Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia,
Carlo Natale Lauro, Marina Marino

UTET

5 Analisi delle relazioni tra due caratteri

Rappresentazione congiunta di una coppia di fenomeni statistici: Distribuzioni doppie di frequenze

Analisi delle relazioni tra due caratteri

Misure di dipendenza

Le relazioni fra variabili quantitative

Le relazioni lineari

Capitolo a cura di
M.G. Grassia, C.N. Lauro, M.Marino

5 Analisi delle relazioni tra due caratteri

Rappresentazione congiunta di una coppia di fenomeni statistici: Distribuzioni doppie di frequenze

Analisi delle relazioni tra due caratteri

Misure di dipendenza

Le relazioni fra variabili quantitative

Le relazioni lineari

Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia, Carlo Natale Lauro, Marina Marino

UTET

Capitolo a cura di
M.G. Grassia, C.N. Lauro, M.Marino

Analisi delle relazioni

Quale indice misura la relazione tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

	2 mutabili	1 variabile 1 mutabile	2 variabili
Approccio simmetrico (interdipendenza)	χ^2, Φ^2, V	χ^2, Φ^2, V	ρ
Approccio asimmetrico (dipendenza)	λ	η^2	Modello di Regressione

Analisi delle relazioni

Quale indice misura la relazione tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

	2 mutabili	1 variabile 1 mutabile	2 variabili
Approccio simmetrico (interdipendenza)			
Approccio asimmetrico (dipendenza)		η^2	

Associazione tra variabili

Le tabelle miste

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Se vogliamo solo determinare se esiste relazione tra i caratteri possiamo calcolare l'indice chi-quadrato o l'indice V-Cramer

Chi-quadrato

	Valore	df
Chi-quadrato di Pearson	128,505[a]	6

Misure simmetriche

	Valore
Phi	1,117
V di Cramer	,790

Associazione tra variabili

Le tabelle miste

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Se invece vogliamo determinare una misura della dipendenza della variabile reddito dalle diverse zone geografiche possiamo confrontare le distribuzioni condizionate del carattere Y (Reddito) in corrispondenza delle diverse modalità del carattere X (Zona Geografica).

Associazione tra variabili

Indipendenza – dipendenza in media

Supponiamo di avere una variabile quantitativa Y ed una variabile X, che può essere sia di natura qualitativa che quantitativa.

Se vogliamo studiare quanto Y dipende in media da X, è opportuno effettuare uno studio sulla **Dipendenza in media**.
In questo caso l'analisi della dipendenza può essere condotta confrontando le distribuzioni condizionate del carattere Y in corrispondenza delle diverse modalità del carattere X.
La **Media Condizionata** di un carattere quantitativo Y rispetto alla i-esima modalità di un carattere X è:

$$M(Y|X=x_i) = \frac{1}{n_{i.}} \sum_{j=1}^{k} y_j \cdot n_{ij}$$

Y è indipendente in media da X se la distribuzione della variabile Y, condizionatamente alle modalità della X, non varia. In altre parole, Y è indipendente in media da X se tutte le medie condizionate di Y sono tra loro uguali e uguali anche alla **Media generale** di Y:

$$M(Y) = \frac{1}{N} \sum_{j=1}^{k} y_j \cdot n_{.j}$$

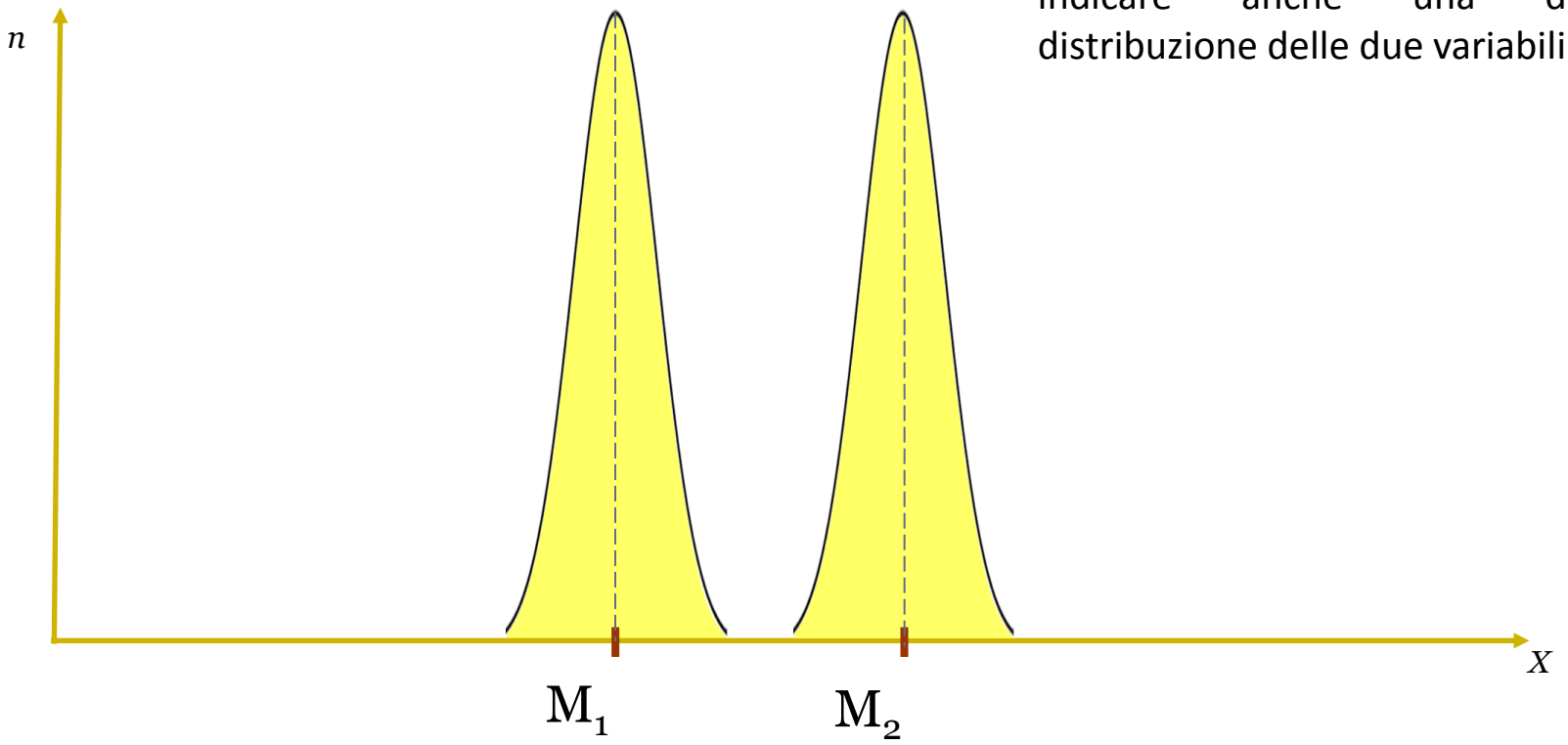
Associazione tra variabili

La valutazione della differenza fra due medie deve tenere conto della variabilità delle distribuzioni attorno ai valori medi stessi.

Associazione tra variabili

La valutazione della differenza fra due medie deve tenere conto della variabilità delle distribuzioni attorno ai valori medi stessi.

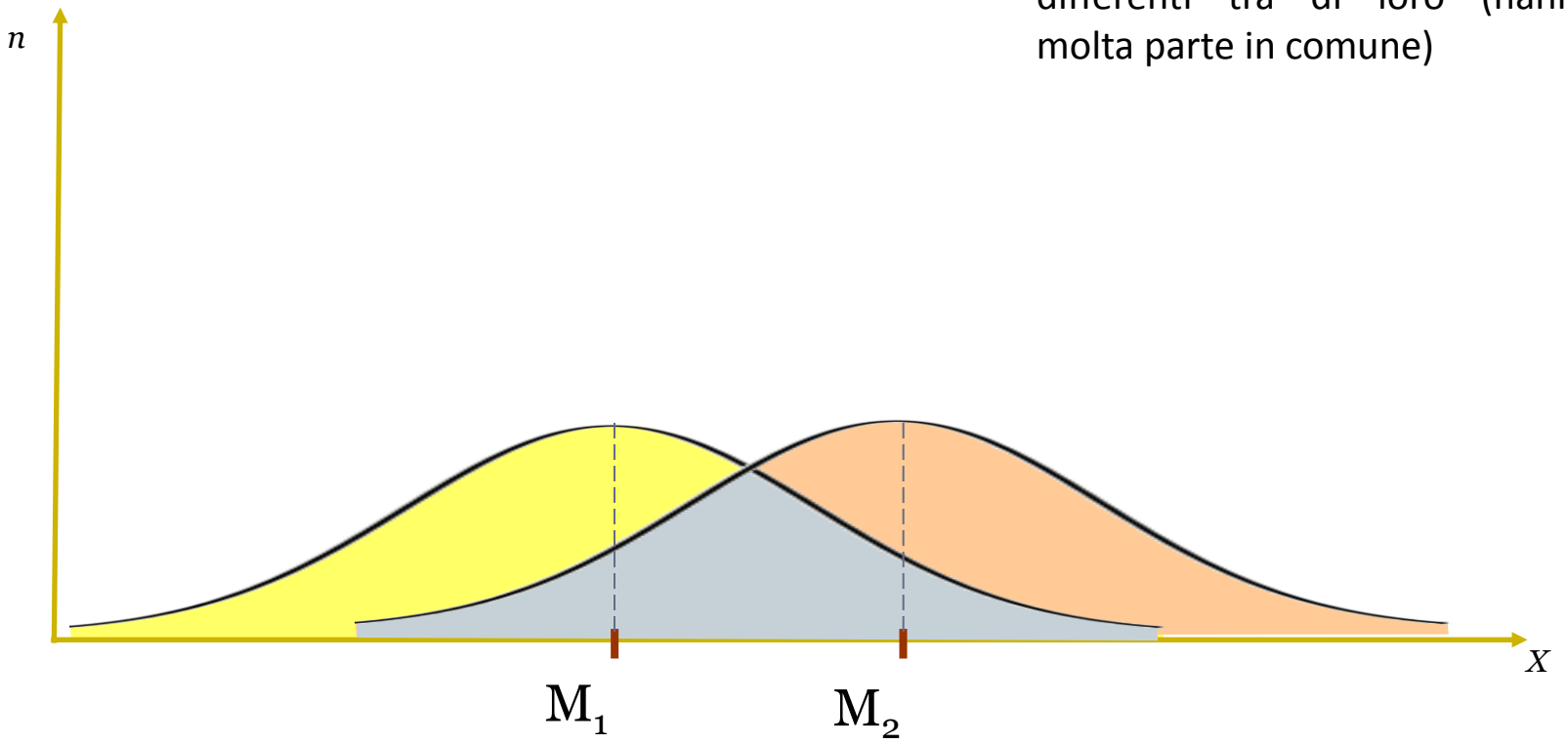
Se c'è poca variabilità, allora il fatto che le medie siano differenti sta ad indicare anche una diversa distribuzione delle due variabili



Associazione tra variabili

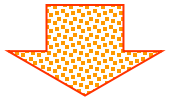
La valutazione della differenza fra due medie deve tenere conto della variabilità delle distribuzioni attorno ai valori medi stessi.

Se c'è molta variabilità, allora le due distribuzioni non sono molto differenti tra di loro (hanno molta parte in comune)

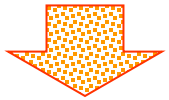


Associazione tra variabili

La valutazione della differenza fra due medie deve tenere conto della variabilità delle distribuzioni attorno ai valori medi stessi.



Per costruire un indice in grado di quantificare l'allontanamento dalla indipendenza in media si deve considerare, e confrontare in qualche modo, la variabilità delle variabile Y dato le diverse modalità della variabile X rispetto alla variabilità totale



Prima di definire un indice che misuri la dipendenza in media di una variabile rispetto a una mutabile (o ad un carattere discreto) è necessario introdurre una proprietà della varianza che prende il nome di ***scomposizione della varianza***

Associazione tra variabili
Rapporto di correlazione eta quadro

Indipendenza – dipendenza in media

Considerando la scomposizione della devianza,

Dev(Y) = Dev(Within) + Dev(Between)

DEV(Y)

esprime la dispersione della nube dei punti attorno alla media generale.

DEV(WITHIN)
o **entro i gruppi:**
esprime la dispersione dei k gruppi attorno alle rispettive medie. Si ottiene sommando le k devianze interne ai k gruppi.

DEV (BETWEEN)
o **fra i gruppi**:
esprime la dispersione delle medie dei k gruppi attorno alla media generale.

si può introdurre il seguente indice di indipendenza in media:

$$\eta^2_{XY} = \frac{Dev(B)}{Dev(Y)}$$

Associazione tra variabili
Rapporto di correlazione eta quadro

Indipendenza – dipendenza in media

$$\eta^2_{XY} = \frac{Dev(B)}{Dev(Y)}$$

$$0 \leq \eta^2_{XY} \leq 1$$

Caso di indipendenza in media

Tutte le medie condizionate sono uguali tra loro

Caso di dipendenza perfetta

Ogni valore di X corrisponde ad un solo valore di Y

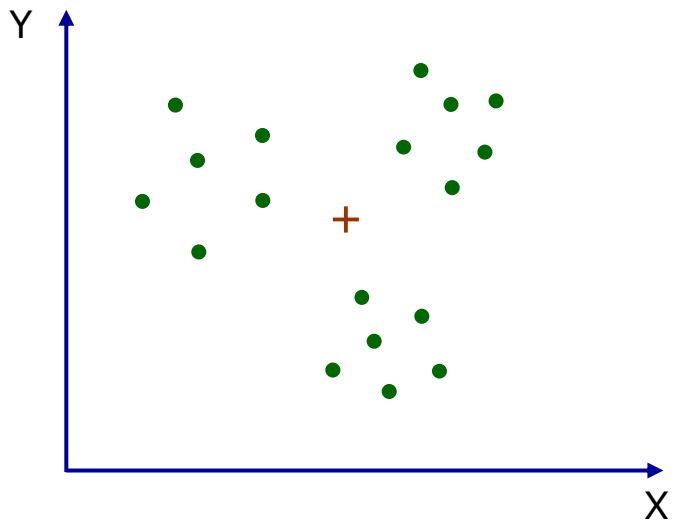
Associazione tra variabili
Proprietà dell'eta quadro

Indipendenza – dipendenza in media

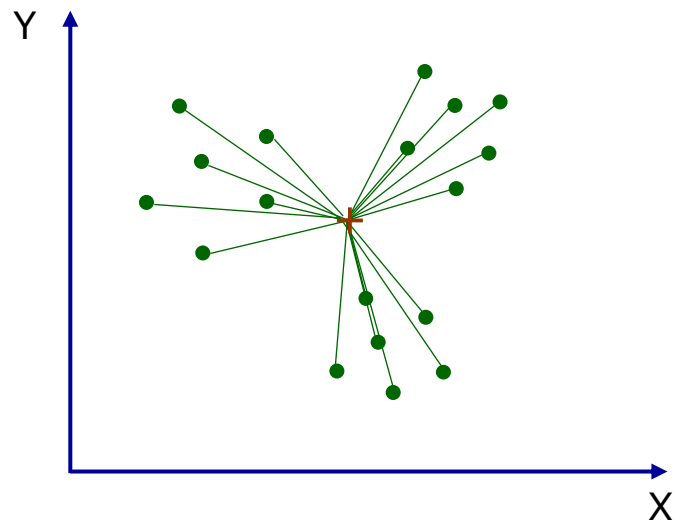
L'indice eta quadro:

- varia tra **0 e 1**: vale 0 quando c'è assoluta indipendenza in media ed 1 quando c'è massima dipendenza in media (ciò si verifica quando ad un solo valore della variabile condizionata corrisponde un solo valore della variabile condizionante)
- l'indice è **asimmetrico** e allorquando vi sono due variabili quantitative è necessario calcolarlo per entrambe
- può essere calcolato per tabelle di correlazione e tabelle miste, misurando in quest'ultimo caso la bontà della divisione effettuata dalla **variabile qualitativa** (variabile indipendente che forma le classi)

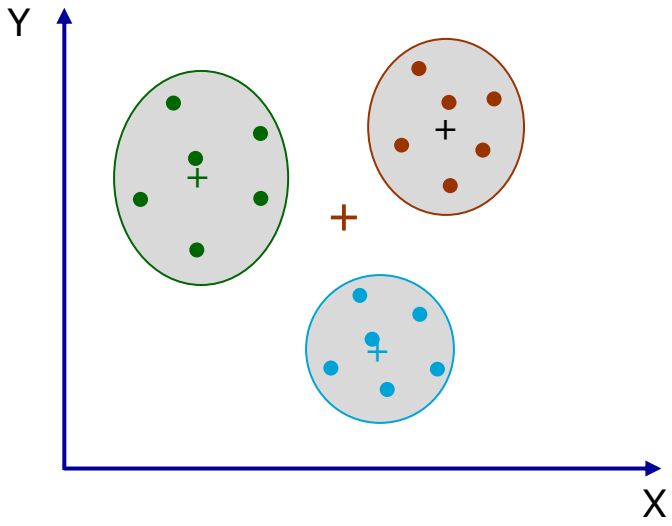
Associazione tra variabili



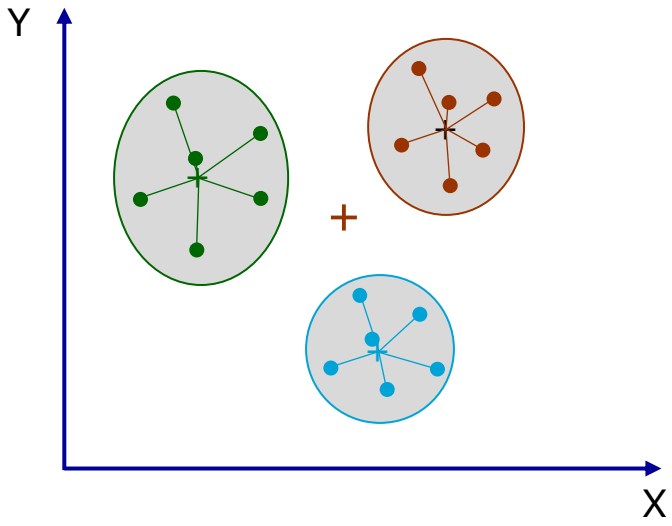
Associazione tra variabili



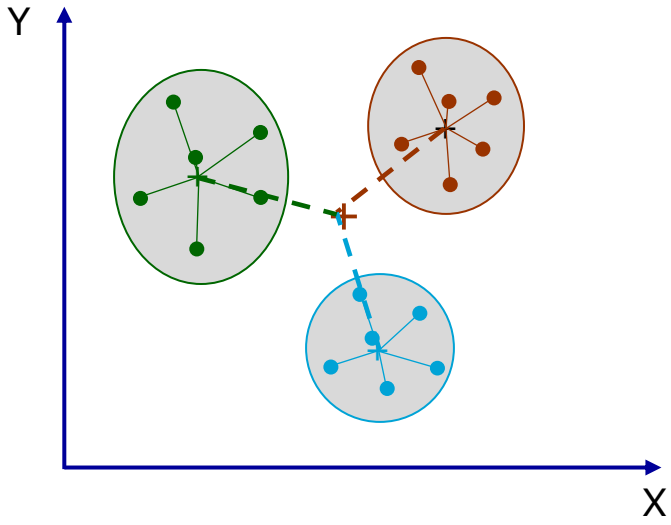
Associazione tra variabili



Associazione tra variabili



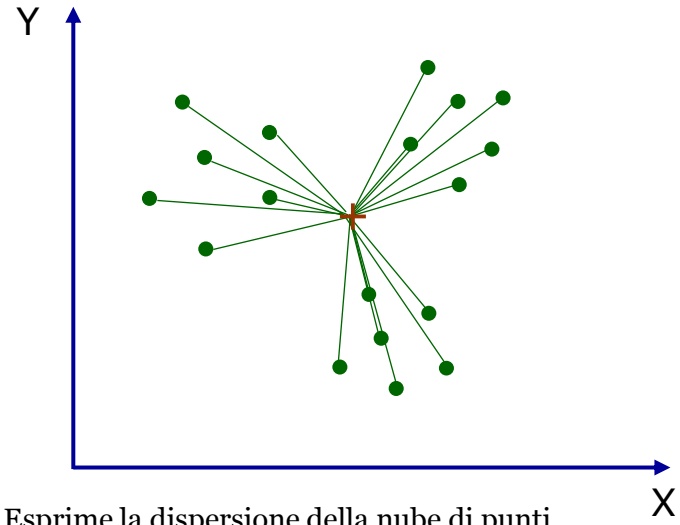
Associazione tra variabili



Associazione tra variabili

Decomposizione della devianza

Dev(Y)



$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \times n_i$$

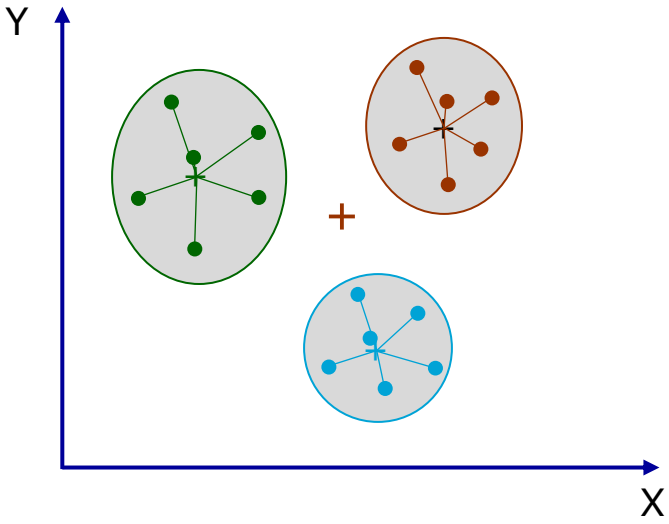
⇨

E' la *devianza totale*. Esprime la dispersione della nube di punti attorno alla media generale.

Associazione tra variabili

Decomposizione della devianza

$$Dev(Y) = Dev(W)$$



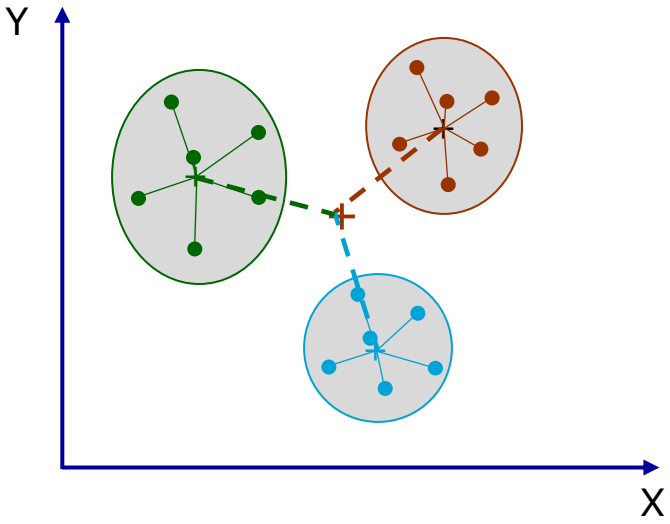
$$Dev(W) = \sum_i \sum_j \left(y_{ij} - \bar{y}_i\right)^2 \cdot n_{ij}$$

E' la *devianza <u>Within</u>*: Esprime la dispersione dei *k* gruppi attorno alle rispettive medie. Si ottiene sommando le *k* devianze interne ai *k* gruppi.

Associazione tra variabili

Decomposizione della devianza

$$Dev(Y) = Dev(W) + Dev(B)$$



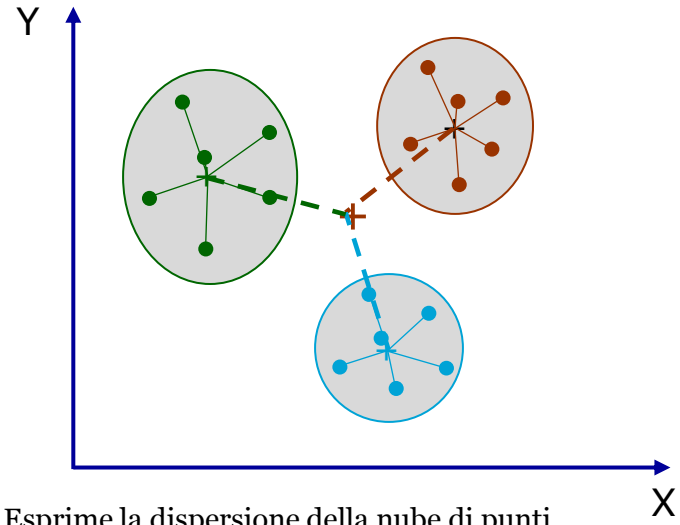
$$Dev(B) = \sum_j (\bar{y}_j - \bar{y})^2 \times n_j$$

E' la devianza *Between*: Esprime la dispersione delle medie dei *k* gruppi attorno alla media generale.

Associazione tra variabili

Decomposizione della devianza

$$Dev(Y) = Dev(W) + Dev(B)$$



$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \times n_i$$

E' la *devianza totale*. Esprime la dispersione della nube di punti attorno alla media generale.

$$Dev(W) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \cdot n_{ij}$$

E' la *devianza Within*: Esprime la dispersione dei *k* gruppi attorno alle rispettive medie. Si ottiene sommando le *k* devianze interne ai *k* gruppi.

$$Dev(B) = \sum_j (\bar{y}_j - \bar{y})^2 \times n_j$$

E' la devianza *Between*: Esprime la dispersione delle medie dei *k* gruppi attorno alla media generale.

Associazione tra variabili

Valori centrali delle classi: 12,5 ; 17,5 ; 22,5 ; 27,5

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Medie condizionate: $M(Y|X=Nord) = \frac{1}{n_{1.}} \sum_{j=1}^{h} y_j \cdot n_{1j}$

$$= \frac{12,5 \times 0 + 17,5 \times 7 + 22,5 \times 34 + 27,5 \times 5}{46}$$

$$= \frac{1.025}{46} = 22,28$$

Associazione tra variabili

Valori centrali delle classi: 12,5 ; 17,5 ; 22,5 ; 27,5

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Medie condizionate: $M(Y|X=Centro) = \frac{1}{n_{2.}} \sum_{j=1}^{h} y_j \cdot n_{2j}$

$$= \frac{12,5 \times 1 + 17,5 \times 18 + 22,5 \times 5 + 27,5 \times 1}{25}$$

$$= \frac{467,5}{25} = 18,7$$

Associazione tra variabili

Valori centrali delle classi: 12,5 ; 17,5 ; 22,5 ; 27,5

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord	7	34	5	46	
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Medie condizionate: $M(Y|X=Sud) = \frac{1}{n_{3.}} \sum_{j=1}^{h} y_j \cdot n_{3j}$

$$= \frac{12,5 \times 31 + 17,5 \times 1 + 22,5 \times 0 + 27,5 \times 0}{32}$$

$$= \frac{405}{32} = 12,66$$

Associazione tra variabili

Medie condizionate:

M(Y|X=Nord) = 22,3

M(Y|X=Centro) = 18,7

M(Y|X=Sud) = 12,66

Y è *indipendente in media* da *X* se al variare delle modalità di *X* le medie condizionate di *Y* sono fra loro uguali e uguali quindi alla media generale

Media generale: $M(Y) = \frac{1}{n}\sum_{j=1}^{h} y_j \cdot n_{.j}$

$$= \frac{(12,5 \times 32 + 17,5 \times 26 + 22,5 \times 39 + 27,5 \times 6)}{103} = 18,4$$

$$M(Y) = 22,3 \times \frac{46}{103} + 18,7 \times \frac{25}{103} + 12,7 \times \frac{32}{103} = 18,4$$

(Media delle medie parziali, ponderata con le frequenze relative dei gruppi)

Dato che le medie condizionate di *Y* non sono fra loro uguali e non sono uguali quindi alla media generale, allora *Y* risulta *dipendente in media* da *X*

Associazione tra variabili

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Associazione tra variabili

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

M(Y) = 18,42

M(Y|X=Nord) = 22,28

M(Y|X=Centro) = 18,70

M(Y|X=Sud) = 12,66

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \times n_i$$

$$= (12,5 - 18,42)^2 \cdot 32 + (17,5 - 18,42)^2 \cdot 26 + (22,5 - 18,42)^2 \cdot 39 + (27,5 - 18,42)^2 \cdot 6$$

$$= 35,05 \cdot 32 + 0,85 \cdot 26 + 16,65 \cdot 39 + 82,45 \cdot 6$$

$$= 1121,60 + 22,10 + 649,35 + 494,70 = 2.287,75$$

Associazione tra variabili

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

M(Y) = 18,42

M(Y|X=Nord) = 22,28

M(Y|X=Centro) = 18,70

M(Y|X=Sud) = 12,66

Dev(Y) = 2287,75

$$Dev(B) = \sum_j (\bar{y}_j - \bar{y})^2 \times n_j$$

$$= (22,28 - 18,42)^2 \cdot 46 + (18,70 - 18,42)^2 \cdot 25 + (12,66 - 18,42)^2 \cdot 32$$

$$= 14,9 \cdot 46 + 0,0784 \cdot 25 + 33,18 \cdot 32$$

$$= 685,40 + 1,96 + 1061,76 = 1.749,12$$

Associazione tra variabili

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

M(Y) = 18,42

M(Y|X=Nord) = 22,28

M(Y|X=Centro) = 18,70

M(Y|X=Sud) = 12,66

Dev(Y) = 2287,75

Dev(B) = 1749,12

$$Dev(W) = \sum_i \sum_j \left(y_{ij} - \bar{y}_i\right)^2 \cdot n_{ij} = Dev(Nord) + Dev(Centro) + Dev(Sud)$$

$$Dev\ (\textbf{Nord}) = (12,5 - 22,3)^2 \times 0 + (17,5 - 22,3)^2 \times 7 + (22,5 - 22,3)^2 \times 34 + (27,5 - 22,3)^2 \times 5$$

$$= 96,04 \times 0 + 23,04 \times 7 + 0,04 \times 34 + 27,04 \times 5$$

$$= 0 + 161,28 + 1,36 + 135,2 = 297,84$$

Associazione tra variabili

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

M(Y) = 18,42

M(Y|X=Nord) = 22,28

M(Y|X=Centro) = 18,70

M(Y|X=Sud) = 12,66

Dev(Y) = 2287,75

Dev(B) = 1749,12

$$Dev(W) = \sum_i \sum_j \left(y_{ij} - \bar{y}_i\right)^2 \cdot n_{ij} = Dev(Nord) + Dev(Centro) + Dev(Sud)$$

$$Dev\ (\textbf{Centro}) = (12,5-18,7)^2 \times 1 + (17,5-18,7)^2 \times 18 + (22,5-18,7)^2 \times 5 + (27,5-18,7)^2 \times 1$$

$$= 38,44 \times 1 + 1,44 \times 18 + 14,44 \times 5 + 77,44 \times 1 = 38,44 + 25,92 + 72,2 + 77,44$$

$$= 214,00$$

Associazione tra variabili

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

M(Y) = 18,42

M(Y|X=Nord) = 22,28

M(Y|X=Centro) = 18,70

M(Y|X=Sud) = 12,66

Dev(Y) = 2287,75

Dev(B) = 1749,12

$$Dev(W) = \sum_i \sum_j \left(y_{ij} - \bar{y}_i\right)^2 \cdot n_{ij} = Dev(Nord) + Dev(Centro) + Dev(Sud)$$

$$Dev(\textbf{Sud}) = (12,5-12,66)^2 \times 31 + (17,5-12,66)^2 \times 1 + (22,5-12,66)^2 \times 0 + (27,5-12,66)^2 \times 0$$

$$= 0,0256 \times 31 + 23,43 \times 1 + 96,83 \times 0 + 220,23 \times 0 = 0,7936 + 23,43 + 0 + 0$$

$$= 24,22$$

Associazione tra variabili

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

M(Y) = 18,42

M(Y|X=Nord) = 22,28

M(Y|X=Centro) = 18,70

M(Y|X=Sud) = 12,66

Dev(Y) = 2287,75

Dev(B) = 1749,12

$$Dev(W) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \cdot n_{ij} = Dev(Nord) + Dev(Centro) + Dev(Sud)$$

$$Dev(W) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \cdot n_{ij} = 297,84 + 214,00 + 24,22 = 536,06$$

Associazione tra variabili

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

M(Y) = 18,42

M(Y|X=Nord) = 22,28

M(Y|X=Centro) = 18,70

M(Y|X=Sud) = 12,66

Dev(Y) = 2287,75

Dev(B) = 1749,12

Dev(W) = 536,06

Indice: η^2

$$\eta^2 = \frac{Dev(B)}{Dev(Y)} = \frac{\sum_{i=1}^{k} \left(\bar{y}_i - \bar{y}\right)^2 \times n_{i.}}{\sum_{j=1}^{h} \left(y_j - \bar{y}\right)^2 \times n_{.j}}$$

$$\eta^2 = \frac{1749,12}{2287,75} = 0,765$$

$$0 \leq \eta^2 \leq 1$$

Statistica per le scienze sociali

Seconda edizione

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia, Carlo Natale Lauro, Marina Marino

UTET