Statistica

Associazione tra variabli categoriche

Andrea Giommi

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze

Scuola di Psicologia
Corso di Studio in Scienze e Tecniche Psicologiche

Analisi dell'associazione tra due variabili qualitative

- Consideriamo due variabili entrambe categoriche (qualitative); consideriamo una delle due come variabile risposta e l'altra come variabile esplicativa

- Diciamo che c'è **associazione** tra le due variabili se, *nella popolazione*, la distribuzione condizionata della variabile risposta è diversa in corrispondenza delle diverse modalità della variabile esplicativa

Esempio I: tavola di contingenza su genere e orientamento politico

	Orientamento politico			
Genere	Democratici	Indipendenti	Repubblicani	Totale
Femmine	573 (38%)	516 (34%)	422 (28%)	1511
Maschi	386 (31%)	475 (38%)	399 (32%)	1260
Totale	959 (35%)	991 (36%)	821 (30%)	2771

Variabile risposta: Orientamento politico

Variabile esplicativa: Genere

Nel campione, la distribuzione condizionata dell'*Orientamento politico* varia al variare del*Genere*. É possibile concludere che questo avviene anche nella popolazione da cui il campione proviene?

Esempio II: tavola di contingenza su: reddito e felicità

Reddito	Felicità			
	Molto	Abbastanza	Poco	Totale
Sopra la media	272 (44%)	294 (48%)	49 (8%)	615
In media	454 (32%)	835 (59%)	131 (9%)	1420
Sotto la media	185 (20%)	527 (57%)	208 (23%)	920

Variabile risposta: Felicità

Variabile esplicativa: Reddito familiare

Anche in questo caso, è possibile concludere che nella popolazione esiste un'associazione tra *Felicità* e categoria di *Reddito*?

Indipendenza e Dipendenza

- **Indipendenza statistica:** nella popolazione la distribuzione condizionata di ciascuna variabile è costante per tutte le modalità dell'altra
- **Dipendenza** (associazione): Nella popolazione le distribuzioni condizionate non sono tutte uguali

Esempio di indipendenza statistica:

	Felicità		
Reddito	Molto	Abbastanza	Poco
Sopra la media	32%	55%	13%
In media	32%	55%	13%
Sotto la media	32%	55%	13%

Test di indipendenza Chi-Quadro

- H_0: Le variabili sono statisticamente indipendenti
- H_a: C'è associazione statistica tra le variabili
- Logica sottostante il test: sintetizzare opportunamente le differenze tra i valori *osservati* nelle varie celle della tavola di contingenza e le frequenze *teoriche*; cioè le frequenze attese nel caso che H_0 sia vera.
- Indichiamo con:

  f_o le frequenze osservate
  f_e le frequenze attese
  r il numero di righe della tabella di contingenza
  c il numero delle colonne

Frequenze teoriche (f_e)

- Le frequenze teoriche (attese) hanno distribuzione condizionate costanti e uguali alla corrispondente distribuzione marginale
- Hanno la stessa distribuzione marginale (sia di riga che di colonna) delle frequenze osservate
- Si possono facilmente calcolare così:

$$f_e = (\text{totale di riga})(\text{totale di colonna})/n$$

Frequenze teoriche per la tavola "Felicità - Reddito"

Reddito	Felicità			
	Molto	Abbastanza	Poco	Totale
Sopra la media	272 (189.6)	294 (344.6)	49 (80.8)	615
In media	454 (437.8)	835 (795.8)	131 (186.5)	1420
Sotto la media	185 (283.6)	527 (515.6)	208 (120.8)	920
Totale	911	1656	388	2955

Esempio: calcolo frequenze attese prima colonna

$f_e = 189.6 = (615)(911)/2955$

$f_e = 437.8 = (1420)(911)/2955$

$f_e = 283.6 = (920)(911)/2955$

Statistica test Chi-Quadro

- Le differenze tra frequenze osservate e teoriche vengono sintetizzate mediante la statistica:

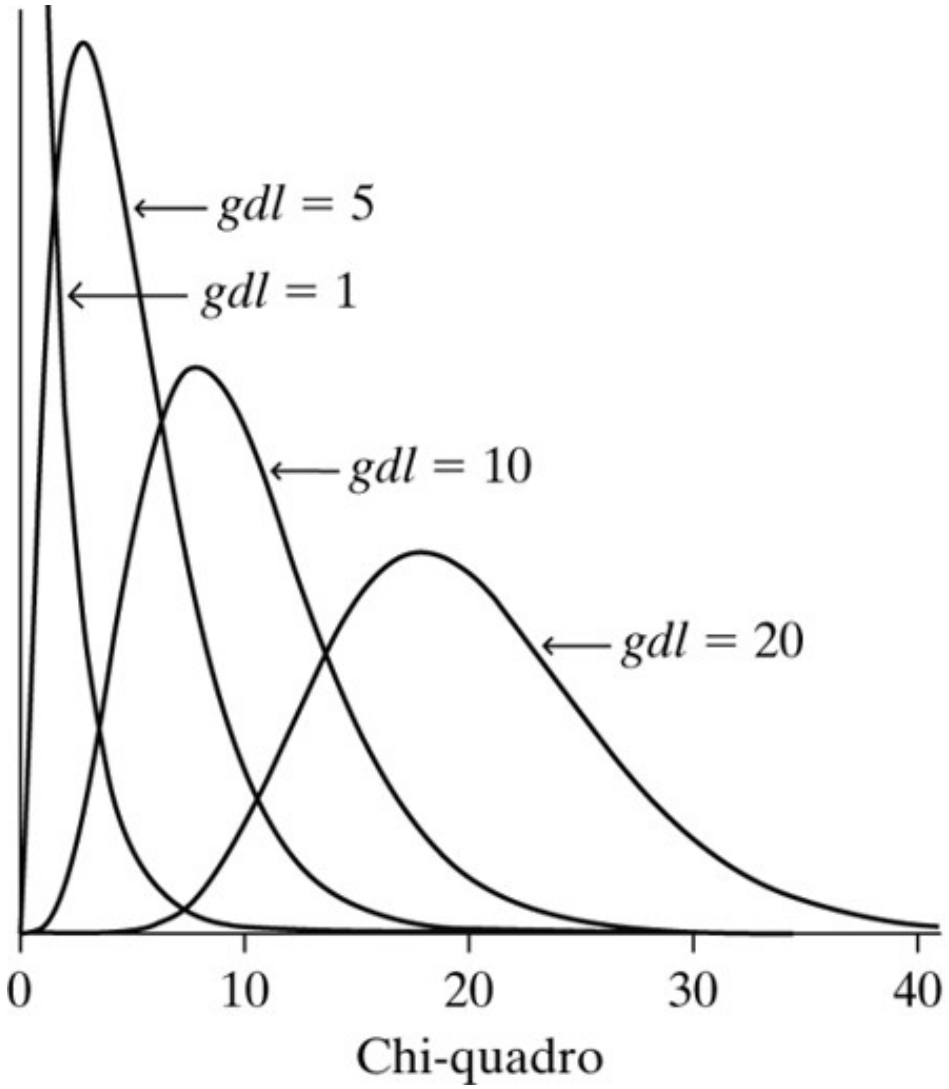
$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

  con la somma estesa a tutte le celle della tabella di contingenza
- Quando H_0 è vera, la di stribuzione campionaria di questa statistica si approssima (per n sufficientemente grande) alla distribuzione di probabilità Chi-Quadro.

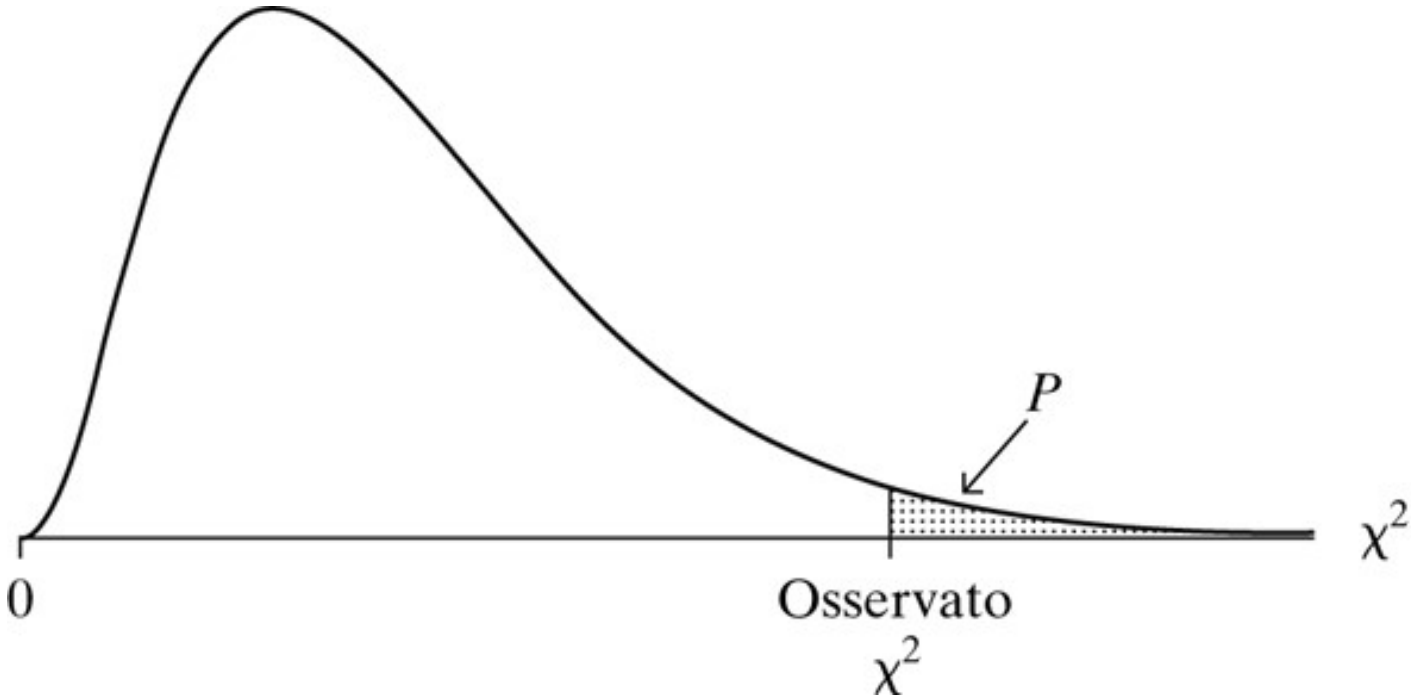
Proprietà della distribuzione Chi-Quadro

- E' definita sulla parte positiva dell'asse reale
- E' asimmetrica positiva (tende a diventare simmetrica al crescere della dimensione campionaria)
- Media e varianza dipendono entrambe dalle dimensioni della tavola di contingenza attraverso i gradi di libertà (gdl):
  $gdl = (r-1)(c-1)$ = media della distribuzione
  $2gdl$ = varianza della distribuzione
  (r = numero delle righe; c = numero delle colonne)
- Valori elevati della statistica χ^2 sono improbabili sotto l'ipotesi H_0; pertanto il P-valore è rappresentato dai valori, nella coda destra della distribuzione, maggiori della statistica test osservata.

Distribuzione chi-quadro



P-valore nel test di indipendenza



Calcolo del chi-quadro sulla tavola"Felicità-Reddito"

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(272 - 189,6)^2}{189,6} + ... = 172,3$$

$gdl = (3-1)(3-1) = 4$ e P-valore $< 0{,}001$

- C'è un'evidenza molto forte contro l'ipotesi nulla H_0: Indipendenza tra le variabili (se H_0 fosse vera, avremmo una probabilità $< 0,001$ di osservare un valore della statistica χ^2 maggiore o uguale a quello osservato, cioè 172,3)
- O è capitato un evento estremamente improbabile o l'ipotesi H_0 non è corretta
- Di conseguenza: al livello di significatività $\alpha = 0,05$ (o $\alpha = 0,01$ o $\alpha = 0,001$), si respinge l'ipotesi H_0 e si conclude che, nella popolazione, c'è associazione tra felicità e reddito

Commenti sul test Chi-Quadro

- L'utilizzazione della distribuzione chi-quadro per approssimare l'effettiva distribuzione della statistica test è adeguata per grandi campioni. Grande significa che per tutte o quasi tutte le celle $f_e \geq 5$
- Per piccoli campioni si applica il **test esatto di Fisher** (non in programma)
- Il Chi-Quadro tratta le variabili come *qualitative nominali*. Per variabili di tipo ordinale o quantitative sono utilizzabili altre tecniche di analisi.

Commenti sul test Chi-Quadro

- $Gdl = (r-1)(c-1)$ significa che dati i marginali di una tabella si possono fissare liberamente solo $(r-1)(c-1)$ valori nella celle; gli altri sono determinati di conseguenza.
- Se z è una statistica con distribuzione normale standardizzata, allora z^2 ha una distribuzione chi-quadro con $gdl = 1$
- La somma di d variabili indipendenti z, con distribuzione normale standardizzata, prese al quadrato, ha una distribuzione chi-quadro con $gdl = d$

Tavole 2×2

- Per le tavole 2×2 il test chi-quadro di indipendenza (con $gdl = 1$) è equivalente a sottoporre a test l'ipotesi:
  $H_0 : \pi_1 = \pi_2$ (confronto tra proporzioni di due popolazioni)

	Variabile risposta	
Gruppo	Risultato 1	Risultato 2
1	π_1	$1 - \pi_1$
2	π_2	$1 - \pi_2$

  $H_0 : \pi_1 = \pi_2$ è equivalente a:

  H_0 : la variabile risposta è indipendente dalla variabile "gruppo"

Pertanto, la statistica χ^2 è il quadrato della statistica test z:

$$z = (\pi_1 - \pi_2)/(es_0)$$

Esempio Studio sull'effetto dell'alcohol, svolto dalla Harvard School of Public Health

- Nel 1993 e, successivamente nel 2001, fu posta ad un campione molto ampio di studenti la seguente domanda: "Hai avuto rapporti sessuali non programmati a seguito dell'assunzione di bevande alcoliche?"

I risultati sono sintetizzati nella seguente tavola 2×2:

	Risposte		
Anno	SI	NO	Totale
1993	$2440(19,2\%)$	$10268(81,8\%)$	12708
2001	$1871(21,3\%)$	$6912(78,7\%)$	8783

- $\chi^2 = 14,3, gdl = 1,\ (P - valore = 0,00016)$
- $z = (\pi_1 - \pi_2)/(es_0) = 3,78;\ (3,78)^2 = 14,3$

Individuare la struttura dell'associazione: **Residui**

- Elevati valori del χ^2 mostrano una forte evidenza che vi sia associazione tra le variabili, ma non forniscono alcun elemento né sulla **struttura** dell'associazione né sulla sua **forza**
- Possiamo indagare sulla *struttura* dell'associazione utilizzando i **residui** (contingenze) in ciascuna cella della tavola
- Residuo $= f_o - f_e$. I residui sono positivi (negativi) se le frequenze osservate sono maggiori (minori) di quelle attese sotto l'ipotesi di indipendenza
- Residui standardizzati: $z = f_o - f_e/es$; (con es = errore standard del residuo). Misurano quanti errori standard la differenza $f_o - f_e$ si allontana da 0 sotto l'ipotesi H_0

Individuare la struttura dell'associazione: **Residui**

- L'errore standard del residuo è definito come:

$$es = \sqrt{f_e(1 - \textit{proporzione di riga})(1 - \textit{proporzione di colonna})}$$

- E, pertanto, il residuo standardizzato $z = f_o - f_e/es$ è dato da:

$$z = \frac{f_o - f_e}{\sqrt{f_e(1 - \textit{proporzione di riga})(1 - \textit{proporzione di colonna})}}$$

Esempio: calcolo dei residui standardizzati sulla tavola felicità vs reddito

- Per la cella con freqenza osservata 272 e frequenza attesa 198,6:
  - Proporzione di riga $= 615/2955 = 0{,}208$
  - Proporzione di colonna $= 911/2955 = 0{,}308$
- E il residuo standardizzato è:

$$z = \frac{(272 - 198,6)}{\sqrt{198,6(1 - 0,208)(1 - 0,308)}} = 8,1$$

- Quindi il numero di persone molto felici con reddito familiare sopra la media, è circa 8 errori standard più grande di quanto ci saremmo aspettati sotto l'ipotesi di indipendenza tra reddito e felicità

Esempio: calcolo dei residui standardizzati sulla tavola felicità vs reddito

- In modo analogo possiamo osservare più persone di quelle attese nella cella "non troppo felice con reddito sotto la media" e un minor numero di persone rispetto all'aspettativa per le celle "molto felice con reddito sotto la media" e "non troppo felice con reddito sopra la media"
- Nelle tavole 2×2 ciascun residuo standardizzato ha lo stesso valore assoluto e soddisfa la seguente uguaglianza:

$$z^2 = \chi^2$$

(poiché $gdl = 1$, calcolato un residuo gli altri sono determinati di conseguenza)

...

Commenti sui residui standardizzati

- Quando l'ipotesi di indipendenza è vera, i residui standardizzati hanno *distribuzione normale standardizzata* con media 0 e varianza 1
- Un residuo standardizzato > 2 o < -2 indica un significativo allontanamento dal valore che ci aspetteremmo per la relativa combinazione di modalità, se l'ipotesi nulla fosse vera
- Questo capiterà infatti, per effetto del caso, solo all'incirca 5 volte su 100
- Se il residuo standardizzato è > 3 o < -3 allora per quella cella (combinazione di modalità) vi è una evidenza molto forte di associazione

Misura dell'associazione

- Il test Chi-quadro risponde alla domanda: "*C'è associazione tra due variabili?*".
- I residui standardizzati ci aiutano a comprendere la struttura dell'associazione e rispondono alla domanda: "*Quanto i dati osservati si allontanano dalla situazione di indipendenza?*"
- Se ci chiediamo: "*Quanto forte è l'associazione tra due variabili?*" possiamo rispondere utilizzando la **differenza tra proporzioni**

Esempio: Opinione sull'operato di G. W. Bush come presidente (sondaggio Gallup del 2008 su circa 1000 adulti)

	Opinione	
Partito politico	Approva	Disapprova
Democratici	3%	97%
Repubblicani	64%	36%

Genere	Approva	Disapprova
Donne	24%	76%
Uomini	27%	73%

La differenza tra proporzioni $0,64 - 0,03 = 0,61$ mostra, tra *opinione* e *partito politico*, un'associazione molto maggiore di quella tra *genere* e *partito politico*: $0,27 - 0,24 = 0,03$

Differenza tra proporzioni

- Più grande la differenza tra $|\hat{\pi}_2 - \hat{\pi}_1|$ maggiore l'associazione
- La differenza tra proporzioni si applica in generale a tavole $r \times c$. Ad esempio:

	Felicità		
Reddito	molto	abbastanza	non troppo
Sopra la media	272 (44%)	294 (48%)	49 (8%)
In media	454 (32%)	835 (59%)	131 (9%)
Sotto la media	185 (20%)	527 (57%)	208 (23%)

- Confrontando quelli che hanno un reddito sopra la media con quelli che sono al di sotto, la differenza tra le proporzioni stimate è, per i molto felici: $0,44 - 0,20 = 0,24$ e, per i non troppo, $0,23 - 0,08 = 0,15$

Confronto mediante rapporti

- Due proporzioni possono essere confrontate mediante un rapporto (rischio relativo) anziché una differenza
- Nell'esempio precedente i due confronti effetuati con differenze danno luogo a:

$$0,44/0,20 = 2,2$$

$$0,23/0,08 = 2,875$$

- Un procedimento alternativo per confrontare proporzioni, tipico delle tavole 2×2 è rappresentato dall' *odds ratio*

Gli Odds

- Per due possibili risultati di una variabile, che definiamo convenzionalmente "successo" e "insuccesso" l'odds è dato dal rapporto:

$$Odds = P(successo)/P(insuccesso) = P(successo)/[1-P(successo)]$$

Per esempio:
se $P(successo) = 0,8, P(insuccesso) = 0,2$
$odds = 0,8/0,2 = 4,0$
se $P(successo) = 0,2, P(insuccesso) = 0,8$
$odds = 0,2/0,8 = \frac{1}{4} = 0,25$

- Da un odds è possibile risalire alla probabilità di successo:

$$\text{Probabilità} = (odds/(odds + 1))$$

Es: $odds = 4,0 \Rightarrow P(successo) = 4/(4 + 1) = 4/5 = 0,8$

Odds Ratio

- In una tavola 2×2:

$$odds\ ratio = \theta = \frac{(odds\ riga\ 1)}{(odds\ riga\ 2)}$$

Esempio: Indagine sugli studenti dell'ultimo anno di scuola superiore

	Alcol	
Sigarette	Sì	No
Sì	1449	46
No	500	281

Odds Ratio

- $\chi^2 = 451,4; gdl = 1, (P - valore = 0,0000)$
- Residui standardizzati pari a +21,2 e -21,2
- Per i fumatori, l'odds di bere alcolici è pari a $1449/46 = 31,50$
- Per i non fumatori l'odds di bere alcolici è pari a $500/281 = 1,78$
- L'*odds ratio* è pari a $31,5/1,78 = 17,7$

Gli odds stimati che un fumatore faccia uso di alcolici è pari a 17,7 volte gli odds che faccia uso di alcolici un non fumatore

Propietà dell'odds ratio

- Il suo valore non dipende da quale variabile è scelta come esplicativa o risposta
  Es: l'odds ratio stimato che chi beve alcolici sia un fumatore è:

$$(1449/500)/(46/281) = 2,90/0,163 = 17,7$$

- Assume valori non negativi ed è pari a 1,0 in assenza di associazione; mentre l'associazione è tanto maggiore quanto più il suo valore si allontana da 1,0
- Può essere calcolato come rapporto dei prodotti in croce nella tavola. Nell'esempio su alcol e fumo:

$$\theta = (1449)(281)/(46)(500) = 17,7$$

- Si noti che l'odds ratio è un rapporto tra *odds* e non tra proporzioni come il *rischio relativo*

Limiti del test chi-quadro

- Il test chi-quadro "misura" esclusivamente l'evidenza a favore della presenza di associazione tra le variabili
- Non ci dice niente riguardo alla struttura dell'associazione (per la quale sono utili i residui standardizzati)
- Non ci dice niente riguardo alla forza dell'associazione (che può essere evidenziata dalle differenza tra proporzioni, i rapporti tra proporzioni o gli odds ratio)

Elevati valori del chi-quadro e piccoli P-valori indicano forte evidenza di associazione, ma non necessariamente una forte associazione

Esempio: Effetto di n sul valore del chi-quadro per un dato grado di associazione

	Risposta							
	1	2	1	2	1	2	1	2
Gruppo 1	15	10	30	20	60	40	600	400
Gruppo 2	10	15	20	30	40	60	400	600
χ^2 :	2		4		8		80	
$(gdl = 1)$								
$P - valore$:	0,16		0,046		0,005		$3,7 \times 10^{-19}$	

- Si noti che $\hat{\pi}_1 - \hat{\pi}_2$= 0,60 - 0,40 = 0,20 per ciascuna tavola
- Con n sufficientemente grande è possibile ottenere un alto valore di chi-quadro (e quindi un piccolo P-valore) anche in corrispondenza di *debole* associazione

Esempio: un piccolo P-valore non implica forte associazione

	Risposta	
	1	2
Gruppo 1	5100	4900
Gruppo 2	4900	5100

- Chi-quadro $\chi^2 = 0,51; (gdl = 1); P - valore = 0,005$
- Si noti che $\hat{\pi}_1 - \hat{\pi}_2$= 0,51 - 0,49 = 0,02 (associazione molto debole)

Nonostante la forte evidenza di associazione, questa sembra essere davvero molto debole

Associazione tra variabili ordinali

- Il chi-quadro si basa esclusivamente su frequenze osservate, f_o e frequenze attese f_e (sotto H_0). Non utilizza in alcun modo le modalità delle variabili in tabella
- Altre misure di analisi so possibili se le variabili sono di natura ordinale o quantitativa
- Per variabili quantitative il Cap. 9 descrive l'analisi della regressione e della correlazione; Per quelle ordinali esistono misure dell'associazione che si basano sui concetti di: **concordanza** e **discordanza**

Concordanza e discordanza

- **Concordanza**: c'è concordanza tra due caratteri per tutti quei soggetti che occupano una posizione elevata per ambedue i caratteri o una posizione non elevata per ambedue

- **Discordanza**: c'è discordanza per tutti i soggetti che occupano una posizione elevata per uno dei due caratteri e non elevata per l'atro

Concordanza e discordanza prefigurano due tipi di associazione: rispettivamente *positiva* e *negativa*. In particolare, possiamo ragionevolmente pensare che due caratteri ordinali in una tavola di contingenza siano associati positivamente se nella tavola le *coppie* di soggetti concordanti sono in prevalenza; associazione negativa se prevalgono le coppie di discordanti; nessuna associazione se le coppie di concordanti e di discordanti si compensano

Esempio: tavola su felicità e reddito

- Dalla tavola su felicità e reddito consideriamo (per semplificare i calcoli) un sottocampione di 296 soggetti
- Ordiniamo le modalità dei due caratteri in senso crescente dall'alto in basso e da sinistra

	Felicità			
Reddito	Poco	Abbastanza	Molto	Totale
Sotto la media	21	53	19	93
In media	13	84	45	142
Sopra la media	5	29	27	61
Totale	39	166	91	296

Analisi "Felicità-Reddito" basata su coppie concordanti e discordanti

- La struttura associativa della tavola è sostanzialmente la stessa di quella con 2955 soggetti
- Da un valore di $\chi^2 =$ 172,3 passiamo ad un valore di chi-quadro di circa dieci volte inferiore: $\chi^2 =$ 17,06 con un corrispondente *P*-valore = 0,00019
- C'è ancora una forte evidenza di associazione tra le variabili ma struttura e forza del'associazione devono essere studiate atraverso i *residui standardizzati* e i *confronto tra proporzioni*
- Sfruttando la scala ordinale delle due variabili possiamo ottenere informazioni sulla struttura e sulla forza dell'associazione costruendo un unico indice statistico

Conteggio delle coppie concordanti e discordanti

Denotiamo con C il numero complessivo di coppie con osservazioni concordanti e con D quelle con osservazioni discordanti. Il procedimento per il calcolo delle coppie concordanti e discordanti è illustrato dalle seguenti tavole:

	PF	*AF*	*MF*
$<$ M	21		
$=$ M		84	45
$>$ M		29	27

	PF	*AF*	*MF*
		53	
			45
			27

	PF	*AF*	*MF*
	13		
		29	27

	PF	*AF*	*MF*
		84	
			27

	PF	*AF*	*MF*
$<$M			19
$=$M	13	84	
$>$M	5	29	

	PF	*AF*	*MF*
	5	29	
			45

	PF	*AF*	*MF*
		53	
	13		
	5		

	PF	*AF*	*MF*
		84	
	5		

Conteggio delle coppie concordanti e discordanti

- Nelle prime 4 tavole si calcola il numero delle coppie concordanti, nelle successive quello delle coppie discordanri:
  - Ciascuno dei 21 soggetti con reddito sotto media ($< M$) e poco felici (PF) è concordante con ciascuno dei soggetti in tabella che abbia un reddito maggiore e un maggiore livello di felicità. Lo stesso vale per i 53, 13, 84, delle altre tre tabelle
  - Si procede in modo del tutto analogo (nelle quattro tabelle successive) partendo dai 19 soggetti molto felici (MF), ma con reddito sotto media e poi proseguendo per gli atri 45, 53, 84:
- Risultato:

$$C = 21 \times (84 + 45 + 29 + 27) + 53 \times (45 + 27) + 13 \times (29 + 27) + 84 \times 27 = 10697$$

$$D = 19 \times (13 + 84 + 5 + 29) + 45 \times (5 + 29) + 53 \times (13 + 5) + 84 \times 5 = 5393$$

Indice Gamma

- Un indice statistico può scaturire dal confronto (differenza o rapporto) tra C e D
- L'indice Gamma è dato dal rapporto tra la differenza tra C e D e la loro somma:

$$\hat{\gamma} = \frac{(C - D)}{(C + D)}$$

- E' facile verificare che:
  - (i) $-1 \leq \gamma \leq 1$; ed è uguale a 0 se $C = D$
  - (ii) Il segno "$+$" o "$-$" mostra se l'associazione è positiva o negativa
  - (iii) Più alto il valore assoluto di gamma e più forte è l'associazione
  - (iv) Se: $\hat{\gamma} = 0$, c'è *assenza di associazione* ma non necessariamente *indipendenza statistica* tra le variabili
- Nell'esempio:

$$\hat{\gamma} = \frac{10697 - 5393}{10697 + 5393} = 0,33$$