ANALISI DELLA DIPENDENZA

Associazione

ANALISI DELLA DIPENDENZA

- Quando le due variabili osservate non sono entrambe quantitative non ha senso parlare di covariazione o di correlazione.
- Come si può fare allora a studiare la dipendenza tra due variabili qualitative ?
- L'analisi si può basare solo sulle distribuzioni di frequenza delle due variabili, con l'obiettivo di trovare una *associazione* tra i comportamenti dei due caratteri, sulla base di come si distribuiscono le frequenze osservate in corrispondenza delle diverse combinazioni di modalità delle due variabili
- Supponiamo di avere intervistato un insieme di soggetti, e rilevato due variabili, ad es. abbiamo chiesto l'età e cosa fanno il sabato sera:
  - la prima variabile è quantitativa, scala rapporto
  - la seconda invece è qualitativa, non ordinabile, cioè nominale
- E' sufficiente che una delle variabili non sia su scala almeno intervallo per non poter applicare le tecniche per l'analisi della dipendenza tra variabili quantitative studiate finora: quel che possiamo fare è trasformare la variabile quantitativa (età) in classi (giovani, adulti, anziani) e procedere come se fossero entrambe qualitative
- I concetti che stiamo per introdurre si applicano quindi anche alle variabili quantitative, per le quali rappresentano una ulteriore possibilità di analisi, mentre per le variabili qualitative sono una via obbligata

ANALISI DELLA DIPENDENZA

- Il punto di partenza dell'analisi è sempre la serie dei dati osservati: nell'es. abbiamo una serie doppia dove per ciascuna unità osservata è riportato
  - un dato quantitativo relativo all'età
  - un dato qualitativo relativo all'attività del sabato sera, codificato con le lettere a, b, c, d ...
- Decidiamo di discretizzare l'età in tre classi (giovani, adulti, anziani) e procediamo quindi allo studio della dipendenza come se avessimo due variabili qualitative
- La prima cosa da fare è organizzare i dati in una tabella di frequenze: avendo due variabili, si tratterà di una tabella di frequenze **congiunte**, osservate per ogni combinazione di modalità della X e della Y
- Abbiamo due possibilità per costruire una tabella di frequenze congiunte:
  - una più adatta per l'elaborazione informatizzata dei dati,
  - l'altra più compatta e immediatamente comprensibile per gli esseri umani

unità	x(i)	y(i)
1	33	a
2	21	b
3	56	c
4	73	a
5	44	d
6	32	c
7	22	b
8	24	a
...	...	...
n	17	a

ANALISI DELLA DIPENDENZA

- **La Tabella di Frequenze Congiunte**
- La forma più utilizzata dai programmi software di elaborazione dei dati è una estensione immediata della tabella di frequenze semplice al caso di due variabili
- In generale, se la variabile X presenta p modalità diverse, e Y ha q modalità diverse :
  - si costruiscono tutte le combinazioni (coppie) di modalità delle due variabili: (x_i, y_j)
  - risultano (p·q) coppie di modalità, che si riportano sulle righe della tabella
  - per ciascuna coppia di modalità, si conta il numero di unità osservate che la presentano, dette **frequenze congiunte n(i,j)**, e si riportano su una terza colonna
- La tabella risultante ha (p q) righe, e la generica riga riporta la frequenza congiunta n(i, j) relativa alla combinazione di modalità (x_i, y_j)

x	y	freq
x_1	y_1	n_{11}
⋮	⋮	⋮
x_1	y_j	n_{1j}
⋮	⋮	⋮
x_1	y_q	n_{1q}
x_2	y_1	n_{21}
⋮	⋮	⋮
x_2	y_j	n_{2j}
⋮	⋮	⋮
x_2	y_q	n_{2q}
...	...	...
x_i	y_j	n_{ij}
...	...	...

ANALISI DELLA DIPENDENZA

- **La Tabella di Frequenze a Doppia Entrata**
- E' il formato preferito per rappresentare una distribuzione di frequenze congiunte, quando dobbiamo ragionare o analizzare manualmente i dati

	y_1	…	y_j	…	y_q	*totale riga*
x_1	n_{11}	…	n_{1j}	…	n_{1q}	$n_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
x_i	n_{i1}	…	n_{ij}	…	n_{iq}	$n_{i\bullet}$
⋮	⋮		⋮		⋮	⋮
x_p	n_{p1}	…	n_{pj}	…	n_{pq}	$n_{p\bullet}$
totale colonna	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet q}$	$n_{\bullet\bullet}$

- Si mettono in riga le p modalità di una variabile, e in colonna le q modalità dell'altra
  - le celle centrali della tabella riportano le frequenze congiunte: n(i, j)
  - l'ultima colonna riporta il totale delle frequenze per riga: n(i, *)
  - l'ultima riga riporta il totale delle frequenze per colonna: n(*, j)

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Le Frequenze Marginali**
- La tabella a doppia entrata, oltre alle frequenze congiunte, riporta le **frequenze marginali**: sono le due serie di frequenze scritte ai margini della tabella
- Frequenze marginali di X: cioè i totali di riga

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$$

- Frequenze marginali di Y: cioè i totali di colonna

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$$

	y_1	…	y_j	…	y_q	*totale riga*
x_1	n_{11}	…	n_{1j}	…	n_{1q}	$n_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
x_i	n_{i1}	…	n_{ij}	…	n_{iq}	$n_{i\bullet}$
⋮	⋮		⋮		⋮	⋮
x_p	n_{p1}	…	n_{pj}	…	n_{pq}	$n_{p\bullet}$
totale colonna	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet q}$	$n_{\bullet\bullet}$

$$n = n_{\bullet\bullet} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = \sum_{j=1}^{q}\sum_{i=1}^{p} n_{ij} = \sum_{j=1}^{q} n_{\bullet j} = \sum_{i=1}^{p} n_{i\bullet}$$

- Le frequenze marginali descrivono la distribuzione di una variabile ignorando l'altra, cioè senza considerare la modalità assunta dall'altra variabile
- Le distribuzioni marginali sono quindi distribuzioni di una sola variabile

ANALISI DELLA DIPENDENZA

- **Tabella a Doppia Entrata di Frequenze Relative**
- Nella tabella a doppia entrata si possono riportare, invece delle frequenze assolute, le **frequenze relative** rispetto al numero totale di osservazioni:

$$f_{ij} = \frac{n_{ij}}{n}$$

- Frequenze relative marginali:

$$f_{i\bullet} = \frac{n_{i\bullet}}{n} \qquad f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

	y_1	…	y_j	…	y_q	*totale riga*
x_1	f_{11}	…	f_{1j}	…	f_{1q}	$f_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
x_i	f_{i1}	…	f_{ij}	…	f_{iq}	$f_{i\bullet}$
⋮	⋮		⋮		⋮	⋮
x_p	f_{p1}	…	f_{pj}	…	f_{pq}	$f_{p\bullet}$
totale colonna	$f_{\bullet 1}$		$f_{\bullet j}$		$f_{\bullet q}$	1

- Osserviamo che il totale di tutte le frequenze relative congiunte, come pure i totali delle due marginali, è pari a 1

$$\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = \sum_{j=1}^{q}\sum_{i=1}^{p} f_{ij} = \sum_{j=1}^{q} f_{\bullet j} = \sum_{i=1}^{p} f_{i\bullet} = 1$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Le Frequenze Condizionate**
- Le frequenze **condizionate di Y rispetto ad X** sono date da:

$$f(y_j \setminus x_i) = \frac{n_{ij}}{n_{i\bullet}}$$

- In pratica, sono le frequenze relative della riga i-esima, rispetto al totale di riga

	y_1	…	y_j	…	y_q	*totale riga*
x_1	n_{11}	…	n_{1j}	…	n_{1q}	$n_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
x_i	n_{i1}	…	n_{ij}	…	n_{iq}	$n_{i\bullet}$
⋮	⋮		⋮		⋮	⋮
x_p	n_{p1}	…	n_{pj}	…	n_{pq}	$n_{p\bullet}$
totale colonna	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet q}$	$n_{\bullet\bullet}$

- Analogamente, le frequenze **condizionate di X rispetto a Y** sono le frequenze relative al totale di colonna:

$$f(x_i \setminus y_j) = \frac{n_{ij}}{n_{\bullet j}}$$

- La frequenza condizionata $f(y_j \setminus x_i)$ ci dice che percentuale di soggetti presentano la modalità y(j) **tra quelli che** presentano la modalità x(i)

ANALISI DELLA DIPENDENZA

- **La Tabella di Frequenze Condizionate f(Y|X)**

$$f(y_j \setminus x_i) = \frac{n_{ij}}{n_{i\bullet}}$$

	y_1 … y_j … y_q	totale riga
x_1	… … … … …	1
⋮		⋮
x_i	$\frac{n_{i1}}{n_{i\bullet}}$ … $\frac{n_{ij}}{n_{i\bullet}}$ … $\frac{n_{iq}}{n_{i\bullet}}$	1
⋮		⋮
x_p	… … … … …	1
totale colonna	$\frac{n_{\bullet 1}}{n}$ … $\frac{n_{\bullet j}}{n}$ … $\frac{n_{\bullet q}}{n}$	1

- La generica cella riporta la frazione di unità della classe Xi che presentano la modalità Yj
- Questa tabella ha senso letta solo nel senso delle righe: ogni riga rappresenta una distribuzione di frequenze relative di Y condizionata ad una modalità di X
- Questa tabella contiene in effetti non una, ma ben p distribuzioni condizionate (più una distribuzione marginale)
- La generica riga f(Y\Xi) descrive la distribuzione di frequenze relative della variabile Y condizionatamente al fatto che i soggetti presentino la modalità Xi, ovvero del sottoinsieme di soggetti che appartengono alla classe Xi
- La riga marginale è semplicemente la distribuzione di frequenze relative della Y

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **La Tabella di Frequenze Condizionate f(X|Y)**

$$f(x_i \setminus y_j) = \frac{n_{ij}}{n_{\bullet j}}$$

	y_1 … y_j … y_q	totale riga
x_1	… … $\frac{n_{1j}}{n_{\bullet j}}$ … …	$\frac{n_{1\bullet}}{n}$
⋮		⋮
x_i	… … $\frac{n_{ij}}{n_{\bullet j}}$ … …	$\frac{n_{i\bullet}}{n}$
⋮		⋮
x_p	… … $\frac{n_{pj}}{n_{\bullet j}}$ … …	$\frac{n_{p\bullet}}{n}$
totale colonna	1 1 1	1

- La generica cella riporta la frazione di unità della classe Yj che presentano la modalità Xi
- Questa tabella ha senso letta nel senso delle colonne: rappresentano altrettante distribuzioni di frequenze di X condizionate alle modalità di Y
- Questa tabella contiene quindi q distribuzioni condizionate, più una marginale
- La generica colonna f(X\Yj) descrive la distribuzione di frequenze relative della variabile X, condizionatamente al fatto che i soggetti presentino la modalità Yj, cioè del sottoinsieme di soggetti che appartengono alla classe Yj
- La colonna marginale è semplicemente la distribuzione marginale della X

ANALISI DELLA DIPENDENZA

- Esempio
  Data la tabella di frequenze a doppia entrata delle variabili "meta preferita della tua vacanza" e "sesso", costruiamo le tabelle di frequenze condizionate f(X|Y) e f(Y|X)
- La tabella f(vacanza\sesso) mette in evidenza le preferenze di maschi e femmine:
  - le distribuzioni condizionate mostrano comportamenti diversi per sesso: c'è una certa *associazione* tra sesso e vacanza
  - la marginale ci fornisce la distribuzione per località preferita del totale dei soggetti intervistati
- La tabella f(sesso\vacanza) mostra la distribuzione per sesso dei soggetti nelle diverse località di vacanza:
  - al mare troveremo più donne, mentre i maschi in viaggio sono il doppio delle donne
  - la marginale ci fornisce la distribuzione per sesso dei soggetti intervistati

	m	f	Totale
mare	182	267	449
montagna	170	112	282
viaggi	174	80	254
casa	132	124	256
Totale	658	583	1241

f(vacanza\sesso)

	m	f	Totale
mare	0,2766	0,4580	0,3618
montagna	0,2584	0,1921	0,2272
viaggi	0,2644	0,1372	0,2047
casa	0,2006	0,2127	0,2063
Totale	1,0000	1,0000	1,0000

f(sesso\vacanza)

	m	f	Totale
mare	0,4053	0,5947	1,0000
montagna	0,6028	0,3972	1,0000
viaggi	0,6850	0,3150	1,0000
casa	0,5156	0,4844	1,0000
Totale	0,5302	0,4698	1,0000

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **L'idea di Indipendenza**
- Supponiamo stavolta che i dati osservati siano i seguenti: come prima costruiamo le due tabelle di frequenze condizionate f(X|Y) e f(Y|X)
- La tabella f(vacanza\sesso) mostra in questo caso due distribuzioni condizionate pressoché identiche: maschi e femmine si comportano allo stesso modo, ovvero la preferenza per il luogo di vacanza è *indipendente* dal sesso
- Tra sesso e preferenza non c'è alcuna relazione: cioè conoscere il sesso di un soggetto non ci può dire nulla sulle sue preferenze circa la meta delle sue vacanze
- La tabella f(sesso\vacanza) mostra, simmetricamente, le distribuzioni condizionate del sesso per località pressoché identiche: la distribuzione dei soggetti in vacanza riproduce esattamente quella dell'intera popolazione intervistata

	m	f	Totale
mare	294	267	561
montagna	123	112	235
viaggi	88	80	168
casa	136	124	260
Totale	641	583	1224

f(vacanza\sesso)

	m	f	Totale
mare	0,4587	0,4580	0,4583
montagna	0,1919	0,1921	0,1920
viaggi	0,1373	0,1372	0,1373
casa	0,2122	0,2127	0,2124
Totale	1,0000	1,0000	1,0000

f(sesso\vacanza)

	m	f	Totale
mare	0,5241	0,4759	1,0000
montagna	0,5234	0,4766	1,0000
viaggi	0,5238	0,4762	1,0000
casa	0,5231	0,4769	1,0000
Totale	0,5237	0,4763	1,0000

ANALISI DELLA DIPENDENZA

- **Indipendenza Statistica (o in Distribuzione)**
- Il concetto di indipendenza si fonda sull'idea che se Y è indipendente da X, allora la modalità assunta da X non influenza il comportamento della Y
- In altri termini, quando Y è indipendente da X, conoscere la modalità assunta dalla X non fornisce nessuna informazione sulla possibile modalità assunta dalla Y: non è possibile fare alcuna previsione sulla Y
- Il concetto di indipendenza statistica (o in distribuzione) fà riferimento alle distribuzioni condizionate, ed è un concetto simmetrico:
  - se nelle diverse mete di vacanza troviamo la stessa distribuzione per sesso, indipendentemente dal tipo di località, cioè le distribuzioni condizionate f(sesso\vacanza) sono tutte uguali …
  - simmetricamente, anche la preferenza per il luogo di vacanza risulta esattamente la stessa per maschi e femmine: quindi anche le distribuzioni condizionate f(vacanza\sesso) sono tutte uguali, cioè la distribuzione delle preferenze è indipendente dal sesso

Y è indipendente da X ⇔ X è indipendente da Y

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Indipendenza Statistica**
- La definizione di **indipendenza statistica** può essere formalizzata come uguaglianza delle distribuzioni condizionate:

$$Y \perp X \quad \Leftrightarrow \quad f(Y \mid x_1) = f(Y \mid x_2) = \ldots = f(Y \mid x_i) = \ldots = f(Y \mid x_p)$$

- Y è indipendente da X se le distribuzioni condizionate di Y\X sono tutte uguali
- Quando tutte le distribuzioni condizionate sono uguali, sono anche uguali alla distribuzione marginale
- L'indipendenza statistica si può quindi formalizzare con la condizione:

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

	y_1 … y_j … y_q	*totale riga*
x_1	… … … … …	1
⋮		⋮
x_i	$\frac{n_{i1}}{n_{i\bullet}}$ … $\frac{n_{ij}}{n_{i\bullet}}$ … $\frac{n_{iq}}{n_{i\bullet}}$	1
⋮		⋮
x_p	… … … … …	1
totale colonna	$\frac{n_{\bullet 1}}{n}$ … $\frac{n_{\bullet j}}{n}$ … $\frac{n_{\bullet q}}{n}$	1

ANALISI DELLA DIPENDENZA

- Quando tutte le distribuzioni condizionate sono uguali, sono anche uguali alla distribuzione marginale :

$$f(y_j \setminus x_1) = f(y_j \setminus x_2) = \ldots = f(y_j \setminus x_p) = f(y_j) \quad \forall j = 1,\ldots,q$$

$$\frac{n_{1j}}{n_{1\bullet}} = \frac{n_{2j}}{n_{2\bullet}} = \ldots = \frac{n_{ij}}{n_{i\bullet}} = \ldots = \frac{n_{pj}}{n_{p\bullet}} = \frac{n_{\bullet j}}{n} \quad \forall j = 1,\ldots,q$$

- Quindi possiamo scrivere la condizione di indipendenza statistica come:

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

	y_1	…	y_j	…	y_q	*totale riga*
x_1	…	…	…	…	…	1
⋮						⋮
x_i	$\frac{n_{i1}}{n_{i\bullet}}$	…	$\frac{n_{ij}}{n_{i\bullet}}$	…	$\frac{n_{iq}}{n_{i\bullet}}$	1
⋮						⋮
x_p	…	…	…	…	…	1
totale colonna	$\frac{n_{\bullet 1}}{n}$	…	$\frac{n_{\bullet j}}{n}$	…	$\frac{n_{\bullet q}}{n}$	1

ANALISI DELLA DIPENDENZA

- L'indipendenza statistica è una relazione simmetrica, infatti :

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

- Dunque se le distribuzioni condizionate f(Y|X) sono uguali, sono uguali anche tutte le distribuzioni condizionate f(X|Y), e sono uguali alla marginale di X: quindi X è indipendente da Y

	y_1	…	y_j	…	y_q	*totale riga*
x_1	…	…	$\frac{n_{1j}}{n_{\bullet j}}$	…	…	$\frac{n_{1\bullet}}{n}$
⋮			⋮			⋮
x_i	…	…	$\frac{n_{ij}}{n_{\bullet j}}$	…	…	$\frac{n_{i\bullet}}{n}$
⋮			⋮			⋮
x_p	…	…	$\frac{n_{pj}}{n_{\bullet j}}$	…	…	$\frac{n_{p\bullet}}{n}$
totale colonna	1		1		1	1

ANALISI DELLA DIPENDENZA

- Dunque, ricapitolando, l'indipendenza statistica tra X e Y prevede che:

$$Y \perp X \quad \Leftrightarrow f(Y \setminus x_1) = f(Y \setminus x_2) = \ldots = f(Y \setminus x_i) = \ldots = f(Y \setminus x_p) = f(Y)$$

ovvero :

$$f(y_j \setminus x_1) = f(y_j \setminus x_2) = \ldots = f(y_j \setminus x_p) = f(y_j) \quad \forall j = 1,\ldots,q$$

$$\frac{n_{1j}}{n_{1\bullet}} = \frac{n_{2j}}{n_{2\bullet}} = \ldots = \boxed{\frac{n_{ij}}{n_{i\bullet}}} = \ldots = \frac{n_{pj}}{n_{p\bullet}} = \boxed{\frac{n_{\bullet j}}{n}} \quad \forall j = 1,\ldots,q$$

Simmetricamente :

$$X \perp Y \quad \Leftrightarrow f(X \setminus y_1) = \ldots = f(X \setminus y_i) = \ldots = f(X \setminus y_q) = f(X)$$

ovvero :

$$f(x_i \setminus y_1) = f(x_i \setminus y_2) = \ldots = f(x_i \setminus y_q) = f(x_i) \quad \forall i = 1,\ldots,p$$

$$\frac{n_{i1}}{n_{\bullet 1}} = \frac{n_{i2}}{n_{\bullet 2}} = \ldots = \boxed{\frac{n_{ij}}{n_{\bullet j}}} = \ldots = \frac{n_{iq}}{n_{\bullet q}} = \boxed{\frac{n_{i\bullet}}{n}} \quad \forall i = 1,\ldots,p$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Indipendenza Statistica: la Condizione di Fattorizzazione**
- Due variabili si dicono statisticamente indipendenti quando le distribuzioni condizionate sono uguali alle distribuzioni di frequenze relative marginali:

$$\left.\begin{array}{l} Y \perp X \;\Leftrightarrow\; f(Y \setminus x_i) = f(Y) \;\; \forall i \;\Leftrightarrow\; \dfrac{n_{ij}}{n_{i\bullet}} = \dfrac{n_{\bullet j}}{n} \\ X \perp Y \;\Leftrightarrow\; f(X \setminus y_j) = f(X) \;\; \forall j \;\Leftrightarrow\; \dfrac{n_{ij}}{n_{\bullet j}} = \dfrac{n_{i\bullet}}{n} \end{array}\right\} \Leftrightarrow \boxed{n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}}$$

- La condizione di indipendenza statistica può essere anche espressa in termini di frequenze relative, e prende il nome di **condizione di fattorizzazione**:

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n \cdot n} \quad \Rightarrow \quad \boxed{f_{ij} = f_{i\bullet} \cdot f_{\bullet j}}$$

- Osserviamo che quando due variabili sono indipendenti, le frequenze congiunte sono completamente specificate dalle frequenze marginali delle due variabili: cioè è sufficiente conoscere le distribuzioni marginali per ricavare l'intera tabella di frequenze a doppia entrata

ANALISI DELLA DIPENDENZA

- **Come si misura la dipendenza statistica ?**
- Quando due variabili non sono indipendenti, si possono verificare molte situazioni diverse, cioè possono essere legate da forme diverse di relazione, e la dipendenza può essere più o meno forte
- Per misurare la forza della dipendenza tra due variabili qualitative, detta **associazione** o **connessione**, si confronta la tabella di frequenze osservate con quella che si avrebbe nel caso di indipendenza tra le due variabili
- Date le distribuzioni delle due variabili prese singolarmente (cioè le loro distribuzioni marginali), la condizione di fattorizzazione permette di costruire la tabella delle frequenze attese, o teoriche, che si avrebbero in caso di indipendenza
- Se partiamo da una tabella di frequenze assolute, costruiremo la tabella delle frequenze assolute attese in caso di indipendenza, con la relazione:

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

- Se partiamo invece dalla tabella di frequenze relative, costruiremo la tabella delle frequenze relative attese in caso di indipendenza, usando la condizione di fattorizzazione:

$$\hat{f}_{ij} = f_{i\bullet} \cdot f_{\bullet j}$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **La Tabella delle Frequenze Teoriche in caso di Indipendenza**
- Date le frequenze marginali, le frequenze congiunte teoriche attese in caso di indipendenza si ottengono cella per cella dalla

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

- Applicando la formula per ogni i e j, costruiamo la tabella delle frequenze teoriche, attese in caso di indipendenza

	y_1	…	y_j	…	y_q	*totale riga*
x_1	…	…	…	…	…	$n_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
x_i	…	…	$\hat{n}_{ij}$	…	…	$n_{i\bullet}$
⋮	⋮		⋮		⋮	⋮
x_p	…	…	…	…	…	$n_{p\bullet}$
totale colonna	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet q}$	$n_{\bullet\bullet}$

- A questo punto possiamo pensare di confrontare la tabella a doppia entrata teorica con quella osservata, per vedere quanto la situazione reale si discosta da quella attesa in caso di indipendenza
- Come primo passo verso la costruzione di un indice di associazione, possiamo calcolare, cella per cella, le differenze tra le frequenze osservate e quelle teoriche

ANALISI DELLA DIPENDENZA

- Esempio
  Riprendiamo la tabella di frequenze a doppia entrata delle variabili "meta di vacanza preferita" e "sesso", e costruiamo, cella per cella, la tabella di frequenze teoriche in caso di indipendenza:

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

$$\hat{n}_{11} = \frac{n_{1\bullet} \cdot n_{\bullet 1}}{n} = \frac{449 * 568}{1241} = 238{,}07$$

$$\hat{n}_{21} = \frac{n_{2\bullet} \cdot n_{\bullet 1}}{n} = \frac{282 * 658}{1241} = 149{,}52$$

...

$$\hat{n}_{42} = \frac{n_{4\bullet} \cdot n_{\bullet 2}}{n} = \frac{256 * 583}{1241} = 120{,}26$$

	m	f	Totale
mare	182	267	449
montagna	170	112	282
viaggi	174	80	254
casa	132	124	256
Totale	658	583	1241

Manteniamo le marginali:

	m	f	Totale
mare			449
montagna			282
viaggi			254
casa			256
Totale	658	583	1241

Frequenze teoriche:

	m	f	Totale
mare	238,07	210,93	449
montagna	149,52	132,48	282
viaggi	134,68	119,32	254
casa	135,74	120,26	256
Totale	658	583	1241

ANALISI DELLA DIPENDENZA

- **La Tabella di Contingenze**
- Le differenze tra frequenze osservate e frequenze teoriche in caso di indipendenza sono dette **contingenze (assolute):**

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

- Se operiamo sulle frequenze relative, avremo invece le **contingenze relative**:

$$c_{ij} = f_{ij} - \hat{f}_{ij}$$

	y_1	…	y_j	…	y_q	*totale riga*
x_1	…	…	…	…	…	0
⋮	⋮		⋮		⋮	⋮
x_i	…	…	C_{ij}	…	…	0
⋮	⋮		⋮		⋮	⋮
x_p	…	…	…	…	…	0
totale colonna	0	…	0	…	0	0

- Tra contingenze assolute e relative vale la ovvia relazione: $c_{ij} = f_{ij} - \hat{f}_{ij} = \frac{n_{ij}}{n} - \frac{\hat{n}_{ij}}{n} = \frac{n_{ij} - \hat{n}_{ij}}{n} = \frac{C_{ij}}{n}$
- La somma delle contingenze per riga e per colonna si annullano: è quindi nulla anche la somma complessiva per l'intera tabella

ANALISI DELLA DIPENDENZA

- **Proprietà delle contingenze**
- La somma algebrica delle contingenze per riga, cioè sommando per j, si annulla:

$$\sum_j C_{ij} = \sum_j (n_{ij} - \hat{n}_{ij}) = \sum_j n_{ij} - \sum_j \frac{n_{i\bullet}\, n_{\bullet j}}{n} = n_{i\bullet} - \frac{n_{i\bullet}}{n} \sum_j n_{\bullet j} =$$

$$= n_{i\bullet} - \frac{n_{i\bullet}}{n} n = 0$$

- La somma delle contingenze per colonna, cioè sommando per tutti gli i, è nulla:

$$\sum_i C_{ij} = \sum_i (n_{ij} - \hat{n}_{ij}) = \sum_i n_{ij} - \sum_i \frac{n_{i\bullet}\, n_{\bullet j}}{n} = n_{\bullet j} - \frac{n_{\bullet j}}{n} \sum_i n_{i\bullet} =$$

$$= n_{\bullet j} - \frac{n_{\bullet j}}{n} n = 0$$

- Ne consegue che è nulla anche la somma algebrica di tutte le contingenze della tabella:

$$\sum_{i,j} C_{ij} = \sum_i \sum_j C_{ij} = \sum_i \sum_j (n_{ij} - \hat{n}_{ij}) = 0$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **L'indice Chi Quadrato (χ^2)**
- In considerazione del fatto che la somma algebrica delle contingenze è nulla, per costruire una misura di associazione si ricorre ancora una volta alla somma dei quadrati:

$$\chi^2 = \sum_i \sum_j \frac{C_{ij}^2}{\hat{n}_{ij}} = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

- Assume solo valori positivi o nulli.
- È nullo quando X e Y sono indipendenti, cioè quando tutte le frequenze osservate sono uguali a quelle teoriche, e positivo se X e Y non sono indipendenti
- Il valore dell'indice aumenta con l'intensità dell'associazione tra le variabili, ma anche con l'aumentare di n, nonostante il denominatore tenti di eliminare l'effetto della numerosità
- L'indice aumenta anche, a parità di n, all'aumentare del numero di righe e di colonne della tabella: si può dimostrare infatti che una maggiorazione dell'indice è data da:

$$\chi^2 \le n(k-1) \quad dove \quad k = \min(p, q)$$

ANALISI DELLA DIPENDENZA

- **Metodo di calcolo indiretto per l'indice χ^2 di Pearson**
- Anche per il Chi Quadrato esiste un metodo di calcolo indiretto, che evita di calcolare tutte le (p q) differenze tra frequenze teoriche e osservate:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_i \sum_j \frac{n_{ij}^2 + \hat{n}_{ij}^2 - 2 n_{ij} \hat{n}_{ij}}{\hat{n}_{ij}} =$$

$$= \sum_i \sum_j \frac{n_{ij}^2}{\hat{n}_{ij}} + \sum_i \sum_j \frac{\hat{n}_{ij}^2}{\hat{n}_{ij}} - 2 \sum_i \sum_j \frac{n_{ij} \hat{n}_{ij}}{\hat{n}_{ij}} =$$

$$= \sum_i \sum_j \frac{n_{ij}^2}{\hat{n}_{ij}} + \sum_i \sum_j \hat{n}_{ij} - 2 \sum_i \sum_j n_{ij} = \sum_i \sum_j \frac{n_{ij}^2}{\hat{n}_{ij}} + n - 2n =$$

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

- In pratica, se già si dispone delle frequenze teoriche si usa la prima espressione, altrimenti la seconda

ANALISI DELLA DIPENDENZA

- **L'indice di contingenza V di Cramér**
- Non è facile valutare l'intensità dell'associazione dal valore del Chi Quadrato, perché non ha un limite superiore (fisso)
- Per avere un indice di associazione più facilmente interpretabile si introduce l'indice V:

$$V = \frac{\chi^2}{n(k-1)} \quad con\ k = \min(p, q)$$

- E' costruito come rapporto tra il Chi Quadrato e il suo massimo
- V assume valori compresi tra 0 e 1: è nullo quando X e Y sono indipendenti, e positivo se X e Y non sono indipendenti
- Il valore dell'indice V aumenta con il livello di associazione tra le variabili, e vale 1 quando c'è massima associazione: cioè quando su ogni riga (o su ogni colonna) della tabella a doppia entrata c'è solo una cella con frequenza diversa da zero

	y_1	y_2	y_3
x_1			
x_2			
x_3			

	y_1	y_2	y_3	y_4
x_1				
x_2				
x_3				
x_4				

	y_1	y_2	y_3	y_4
x_1				
x_2				
x_3				

ANALISI DELLA DIPENDENZA

- Esempio. Calcoliamo il χ^2 utilizzando la definizione
- Per prima cosa calcoliamo, cella per cella, le contingenze assolute
- Calcoliamo poi, cella per cella, i rapporti:

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

	m	f	Totale
mare	13,20	14,90	28,11
montagna	2,80	3,17	5,97
viaggi	11,48	12,96	24,44
casa	0,10	0,12	0,22
Totale	27,59	31,14	58,74

$$\chi^2 = \sum_i \sum_j \frac{C_{ij}^2}{\hat{n}_{ij}} = 58{,}74$$

$$V = \frac{\chi^2}{n \min(p-1, q-1)} = \frac{58{,}74}{1241 * 1} = 0{,}0473$$

	m	f	Totale
mare	182	267	449
montagna	170	112	282
viaggi	174	80	254
casa	132	124	256
Totale	658	583	1241

Frequenze teoriche:

	m	f	Totale
mare	238,07	210,93	449
montagna	149,52	132,48	282
viaggi	134,68	119,32	254
casa	135,74	120,26	256
Totale	658	583	1241

Contingenze:

	m	f	Totale
mare	-56,07	56,07	0
montagna	20,48	-20,48	0
viaggi	39,32	-39,32	0
casa	-3,74	3,74	0
Totale	0	0	0

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- Esempio. Calcoliamo il χ^2 utilizzando il metodo indiretto
- Avendo già calcolato le frequenze teoriche, utilizziamo l'espressione:

$$\chi^2 = \sum_i \sum_j \frac{n_{ij}^2}{\hat{n}_{ij}} - n$$

quindi otteniamo velocemente:

$$\chi^2 = 1299{,}74 - 1241 = 58{,}74$$

	m	f	Totale
mare	182	267	449
montagna	170	112	282
viaggi	174	80	254
casa	132	124	256
Totale	658	583	1241

Frequenze teoriche:

	m	f	Totale
mare	238,07	210,93	449
montagna	149,52	132,48	282
viaggi	134,68	119,32	254
casa	135,74	120,26	256
Totale	658	583	1241

Metodo indiretto:

	m	f	Totale
mare	139,14	337,97	477,11
montagna	193,28	94,69	287,97
viaggi	224,81	53,64	278,44
casa	128,37	127,85	256,22
Totale	685,59	614,14	1299,74

ANALISI DELLA DIPENDENZA

- Esempio. Calcolare gli indici χ^2 e V sulla seguente tabella
- Riconosciamo subito una situazione di massima associazione tra le due variabili: infatti le frequenze sono concentrate su una sola cella per riga
- Usando il procedimento diretto, abbiamo:

$$\frac{C_{ij}^2}{\hat{n}_{ij}} =$$

	m	f	Totale
mare	193,93	147,44	341,37
montagna	210,71	160,20	370,91
viaggi	189,79	144,29	334,09
casa	110,57	84,06	194,63
Totale	705,00	536,00	1241,00

$$\chi^2 = \sum_i \sum_j \frac{C_{ij}^2}{\hat{n}_{ij}} = 1241$$

$$V = \frac{\chi^2}{n \min(p-1, q-1)} = \frac{1241}{1241 * 1} = 1$$

	m	f	Totale
mare	0	449	449
montagna	282	0	282
viaggi	254	0	254
casa	0	256	256
Totale	536	705	1241

Frequenze teoriche:

	m	f	Totale
mare	193,93	255,07	449
montagna	121,80	160,20	282
viaggi	109,71	144,29	254
casa	110,57	145,43	256
Totale	536	705	1241

Contingenze:

	m	f	Totale
mare	-193,93	193,93	0
montagna	160,20	-160,20	0
viaggi	144,29	-144,29	0
casa	-110,57	110,57	0
Totale	0	0	0

ANALISI DELLA DIPENDENZA

- **Indipendenza Statistica (o in Distribuzione) e Incorrelazione**
- Il concetto di indipendenza statistica e le misure di associazione che ne derivano sono applicabili a qualunque tipo di variabili, quindi anche a quelle quantitative, dopo averle discretizzate in classi di valori
- Se stiamo analizzando la dipendenza tra due variabili quantitative, ci possiamo quindi domandare come si relazionano il concetto di indipendenza statistica e quello di incorrelazione
- L'indipendenza statistica, come abbiamo appena visto, riguarda la *distribuzione* congiunta delle due variabili, ed è detta infatti anche indipendenza in distribuzione.
- La correlazione è invece un particolare tipo di relazione di covariazione *lineare* tra i valori assunti dalle due variabili
- (In-)dipendenza statistica e (in-)correlazione sono solo due delle tante forme di relazione che si possono definire: per capire il collegamento esistente tra queste due forme di relazione, ne dobbiamo introdurre una terza, la **dipendenza in media**
- La dipendenza in media è una relazione che fa riferimento alle medie delle distribuzioni condizionate di una variabile rispetto ad un'altra (es. di Y\X), dette **medie condizionate: M(Y|X)**

ANALISI DELLA DIPENDENZA

- **Le Medie Condizionate**
- Riprendiamo la tabella delle distribuzioni condizionate Y\X: ogni riga rappresenta una distribuzione di frequenze (relative) di Y condizionate a x(i)
- Se Y è quantitativa, è possibile calcolarne la media (complessiva) come media *ponderata* con le frequenze marginali

$$M(Y) = \sum_{j=1}^{q} y_j \frac{n_{\bullet j}}{n}$$

	y_1 … y_j … y_q	*totale riga*
x_1	… … … … …	1
⋮		⋮
x_i	$\frac{n_{i1}}{n_{i\bullet}}$ … $\frac{n_{ij}}{n_{i\bullet}}$ … $\frac{n_{iq}}{n_{i\bullet}}$	1
⋮		⋮
x_p	… … … … …	1
totale colonna	$\frac{n_{\bullet 1}}{n}$ … $\frac{n_{\bullet j}}{n}$ … $\frac{n_{\bullet q}}{n}$	1

- E' inoltre possibile calcolare le medie della Y per *ciascuna* delle p distribuzioni condizionate, dette **medie condizionate di Y\X**:

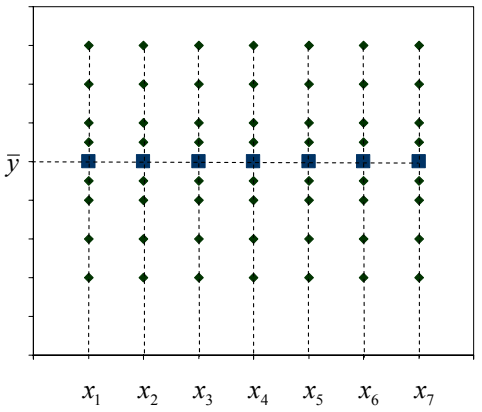
$$M(Y \mid x_i) = \sum_{j=1}^{q} y_j \, f(y_j \mid x_i) = \sum_{j=1}^{q} y_j \frac{n_{ij}}{n_{i\bullet}}$$

ANALISI DELLA DIPENDENZA

- **Dipendenza in Media**
- Si dice che tra Y e X c'è **indipendenza in media** quando tutte le medie condizionate M(Y\X) sono uguali:

$$M(Y \mid x_1) = M(Y \mid x_2) = \ldots = M(Y \mid x_i) = \ldots = M(Y \mid x_p)$$

- Se le medie condizionate sono uguali, sono anche uguali alla media complessiva della Y
- Non c'è variabilità *tra* le medie condizionate (sono tutte uguali): tutta la variabilità è *interna* alle distribuzioni condizionate
- Significato dell'indipendenza in media: se Y è indipendente in media da X, conoscere il valore di X non fornisce nessuna informazione aggiuntiva sul valore assunto (in media) dalla Y



ANALISI DELLA DIPENDENZA

- **Relazione tra Indipendenza Statistica e Indipendenza in Media**
- Se due variabili sono statisticamente indipendenti, sono anche indipendenti in media
- Infatti se X e Y sono statisticamente indipendenti, per definizione, le distribuzioni condizionate di Y|X sono uguali:

$$f(Y \mid x_1) = f(Y \mid x_2) = \ldots = f(Y \mid x_i) = \ldots = f(Y \mid x_p)$$

- di conseguenza saranno uguali anche le medie di tali distribuzioni:

$$M(Y \mid x_1) = M(Y \mid x_2) = \ldots = M(Y \mid x_i) = \ldots = M(Y \mid x_p)$$

- perché sono date da:

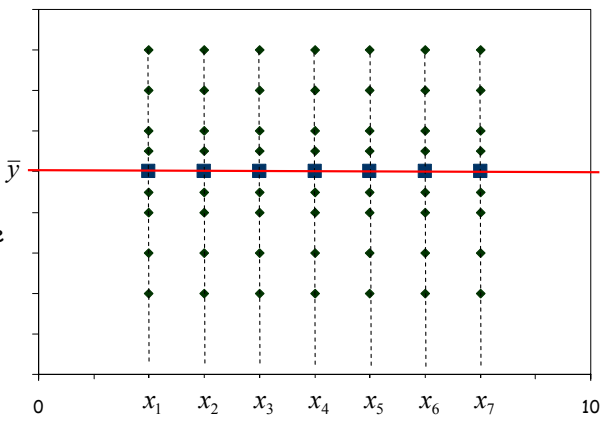
$$\bar{y}_i = M(Y \mid x_i) = \sum_{j=1}^{q} y_j \; f(y_j \mid x_i)$$

- Dunque l'indipendenza in distribuzione implica l'indipendenza in media

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Relazione tra Indipendenza in Media e Incorrelazione**
- Se X e Y sono indipendenti in media, per definizione, le medie condizionate di Y|X sono tutte uguali
- Stimando ai minimi quadrati la retta di regressione, troviamo che passa (necessariamente) per le medie condizionate e risulta quindi parallela all'asse X
- Una situazione per noi ormai ben nota, in cui sappiamo che il coefficiente di correlazione è nullo



$$\hat{y} = a + b\,x$$

$$\begin{cases} \hat{a} = \bar{y} \\ \hat{b} = 0 \end{cases} \Rightarrow \boxed{\rho = 0}$$

- Dunque l'indipendenza in media implica l'incorrelazione

ANALISI DELLA DIPENDENZA

- **Relazioni tra Indipendenza Statistica, Indipendenza in Media e Incorrelazione**
- Dunque ricapitolando: se due variabili sono statisticamente indipendenti, sono anche indipendenti in media; e se sono indipendenti in media, sono anche incorrelate

$$indip.\ stat. \Rightarrow indip.\ in\ media \Rightarrow incorrelazione$$

- Quindi si può concludere che se due variabili sono statisticamente indipendenti, sono anche incorrelate:

$$\chi^2 = 0 \quad \Rightarrow \quad \rho = 0$$

- Non vale invece il viceversa: due variabili possono essere incorrelate, ma dipendenti in media e in distribuzione:

$$\rho = 0 \quad \not\Rightarrow \quad \chi^2 = 0$$

- Rovesciando il discorso, è possibile affermare che se due variabili sono correlate, sono sicuramente dipendenti in media e in distribuzione:

$$\rho \neq 0 \quad \Rightarrow \quad \chi^2 > 0$$

ANALISI DELLA DIPENDENZA

- **Scomposizione della Devianza (e della Varianza) di Y**
- La devianza della variabile Y può essere scomposta in due componenti:

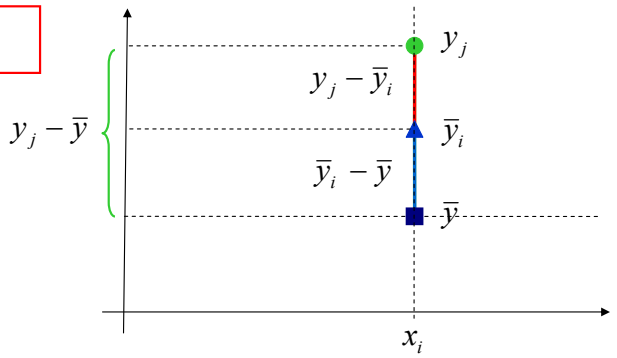
$$\sum_j (y_j - \bar{y})^2 n_{\bullet j} = \sum_i (\bar{y}_i - \bar{y})^2 n_{i\bullet} + \sum_i \sum_j (y_j - \bar{y}_i)^2 n_{ij}$$

cioè come somma della **devianza delle medie condizionate** e della **devianza intorno alle medie condizionate**

- Lo stesso vale anche per la varianza, basta dividere tutto per n:

$$\sigma_y^2 = \sigma_{medie}^2 + \sigma_{cond}^2$$

- Capiamo il significato della scomposizione:



ANALISI DELLA DIPENDENZA

Dimostrazione:

$$V(y) = \sum_j (y_j - \bar{y})^2 n_{\bullet j} = \sum_j (y_j - \bar{y})^2 \sum_i n_{ij} = \sum_i \sum_j (y_j - \bar{y})^2 n_{ij} =$$

$$= \sum_i \sum_j (y_j - \bar{y}_i + \bar{y}_i - \bar{y})^2 n_{ij} = \sum_i \sum_j [(y_j - \bar{y}_i) + (\bar{y}_i - \bar{y})]^2 n_{ij} =$$

$$= \sum_i \sum_j (y_j - \bar{y}_i)^2 n_{ij} + \sum_i \sum_j (\bar{y}_i - \bar{y})^2 n_{ij} + 2 \sum_i \sum_j (y_j - \bar{y}_i)(\bar{y}_i - \bar{y}) n_{ij} =$$

$$= \sum_i \sum_j (y_j - \bar{y}_i)^2 n_{ij} + \sum_i (\bar{y}_i - \bar{y})^2 \sum_j n_{ij} + 2 \sum_i (\bar{y}_i - \bar{y}) \sum_j (y_j - \bar{y}_i) n_{ij} =$$

$$= \sum_i \sum_j (y_j - \bar{y}_i)^2 n_{ij} + \sum_i (\bar{y}_i - \bar{y})^2 n_{i\bullet}$$

= 0 in quanto somma degli scarti dei valori dalla propria media condizionata

devianza "residua" intorno alle medie condizionate +
devianza "spiegata" dalle medie condizionate

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Il rapporto di correlazione η^2**
- Il rapporto di correlazione $\eta^2(Y|X)$ è dato dal rapporto tra la varianza delle medie condizionate e la varianza totale di Y:

$$\eta^2_{y|x} = \frac{\sigma^2_{medie}}{\sigma^2_y} = \frac{\sum (\bar{y}_i - \bar{y})^2 n_{i\bullet}}{\sum \sum (y_j - \bar{y})^2 n_{ij}}$$

- Misura la frazione di varianza (o di devianza) della Y spiegata dalle medie condizionate M(Y|X)
- $\eta^2(Y|X)$ si può scrivere anche come:

$$\eta^2_{y|x} = \frac{\sigma^2_{medie}}{\sigma^2_y} = \frac{\sigma^2_y - \sigma^2_{cond}}{\sigma^2_y} = 1 - \frac{\sigma^2_{cond}}{\sigma^2_y}$$

- In generale, η^2 non è una quantità simmetrica: $\eta^2_{y|x} \neq \eta^2_{x|y}$

$$\eta^2_{x|y} = \frac{\sum (\bar{x}_j - \bar{x})^2 n_{\bullet j}}{\sum \sum (x_i - \bar{x})^2 n_{ij}}$$

ANALISI DELLA DIPENDENZA

- **Proprietà del rapporto di correlazione η^2**

$$\eta^2_{y|x} = \frac{\sigma^2_{medie}}{\sigma^2_y} = \frac{\sum(\bar{y}_i - \bar{y})^2 n_{i\bullet}}{\sum(y_j - \bar{y})^2 n_{\bullet j}} = 1 - \frac{\sigma^2_{cond}}{\sigma^2_y} = 1 - \frac{\sum\sum(y_j - \bar{y}_i)^2 n_{ij}}{\sum(y_j - \bar{y})^2 n_{\bullet j}}$$

- Assume valori compresi tra 0 e 1:
  - vale 0 quando la varianza descritta dalle medie è nulla:
    questo accade quando le medie condizionate sono tutte uguali

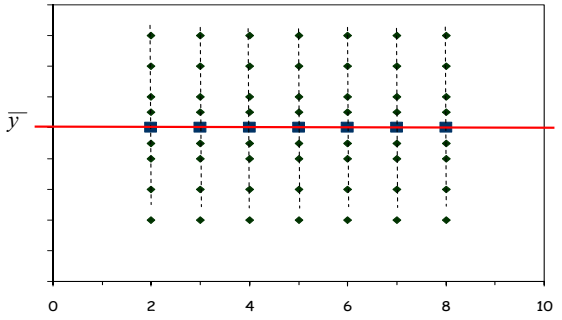
$$\bar{y}_i = \bar{y} \;\Rightarrow\; \sum(\bar{y}_i - \bar{y})^2 n_{i\bullet} = 0$$

  - vale 1 quando le medie condizionate spiegano completamente la varianza della Y:
    questo accade quando tutti i punti si concentrano sulle medie condizionate,
    ovvero le frequenze di ciascuna distribuzione condizionata sono concentrate in un'unica cella

ANALISI DELLA DIPENDENZA

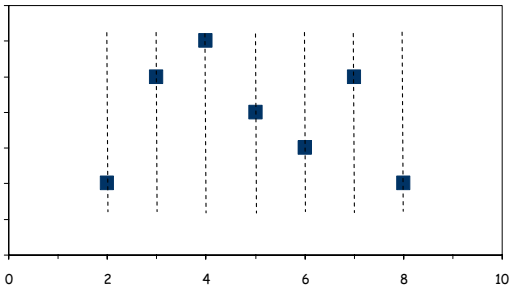
- Indipendenza in media:

$$\eta^2_{x|y} = 0$$



- Massima dipendenza in media:

$$\eta^2_{x|y} = 1$$



ANALISI DELLA DIPENDENZA

- **Relazione tra η^2 e ρ^2**
- Il coefficiente di correlazione al quadrato è sempre minore del rapporto di correlazione, o meglio di entrambi i rapporti di correlazione:

$$\rho_{xy}^2 \leq \eta_{y|x}^2 \quad e \quad \rho_{xy}^2 \leq \eta_{x|y}^2$$

- L'uguaglianza si verifica solo quando le medie condizionate giacciono tutte sulla retta di regressione, cioè quando sono perfettamente allineate:

$$\bar{y}_i = \hat{y}_i \quad \forall i$$

- La disuguaglianza si può scrivere anche nella forma del tutto equivalente:

$$1 - \eta_{y|x}^2 \leq 1 - \rho^2$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DELLA DIPENDENZA

- **Relazioni tra Indipendenza Statistica, Indipendenza in Media e Incorrelazione**
- Dunque ricapitolando: se due variabili sono statisticamente indipendenti, sono anche indipendenti in media; ma se sono indipendenti in media, sono anche incorrelate

$$indip.\,stat. \Rightarrow indip.\,in\,media \Rightarrow incorrelazione$$

- Quindi si può concludere che se due variabili sono statisticamente indipendenti, sono anche incorrelate:

$$\chi^2 = 0 \Rightarrow \eta^2 = 0 \Rightarrow \rho = 0$$

- Non vale invece il viceversa: due variabili possono essere incorrelate, ma dipendenti in media e in distribuzione:

$$\rho = 0 \not\Rightarrow \eta^2 = 0 \not\Rightarrow \chi^2 = 0$$

- Rovesciando il discorso, è possibile affermare che se due variabili sono correlate, sono sicuramente dipendenti in media e in distribuzione:

$$\rho \neq 0 \Rightarrow \eta^2 > 0 \Rightarrow \chi^2 > 0$$