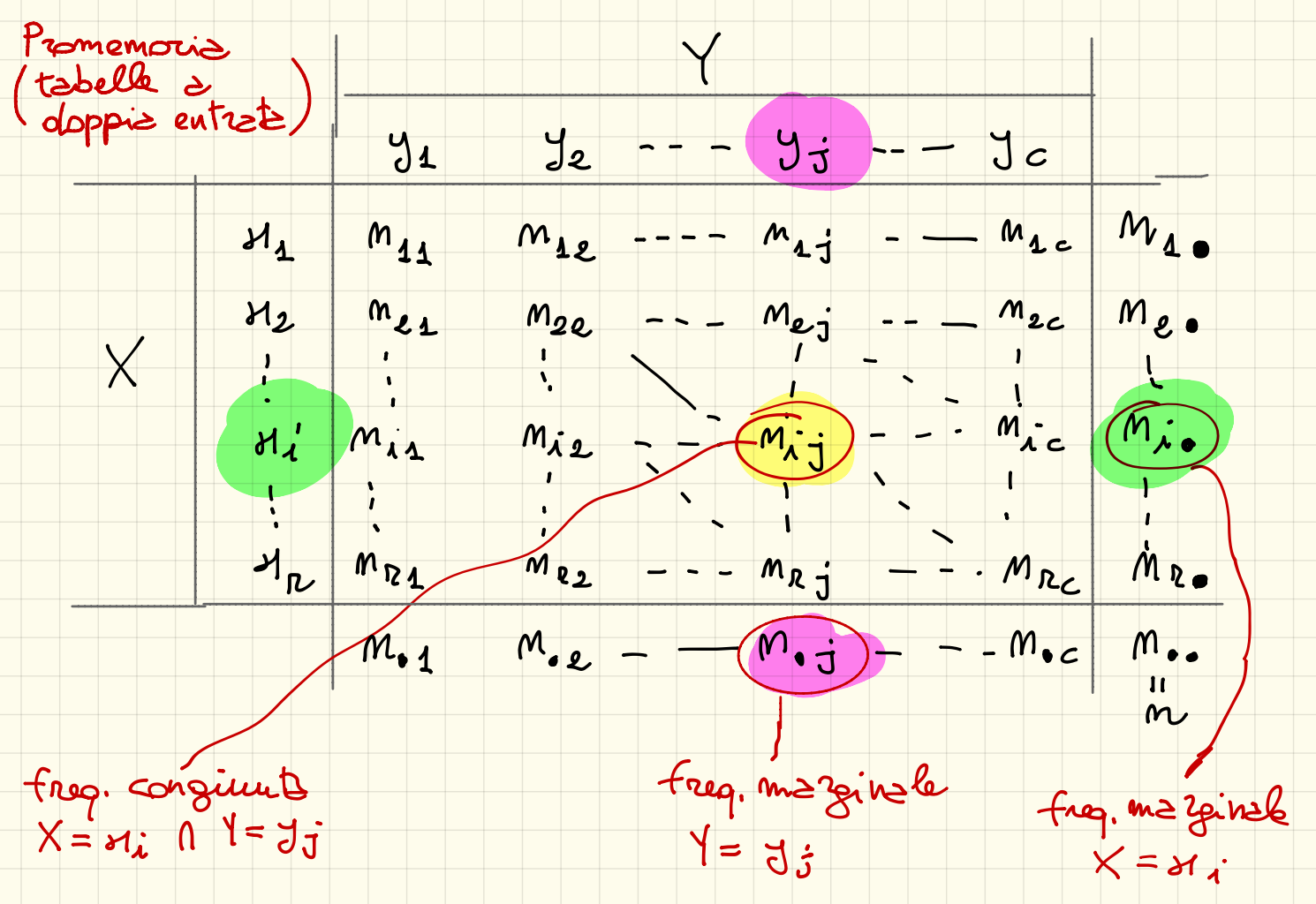
Studio dell'associazione tra due variabili qualitative

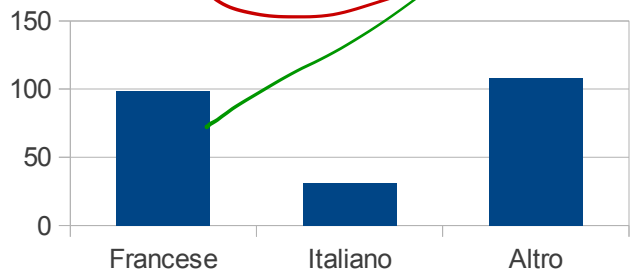
	Task Description	Due Date	Status
☐	Studio dell'associazione tra due variabili qualitative: l'indipendenza assoluta; confronto tra distribuzioni condizionate e distribuzione marginale; condizione necessaria e sufficiente per la sussistenza di indipendenza assoluta; frequenze teoriche sotto l'ipotesi di indipendenza assoluta; l'indice χ^2 come distanza normalizzata tra la tabella delle frequenze osservate e la tabella delle frequenze teoriche; normalizzazione dell'indice χ^2: l'indice Φ^2 e l'indice V di Cramer		
☐			
☐			
☐			
☐			
☐			
☐			
☐			
☐			
☐			

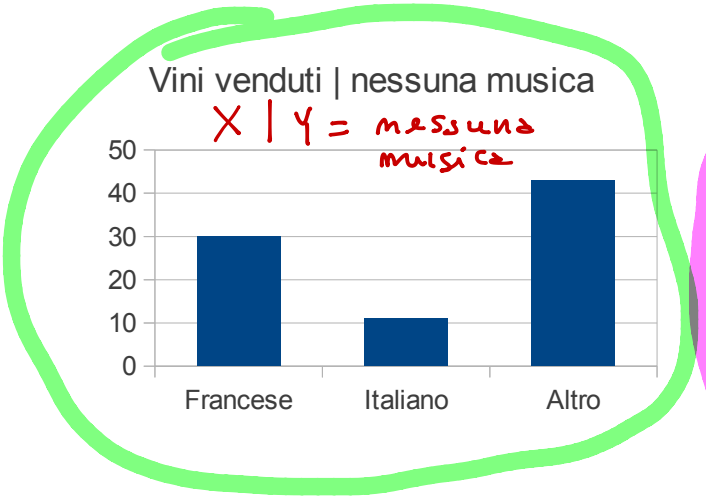

Promemoria (tabella a doppia entrata)
Y
y_1 y_2 y_j y_c
X
x_1 x_2 x_i x_r
n_{11} n_{12} n_{1j} n_{1c}
n_{21} n_{22} n_{2j} n_{2c}
n_{i1} n_{i2} n_{ij} n_{ic}
n_{r1} n_{r2} n_{rj} n_{rc}
$n_{1\bullet}$ $n_{2\bullet}$ $n_{i\bullet}$ $n_{r\bullet}$
$n_{\bullet 1}$ $n_{\bullet 2}$ $n_{\bullet j}$ $n_{\bullet c}$
$n_{\bullet\bullet} = n$
freq. congiunta
$X = x_i \cap Y = y_j$
freq. marginale
$Y = y_j$
freq. marginale
$X = x_i$

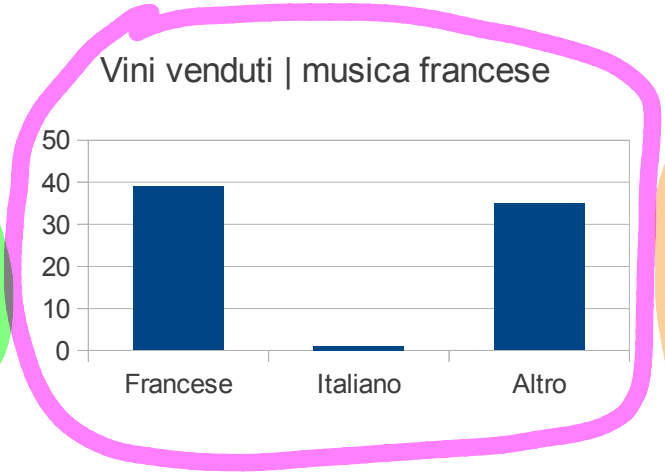
Sia X che Y sono qualitative
(si può lavorare allo stesso modo anche in caso di variabili miste o di due variabili numeriche)

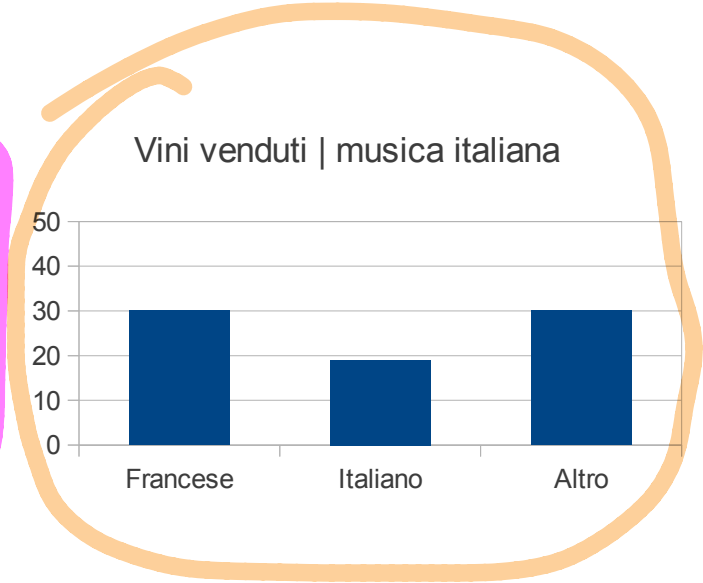
Vino	Musica sottofondo: Nessuna	Francese	Italia	TOTALE
Francese	30	39	30	99
Italiano	11	1	19	31
Altro	43	35	30	108
TOTALE	84	75	79	238

n_{ij} — X: paese provenienza — Y: musica sottofondo — N = 238









	n_{ij}	Musica sottofondo			
		Nessuna	Francese	Italia	TOTALE
Vino	Francese	30	39	30	**99**
	Italiano	11	1	19	**31**
	Altro	43	35	30	**108**
	TOTALE	**84**	**75**	**79**	238

profili colonna → $\frac{n_{ij}}{n_{\cdot j}}$

		Musica sottofondo			
		Nessuna	Francese	Italia	TOTALE
Vino	Francese	0.36	0.52	0.38	**0.42**
	Italiano	0.13	0.01	0.24	**0.13**
	Altro	0.51	0.47	0.38	**0.45**
	TOTALE	**1**	**1**	**1**	1

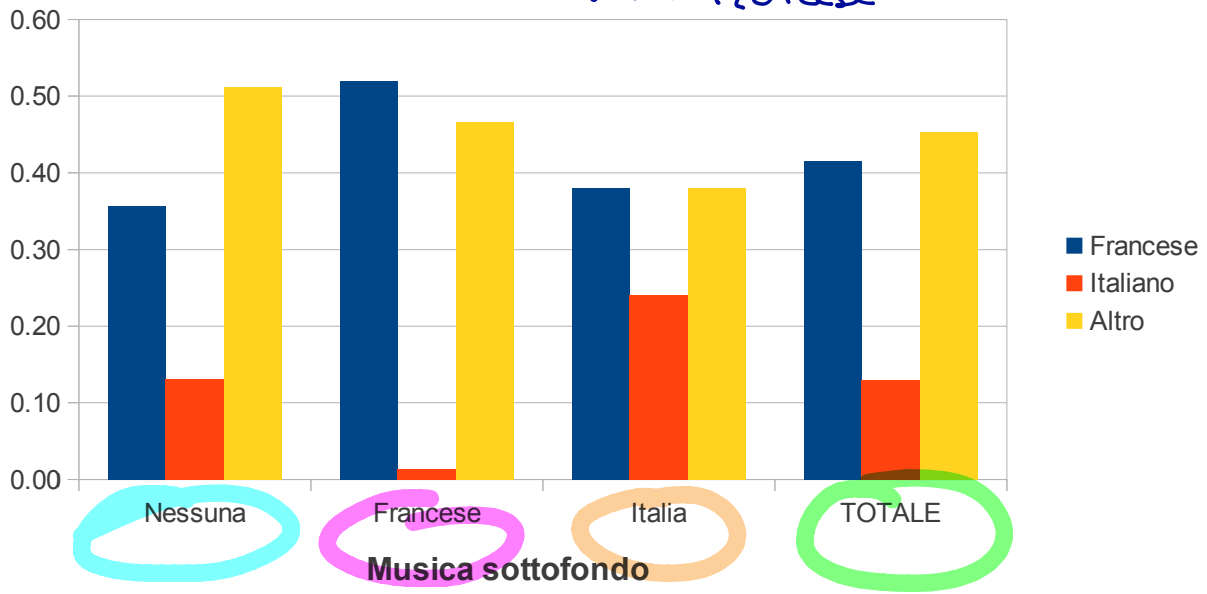
$\frac{30}{84}$ $\frac{11}{84}$ $\frac{43}{84}$

media ponderata di 0,36, 0,52 e 0,38

$\frac{0{,}36 \times 84 + 0{,}52 \times 75 + 0{,}38 \times 79}{238}$

il profilo marginale può essere interpretato come profilo medio

il 42% dei clienti acquista una bottiglia di vino francese



Se tra le due variabili X e Y non vi fosse associazione (connessione o legame) dovrei aspettarmi distribuzioni condizionate uguali tra loro (uguali anche alla distribuzione marginale)

↳ condizione di indipendenza assoluta

Frequenze osservate n_{ij}

		Musica			
		N	F	I	
VINO	F	30	39	30	99
	I	11	1	19	31
	A	43	35	30	108
		84	75	79	238

$\frac{n_{ij}}{n_{\bullet j}}$

		Musica			
		N	F	I	
VINO	F	0,36	0,52	0,38	0,42
	I	0,13	0,01	0,24	0,13
	A	0,51	0,47	0,38	0,45
		1	1	1	1

CONDIZIONE DI INDIPENDENZA ASSOLUTA

Se non vi fosse alcun legame tra scelta di acquisto e stereotipo di acquisto, ciascun profilo colonna dovrebbe essere uguale al profilo marginale

Frequenze osservate n_{ij}

		Musica N	Musica F	Musica I	
VINO	F	30	39	30	99
	I	11	1	19	31
	A	43	35	30	108
		84	75	79	238

$\frac{n_{ij}}{n_{\bullet j}}$

		Musica N	Musica F	Musica I	
VINO	F	0,36	0,52	0,38	0,42
	I	0,13	0,01	0,24	0,13
	A	0,51	0,47	0,38	0,45
		1	1	1	1

frequenze teoriche o attese sotto l'ipotesi di indipendenza

$l_{ij} = \hat{n}_{ij} = n^*_{ij} = n'_{ij}$

		Musica N	Musica F	Musica I	
VINO	F	0,42×84	0,42×75	0,42×79	99
	I	0,13×84	0,13×75	0,13×79	31
	A	0,45×84	0,45×75	0,45×79	108
		84	75	79	238

questi devono essere uguali alla tabella n_{ij}

frequenze teoriche o attese sotto l'ipotesi di indipendenza

$\ell_{ij} = \hat{m}_{ij} = m^{*}_{ij} = m'_{ij}$

VINO \ Musica	N	F	I	
F	0,42×84	0,42×75	0,42×79	99
I	0,13×84	0,13×75	0,13×79	31
A	0,45×84	0,45×75	0,45×79	108
	84	75	79	238

$0{,}42 = \frac{99}{238} = \frac{m_{1\cdot}}{n}$

$0{,}13 = \frac{31}{238} = \frac{m_{2\cdot}}{n}$

$0{,}45 = \frac{108}{238} = \frac{m_{3\cdot}}{n}$

$0{,}42 \times 84 = \frac{m_{1\cdot}}{n} m_{\cdot 1} = \frac{m_{1\cdot} m_{\cdot 1}}{n} = \hat{m}_{11}$

$0{,}13 \times 79 = \frac{m_{2\cdot}}{n} m_{\cdot 3} = \frac{m_{2\cdot} m_{\cdot 3}}{n} = \hat{m}_{23}$

$0{,}45 \times 75 = \frac{m_{3\cdot}}{n} m_{\cdot 2} = \frac{m_{3\cdot} m_{\cdot 2}}{n} = \hat{m}_{32}$

	y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	
x_1							
x_2							
$\vdots$							
x_i				n_{ij}			$n_{i\bullet}$
$\vdots$							
x_r							
				$n_{\bullet j}$			n

	y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	
x_1							
x_2							
$\vdots$							
x_i				$\hat{n}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$			$n_{i\bullet}$
$\vdots$							
x_r							
				$n_{\bullet j}$			n

CONDIZIONE DI INDIPENDENZA ASSOLUTA

↳ i profili colonna (riga) devono essere tutte uguali

$$\frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n} \quad \forall i, j$$

$$\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n} \quad \forall i, j$$

CONDIZIONE INDIPENDENZA SUI PROFILI COLONNA

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{n} \quad \forall i, j$$

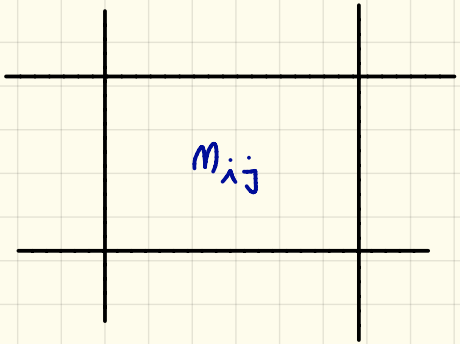
CONDIZIONE INDIPENDENZA SUI PROFILI RIGA

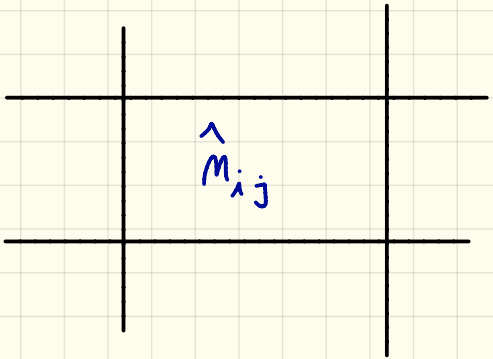
$$\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n} \quad \forall i, j$$

$$\hat{n}_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

CONDIZIONE NECESSARIA E SUFFICIENTE

frequenze attese o teoriche sotto l'ipotesi di indipendenza
(per evitare confusione le indichiamo con $\hat{n}_{ij}$)





l'indice di associazione è una distanza tra la tabella della n_{ij} e la tabella della n_{ij}

n_{ij}

		Musica sottofondo			
		Nessuna	Francese	Italia	TOTALE
Vino	Francese	30	39	30	**99**
	Italiano	11	1	19	**31**
	Altro	43	35	30	**108**
	TOTALE	**84**	**75**	**79**	238

$34.94 = \frac{99 \times 84}{238} = \frac{n_{i\cdot}\, n_{\cdot j}}{N}$

$\frac{n_{i\cdot}\, n_{\cdot j}}{n} = \hat{n}_{ij}$

		Musica sottofondo			
		Nessuna	Francese	Italia	TOTALE
Vino	Francese	34.94	31.20	32.86	**99**
	Italiano	10.94	9.77	10.29	**31**
	Altro	38.12	34.03	35.85	**108**
	TOTALE	**84**	**75**	**79**	238

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

		Musica sottofondo			
		Nessuna	Francese	Italia	TOTALE
Vino	Francese	-4.94	7.80	-2.86	**0**
	Italiano	0.06	-8.77	8.71	**0**
	Altro	4.88	0.97	-5.85	**0**
	TOTALE	**0**	**0**	**-1.0658E-014**	0

$= 0$

$\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$

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

		Musica sottofondo			
		Nessuna	Francese	Italia	
Vino	Francese	0.70	1.95	0.25	
	Italiano	0.00	7.87	7.37	
	Altro	0.63	0.03	0.95	
					19.75

somma di tutte le distanze (normalizz.)

indice del chi-quadro (χ^2)

la divisione per $\hat{n}_{ij}$ ha l'effetto di normalizzare tutte le celle

trattare tutte le celle allo stesso modo: senza questa operazione le celle con $n_{i\cdot}$ e/o $n_{\cdot j}$ avrebbero un peso maggiore

contingenza

queste sono le due celle che contribuiscono di più a spiegare l'associazione tra X e Y

Ricapitoliamo---

Lo studio dell'indipendenza si basa sul confronto tra i profili

↓

nell'es. sul confronto tra i profili colonna e il profilo marginale

[distribuz. condizionate $X \mid Y = y_j$ $j = 1, \dots, c$] [X]

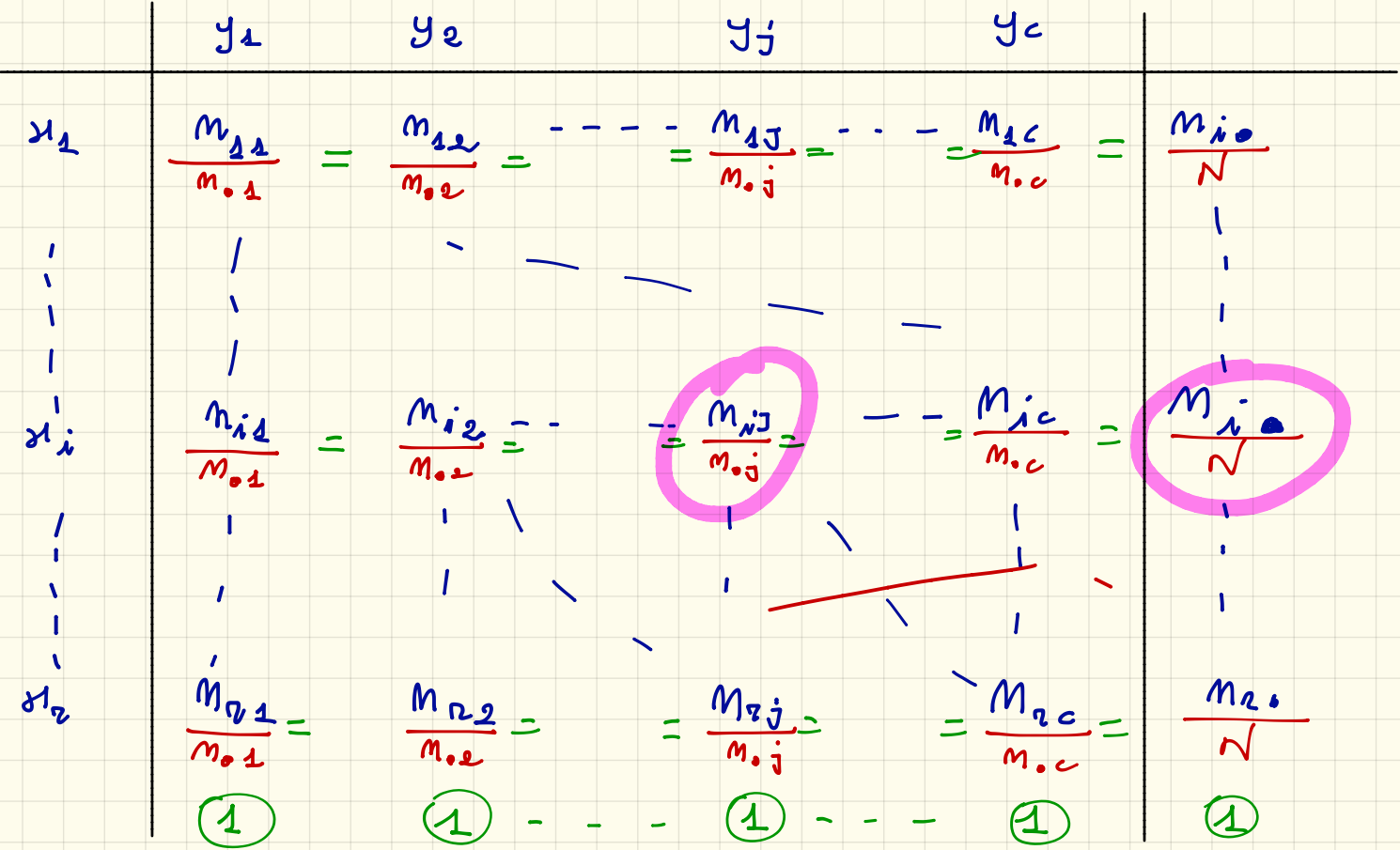
Lo studio è di tipo simmetrico

↓

confrontare i profili colonna equivale a confrontare i profili riga da un punto di vista numerico

↓

non è così rispetto all'interpretazione dei risultati


y_1 y_2 y_j y_c
x_1
x_i
x_n
$\frac{n_{11}}{n_{\cdot 1}} = \frac{n_{12}}{n_{\cdot 2}} = \frac{n_{1j}}{n_{\cdot j}} = \frac{n_{1c}}{n_{\cdot c}} = \frac{n_{i\cdot}}{N}$
$\frac{n_{i1}}{n_{\cdot 1}} = \frac{n_{i2}}{n_{\cdot 2}} = \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{ic}}{n_{\cdot c}} = \frac{n_{i\cdot}}{N}$
$\frac{n_{n1}}{n_{\cdot 1}} = \frac{n_{n2}}{n_{\cdot 2}} = \frac{n_{nj}}{n_{\cdot j}} = \frac{n_{nc}}{n_{\cdot c}} = \frac{n_{n\cdot}}{N}$
1 1 1 1 1

formulazione compatta condizione di indipendenza

$$\frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{N}$$

$\forall\ i = 1, \dots, r$

$\forall\ j = 1, \dots, c$

moltiplicando ambo i membri per $n_{\bullet j}$

$$n_{ij} = \frac{n_{i\bullet}\ n_{\bullet j}}{N}$$

formulaz. equivalente

freq. che osserei in ciascuna cella se vi fosse indipendenza

per distinguere queste freq. teoriche o attese sotto l'ipotesi di indipendenza di solito le si indica con $\hat{n}_{ij}$ oppure n^*_{ij} oppure e_{ij}

	Y
X	n_{ij} freq. osservate

	Y
X	$\hat{n}_{ij}$ freq. teoriche o attese sotto l'ipotesi di indipendenza

distanza tra le n_{ij} e le $\hat{n}_{ij}$ come misura dell'indipendenza

↓

se ci fosse indipendenza le $\hat{n}_{ij}$ dovrebbero essere uguali alle n_{ij} (le due tabelle dovrebbero essere identiche)

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

- 0 → se solo se $n_{ij} = \hat{n}_{ij}$ $\forall i, j$
- $\max$ → dipende dalla struttura delle tabelle (r, c) ma anche dal n. di dati (N)

Si definisce: $\Phi^2 = \frac{\chi^2}{N}$ → elimina la dipendenza da N

indice di contingenza media

$$V = \sqrt{\frac{\Phi^2}{\min(r, c) - 1}}$$

(Cramer)

→ alcuni autori estraggono la radice

elimina l'effetto anche della struttura

→ è un indice normalizzato tra 0 e 1

χ^2 { 0 minimo

$N \times [\min(r, c) - 1]$ massimo

$$\frac{\chi^2 - \min}{\max - \min} = \frac{\chi^2}{N[\min(r,c)-1]} = \frac{\Phi^2}{\min(r,c)-1}$$

$$\Phi^2$$

quando si verifica $\chi^2 = \max$?

Vino	Musica: Nessuna	Musica: Francese	Musica: Italiana	
Francese	0	99	0	99
Italiano	0	0	31	31
Altro	108	0	0	108
	108	99	31	238

$$\chi^2 = 238 \times [\min(3,3) - 1]$$

$$\Phi^2 = \min(3,3) - 1$$

$$V = 1$$

nel caso di associazione 1 ad 1 riga - colonna

↓

massima interdipendenza

↓

il max assoluto è osservabile solo nel caso di tabelle quadrate ($r = c$)

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

ragion. profili colonna

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

ragionam. profili riga

$$\hat{n}_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

indice X^2 è simmetrico

si arriva alla stessa condizione di indipendenza sia che si parte dai profili riga che dai profili colonna

studio associazione è simmetrico

X e Y giocano lo stesso ruolo

INTERDIPENDENZA

dipendenza reciproca

Schema riepilogativo di calcolo dell'indice χ^2

① tabella frequenze osservate ← input

$$n_{ij} \qquad i = 1, \dots, r \qquad j = 1, \dots, c$$

② tabella frequenze teoriche o attese (sotto le ipotesi di indipendenza assoluta)

$$\hat{n}_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

③ distanza (normalizzata) tra le n_{ij} e $\hat{n}_{ij}$

$$\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

④ sintetizzo la tabella delle distanze con il seguente indice

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

Formula alternativa (ridotta) del x^2

$$x^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}} =$$

$$= \sum_i \sum_j \frac{n_{ij}^2 + \left(\frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2 - 2 n_{ij} \frac{n_{i\cdot} n_{\cdot j}}{n}}{\frac{n_{i\cdot} n_{\cdot j}}{n}} =$$

$$= \sum_i \sum_j \frac{n_{ij}^2 + \left(\frac{n_{i\cdot} n_{\cdot j}}{n}\right)\left(\frac{n_{ij} n_{\cdot j}}{n}\right) - 2 n_{ij} \frac{n_{i\cdot} n_{\cdot j}}{n}}{\frac{n_{i\cdot} n_{\cdot j}}{n}} =$$

$$= \sum_i \sum_j \left\{ \frac{n_{ij}^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}} + \frac{\left(\frac{n_{i\cdot} n_{\cdot j}}{n}\right)\left(\frac{n_{i\cdot} n_{\cdot j}}{n}\right)}{\frac{n_{i\cdot} n_{\cdot j}}{n}} - \frac{2\, n_{ij} \frac{n_{i\cdot} n_{\cdot j}}{n}}{\frac{n_{i\cdot} n_{\cdot j}}{n}} \right\} =$$

$$= n \sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} + \sum_i \sum_j \frac{n_{i\cdot} n_{\cdot j}}{n} - 2 \sum_i \sum_j n_{ij} =$$

$$= n \sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} + \frac{\overbrace{\sum_i n_{i\cdot}}^{(=n)} \overbrace{\sum_j n_{\cdot j}}^{(=n)}}{n} - 2n =$$

$$= n \sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - n = n \left[\sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right]$$