Metodi statistici per le ricerche di mercato

Prof.ssa Isabella Mingo
A.A. 2016-2017

SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Scienze Politiche, Sociologia, Comunicazione
Corso di laurea Magistrale in «Organizzazione e marketing per la comunicazione d'impresa»

Caratteri qualitativi
Indici di eterogeneità o mutabilità

- Evidenziano e quantificano la presenza di eterogeneità nella distribuzione di un carattere qualitativo.
- Un indice di mutabilità deve soddisfare le seguenti condizioni:
  - assumere valore 0 se e solo se il collettivo è omogeneo rispetto al carattere considerato;
  - crescere, assumendo valori maggiori di 0, all'aumentare dell'eterogeneità tra le modalità del carattere.
- Si possono distinguere due situazioni estreme:
  - *mutabilità nulla*, tutte le unità presentano la medesima modalità del carattere
  - *mutabilità massima* quando tutte le unità presentano modalità differenti del carattere oppure le diverse modalità del carattere hanno le stesse frequenze.

Reclami verso due gestori dì telefonia per motivo			
	Gestore A	Gestore B	Gestore C
Copertura nazionale	200	600	200
Costi	200	0	150
Piano tariffario	200	0	250
Totale	600	600	600

Caratteri qualitativi
Indice di eterogeneità di Gini

Reclami verso tre gestori dì telefonia per motivo (freq. Assolute)			
	Gestore A	Gestore B	Gestore C
Copertura nazionale	200	600	200
Costi	200	0	150
Piano tariffario	200	0	250
Totale	600	600	600

Reclami verso tre gestori dì telefonia per motivo (freq. Relative)			
	Gestore A	Gestore B	Gestore C
Copertura nazionale	0,33	1,00	0,33
Costi	0,33	0,00	0,25
Piano tariffario	0,33	0,00	0,42
Totale	1,00	1,00	1,00

$$G = 1 - \sum_{i=1}^{K} \left(\frac{n_i}{n}\right)^2$$

$G1 = 1 - (0{,}33^2 + 0{,}33^2 + 0{,}33^2) = 0{,}67$

$G2 = 1 - (1^2 + 0^2 + 0^2) = 0$

$G3 = 1 - (0{,}33^2 + 0{,}25^2 + 0{,}42^2) = 0{,}65$

Caratteri qualitativi
Indice di eterogeneità di Gini normalizzato

Assume sempre valori compresi tra 0 (caso di eterogeneità nulla) e 1 (caso di eterogeneità massima)
Può essere utilizzato per operare confronti del medesimo carattere osservato su collettivi differenti, con numerosità diversa, o tra distribuzioni di caratteri differenti osservati sul medesimo collettivo

$$G_{rel} = \frac{G}{\max(G)} = \frac{G}{1 - \frac{1}{K}} = \frac{K \times G}{K - 1}$$

K = numero di modalità

Nel nostro esempio K = 3

max (G)=(1-1/3)=0,67

$G1_N$=0,67/0,67=1

$G2_N$=0/0,67=0

$G3_N$=0,65/0,67=0,97

Caratteri qualitativi
Indice di eterogeneità esercizio

Date le seguenti distribuzioni di frequenze riguardanti la rilevazione delle vendite degli stessi prodotti in due supermercati differenti, indicare in quale supermercato la clientela è più eterogenea rispetto all'acquisto dei prodotti considerati.

Numero di prodotti venduti per marca		
	Supermercato A n_i	Supermercato B n_j
Barilla	1200	360
Buitoni	870	230
Divella	360	220
Voiello	580	230
Totale	3010	1040

Calcolo

Analisi bivariata

Esistono diverse valutazioni per categorie di utenti?

- Operativamente, per ottenere informazioni più specifiche per sottogruppi di popolazione è necessario applicare tecniche di *analisi bivariata*:
  - Tabelle a doppia entrata se le variabili sono nominali o ordinali
  - Statistiche descrittive (ad esempio indici medi) della variabile quantitativa per ogni sottogruppo individuato dalle modalità della variabile qualitativa.

Che cosa è l'analisi bivariata?

E' lo studio congiunto di due caratteri

- Esempio nel casi di caratteri qualitativi:

Sesso

		Frequenza
Valido	Maschio	750
	Femmina	750
	Totale	1500

Soddisfazione ASL

		Frequenza
Valido	Per niente	176
	Poco	412
	Abbastanza	838
	Molto	74
	Totale	1500

Tavola di contingenza Sesso * Soddisfazione ASL

Conteggio

		Soddisfazione ASL				
		Per niente	Poco	Abbastanza	Molto	Totale
Sesso	Maschio	84	232	398	36	750
	Femmina	92	180	440	38	750
Totale		176	412	838	74	1500

I valori delle celle derivano dall'analisi della tabella unitaria!

Distribuzione doppia di frequenze: caratteristiche

- ✓ Tabella che consente di sintetizzare l'informazione disponibile su due caratteri osservati contemporaneamente sul medesimo collettivo di ***n*** u.s.

Tavola di contingenza Sesso * Soddisfazione ASL

Conteggio

		Soddisfazione ASL				Totale
		Per niente	Poco	Abbastanza	Molto	
Sesso	Maschio	84	232	398	36	750
	Femmina	92	180	440	38	750
Totale		176	412	838	74	1500

- ✓ In colonna :
  - ✓ Lista di modalità del carattere 1
- ✓ *In riga :*
  - ✓ Lista di modalità del carattere 2
- ✓ Nella tabella si considerano tutte le possibili coppie di modalità (una del car. 1 ed una del car. 2).
- ✓ I valori rappresentati sono il conteggio, ossia le frequenze assolute, del numero di u.s. del collettivo considerato che presentano una coppia di modalità dei 2 caratteri.

Tabella a doppia entrata

Frequenze n_{ij} delle unità del collettivo che presentano congiuntamente la modalità i-esima di un carattere e la modalità J-esima di un secondo carattere.

Ha un numero di righe maggiore o uguale al numero di modalità della variabile rappresentata in riga e un numero di colonne maggiore o uguale a quello delle modalità della variabile rappresentata in colonna.

Tavola di contingenza Sesso * Soddisfazione ASL

Conteggio

		Soddisfazione ASL				Totale
		Per niente	Poco	Abbastanza	Molto	
Sesso	Maschio	84	232	398	36	750
	Femmina	92	180	440	38	750
Totale		176	412	838	74	1500

Distribuzioni condizionate

Distribuzioni marginali di riga e di colonna

Dalla distribuzione unitaria multipla alla distribuzione doppia di frequenza: esempio

	Sesso	Soddisfazione ASL
1	Maschio	Abbastanza
2	Maschio	Molto
3	Femmina	Molto
4	Femmina	Abbastanza
5	Maschio	Per niente
6	Maschio	Per niente
7	Femmina	Abbastanza
8	Femmina	Abbastanza
9	Femmina	Molto
10	Femmina	Molto
11	Femmina	Abbastanza
12	Femmina	Poco
13	Femmina	Abbastanza
14	Femmina	Abbastanza
15	Maschio	Molto
16	Maschio	Abbastanza
17	Femmina	Poco
18	Femmina	Poco
19	Femmina	Per niente
20	Femmina	Per niente
Totale N	20	20

1- Costruiamo un tabella che ha:

- un numero di righe uguale al numero di modalità della variabile che vogliamo rappresentare in riga più 1 per i totali di colonna
- un numero di colonne uguale a quello delle modalità della variabile che vogliamo rappresentare in colonna più 1 per i totali di riga.

		Soddisfazione ASL				Totale
		Per niente	Poco	Abbastanza	Molto	
Sesso	Maschio					
	Femmina					
Totale						

FSSC

Dalla distribuzione unitaria multipla alla distribuzione doppia di frequenza: esercizio (segue)

	Sesso	Soddisfazione ASL
1	Maschio	Abbastanza
2	Maschio	Molto
3	Femmina	Molto
4	Femmina	Abbastanza
5	Maschio	Per niente
6	Maschio	Per niente
7	Femmina	Abbastanza
8	Femmina	Abbastanza
9	Femmina	Molto
10	Femmina	Molto
11	Femmina	Abbastanza
12	Femmina	Poco
13	Femmina	Abbastanza
14	Femmina	Abbastanza
15	Maschio	Molto
16	Maschio	Abbastanza
17	Femmina	Poco
18	Femmina	Poco
19	Femmina	Per niente
20	Femmina	Per niente
Totale N	20	20

2 – Contiamo per ciascun carattere le unità che presentano una stessa modalità e scriviamo i totali nelle rispettive celle marginali della tabella.

3- Contiamo le unità statistiche che presentano congiuntamente le modalità a due a due e scriviamo le frequenze nelle rispettive celle condizionate

4- verifichiamo che le somme dei valori siano coerenti.

		Soddisfazione ASL				Totale
		Per niente	Poco	Abbastanza	Molto	
Sesso	Maschio					somma maschi
	Femmina					somma femmine
Totale		somma per niente soddisfatti	somma abbastanza soddisfatti	somma molto soddisfatti	somma poco soddisfatti	totale

Tabelle a doppia entrata : profili di riga e distribuzioni marginali percentuali

Tavola di contingenza Sesso * Soddisfazione ASL

			Soddisfazione ASL				Totale
			Per niente	Poco	Abbastanza	Molto	
Sesso	Maschio	Conteggio	84	232	398	36	750
		% entro Sesso	11,2%	30,9%	53,1%	4,8%	100,0%
	Femmina	Conteggio	92	180	440	38	750
		% entro Sesso	12,3%	24,0%	58,7%	5,1%	100,0%
Totale		Conteggio	176	412	838	74	1500
		% entro Sesso	11,7%	27,5%	55,9%	4,9%	100,0%

- Le **distribuzioni marginali percentuali** si ottengono dividendo le frequenze assolute marginali per il totale:
- $f_{i.}=n_{i.}/n_{..}*100$; nella tabella precedente non sono calcolate
- $f_{.j}=n_{.j}/n..*100$ 176/1500*100=11,7% ; 412/1500*100=27,5%; 838/1500*100=55,9%; 74/1500*100=4,9%

- Nell'esempio le **distribuzioni percentuali condizionate** (profili di riga) della variabile "Sesso" e della variabile soddisfazione si ottengono rispettivamente rapportando le distribuzioni condizionate ai corrispondenti totali di riga e moltiplicando per 100.
- Per i maschi: 84/750*100=11,2 ; 232/750*100=30,9%.....
- Per le femmine 92/750*100=12,3%; 180/750*100=24% …….

Tabelle a doppia entrata : profili di colonna e distribuzioni marginali percentuali

Tavola di contingenza Sesso * Soddisfazione ASL

			Soddisfazione ASL				Totale
			Per niente	Poco	Abbastanza	Molto	
Sesso	Maschio	Conteggio	84	232	398	36	750
		% entro Soddisfazione ASL	47,7%	56,3%	47,5%	48,6%	50,0%
	Femmina	Conteggio	92	180	440	38	750
		% entro Soddisfazione ASL	52,3%	43,7%	52,5%	51,4%	50,0%
Totale		Conteggio	176	412	838	74	1500
		% entro Soddisfazione ASL	100,0%	100,0%	100,0%	100,0%	100,0%

- Le **distribuzioni marginali percentuali di riga**
- $f_{i.}=n_{i.}/n_{..}*100$; 750/1500*100;
- Nell'esempio le **distribuzioni percentuali condizionate** (profili di colonna della variabile "Sesso" e della variabile soddisfazione si ottengono rispettivamente rapportando le distribuzioni condizionate ai corrispondenti totali di riga e moltiplicando per 100.
- Per gli utenti per niente soddisfatti : 84/176*100=47,7%; 92/176*100=52,3%
- ……
- Per gli utenti molto soddisfatti: 36/74*100= 48,6%; 38/74*100= 51,4%

Profili riga e profili colonna: formalizzazione

- Profili Riga

		X						Tot.
		x_1	x_2	…	x_j	…	x_K	
Y	y_1	$n_{11}/n_{1.}$	$n_{12}/n_{1.}$	…	$n_{1j}/n_{1.}$	…	$n_{1K}/n_{1.}$	1
	y_2	$n_{21}/n_{2.}$	$n_{22}/n_{2.}$	…	$n_{2j}/n_{2.}$	…	$n_{2K}/n_{2.}$	1
	:	:	:	…	:	…	:	:
	y_i	$n_{i1}/n_{i.}$	$n_{i2}/n_{i.}$	…	$n_{ij}/n_{i.}$	…	$n_{iK}/n_{i.}$	1
	:	:	:	…	:	…	:	:
	y_H	$n_{H1}/n_{H.}$	$n_{H2}/n_{H.}$	…	$n_{Hj}/n_{H.}$	…	$n_{HK}/n_{H.}$	1
		$n._1/n$	$n._2/n$	…	$n._j/n$	…	$n._K/n$	

- Profili Colonna

		X						
		x_1	x_2	…	x_j	…	x_K	
Y	y_1	$n_{11}/n_{.1}$	$n_{12}/n_{.2}$	…	$n_{1j}/n_{.j}$	…	$n_{1K}/n_{.K}$	$n_{1.}/n$
	y_2	$n_{21}/n_{.1}$	$n_{22}/n_{.2}$	…	$n_{2j}/n_{.j}$	…	$n_{2K}/n_{.K}$	$n_{2.}/n$
	:	:	:	…	:	…	:	:
	y_i	$n_{i1}/n_{.1}$	$n_{i2}/n_{.2}$	…	$n_{ij}/n_{.j}$	…	$n_{iK}/n_{.K}$	$n_{i.}/n$
	:	:	:	…	:	…	:	:
	y_H	$n_{H1}/n_{.1}$	$n_{H2}/n_{.2}$	…	$n_{Hj}/n_{.j}$	…	$n_{HK}/n_{.K}$	$n_{H.}/n$
Tot.		1	1	1	1	1	1	

RMer

Tab.1

Tavola di contingenza Classi di età * Utenza dicotomica

Conteggio

		Utenza dicotomica		Totale
		utenza tradizionale	utenza non tradizionale	
Classi di età	15-18	43	19	62
	19-34	413	264	677
	35-60	269	137	406
	oltre 60	61	61	122
Totale		786	481	1267

Esercizio

A partire dalle frequenze assolute della tabella 1, calcolare i profili % di riga e di colonna.
Commentare le due tabelle ottenute.

Profili % di riga

Tavola di contingenza Classi di età * Utenza dicotomica

% in Classi di età

		Utenza dicotomica		Totale
		utenza tradizionale	utenza non tradizionale	
Classi di età	15-18	69,4%	30,6%	100,0%
	19-34	61,0%	39,0%	100,0%
	35-60	66,3%	33,7%	100,0%
	oltre 60	50,0%	50,0%	100,0%
Totale		62,0%	38,0%	100,0%

Profili % di colonna

Tavola di contingenza Classi di età * Utenza dicotomica

% in Utenza dicotomica

		Utenza dicotomica		Totale
		utenza tradizionale	utenza non tradizionale	
Classi di età	15-18	5,5%	4,0%	4,9%
	19-34	52,5%	54,9%	53,4%
	35-60	34,2%	28,5%	32,0%
	oltre 60	7,8%	12,7%	9,6%
Totale		100,0%	100,0%	100,0%

RMer

Percentuali di riga, di colonna e sul totale … informazioni diverse

Tavola di contingenza Classi di età * Utenza dicotomica

% in Classi di età

		Utenza dicotomica		Totale
		utenza tradizionale	utenza non tradizionale	
Classi di età	15-18	69,4%	30,6%	100,0%
	19-34	61,0%	39,0%	100,0%
	35-60	66,3%	33,7%	100,0%
	oltre 60	50,0%	50,0%	100,0%
Totale		62,0%	38,0%	100,0%

Tavola di contingenza Classi di età * Utenza dicotomica

% in Utenza dicotomica

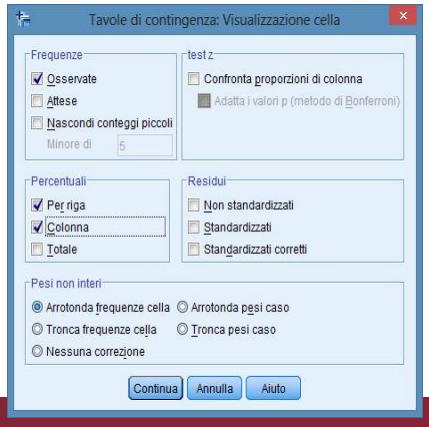
		Utenza dicotomica		Totale
		utenza tradizionale	utenza non tradizionale	
Classi di età	15-18	5,5%	4,0%	4,9%
	19-34	52,5%	54,9%	53,4%
	35-60	34,2%	28,5%	32,0%
	oltre 60	7,8%	12,7%	9,6%
Totale		100,0%	100,0%	100,0%

Tavola di contingenza Classi di età * Utenza dicotomica

% del totale

		Utenza dicotomica		Totale
		utenza tradizionale	utenza non tradizionale	
Classi di età	15-18	3,4%	1,5%	4,9%
	19-34	32,6%	20,8%	53,4%
	35-60	21,2%	10,8%	32,0%
	oltre 60	4,8%	4,8%	9,6%
Totale		62,0%	38,0%	100,0%

Uso del software
Tavole di contingenza



Approfondimento:
Le relazioni Statistiche

L'analisi congiunta di due o più caratteri è utile per studiare le relazioni tra di essi.

- Analisi dell'associazione
  - Indipendenza
  - Interdipendenza
  - Dipendenza

Tipi di relazioni tra caratteri

- **Indipendenza statistica (**relazione simmetrica):
  - Due caratteri sono statisticamente indipendenti quando la conoscenza delle modalità di uno non consente di prevedere le modalità dell'altro
- **Dipendenza** (relazione asimmetrica):
  - Due caratteri sono dipendenti quando si può stabilire un legame unidirezionale tra le modalità di un carattere e quelle di un altro
- **Interdipendenza** (relazione simmetrica) **:**
  - Due caratteri sono interdipendenti quando si può stabilire un legame bidirezionale tra le modalità di un carattere e quelle di un altro.

Indipendenza Statistica

- Due caratteri sono *statisticamente indipendenti* quando la conoscenza di uno dei due caratteri non migliora la "previsione" della modalità dell'altro
- Assenza di qualsiasi legame tra i due caratteri
- Relazione simmetrica: se *X* è indipendente da *Y* allora *Y* è indipendente da *X*

Indipendenza Statistica in una tabella doppia

In una tabella a doppia entrata si ha indipendenza tra i due caratteri X e Y se le distribuzioni relative condizionate di X rispetto alle modalità di Y sono uguali tra loro e alla distribuzione relativa marginale

- Matrice profili riga ha tutte le righe uguali
- Matrice profili colonna ha tutte le colonne uguali

Indipendenza statistica:esempio

Numero di giudizi positivi				
Canale	Spot A	Spot B	Spot C	Totale
RAI1	10	5	15	30
CAN5	14	7	21	42
Totale	24	12	36	72

- *Profili riga %*

*10/30*100=33*
*5/30*100=17*
*15/30*100=50*
*14/42*100=33*
*7/42*100=17*
*21/42*100=50*

Profili di riga %				
Canale	Spot A	Spot B	Spot C	Totale
RAI1	33,33	16,67	50,00	30
CAN5	33,33	16,67	50,00	42
Totale	33,33	16,67	50,00	72

- *Profili colonna %*

*10/24*100=42*
*14/24*100=58*
*5/12*100=42*
*7/12*100=58*
*15/36*100=42*
*21/36*100=58*

Profili di colonna %				
Canale	Spot A	Spot B	Spot C	Totale
RAI1	41,67	41,67	41,67	41,67
CAN5	58,33	58,33	58,33	58,33
Totale	24,00	12,00	36,00	72

2016-2017

Dipendenza perfetta di due caratteri

- In una tabella doppia il carattere Y dipende perfettamente da X se ad ogni modalità di X è associata una sola modalità di Y.
- Se i due caratteri sono perfettamente dipendenti la tabella doppia avrà per ogni riga di X solo una colonna di Y in cui $n_{ij} \neq 0$

Y dipende perfettamente da X				
X | Y	1	2	3	totale
1	0	0	30	30
2	0	20	0	20
3	10	0	0	10
4	0	10	0	10
totale	10	30	30	70

2016-2017

Interdipendenza perfetta di due caratteri

- In una tabella doppia sussiste perfetta interdipendenza se ad ogni modalità di X è associata una sola modalità di Y e viceversa.

Interdipendenza perfetta tra X e Y				
X | Y	1	2	3	totale
1	0	0	30	30
2	0	20	0	20
3	10	0	0	10
totale	10	20	30	60

Esempi di dipendenza perfetta

- Interdipendenza perfetta tra X e Y

	X = Prodotto			
Y = medium	Tablet	Auto	CD audio	Tot
web	82	0	0	82
TV	0	37	0	37
Radio	0	0	5	5
Tot	82	37	5	124

- X dipende perfettamente da Y

	X = Prodotto		
Y = Medium	Tablet	Auto	Tot
Web	82	0	82
TV	0	37	37
Radio	0	51	51
Tot	82	88	170

- Y dipende perfettamente da X

	X = Prodotto			
Y = Canale acquist	Divano	Viaggio	Pasta	Totale
WEB	0	23	0	23
NEGOZIO	41	0	8	49
Totale	41	23	8	72

Situazioni intermedie tra indipendenza e perfetta associazione

Tavola di contingenza titolo di studio * lettura libri negli ultimi 12 mesi

			ttura libri negli ultim 12 mesi		
			no	si	Totale
titolo di studio	laurea	Conteggio	6	46	52
		Conteggio atte	28,6	23,4	52,0
		Residui	-22,6	22,6	
	dipl. univ.	Conteggio	1	17	18
		Conteggio atte	9,9	8,1	18,0
		Residui	-8,9	8,9	
	diploma m. superi	Conteggio	111	177	288
		Conteggio atte	158,2	129,8	288,0
		Residui	-47,2	47,2	
	diploma m. inferio	Conteggio	149	132	281
		Conteggio atte	154,3	126,7	281,0
		Residui	-5,3	5,3	
	licenza elementar	Conteggio	193	62	255
		Conteggio atte	140,1	114,9	255,0
		Residui	52,9	-52,9	
	nessun titolo	Conteggio	81	10	91
		Conteggio atte	50,0	41,0	91,0
		Residui	31,0	-31,0	
Totale		Conteggio	541	444	985
		Conteggio atte	541,0	444,0	985,0

Il grado di associazione (dipendenza o interdipendenza) è tanto maggiore quanto più la tabella osservata si discosta da quella di indipendenza.

Vedremo in seguito come valutare la significatività di questi scostamenti.

Frequenze osservate n_{ij}

Frequenze teoriche di indipendenza n^*_{ij}

Differenze tra Freq. Osserv e freq. teoriche (c_{ij})

Indipendenza o interdipendenza? Esempio

- Le frequenze assolute nell'ipotesi di indipendenza tra i 2 caratteri sono date da

$$n^*_{ij} = \frac{\text{Totale riga} \times \text{Totale colonna}}{\text{Totale us}} = \frac{n_{i.} \times n_{.j}}{n}$$

Frequenza Teorica di Indipendenza

Situazione osservata

Sesso	Liv.Soddisfazione			Totale
	Basso	Medio	Alto	
f	19	5	0	24
m	6	6	4	16
Totale	25	11	4	40

n^*_{11} = 24x25/40=15
n^*_{12} = 24x11/40=6.6
n^*_{13} = 24x4/40=2.4

n^*_{21}=16x25/40=10
n^*_{22}=16x11/40=4.4
n^*_{23}=16x4/40=1.6

Situazione teorica di indipendenza

Sesso	Liv.Soddisfazione			Totale
	Basso	Medio	Alto	
f	15	6,6	2,4	24
m	10	4,4	1,6	16
Totale	25	11	4	40

Differenza tra situazione osservata e situazione teorica : le contingenze

Situazione osservata

Sesso	Liv.Soddisfazione			Totale
	Basso	Medio	Alto	
f	19	5	0	24
m	6	6	4	16
Totale	25	11	4	40

Situazione teorica di indipendenza

Sesso	Liv.Soddisfazione			Totale
	Basso	Medio	Alto	
f	15	6,6	2,4	24
m	10	4,4	1,6	16
Totale	25	11	4	40

Contingenze o Residui

$$n_{ij} - n_{ij}^*$$

c_{11} = 19-15=4
c_{12} = 5-6.6=-1.6
c_{13} =0-2.4=-2.4

C_{21}=6-10=-4
c_{22}=6-4.4=1.6
c_{23}=4-1.6=2.4

Tabella delle Contingenze

Sesso	Liv.Soddisfazione		
	Basso	Medio	Alto
f	4	-1,6	-2,4
m	-4	1,6	2,4

Misura di associazione: il Chi-Quadrato di Pearson

$$\chi^2 = \sum_{i=1}^{H} \sum_{j=1}^{K} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$$

$$\left(n_{ij} - n_{ij}^*\right) = c_{ij}$$

- ✓ Proprietà
  - ✓ Assume valore 0 se X e Y sono perfettamente indipendenti
  - ✓ Assume valore positivo se esiste un legame di dipendenza o interdipendenza tra X e Y
  - ✓ Ha le dimensioni di una frequenza assoluta

Esempio di calcolo del Chi quadrato

$$\chi^2 = \sum_{i=1}^{H}\sum_{j=1}^{K}\frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}}$$

Contingenze

$$\left(n_{ij} - n^*_{ij}\right) = c_{ij}$$

Tabella delle Contingenze

Sesso	Liv.Soddisfazione			Totale
	Basso	Medio	Alto	
f	15	6,6	2,4	24
m	10	4,4	1,6	16
Totale	25	11	4	40

Sesso	Liv.Soddisfazione		
	Basso	Medio	Alto
f	4	-1,6	-2,4
m	-4	1,6	2,4

$$\chi^2 = \frac{4^2}{15} + \frac{(-1.6)^2}{6.6} + \frac{(-2.4)^2}{2.4} + \frac{(-4)^2}{10} + \frac{(1.6)^2}{4.4} + \frac{(2.4)^2}{1.6} =$$
$$= 1.067 + 0.39 + 2.4 + 1.6 + 0.58 + 3.6 = 9.64$$

Come si interpreta il Chi quadrato χ^2

- La differenza fra i valori corrispondenti n_{ij} e n^*_{ij} (valori osservati e valori attesi nell'ipotesi di indipendenza fra le variabili studiate) indica quanto la situazione osservata si discosta da quella di indipendenza:
  - se la differenza è nulla, o è piccola, non c'è relazione tra i caratteri
  - se i valori sono grandi allora si può ipotizzare che c'è una relazione .
- Ma quando questa differenza può essere considerata piccola o grande?
- Per rispondere a questo quesito bisogna conoscere la distribuzione del test statistico del Chi Quadrato, di cui parleremo nelle prossime lezioni sulla statistica inferenziale.

Caratteristiche del Chi quadrato

- Nel calcolo del Chi quadrato il ruolo delle variabili è simmetrico.
- Il Chi quadrato non cambia se le modalità sono ordinate in modo diverso: è un test in cui le variabili sono sempre trattate come qualitative non ordinabili .
- Il Chi quadrato ci dice quanta evidenza c'è a favore della interdipendenza, ma non misura la forza di questa relazione.
- Il valore del Chi quadrato dipende dal numero di unità statistiche, tende a crescere all'aumentare del numero delle righe e delle colonne della tabella di contingenza.

Indici di associazione per tabelle doppie di frequenze

Misurano l'associazione tra due caratteri analizzando la distribuzione congiunta delle frequenze.

- I più comunemente usati sono:
  - L'indice di contingenza quadratica media
  - L'indice V di Cramer
  - L'indice P di Pearson

Glossario

Indice PHI ^{2} (contingenza quadratica media)

- $$\Phi^2 = \frac{\chi^2}{n}$$

✓E' un indice normalizzato (non dipendente dalla numerosità del collettivo)

✓Assume il suo valore minimo, 0, in caso di perfetta indipendenza ossia quando frequenze osservate e frequenze teoriche coincidono

✓Il valore massimo è pari a 1 solo nel caso di tabelle quadrate 2x2, altrimenti è maggiore di 1.

✓Viene spesso utilizzata la sua radice quadrata (PHI).

Glossario

Indice P di Pearson

- $$P = \sqrt{\frac{\Phi^2}{\Phi^2 + 1}}$$

- Varia tra 0 e 1 .
  - Vale 0 nel caso di indipendenza
  - Nel package Spss è definito coefficiente di contingenza

Glossario
Indice V di Cramér

- $$V = \sqrt{\frac{\Phi^2}{\min(R-1;\ C-1)}}$$
  - varia tra 0 e 1
  - Vale 0 nel caso di indipendenza
  - Vale 1 nei seguenti casi:
    - I due caratteri sono perfettamente associati e la tabella è quadrata
    - X dipende perfettamente da Y e il numero di righe è minore di quello delle colonne
    - Y dipende perfettamente da X e il numero di righe è maggiore di quello delle colonne

Esercizio

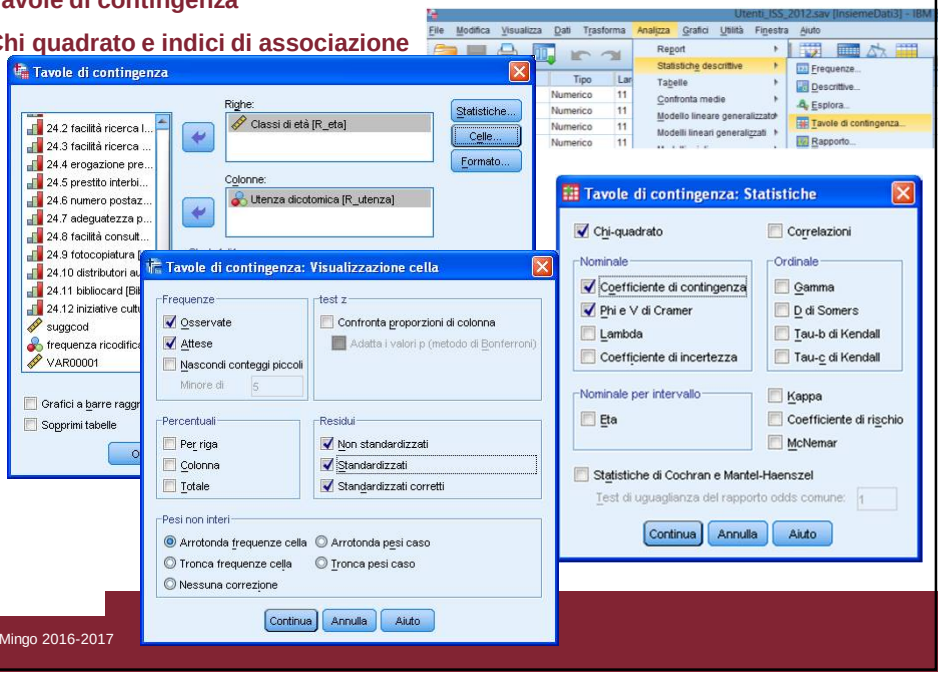
Sapendo che su una tabella di contingenza in cui si riporta la distribuzione doppia di 1000 clienti, incrociando in riga il tipo di Banca utilizzata (modalità: Unicredit, Credito Cooperativo, Banca Agricola Popolare) e la condizione professionale dei clienti (modalità: Imprenditore, Artigiano, Lavoratore dipendente, Libero Professionista) si è ottenuto :
$\chi^2 = 988{,}07$

Calcolare :
- PHI^2 e V di Cramer
- L'indice P di Pearson

calcoli

Uso del software
Tavole di contingenza

Chi quadrato e indici di associazione



I. Mingo 2016-2017

Calcolare le contingenze o residui

Tavola di contingenza Classi di età * Utenza dicotomica

			Utenza dicotomica		
			1,00	2,00	Totale
Classi di età	1,00 15-18	Conteggio	43	19	62
		Conteggio atteso	38,5	23,5	62,0
		Residui	4,5	-4,5	
		Residui stand.	,7	-,9	
		Residui corretti	1,2	-1,2	
	2,00 19-34	Conteggio	413	264	677
		Conteggio atteso	420,0	257,0	677,0
		Residui	-7,0	7,0	
		Residui stand.	-,3	,4	
		Residui corretti	-,8	,8	
	3,00 35-60	Conteggio	269	137	406
		Conteggio atteso	251,9	154,1	406,0
		Residui	17,1	-17,1	
		Residui stand.	1,1	-1,4	
		Residui corretti	2,1	-2,1	
	4,00 oltre 60	Conteggio	61	61	122
		Conteggio atteso	75,7	46,3	122,0
		Residui	-14,7	14,7	
		Residui stand.	-1,7	2,2	
		Residui corretti	-2,9	2,9	
Totale		Conteggio	786	481	1267
		Conteggio atteso	786,0	481,0	1267,0

c_{11}= 43- 38,5=4,5

$s_{11}= 4,5/\sqrt{38,5}=0,7$

$z_{11}= 0,7/\sqrt{(1-62/1267)(1-786/1267)}=1,2$

Residuo = Conteggio - conteggio atteso
Residuo Standardizzato= Residuo/√conteggio atteso
Residuo Corretto= Residuo standardizzato/ √(1- (ni./n..) (1- (n.j ./n..))

Tavola di contingenza Classi di età * Utenza dicotomica

			Utenza dicotomica		
			1,00	2,00	Totale
Classi di età	1,00 15-18	Conteggio	43	19	62
		Conteggio atteso	38,5	23,5	62,0
		Residui	4,5	-4,5	
		Residui stand.	,7	-,9	
		Residui corretti	1,2	-1,2	
	2,00 19-34	Conteggio	413	264	677
		Conteggio atteso	420,0	257,0	677,0
		Residui	-7,0	7,0	
		Residui stand.	-,3	,4	
		Residui corretti	-,8	,8	
	3,00 35-60	Conteggio	269	137	406
		Conteggio atteso	251,9	154,1	406,0
		Residui	17,1	-17,1	
		Residui stand.	1,1	-1,4	
		Residui corretti	2,1	-2,1	
	4,00 oltre 60	Conteggio	61	61	122
		Conteggio atteso	75,7	46,3	122,0
		Residui	-14,7	14,7	
		Residui stand.	-1,7	2,2	
		Residui corretti	-2,9	2,9	
Totale		Conteggio	786	481	1267
		Conteggio atteso	786,0	481,0	1267,0

Chi-quadrato

	Valore	df	Sig. asint. (2 vie)
Chi-quadrato di Pearson	12,291[a]	3	,006

Misure simmetriche

		Valore	Sig. appross.
Nominale per nominale	Phi	,098	,006
	V di Cramer	,098	,006
	Coefficiente di contingenza	,098	,006
N. di casi validi		1267	

a. Senza assumere l'ipotesi nulla.

b. Viene usato l'errore standard asintotico in base all'assunzione dell'ipotesi

Uso del software: interpretare l'output

Analizzando i residuo corretti, si può affermare che si registra un'associazione statisticamente significativa, [con un livello di probabilità del 95%] , per le modalità che presentano un residuo corretto z>=|1,96| :

- se z è positivo la cella presenta un numero di casi significativamente più elevato di quello che si otterrebbe nel caso di indipendenza tra le modalità.

- se z è negativo, la cella presenta un numero di casi significativamente più ridotto di quello che si otterrebbe nel caso di indipendenza tra le modalità.

Spiegheremo più avanti perché!

Esercizio

Sulla base dei dati della tabella seguente, calcolare il Chi quadrato tra le variabili sesso e valutazione della completezza del patrimonio librario degli utenti delle biblioteche .
Quale indicazione si può trarre dai residui corretti?

Tavola di contingenza 23.1 completezza libri * sesso

			sesso		
			maschi	femmine	Totale
23.1 completezza libri	per niente gradito	Conteggio	15	16	31
		Conteggio atteso	15,2	15,8	31,0
		Residui	-,2	,2	
		Residui corretti	-,1	,1	
	poco gradito	Conteggio	119	125	244
		Conteggio atteso	119,4	124,6	244,0
		Residui	-,4	,4	
		Residui corretti	-,1	,1	
	abbastanza gradito	Conteggio	267	297	564
		Conteggio atteso	276,1	287,9	564,0
		Residui	-9,1	9,1	
		Residui corretti	-1,1	1,1	
	molto gradito	Conteggio	112	97	209
		Conteggio atteso	102,3	106,7	209,0
		Residui	9,7	-9,7	
		Residui corretti	1,5	-1,5	
Totale		Conteggio	513	535	1048
		Conteggio atteso	513,0	535,0	1048,0

$$\chi^2 = \sum_{i=1}^{H} \sum_{j=1}^{K} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$$

calcolo

Indici di associazione per caratteri ordinati

- Se la tabella si riferisce a caratteri ordinati è possibile costruire indici che oltre a misurare l'intensità dell'associazione ne misurano il verso.
- Tra due caratteri ordinati possono sussistere due tipi di relazioni:
  - Relazione diretta (concordanza): a modalità di ordine elevato di un carattere corrispondono più frequentemente modalità di ordine elevato dell'altro carattere.
  - Relazione inversa (discordanza): a modalità elevate di un carattere corrispondono modalità di ordine basso dell'altro carattere e viceversa.

Indici di concordanza e discordanza

- **Possono assumere** :
  - valori positivi , nel caso di concordanza tra i caratteri
  - valori negativi , nel caso di discordanza
- **I più noti**:
  - Indice Gamma di Goodman e Kruskal
  - Indice τ_b di Kendall
  - Indice d di Sommer
  - **Indice rho di Spearman**

Tali indici variano fra -1 e 1
- zero indica assenza di associazione
- +1 indica che l'ordinamento dei due caratteri è sempre concorde
- -1 indica che l'ordinamento è sempre discorde.
- valori prossimi a 1 in valore assoluto indicano forte relazione

Indice rho di Spearman

- E' un indice di cograduazione tra graduatorie, particolarmente indicato quando i caratteri ordinati presentano un numero elevato di modalità.
- Per calcolare l'indice è necessario ordinare gli individui in senso decrescente per ognuno dei due caratteri e attribuire il rango.
- L'indice si definisce come:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

dove d indica la differenza tra i ranghi cioè i posti nelle due graduatorie ordinate.

- L'indice assume valori tra -1 e + 1
  - Il valore 0 implica indipendenza tra x e y
  - L'opposta graduatoria ($\rho = -1$) implica discordanza tra x e y .
  - E' uguale ad 1 quando le unità presentano lo stesso rango in entrambe le graduatorie cioè nel caso di perfetta cograduazione.

Cograduazione: esempio

	Livello territorale	grad. attiv. fem m.	grad. Tasso disocc.	d	d2
1	Piemonte	5	12	-7	49
2	Valle d'Aosta	1	18	-17	289
3	Liguria	12	9	3	9
4	Lombardia	4	16	-12	144
5	Trentino Alto Adige	3	20	-17	289
6	Friuli Venezia Giulia	9	15	-6	36
7	Veneto	6	19	-13	169
8	Emilia Romagna	2	17	-15	225
9	Marche	7	14	-7	49
10	Toscana	8	13	-5	25
11	Umbria	10	11	-1	1
12	Lazio	11	8	3	9
13	Campania	18	3	15	225
14	Abruzzo	14	10	4	16
15	Molise	13	7	6	36
16	Puglia	19	5	14	196
17	Basilicata	16	6	10	100
18	Calabria	17	1	16	256
19	Sicilia	20	2	18	324
20	Sardegna	15	4	11	121
					2568

$\rho = 1 - [6*2568)/20*(400-1)] = -0,931$

Misure simmetriche

		Valore
Ordinale per ordinale	Tau-b di Kendall	-,779
	Gamma	-,779
	Correlazione di Spearman	-,931
N. di casi validi		20

Esercizio

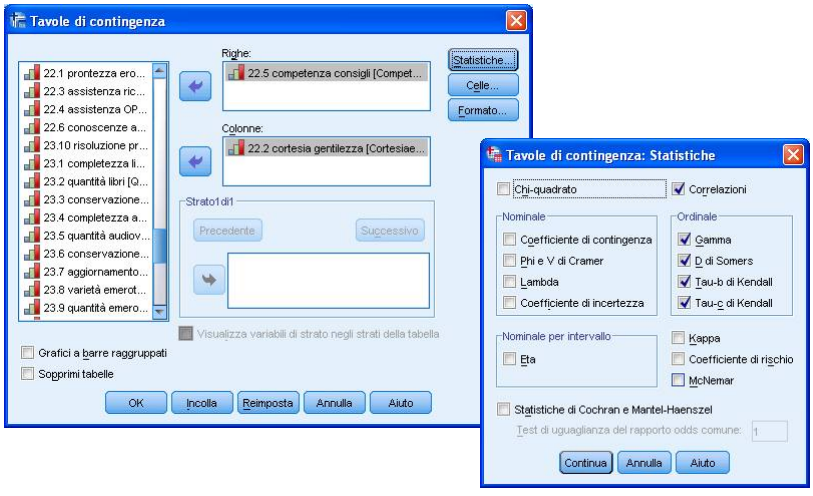
Calcolare il coefficiente di graduazione tra le valutazioni dei clienti riguardo all' assistenza post vendita e alla consulenza alla vendita rilevate per ripartizione geografica

Valutazioni medie (punteggi 1 min -10 max)		
	Assistenza post vendita	Consulenza alla vendita
Nord-ovest	7	9
Nord-est	9,4	7,8
Centro	8	8,5
Sud	7,5	7
Isole	8,5	7,5

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

calcoli

Indici di concordanza e discordanza : uso del software



Indici di concordanza e discordanza: uso del software

Tavola di contingenza 22.5 competenza consigli * 22.2 cortesia gentilezza

Conteggio

		22.2 cortesia gentilezza				
		1 per niente gradito	2 poco gradito	3 abbastanza gradito	4 molto gradito	Totale
22.5 competenza consigli	1 per niente gradito	10	14	10	3	37
	2 poco gradito	10	20	43	25	98
	3 abbastanza gradito	4	17	161	129	311
	4 molto gradito	2	4	36	336	378
Totale		26	55	250	493	824

Misure simmetriche

		Valore	E.S.
Ordinale per ordinale	Tau-b di Kendall	,552	
	Tau-c di Kendall	,431	
	Gamma	,786	
	Correlazione di Spearman	,588	
Intervallo per intervallo	R di Pearson	,594	
N. di casi validi		824	

Anche per la correlazione di Spearman esistono dei test statistici per verificare se la correlazione calcolata è stata casualmente estratta da una popolazione con correlazione nulla. Di questi test ci occuperemo nella parte sulla statistica inferenziale.