delle due variabili, mentre le sue colonne elencano le categorie dell'altra.
- Ogni cella della tabella contiene il numero di osservazioni che, nel campione, sono da riferirsi a una determinata combinazione di categorie delle due variabili qualitative

Frequenze congiunte e marginali

Le frequenze n_{ij} sono le cd. **frequenze congiunte** e rappresentano il numero (assoluto) di unità che presentano congiuntamente la modalità i -esima del carattere X la modalità j -esima del carattere Y .

L'ultima colonna fornisce la distribuzione *marginale* (di frequenze assolute) del carattere X , mentre l'ultima riga fornisce la distribuzione marginale del carattere Y .

Le **frequenze marginali** sono fornite da:

$$n_{i.} = \sum_{j=1}^K n_{ij}, \quad n_{.j} = \sum_{i=1}^H n_{ij},$$

$$n = \sum_i \sum_j n_{ij} = \sum_i n_{i.} = \sum_j n_{.j}$$

Frequenze condizionate

La j -esima colonna fornisce la distribuzione del carattere X *condizionata* alla modalità y_j del carattere Y . L' i -esima riga rappresenta la distribuzione del carattere Y *condizionata* alla modalità x_i del carattere X :

y_1	$\cdots$	y_j	$\cdots$	y_K	Tot.
n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iK}	$n_{i.}$

Frequenze relative congiunte

Dividendo per il numero totale delle unità si ottiene la tabella delle frequenze relative:

$$f_{ij} = \frac{n_{ij}}{n}, i = 1 \dots, H, j = 1, \dots, K.$$

Questa è tale che la somma dei suoi elementi è pari all'unità:

$$\sum_i \sum_j f_{ij} = 1.$$

Le **frequenze relative marginali** di riga e colonna sono rispettivamente $f_{i.}$ e $f_{.j}$

Relazione tra le frequenze

$$n_{i\bullet} = \sum_{j=1}^J n_{ij} \quad n_{\bullet j} = \sum_{i=1}^I n_{ij}$$

$$n = \sum_{j=1}^J n_{\bullet j} \quad n = \sum_{i=1}^I n_{i\bullet} \quad n = \sum_{i=1}^I \sum_{j=1}^J n_{ij} = \sum_{j=1}^J \sum_{i=1}^I n_{ij}$$

$$f_{i\bullet} = \sum_{j=1}^J f_{ij} \quad f_{\bullet j} = \sum_{i=1}^I f_{ij}$$

$$\sum_{j=1}^J f_{\bullet j} = \sum_{i=1}^I f_{i\bullet} = \sum_{i=1}^I \sum_{j=1}^J f_{ij} = 1$$

$$\sum_{i=1}^I f_{i|j} = \sum_{j=1}^J f_{j|i} = 1$$

Laureati in Corso e Genere: distribuzioni condizionate

- Esiste una differenza nella proporzione di maschi e femmine tra i laureati in corso o laureati fuori corso?
- RISPOSTA: frequenze condizionate

Laureati in Corso e Genere: distribuzioni condizionate

	Femmine	Maschi	Totale
Laureati in corso	487	505	992
Laureati fuori corso	377	504	881
Totale	864	1009	

Tabella Condizionata allo stato di Laurea

	Femmine	Maschi	
Laureati in corso	42.79%	57.21%	100%
Laureati fuori corso	49.09%	50.91%	100%

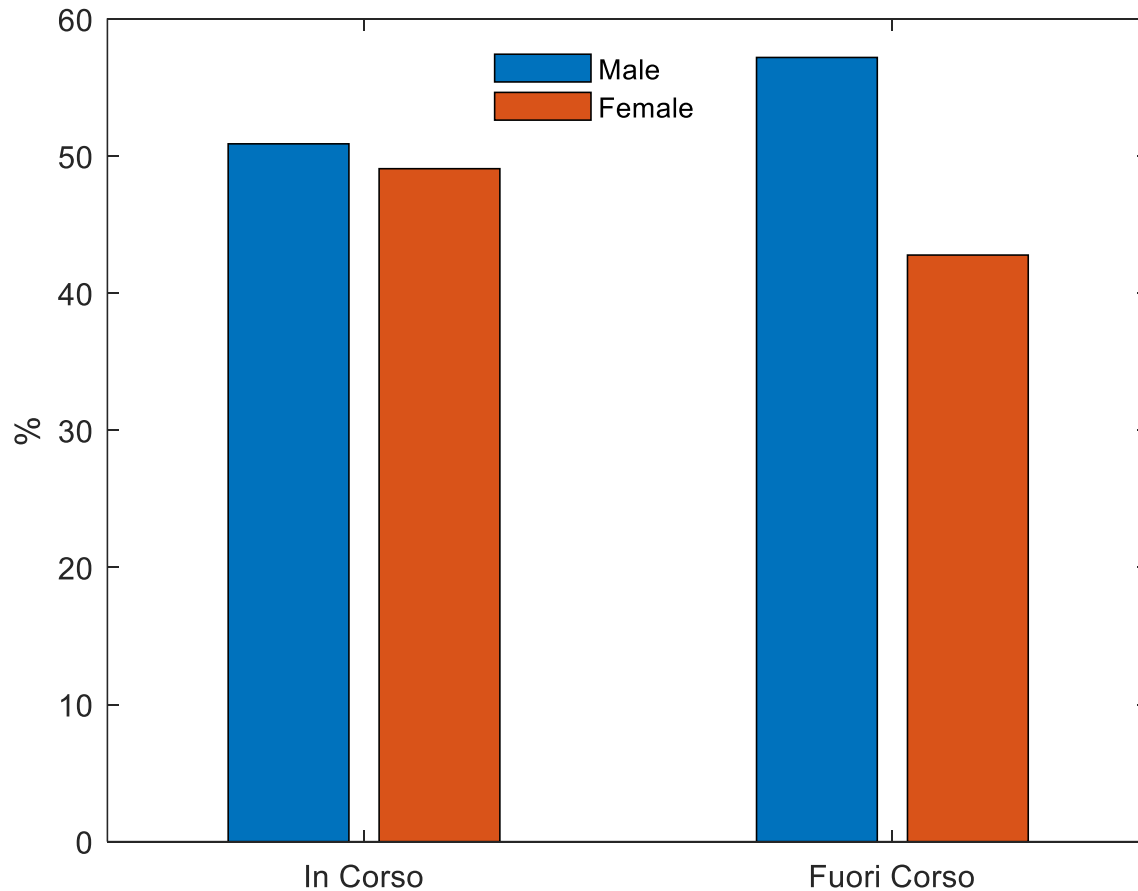
Laureati in Corso e Genere: distribuzioni condizionate

	Femmine	Maschi	Totale
Laureati in corso	487	505	992
Laureati fuori corso	377	504	881
Totale	864	1009	

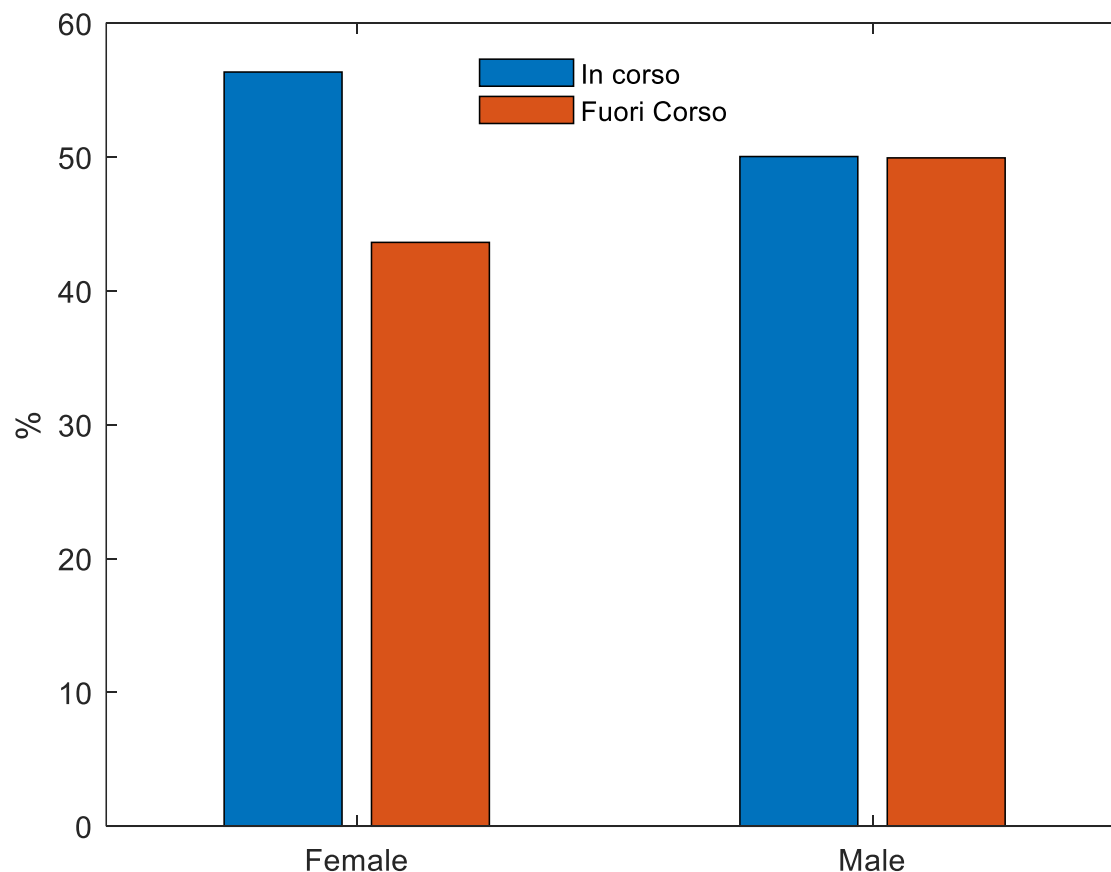
Tabella Condizionata al Genere

	Femmine	Maschi	
Laureati in corso	56.37%	50.05%	
Laureati fuori corso	43.63%	49.95%	
	100%	100%	

Laureati in Corso e Genere: distribuzioni condizionate al tempo di laurea



Laureati in Corso e Genere: distribuzioni condizionate al genere



Esempio: orientamento politico e facoltà

	Democrat	Independent	Republican	Total
Business	440	140	420	1000
Engineering	44	14	42	100
Health Professions	110	35	105	250
Total	594	189	567	1350



Esempio



Relative frequency of:

- Major in Business AND republican voters
- Major in Engineering AND vote democratic
- Persons vote Independent AND choose major in Health Professions
- Major in Business
- Democratic voters
- Democratic voters given Major in Business
- Have a Major in business conditional on vote democratic

Esempio

Relative frequency of:

$$420/1350=0.31$$

$$44/1350=0.033$$

$$35/1350=0.026$$

$$1000/1350=0.74$$

$$594/1350=0.44$$

$$440/1000=0.44$$

$$440/594=0.74$$



Esempio: orientamento politico e facoltà

	Democrat	Independent	Republican	Total
Business	33%	10%	31%	74%
Engineering	3%	1%	3%	7%
Health Professions	8%	3%	8%	19%
Total	44%	14%	42%	100%



	Democrat	Independent	Republican	Total
Business	74%	74%	74%	74%
Engineering	7%	7%	7%	7%
Health Professions	19%	19%	19%	19%
Total	100%	100%	100%	100%

Percentuali condizionate: rapporto tra frequenze marginali di colonna e totale

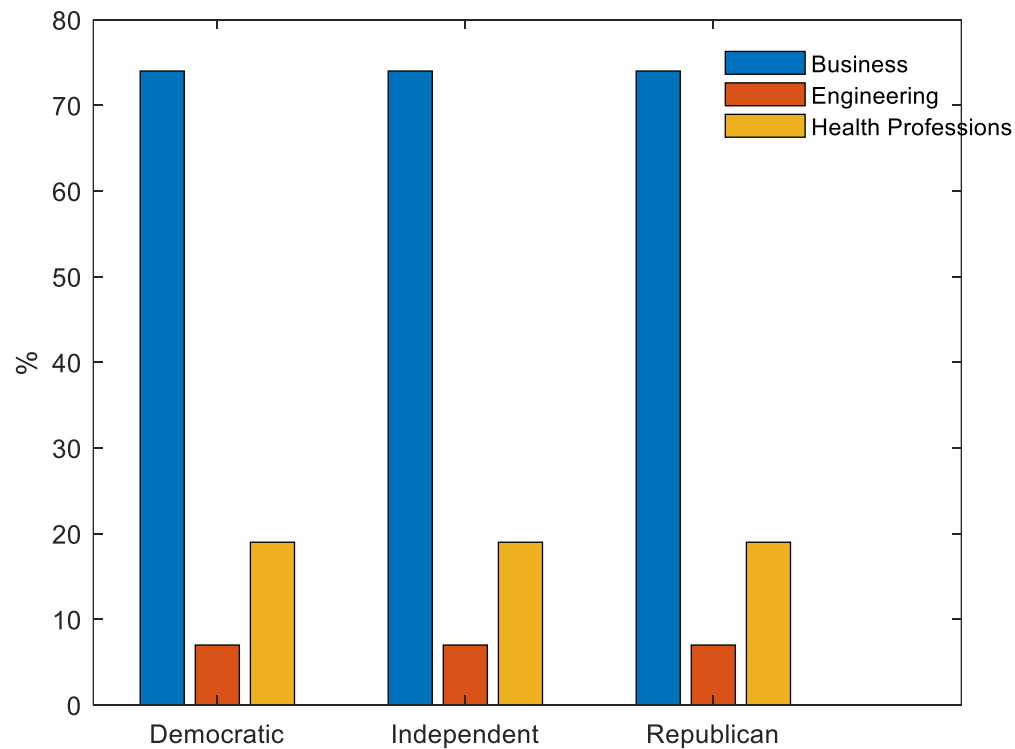


	Democrat	Independent	Republican	Total
Business	44%	14%	42%	100%
Engineering	44%	14%	42%	100%
Health Professions	44%	14%	42%	100%
Total	44%	14%	42%	100%

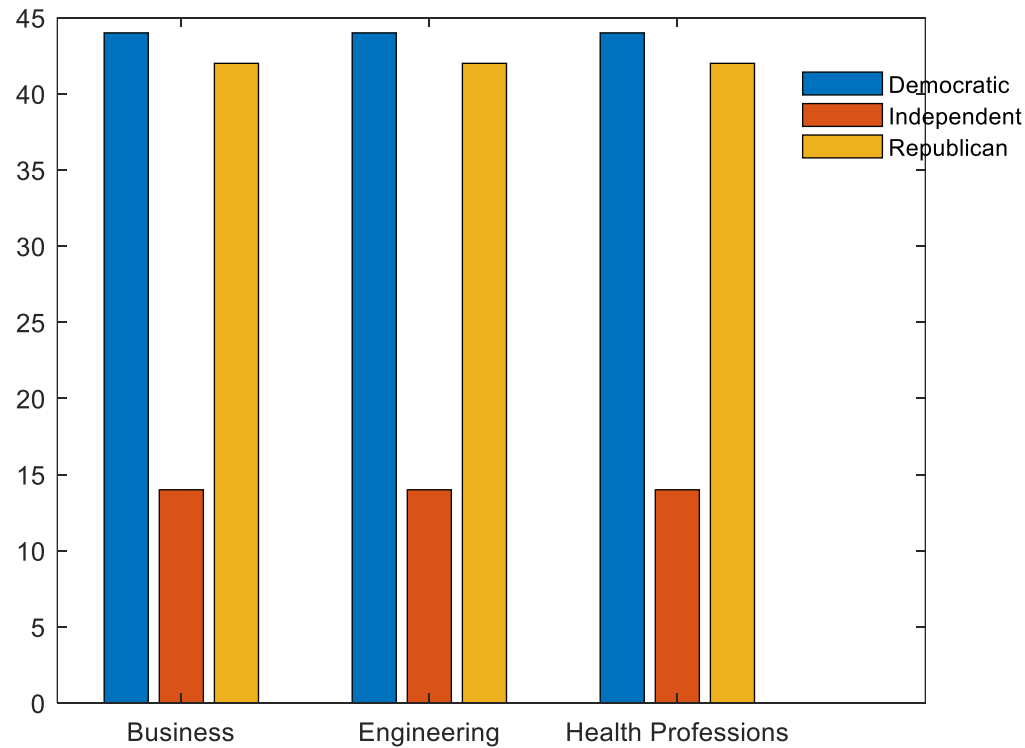
Percentuali condizionate: rapporto tra frequenze marginali di riga e totale



Esempio: orientamento politico e facoltà



Esempio: orientamento politico e facoltà



Indipendenza statistica tra caratteri

Due caratteri si dicono indipendenti se la distribuzione di uno dei due caratteri non varia al variare dell'altro.

In tal caso, le distribuzioni di frequenze relative condizionate rispetto alle modalità dell'altro sono tutte uguali tra di loro.

Indipendenza

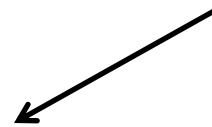
Definizione Due caratteri sono **indipendenti** se le distribuzioni condizionate sono tutte uguali tra di loro

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}} \quad f_{i|j'} = \frac{n_{ij'}}{n_{\bullet j'}}$$

$$f_{i|j} = f_{i|j'} = \dots = f_{i\bullet}$$

$$\frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n} \Leftrightarrow n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

Condizione di
indipendenza



Esempio



Esempio

- I tatuaggi tra gli adolescenti sono sempre più diffusi nella nostra società.
- C'è una relazione tra la decisione di farsi un tatuaggio e il genere? E' più facile per un adolescente maschio avere un tatuaggio rispetto ad una coetanea femmina?
- In un articolo sulla rivista *Pediatrics* Carroll et al. ha presentato i risultati di un sondaggio di 442 adolescenti, condotto tra il dicembre 2000 e l'aprile 2001 presso la Clinica Adolescent, Naval Medical Center di San Diego.

Esempio

Il questionario del sondaggio contiene domande su diversi comportamenti tra cui l'uso di droghe, attività sessuale, e la presenza di tatuaggi. La tabella che segue mostra la la presenza o meno di tatuaggi e il sesso

	Maschi	Femmine	Totale
Tattoo	15	44	59
No Tattoo	171	212	383
Total	186	256	442

Se la presenza o meno di tatuaggi fosse indipendente dal sesso, la frequenze condizionate sarebbero uguali...

	Maschi	Femmine	Totale
Tattoo	0.03	0.10	0.13
No Tattoo	0.39	0.48	0.87
Total	0.42	0.58	1

Frequenze condizionate

	Male	Female	Total
Tattoo	0.08	0.17	0.13
No Tattoo	0.92	0.83	0.87
Total	1	1	1

Frequenze dei tatuaggi e Non
tatuaggio, condizionate ai maschi

	Male	Female	Total
Tattoo	0.25	0.75	1
No Tattoo	0.45	0.55	1
Total	0.42	0.58	1

Frequenze di maschi e
femmine, condizionate a
Tatto=sì

Frequenze Condizionate

- 8% dei maschi ha il tatuaggio, rispetto al 17% delle femmine
- Solo il 25% degli adolescenti con tatuaggio sono maschi
- Il 45% degli adolescenti senza tatuaggio sono maschi

Frequenze sotto ipotesi di indipendenza

- $\hat{n}_{ij}$ = numero atteso per ogni cella, sotto ipotesi di indipendenza, coincide con prodotto marginali diviso per il totale

Maschi con tatuaggio
 $(186 \cdot 59) / 442 = 24.83$

Esempio

Il questionario del sondaggio contiene domande su diversi comportamenti tra cui l'uso di droghe, attività sessuale, e la presenza di tatuaggi. La tabella che segue mostra la la presenza o meno di tatuaggi e il sesso

	Maschi	Femmine	Totale
Tattoo	$186 \cdot 59 / 442$	$256 \cdot 59 / 442$	59
No Tattoo	$186 \cdot 383 / 442$	$256 \cdot 383 / 442$	383
Total	186	256	442

Valori attesi

Maschi senza tatuaggio

$$(186 \times 383) / 442 = 161.17$$

Femmine con tatuaggio

$$(256 \times 59) / 442 = 34.17$$

Femmine senza tatuaggio

$$(256 \times 383) / 442 = 221.83$$

Valori attesi

	Maschi	Femmine	Totale
Tattoo	24.83	34.17	59
No Tattoo	161.17	221.83	383
Totale	186	256	442

Frequenze attese versus Frequenze osservate

	Male	Female	Total
Tattoo	15	44	59
No Tattoo	171	212	383
Total	186	256	442

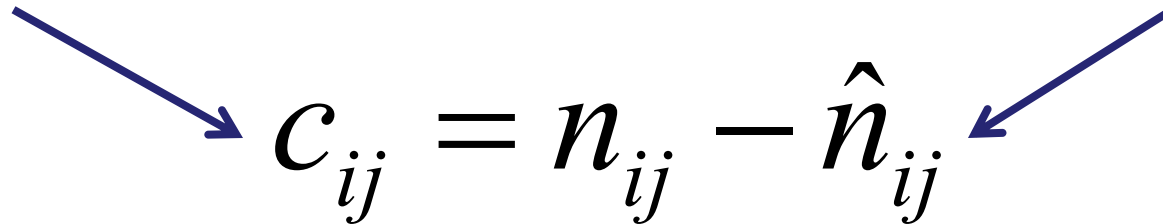
	Male	Female	Total
Tattoo	24.83	34.17	59
No Tattoo	161.17	221.83	383
Total	186	256	442

Quanto sono differenti le due tabelle?

Contingenze

Frequenze
osservate

Frequenze
attese


$$c_{ij} = n_{ij} - \hat{n}_{ij}$$

$$\sum_{j=1}^J \hat{n}_{ij} = n_{i\bullet} \quad \sum_{i=1}^I \hat{n}_{ij} = n_{\bullet j} \quad \sum_{i=1}^I \sum_{j=1}^J \hat{n}_{ij} = n$$

$$\sum_{j=1}^J c_{ij} = \sum_{i=1}^I c_{ij} = \sum_{i=1}^I \sum_{j=1}^J c_{ij} = 0$$

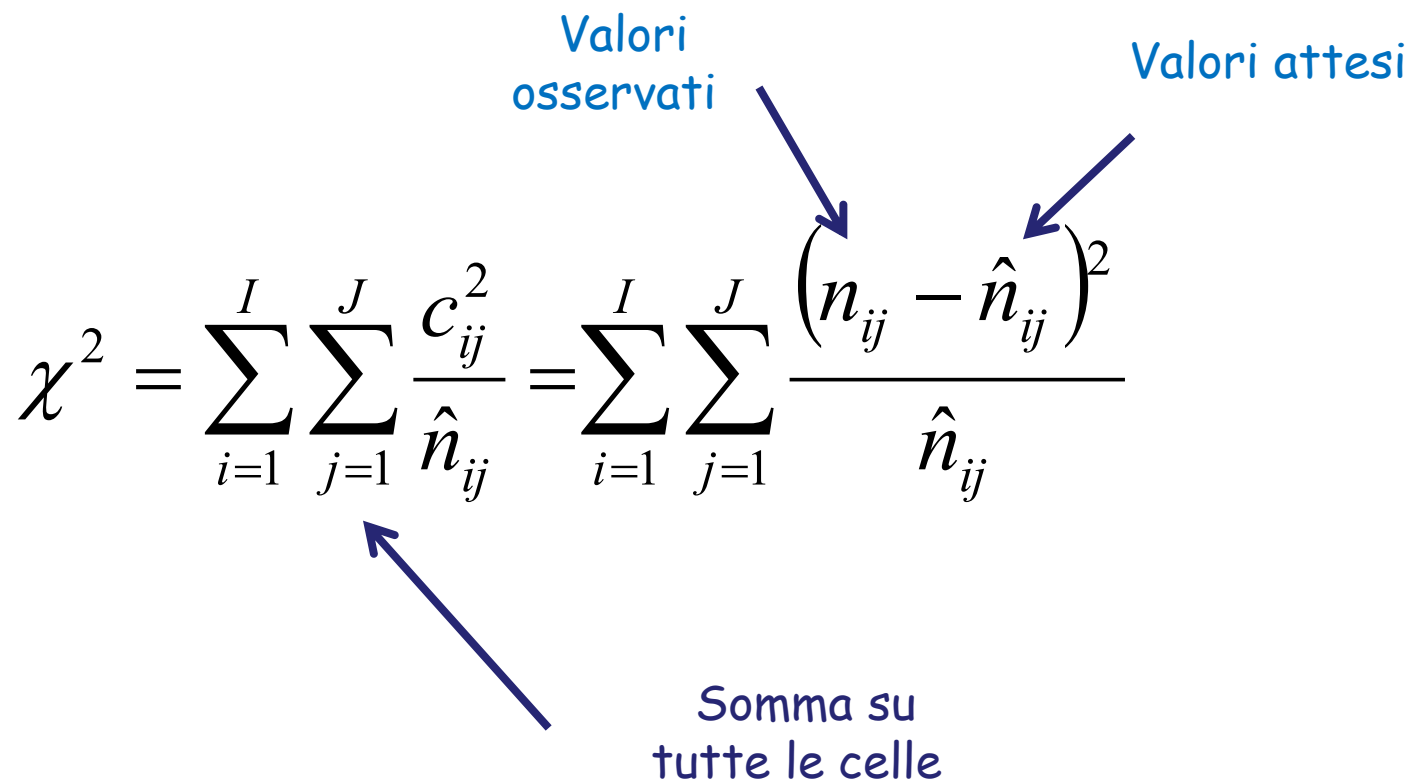
Misurare la differenza tra le due celle

Valori osservati

Valori attesi

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

Somma su tutte le celle



Misurare l'associazione

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

Chi-quadro di Pearson

Viene utilizzata per ogni coppia di caratteri X e Y con poche modalità

Non considera valori di X e Y ma solo Frequenze congiunte

Come valutare χ^2 ?

- Non può essere negativo: somma di quadrati
- Può essere molto grande: maggiore è la differenza tra osservato e atteso, maggiore ci aspettiamo che sia il valore di χ^2 .
- $\chi^2 = 0$, quando osserviamo ciò che ci aspettiamo nell'ipotesi di indipendenza tra le variabili

$\chi^2=0$  INDIPENDENZA

Indici di associazione

Pearson
Chi-square

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

Mean-square
contingency
coefficient

$$\phi^2 = \frac{\chi^2}{n} = \sum_{i=1}^I \sum_{j=1}^J \frac{f_{ij}^2}{f_{i\bullet} f_{\bullet j}}$$

Cramer V

$$V = \sqrt{\frac{\phi^2}{\min(I-1, J-1)}}$$

- Il genere e l'avere un tatuaggio sono indipendenti?
- Il genere e l'avere un tatuaggio sono dipendenti?

Esempio

$$\begin{aligned}\chi^2 &= \frac{(15 - 24.83)^2}{24.83} + \frac{(44 - 34.17)^2}{34.17} \\ &\quad + \frac{(171 - 161.17)^2}{161.17} + \frac{(212 - 221.83)^2}{221.83} \\ &= 3.89 + 2.83 + 0.60 + 0.44 = 7.76\end{aligned}$$

Esempio

$\chi^2 > 0$ Il genere e l'avere un tatuaggio sono **dipendenti**

Quanto è forte l'associazione?

$$V = \sqrt{\frac{\chi^2}{n \times \min(I - 1, J - 1)}} = 0.13$$

Associazione debole!

Esempio



50 studenti del secondo anno intervistati:

		Esame di matematica?		
		YES	NO	
Esame di economia aziendale?	YES	20	10	30
	NO	10	10	20
	Total	30	20	50

L'esito dell'esame di matematica è associato all'esito dell'esame di economia aziendale ?

Expected frequencies

		Esame di matematica?		
		YES	NO	
		YES	18	12
Esame di economia aziendale?	NO	12	8	20
	Total	30	20	50

Table of contingencies



contingencies

		Esame di matematica?		
		YES	NO	
Esame di economia aziendale?	YES	2	-2	0
	NO	-2	2	0
	Total	0	0	0

Example

$$\chi^2 = \frac{(20-18)^2}{18} + \frac{(10-12)^2}{12} + \frac{(10-12)^2}{12} + \frac{(10-8)^2}{8} = 1.389$$

$$V = \sqrt{\frac{\chi^2}{n \times \min(I-1, J-1)}} = 0.15$$

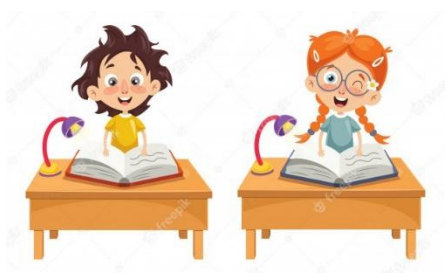
Weak association!





Esempio excel: student_survey

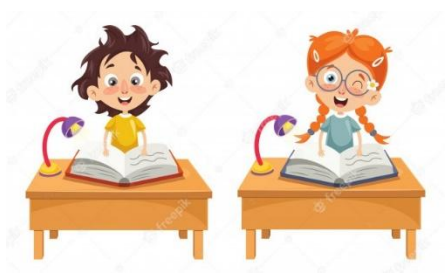
	Maschi	Femmine	Totale
Vegetarian	3	6	9
No Vegetarian	26	25	51
Totale	29	31	60



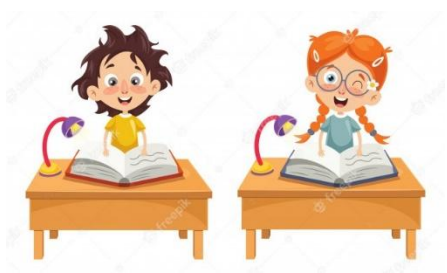
		sesso		
		maschi	femmine	Totale complessivo
Titolo di studio	elementari	4	1	5
	media	15	10	25
	superiori	40	10	50
	laurea	1	19	20
Totale complessivo		60	40	100



		sesso		
		maschi	femmine	Totale complessivo
Titolo di studio	elementari	3	2	5
	media	15	10	25
	superiori	30	20	50
	laurea	12	8	20
Totale complessivo		60	40	100



		sesso		
		maschi		
Titolo di studio	elementari	3	0.05	
	media	15	0.25	
	superiori	30	0.5	
	laurea	12	0.2	
Totale complessivo		60	1	



		seesso		
		femmine		
Titolo di studio	elementari	2	0.05	
	media	10	0.25	
	superiori	20	0.5	
	laurea	8	0.2	
	Totale complessivo	40	1	

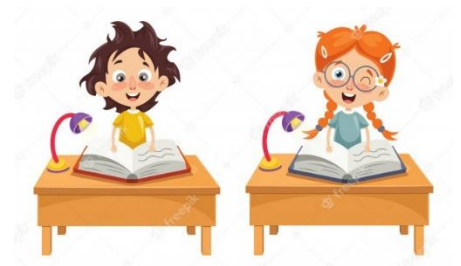


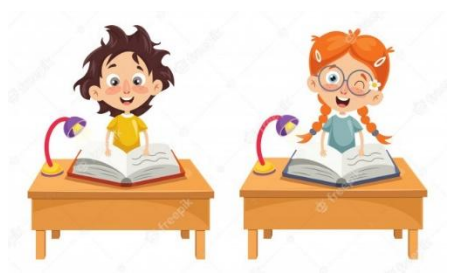
distribuzione sesso condizionato a titolo di studio

		sesso		
		maschi	femmine	
Titolo di studio	elementari	0.6	0.4	1
	media	0.6	0.4	1
	superiori	0.6	0.4	1
	laurea	0.6	0.4	1

distribuzione titolo di studio condizionato a sesso

		sesso		
		maschi	femmine	
Titolo di studio	elementari	0.05	0.05	
	media	0.25	0.25	
	superiori	0.5	0.5	
	laurea	0.2	0.2	
	Totale complessivo	1	1	





Se invece la distribuzione fosse così?
Sono indipendenti i caratteri?

		sesso		
		maschi	femmine	Totale complessivo
Titolo di studio	elementari	4	1	5
	media	15	10	25
	superiori	40	10	50
	laurea	1	19	20
	Totale complessivo	60	40	100

Esercizio

A 400 dipendenti di una grande multinazionale viene chiesto se fumano o se non fumano. Il tipo di lavoro all'interno dell'azienda risulta indipendente dalle abitudini al fumo. Il 50% dei dipendenti non fuma. Tra le persone intervistate, 240 sono amministrativi. La quota di venditori corrisponde invece a metà della quota degli amministrativi. 30 venditori fumano poco.

Lavoro	Amministr azione	Venditori	Esecutivi	Totali
Fumo				
Non fumatori				200
Poco fumatori		30		
Forti fumatori				
totali	240	120		400

Lavoro	Amministr azione	Venditori	Esecutivi	Totali
Fumo				
Non fumatori	120	60	20	200
Poco fumatori	60	30	10	100
Forti fumatori	69	30	10	100
totali	240	120	40	400

Lavoro	Amministr azione	Venditori	Esecutivi	Totali
Fumo				
Non fumatori	120	60	20	200
Poco fumatori	60	30	10	100
Forti fumatori	60	30	10	100
totali	240	120	40	400

Sono indipendenti i caratteri X e Y?

		Y			Totale
		A	B	C	
x	1	30	0	0	30
	2	0	0	40	40
	3	0	30	0	30
Totale		30	30	40	100

Sono indipendenti i caratteri X e Y?

		Y			Totale
		A	B	C	
X	1	15	10	5	30
	2	0	10	30	40
	3	15	10	5	30
Totale		30	30	40	100

Scrivere le frequenze congiunte
Sotto l'ipotesi X e Y siano indipendenti

		Y			Totale
		A	B	C	
X	1				30
	2				40
	3				30
Totale		30	30	40	100

Scrivere le frequenze congiunte
Sotto l'ipotesi X e Y siano indipendenti

		Y			Totale
		A	B	C	
X	1	9	9	12	30
	2	12	12	16	40
	3	9	9	12	30
Totale		30	30	40	100

Scrivere le frequenze congiunte

Sotto l'ipotesi X e Y siano perfettamente dipendenti

		Y			Totale
		A	B	C	
X	1	0	30	0	30
	2	0	0	40	40
	3	30	0	0	30
Totale		30	30	40	100

Sono indipendenti i caratteri X e Y?

		Y			
		A	B	C	Totale
X	1	20	20	20	60
	2	60	80	0	140
	3	40	80	40	160
Totale		120	180	60	360

Sono indipendenti i caratteri X e Y?

		Y			
		A	B	C	Totale
X	1	20	20	20	60
	2	60	80	20	160
	3	40	80	10	130
Totale		120	180	50	350

Sono indipendenti i caratteri X e Y?

		Y			Totale
		A	B	C	
X	1	20	30	10	60
	2	60	90	30	180
	3	40	60	20	120
Totale		120	180	60	360