01 – Introduzione al problema

Esistono diverse tipologie di relazioni tra due caratteri: alcune relazioni sono osservabili praticamente su tutti i tipi di variabile doppia (quantitativa, qualitativa o mista), altre invece solo su alcuni tipi di variabile

In generale possiamo pensare che presi due caratteri ci sia una relazione di **dipendenza** o **indipendenza logica**, in termini di causa/effetto: l'altezza di un uomo dipende ad esempio dall'età, dall'alimentazione, dal patrimonio genetico dei genitori, ma è indipendente dalla marca di automobile o di cellulare preferita

➡ **Noi siamo interessati a studiare la dipendenza o l'indipendenza da un punto di vista statistico**

	y_1	…	y_j	…	y_c	TOT
x_1	n_{11}	…	n_{1j}	…	n_{1c}	$n_{1.}$
…	…	…	…	…	…	…
x_i	n_{i1}	…	n_{ij}	…	n_{ic}	$n_{i.}$
…	…	…	…	…	…	…
x_r	n_{r1}	…	n_{rj}	…	n_{rc}	$n_{r.}$
TOT	$n_{.1}$	…	$n_{.j}$	…	$n_{.c}$	$n_{..}$

Consideriamo una variabile doppia **(X,Y)** e supponiamo che sia stata organizzata in una tabella: sulle righe le **r** modalità di X e sulle colonne le **c** modalità di Y

Siamo interessati a studiare se esiste una relazione tra la variabile in riga e la variabile in colonna

Una volta verificata l'esistenza della relazione, in caso positivo, vogliamo misurarne l'intensità

02 – Tipi di relazione

Possiamo osservare, a seconda del problema trattato, diversi tipologie di relazione:

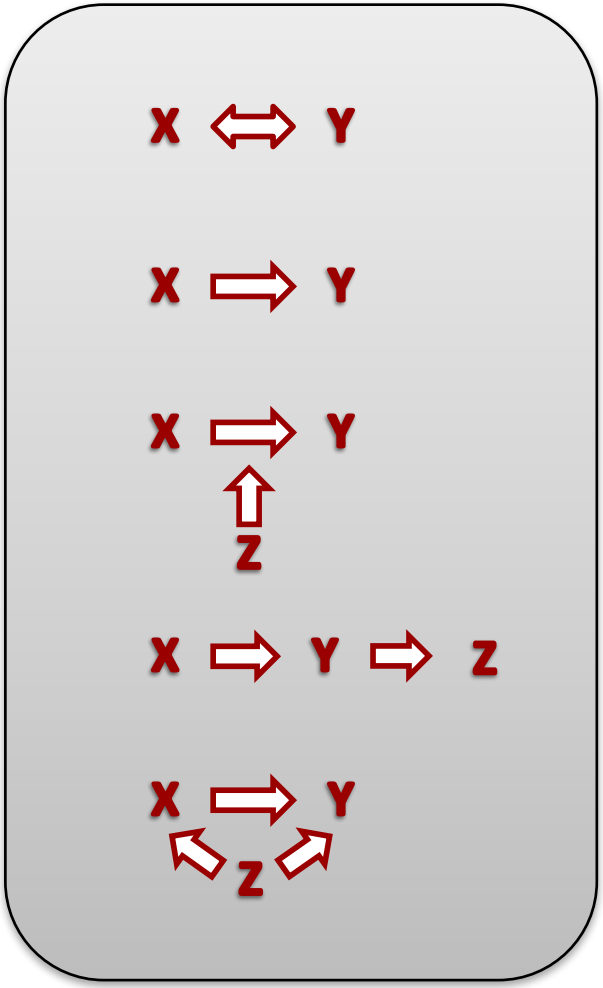
relazione di INTERDIPENDENZA
ES.: ore dedicate allo studio <-> ore dedicate al tempo libero

relazione DIRETTA (di una variabile da un'altra, e non viceversa)
ES.: reddito -> consumi

relazione CONDIZIONATA (da un terzo fenomeno)
ES.: età -> ascolto musica classica (condizionata da istruzione)

relazione INDIRETTA
ES.: istruzione -> reddito -> tenore di vita

relazione SPURIA
ES.: stato civile <- età -> consumo di dolci



03 – Studio della connessione tra due variabili — Unità n° 07

Supponiamo di voler studiare la relazione tra due variabili di qualsiasi natura e consideriamo a tal proposito le distribuzioni condizionate della X rispetto alle diverse modalità di Y:

se la distribuzione condizionata X|Y non si modifica al variare delle modalità di Y allora si dice che la variabile X è ***indipendente in distribuzione*** da Y, o semplicemente che le variabili sono **sconnesse**

$$f(X=x_i|Y=y_j) = f(X=x_i) \quad \text{per } i = 1, \ldots, r \text{ e } j = 1, \ldots, c$$

L'indipendenza in distribuzione è una relazione simmetrica: se X è indipendente da Y allora anche Y risulta essere indipendente da X

$$f(Y=y_j|X=x_i) = f(Y=y_j) \quad \text{per } i = 1, \ldots, r \text{ e } j = 1, \ldots, c$$

Ovviamente la relazione di dipendenza o indipendenza è determinata sulla base delle osservazioni dei caratteri su un certo collettivo: potrebbe capitare che in base a quanto è stato empiricamente osservato risulti che tra due caratteri logicamente indipendenti sia invece evidenziata una relazione: in questi casi si parla generalmente di ***associazione spuria***

04 – Condizione di indipendenza (in distribuzione) — Unità n° 07

In generale se due variabili sono indipendenti in distribuzione è possibile allora ricostruire la tabella doppia a partire dalle distribuzioni marginali di riga e colonna, avendo che:

$$f\left(X = x_i \middle| Y = y_j\right) = f(X_i) \Rightarrow \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} \Rightarrow \boxed{n_{ij} = \frac{n_{i.} * n_{.j}}{n}}$$

CONDIZIONE DI INDIPENDENZA

sono anche dette
frequenze teoriche

Quando anche per un solo elemento della tabella non è possibile ottenere la frequenza congiunta a partire dalle frequenze marginali allora cade l'ipotesi iniziale di indipendenza

In tal caso c'è un certo legame di dipendenza che dobbiamo misurare: le due variabili sono dette ***connesse*** => Per tale motivo si parla anche di studio della connessione tra due variabili

05 – Esempio

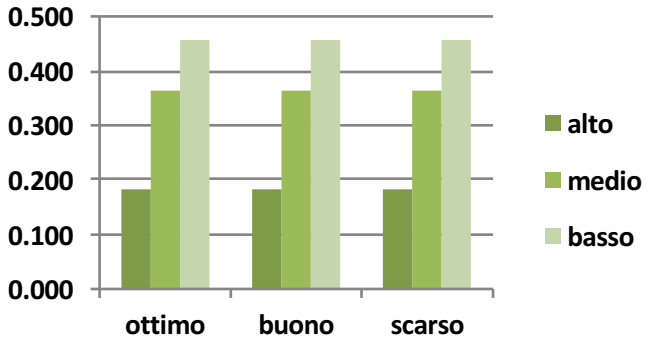
Supponiamo di voler studiare l'associazione tra il rendimento scolastico e il reddito familiare di un collettivo di bambini

Confrontando i profili riga e colonna si rileva che sono sostanzialmente identici

Rendimento	Reddito: *alto*	*medio*	*basso*	
ottimo	16	32	40	88
buono	8	16	20	44
scarso	24	48	60	132
	48	96	120	264

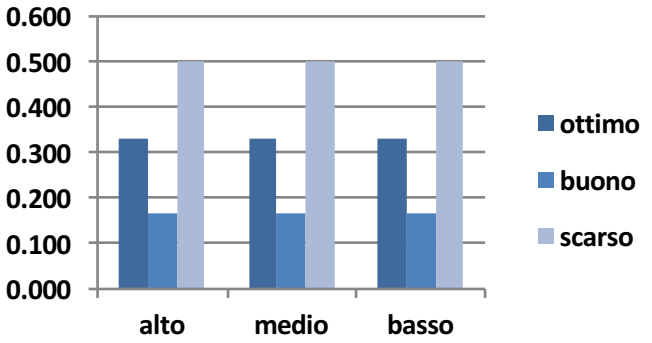
	alto	medio	basso	
ottimo	0,182	0,364	0,454	1,000
buono	0,182	0,364	0,454	1,000
scarso	0,182	0,364	0,454	1,000
	0,182	0,364	0,454	1,000

Profili riga



	alto	medio	basso	
ottimo	0,333	0,333	0,333	0,333
buono	0,167	0,167	0,167	0,167
scarso	0,500	0,500	0,500	0,500
	1,000	1,000	1,000	1,000

Profili colonna



$$\frac{96 \cdot 132}{264} = 48$$

frequenze osservate *e* ***frequenze teoriche*** *coincidono per ogni elemento*

06 – Massima dipendenza (connessione) — Unità n° 07

Tra la variabile Y e la variabile X esiste la **MASSIMA CONNESSIONE** se nota una qualsiasi modalità di Y è univocamente determinata la corrispondente modalità di Y

Se la tabella non è quadrata non è più possibile parlare di reciprocità nel legame tra le variabili

	y_1	y_2	y_3
x_1	n_{11}	0	0
x_2	0	0	n_{23}
x_3	0	n_{32}	0

Se scegliamo una modalità qualsiasi della Y allora sappiamo qual è la modalità osservata della X, e viceversa: c'è una **perfetta interdipendenza** di Y rispetto a X e di Y rispetto a X

	y_1	y_2	y_3
x_1	n_{11}	0	0
x_2	0	0	n_{23}
x_3	n_{31}	0	0
x_4	0	n_{42}	0

Se scegliamo una modalità qualsiasi della X allora sappiamo qual è la modalità osservata della Y, ma non è vero il contrario: c'è una **perfetta dipendenza** di Y rispetto a X

07 – Misura del grado di connessione

Unità n° 07

La misura del grado di connessione si basa sullo scarto tra la frequenza osservata (in una cella della tabella) e la frequenza teorica che si osserverebbe se tra le variabili ci fosse perfetta indipendenza

contingenza ← $c_{ij} = n_{ij} - n^{*}_{ij}$ → *frequenza teorica* $\frac{n_{i.} * n_{.j}}{n}$

n_{ij}: *frequenza osservata*

In caso di indipendenza le contingenze sono tutte nulle

$c_{ij} > 0$	nella cella "i,j" si riscontra un addensamento di frequenze rispetto alla situazione di indipendenza dei due fenomeni.	→ **CONNESSIONE POSITIVA**
$c_{ij} < 0$	nella cella "i,j" si riscontra una rarefazione di frequenze rispetto alla situazione di indipendenza dei due fenomeni.	→ **CONNESSIONE NEGATIVA**

La condizione di indipendenza è molto stringente: è sufficiente che per una sola cella ci sia una contingenza diversa da 0 affinché si rilevi una connessione tra le due variabili oggetto di studio

Dobbiamo costruire un indice che ci permetta di misurare il livello di connessione tra due variabili

08 – Esempio

Vogliamo studiare se c'è una connessione tra il titolo di studio di marito e moglie analizzando un collettivo di 14118 coppie sposate in un certo anno, per verificare se tale aspetto influenza la scelta del partner

FREQUENZE OSSERVATE

istruzione marito \ *istruzione moglie*	laurea	diploma	lic. media	lic. elem.	nessun titolo	*Totale*
laurea	236	327	119	48	2	732
diploma	134	1038	701	430	20	2323
lic.media	30	492	1199	1222	110	3053
lic.elem.	9	218	666	4523	951	6367
nessun titolo	3	20	33	405	1195	1643
Totale	412	2082	2718	6628	2278	14118

FREQUENZE TEORICHE

istruzione marito \ *istruzione moglie*	laurea	diploma	lic. media	lic. elem.	nessun titolo	*Totale*
laurea	21.36	107.95	140.92	343.65	118.11	732
diploma	67.79	342.58	447.22	1090.58	374.83	2323
lic.media	89.09	450.23	587.76	1433.30	492.61	3053
lic.elem.	185.81	938.95	1225.78	2989.13	1027.34	6367
nessun titolo	47.95	242.30	316.31	771.34	265.11	1643
Totale	412	2082	2718	6628	2278	14118

I totali della tabella osservata e di quella ricostruita sono sempre uguali: nelle celle della tabella ricostruita possiamo anche trovare valori non interi, perché sono ottenuti per costruzione sotto l'ipotesi di indipendenza

istruzione marito \ *istruzione moglie*	laurea	diploma	lic. media	lic. elem.	nessun titolo	*Totale*
Laurea	214.64	219.05	-21.92	-295.65	-116.11	0
diploma	66.21	695.42	253.78	-660.58	-354.83	0
lic.media	-59.09	41.77	611.24	-211.30	-382.61	0
lic.elem.	-176.81	-720.95	-559.78	1533.87	-76.34	0
nessun titolo	-44.95	-222.30	-283.31	-366.34	929.89	0
Totale	0	0	0	0	0	0

CONTINGENZE

È possibile osservare una connessione positiva tra titoli di studio uguali per moglie e marito, e una connessione negativa tra titoli bassi e titoli alti. Si osserva inoltre una connessione positiva anche tra i titoli più alti (laurea/diploma)

09 – L'indice Chi-quadro (χ^2)

Nella tabella potremmo trovare come nel caso dell'esempio precedente delle contingenze positive e negative: abbiamo bisogno di un indice che valuti complessivamente l'indipendenza tra le variabili e che allo stesso tempo annulli l'effetto compensativo dei segni positivi e negativi

L'indice ***Chi-quadro*** è ottenuto come somma delle contingenze al quadrato sulle frequenze teoriche

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

Quando si ha perfetta indipendenza l'indice assume valore 0 (tutte le contingenze sono nulle)

Nel caso invece di massima connessione, il valore (massimo) che l'indice può assumere dipende dalla dimensione della tabella e dalla numerosità del collettivo. Si dimostra infatti che:

$$\max\chi^2 = n \cdot \min\left[(r-1);(c-1)\right]$$

→ **il massimo del Chi-quadro è uguale alla numerosità del collettivo studiato per il valore più piccolo tra il numero delle righe meno 1 e il numero delle colonne meno 1**

10 – Esempio

Supponiamo di voler studiare l'associazione tra la spesa mensile per ricariche telefoniche e l'età in un certo collettivo

		Spesa per ric. telefoniche			
		0 -| 25	25 -| 50	50 -| 100	
Età	18 - 22	32	34	19	85
	23 - 27	12	42	26	80
	28 - 32	21	44	15	80
		65	120	60	245

1 ***Possiamo immediatamente verificare se l'ipotesi di indipendenza è ammissibile calcolando, ad es., la frequenza teorica di consumatori tra 18 e 22 anni che ricaricano fino a 25 € al mese***

$$\frac{65 \cdot 85}{245} = 22,551 \neq 32$$

fr. teoriche

		Spesa per ric. telefoniche			
		0 -| 25	25 -| 50	50 -| 100	
Età	18 - 22	22.551	41.633	20.816	85
	23 - 27	21.224	39.184	19.592	80
	28 - 32	21.224	39.184	19.592	80
		65	120	60	245

contingenze

		Spesa per ric. telefoniche			
		0 -| 25	25 -| 50	50 -| 100	
Età	18 - 22	9.449	-7.633	-1.816	0
	23 - 27	-9.224	2.816	6.408	0
	28 - 32	-0.224	4.816	-4.592	0
		0	0	0	0

Cade l'ipotesi di indipendenza in distribuzione tra i caratteri

2 ***Calcoliamo la tabella delle frequenze teoriche e la tabella della contingenze, da cui possiamo dedurre in quali casi c'è connessione positiva e in quali connessione negativa***

3 $$\chi^2 = \frac{9,449^2}{22,551} + \frac{-9,224^2}{21,224} + \ldots = 13,493$$

Poiché il Chi-quadro è diverso da 0 allora possiamo concludere che c'è associazione

11 – Un indice normalizzato: la V di Cramer

Unità n° 07

Poiché il valore dell'indice χ^2 è influenzato dalle dimensioni della tabella e dalla numerosità del collettivo analizzato non è possibile effettuare dei confronti diretti tra il grado di connessione di due variabili in collettivi diversi, o tra coppie diverse di variabili in uno stesso collettivo

Così come visto in altre situazioni è utile ricorrere ad un indice normalizzato, che ha il vantaggio di variare tra 0 e 1 e può quindi essere espresso in percentuale

$$V=\sqrt{\frac{\chi^2}{\max \chi^2}}=\sqrt{\frac{\chi^2}{n \cdot \min\left[(r-1);(c-1)\right]}}$$

→ **V di Cramer**

Quando la V di Cramer vale **0** allora abbiamo perfetta indipendenza (il numeratore è nullo perché le contingenze sono tutte nulle)

Quando la V di Cramer vale **1** allora abbiamo massima connessione (il numeratore è esattamente pari al denominatore, cioè il suo valore massimo)

12 – Analisi e interpretazione dei dati — Unità n° 07

Una volta effettuato un veloce controllo sull'ipotesi di indipendenza, calcolando una frequenza teorica e confrontandola con la corrispondente frequenza osservata, si procede alla costruzione della tabella delle frequenze teoriche e quindi delle contingenze e si fanno le prime valutazioni

A questo punto si procede al calcolo dell'indice Chi-quadro, con il quale siamo soltanto in grado di confermare se c'è indipendenza (l'indice vale 0) o se c'è dipendenza (l'indice è diverso da 0)

In caso di dipendenza ne misuriamo l'intensità attraverso l'indice V di Cramer:

- *da 0 a 0,25 la connessione è bassa**
(es. V=0,18 -> 18% della max connessione osservabile, quindi si ha una debole connessione tra le variabili)

- *da 0,25 a 0,5 la connessione è medio-bassa**
(es. V=0,36 -> 36% della max connessione osservabile, quindi si ha un livello medio-basso di connessione)

- *da 0,5 a 0,75 la connessione è medio-alta**
(es. V=0,69 -> 69% della max connessione osservabile, quindi si ha un livello medio-alto di connessione)

- *da 0,75 a 1 la connessione è alta**
(es. V=0,83 -> 83% della max connessione osservabile, quindi si ha una forte connessione)

**** I valori riportati sono solo esemplificativi e non rappresentano degli intervalli rigidi: l'esperienza vale più di ogni regola!***

13 – Esempio

Nel corso del 2006 sono stati rilevati giornalmente in una città del Sud CONDIZIONE METEOROLOGICA e LIVELLO DEL TRAFFICO URBANO:
vogliamo studiare se c'è associazione tra i due caratteri

	BASSO	MEDIO	ALTO
SERENO	100	13	5
VARIABILE	21	105	10
PIOGGIA	6	15	90

Verifichiamo se l'ipotesi di indipendenza è ammissibile:

$$\frac{n_{1.} \cdot n_{.1}}{n} = \frac{118 \cdot 127}{365} = 41,05 \neq n_{11} = 100$$

fr. teoriche

	BASSO	MEDIO	ALTO	TOT
SERENO	41,05	43,00	33,95	118
VARIABILE	47,32	49,56	39,12	136
PIOGGIA	38,62	40,45	31,93	111
TOT	127	133	105	365

contingenze

	BASSO	MEDIO	ALTO	TOT
SERENO	58,95	-30,00	-28,95	0,00
VARIABILE	-26,32	55,44	-29,12	0,00
PIOGGIA	-32,62	-25,45	58,07	0,00
TOT	0,00	0,00	0,00	0,00

Sulla diagonale principale della tabella delle contingenze si nota un addensamento di casi osservati superiore alla situazione teorica di indipendenza

Essendo tutte le contingenze diverse da zero escludiamo l'ipotesi di indipendenza e calcoliamo l'indice Chi quadro

$$\chi^2 = 377,74$$

Calcolando il grado di connessione con la V di Cramer:

$$V = (377,74/730)^{1/2} = 0,52^{1/2} \Rightarrow 72\%$$

alta associazione tra traffico e cond. meteorologica

14 – Un altro tipo di indipendenza… Unità n° 07

Supponiamo di considerare una variabile quantitativa Y (*es. la classe di voto*) e di voler studiare la relazione con una variabile quantitativa X (*es. il sesso degli studenti*)

Se le medie condizionate di Y rispetto a X sono tutte uguali tra di loro e uguali anche alla media generale allora si dice che la variabile Y è ***indipendente in media*** da X

$$M(Y \mid x_1) = M(Y \mid x_2) = \ldots\ldots\ldots = M(Y \mid x_r) = \bar{y}_a$$

In questo caso non ci interessa se il variare di un carattere influisce o no sulla distribuzione dell'altro: la nostra attenzione è limitata alla media

condizione importante:

è possibile verificare l'indipendenza in media solo se almeno una delle due variabili in gioco è di tipo quantitativo, se le due variabili sono qualitative possiamo studiare solo l'indipendenza in distribuzione

A differenza dell'indipendenza in distribuzione l'indipendenza in media non è simmetrica: ciò implica che se Y è indipendente in media da X non è automaticamente verificato il contrario

15 – Media generale e medie condizionate

	y_1	...	y_j	...	y_c	TOT
x_1	n_{11}	...	n_{1j}	...	n_{1c}	$n_{1.}$
...	...	...	...	...	...	...
x_i	n_{i1}	...	n_{ij}	...	n_{ic}	$n_{i.}$
...	...	...	...	...	...	...
x_r	n_{r1}	...	n_{rj}	...	n_{rc}	$n_{r.}$
TOT	$n_{.1}$	...	$n_{.j}$	...	$n_{.c}$	n

Consideriamo una variabile doppia **(X,Y)** e supponiamo che sia stata organizzata in una tabella che contiene sulle righe le r modalità di X e sulle colonne le c modalità di Y

Supponiamo che la variabile **(X,Y)** sia mista o quantitativa (cioè che almeno una delle due variabili sia quantitativa): in particolare sia **Y** una variabile quantitativa e **X** qualitativa

MEDIA DEL GRUPPO i

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

$$\bar{y}_a = \frac{\sum_{j=1}^{c} y_j n_{.j}}{n}$$

MEDIA GENERALE

La media generale può essere ottenuta come media ponderata delle medie condizionate (medie del carattere per i singoli gruppi), con pesi pari alla numerosità di ogni sottopopolazione

16 – Scomposizione della devianza

Unità n° 07

La variabilità di un fenomeno può essere valutata tramite la ***varianza***, ma anche tramite il suo numeratore, noto come ***devianza***

	y_1	…	y_j	…	y_c	
x_1	n_{11}	…	n_{1j}	…	n_{1c}	$n_{1.}$
…	…	…	…	…	…	…
x_r	n_{r1}	…	n_{rj}	…	n_{rc}	$n_{r.}$
	$n_{.1}$	…	$n_{.j}$	…	$n_{.c}$	n

$$VAR(Y) = \frac{\sum_{j=1}^{c}\left(y_j - \bar{y}_a\right)^2 n_{.j}}{n}$$

DEV(Y)

In questo caso valutiamo la variabilità del fenomeno espresso da Y senza considerare la variabile X

La variabilità totale può essere però scomposta in due elementi

$$DEV(T) = DEV(W) + DEV(B)$$

- $DEV(T)$: **devianza totale**
- $DEV(W)$: **devianza nei gruppi**
- $DEV(B)$: **devianza tra i gruppi**

$$DEV(W) = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(y_j - M(Y \mid x_i)\right) n_{ij} \qquad DEV(B) = \sum_{i=1}^{r}\left(M(Y|x_i) - \bar{y}_a\right)^2 n_{i.}$$

17 – Devianza tra i gruppi e nei gruppi

Da quanto detto la variabilità complessiva del fenomeno oggetto di studio può essere di fatto scomposta in due parti distinte:

la **DEVIANZA TOTALE** è la misura della variabilità misurata su tutto il collettivo

la **DEVIANZA TRA I GRUPPI** (o devianza "between") esprime la parte di variabilità dovuta alle differenze tra i diversi gruppi di unità -> gli studenti frequentanti sono diversi dagli studenti non frequentanti

la **DEVIANZA NEI GRUPPI** (o devianza "within") esprime la parte di variabilità dovuta alle differenze presenti in ciascuno dei gruppi -> non tutti gli studenti frequentanti si comportano allo stesso modo, così come non tutti gli studenti non frequentanti si comportano ugualmente

N.B.: la DEV(W) è data dalla somma delle devianze calcolate su ogni singolo gruppo (sono i numeratori delle varianze condizionate…)

18 – Esempio

Si osservano 13 studenti in base al voto conseguito all'esame di statistica e la frequenza al corso, calcolare quanto la variabilità dei voti dipende dalla frequenza.

- studenti frequentanti: 23,25,25,27,28,29,29,30;
- studenti non frequentanti: 18,19,20,22,30

$$M(voto|F) = \frac{23+(2\bullet 25)+27+28+(29\bullet 2)+30}{8} = 27$$

➡ *Voto medio degli studenti che hanno seguito il corso*

$$M(voto|NF) = \frac{18+19+20+22+30}{5} = 21{,}8$$

➡ *Voto medio degli studenti che non hanno seguito il corso*

$$M(voto) = \frac{(27\bullet 8)+(21{,}8\bullet 5)}{13} = 25$$

➡ *Voto medio di tutti gli studenti (media generale)*

$$\begin{aligned} Dev(Voto) = {} & (18-25)^2+(19-25)^2+(20-25)^2+(22-25)^2+ \\ & +(23-25)^2+(25-25)^2\,2+(27-25)^2+(28-25)^2+ \\ & +(29-25)^2\,2+(30-25)^2\,2= \\ & =218 \end{aligned}$$

$$\begin{aligned} Dev(Voto\ tra\ gruppi) &= (27-25)^2\,8+(21{,}8-25)^2\,5 \\ &= 83{,}2 \end{aligned}$$

➡ $218 = 134{,}8 + 83{,}2$

Il 38% della variabilità del voto dipende dalla differenza tra frequentanti e non frequentanti

19 – Misurare l'indipendenza in media

Abbiamo visto come l'indipendenza in media di una variabile Y da una variabile X è verificata se è verificata la relazione

$$M(Y \mid x_1) = M(Y \mid x_2) = \ldots\ldots = M(Y \mid x_r) = \bar{y}_a$$

È sufficiente che una sola media sia diversa perché sia violata la condizione di indipendenza e ci sia quindi dipendenza in media di Y rispetto a X

Per misurare il livello di dipendenza (indipendenza) in media si ricorre alla scomposizione della devianza e in particolare si utilizza un particolare indice, il ***rapporto di correlazione di Pearson***

$$\eta^2_{y|x} = \frac{\sum_{i=1}^{r}\left(M(Y|x_i) - \bar{y}_a\right)^2 n_{i.}}{\sum_{j=1}^{c}\left(y_j - \bar{y}_a\right)^2 n_{.j}}$$

"eta quadro"

$\sum_{i=1}^{r}\left(M(Y|x_i) - \bar{y}_a\right)^2 n_{i.} \rightarrow DEV(B)$

$\sum_{j=1}^{c}\left(y_j - \bar{y}_a\right)^2 n_{.j} \rightarrow DEV(T)$

l'indice varia tra 0 e 1: assume valore 0 se si ha indipendenza in media e valore 1 se c'è massima dipendenza in media

i valori intermedi si leggono nel solito modo

Di fatto può essere utilizzato per valutare la bontà della divisione in gruppi di un certo collettivo

20 – Analisi e interpretazione dei dati — Unità n° 07

Una volta effettuato il controllo sull'ipotesi di indipendenza, calcolando le medie condizionate e confrontandole con la media generale, si procede al calcolo del rapporto di correlazione, avendo cura di selezionare la variabile dipendente e la variabile indipendente

Di seguito è riportato un esempio di lettura e interpretazione dei risultati ottenuti:

- *da 0 a 0,25 la dipendenza in media è bassa*
(es. η^2=0,18 -> 18% della max dipendenza in media osservabile, quindi si ha una intensità debole)

- *da 0,25 a 0,5 la dipendenza in media è medio-bassa*
(es. η^2=0,36 -> 36% della max dipendenza in media osservabile, quindi si ha una intensità medio-bassa)

- *da 0,5 a 0,75 la dipendenza in media è medio-alta*
(es. η^2=0,69 -> 69% della max dipendenza in media osservabile, quindi si ha una intensità medio-alta)

- *da 0,75 a 1 la dipendenza in media è alta*
(es. η^2=0,83 -> 83% della max dipendenza in media osservabile, quindi si ha una intensità forte)

Attenzione! *Se invertiamo il ruolo delle variabili abbiamo uno stesso criterio di lettura ma una interpretazione completamente differente del fenomeno*

21 – Esempio

MARCA CELLULARE	MINUTI DI CONVERSAZIONE 0 -| 30	30 -| 60	60 -| 90	
NOKIA	12	8	5	25
APPLE	6	14	12	32
SAMSUNG	11	15	17	43
	29	37	34	100

Sono stati intervistati 100 viaggiatori del treno Napoli – Roma chiedendo la marca preferita di cellulare e i minuti di conversazione effettuati in una giornata: c'è dipendenza in media tra i due caratteri? ***(quanto parlo al cellulare dipende mediamente dalla marca che preferisco?)***

Calcoliamo le medie condizionate per sapere quanti minuti di conversazione effettua in media chi preferisce NOKIA, APPLE o SAMSUNG

M(min|N) = 36,60 minuti
M(min|A) = 50,63 minuti
M(min|S) = 49,19 minuti

Le medie condizionate sono diverse tra loro e dalla media generale, quindi l'ipotesi di indipendenza in media non è ammissibile

DEV(TOT) = 56475 DEV(B) = 3305

$$M(min) = \frac{36{,}60 \cdot 25 + 50{,}63 \cdot 32 + 49{,}19 \cdot 43}{100} = 46{,}50 \text{ minuti}$$

$$\eta^2_{Y|X} = \frac{DEV(B)}{DEV(TOT)} = \frac{3305}{56475} = 0{,}06$$

Avendo osservato il 6% della massima dipendenza in media osservabile, possiamo concludere che il grado di dipendenza tra marca e minuti di conversazione è basso

22 – Relazione tra indipendenza in distribuzione e in media — Unità n° 07

L'indipendenza in distribuzione di una variabile quantitativa Y da una variabile (qualsiasi) X implica anche l'indipendenza in media ma ciò non è sempre vero per il contrario

Supponiamo di voler studiare l'associazione tra il genere musicale preferito e l'età

		Età			
		14 - 16	*17 - 19*	*20 - 22*	
Musica	*POP*	8	10	4	22
	PUNK	24	30	12	66
	ROCK	16	20	8	44
		48	60	24	132

		Età			
		14 – 16	*17 – 19*	*20 – 22*	
Musica	*POP*	0,36	0,45	0,18	1
	PUNK	0,36	0,45	0,18	1
	ROCK	0,36	0,45	0,18	1
		0,36	0,45	0,18	1

Dal confronto dei profili riga si deduce che tra i caratteri c'è ind. in distribuzione

(si può verificare che in questo caso $\chi^2 = 0$)

Se calcoliamo l'età media di coloro che preferisco il POP, il PUNK e il ROCK vediamo che:

M(Età|POP) = M(Età|PUNK) = M(Età|ROCK) =M(Età) = 17,45 anni

C'è ind. in media

Il tempo di attesa dipende dal tipo di auto acquistato?

		Tempo di attesa (mesi)				
		0	**1**	**2**	**3**	
Auto	*s.wagon*	2	6	40	10	58
	berlina	0	2	4	2	8
		2	8	44	12	66

M(tempo|s.wagon) = M(tempo|berlina) =M(tempo)= 2

$$\frac{n_{1\cdot} \cdot n_{\cdot 1}}{n} = \frac{2 \cdot 58}{66} = 1,76 \neq n_{11} = 2$$

C'è ind. in media ma non c'è ind. in distribuzione

23 – Analisi della interdipendenza lineare

Quando si analizzano due o più caratteri quantitativi si può cercare di individuare una funzione che descriva in modo dettagliato la relazione che emerge dai dati, con uno scopo *descrittivo*, *interpretativo* o *previsivo*

Consideriamo due variabili quantitative **Y** e **X**, e supponiamo di essere interessati a comprendere come la **Y** (var. *di risposta* o *dipendente*) sia influenzata dalla **X** (var. *esplicativa* o *indipendente*)

Una variabile **Y** è una funzione di **X** se ad ogni valore di **X** corrisponde uno e un solo valore di **Y**: in tale caso si dice che tra le due variabili c'è una ***relazione funzionale***

esempio

Se X è la lunghezza del lato di un quadrato e Y la sua area allora => $Y = X^2$

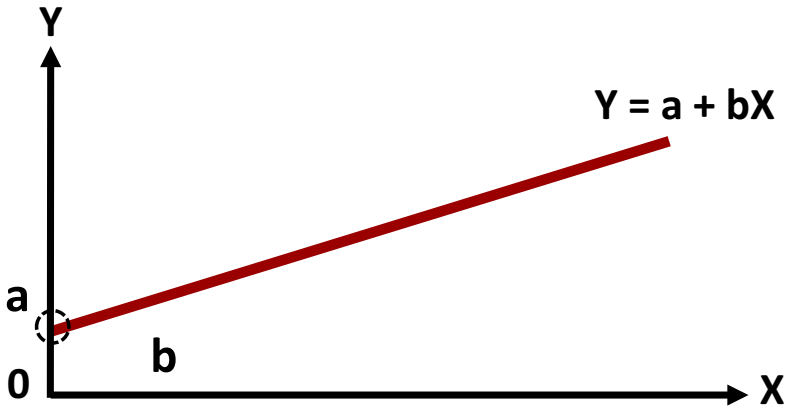
Se X è la temperatura in gradi Celsius e Y quella in gradi Fahrenheit allora => $Y = 32 + 1{,}8X$

Il secondo caso è un esempio di relazione funzionale lineare (=> *equazione di una retta*)

$$Y = a + bX$$

coefficiente angolare: *inclinazione della retta sul piano cartesiano*

intercetta: *intersezione tra la retta e l'asse verticale (X = 0)*



24 – Concordanza

Uno degli aspetti fondamentali dello studio della relazione tra due variabili quantitative è legato al concetto di ***concordanza***, ossia la ricerca della direzione della dipendenza tra **Y** e **X**:

→ *ci si chiede se a valori inferiori (superiori) della media di una variabile si accompagnano valori superiori (inferiori) della media dell'altra*

Per ciascuna delle possibili combinazioni di valori di **Y** e **X** è possibile avere una indicazione sulla direzione della dipendenza calcolando gli *scarti misti*

$$s_{ij}=[x_i-M(X)][y_j-M(Y)]$$ → il segno dello scarto misto indica la direzione della relazione tra **Y** e **X**

→ Quando $s_{ij} > 0$ si ha che $x_i > M(X)$ e $y_j > M(Y)$ oppure $x_i < M(X)$ e $y_j < M(Y)$ **CONCORDANZA**

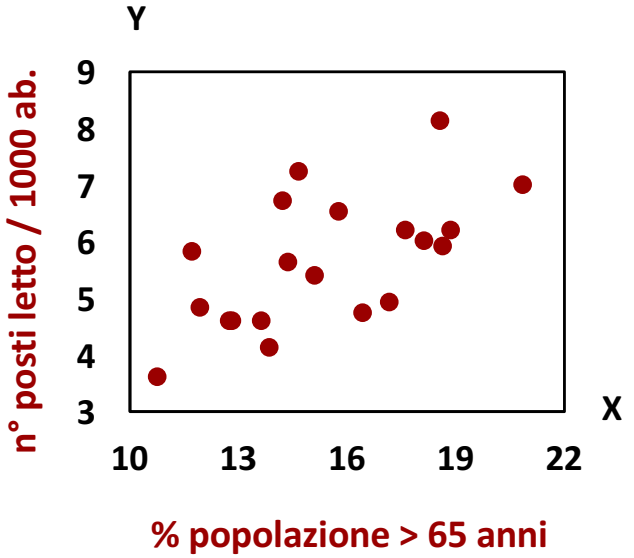
→ Quando $s_{ij} < 0$ si ha che $x_i > M(X)$ e $y_j < M(Y)$ oppure $x_i < M(X)$ e $y_j > M(Y)$ **DISCONCORDANZA**

Poiché non è possibile cogliere il segno della concordanza (o discordanza) osservando tutti gli scarti misti uno per volta, è necessario costruire una misura di sintesi => **COVARIANZA**

$$COV(X,Y)=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}[x_i-M(X)][y_j-M(Y)]}{n}$$

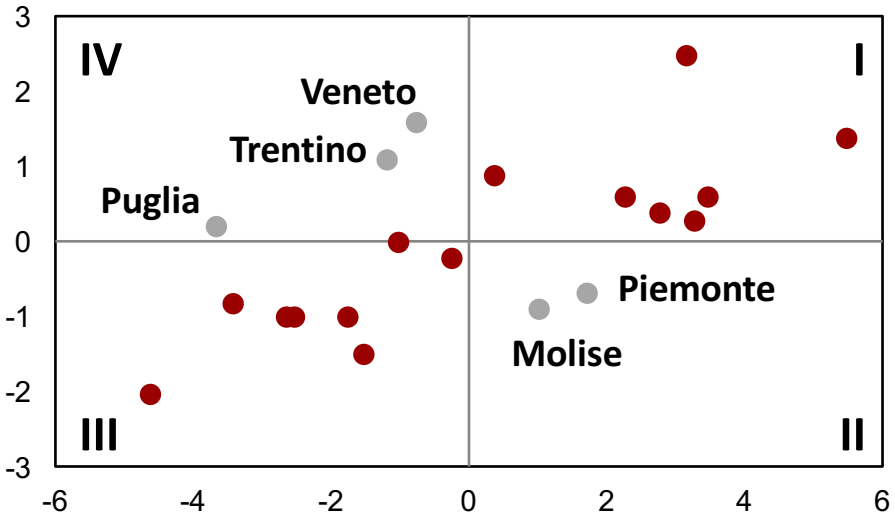
25 – Esempio

Regione	X	Y	Regione	X	Y
Piemonte	17.2	4.9	*Marche*	17.7	6.2
Valle d'Aosta	15.2	5.4	*Lazio*	13.7	4.6
Lombardia	14.4	5.6	*Abruzzo*	15.8	6.5
Trentino-Alto Ad.	14.3	6.7	*Molise*	16.5	4.7
Veneto	14.7	7.2	*Campania*	10.8	3.6
Friuli-Venez.-Giul.	18.6	8.1	*Puglia*	11.8	5.8
Liguria	20.9	7.0	*Basilicata*	13.9	4.1
Emilia Romagna	18.9	6.2	*Calabria*	12.8	4.6
Toscana	18.7	5.9	*Sicilia*	12.9	4.6
Umbria	18.2	6.0	*Sardegna*	12.0	4.8



Dall'analisi degli scostamenti si rileva come Molise e Piemonte hanno una percentuale di popolazione anziana superiore alla media mentre Puglia, Trentino e Veneto hanno un n° di posti letto in strutture ospedaliere ogni 1000 ab. superiore alla media

Quadranti I e III : scostamenti concordi
Quadranti II e IV : scostamenti discordi



COV(X,Y) = 1.86675

Tra i due caratteri c'è concordanza

26 – Proprietà della covarianza

In generale si ha sempre che **COV(X,Y) = COV(Y,X)** e che **σ^2_X = COV(X,X)**

- Se la covarianza è maggiore di 0 allora c'è una predominanza di scostamenti positivi e quindi tra i due caratteri c'è concordanza
- Se la covarianza è minore di 0 allora c'è una predominanza di scostamenti negativi e quindi tra i due caratteri c'è discordanza
- Se la covarianza è nulla allora gli scostamenti positivi e negativi si bilanciano e quindi si dice che i due caratteri sono **incorrelati**

NB

Se i due caratteri sono statisticamente indipendenti allora sono anche incorrelati (la covarianza è nulla) ma non è vero il contrario: infatti la covarianza si annulla se i prodotti degli scostamenti si compensano, ma ciò può avvenire anche se tra i due caratteri c'è una relazione di tipo non lineare

esempio

X	-2	-1	1	2
Y	8	2	2	8

Se calcoliamo le medie risulta che **M(X) = 0** e **M(Y) = 5** e quindi che la **COV(X,Y) = 0** => tra le due variabili sussiste però una relazione del tipo **$Y = 2X^2$**, quindi c'è perfetta dipendenza ma non lineare

27 – Covarianza per distribuzioni doppie

	y_1	…	y_j	…	y_c	TOT
x_1	n_{11}	…	n_{1j}	…	n_{1c}	$n_{1.}$
…	…	…	…	…	…	…
x_i	n_{i1}	…	n_{ij}	…	n_{ic}	$n_{i.}$
…	…	…	…	…	…	…
x_r	n_{r1}	…	n_{rj}	…	n_{rc}	$n_{r.}$
TOT	$n_{.1}$	…	$n_{.j}$	…	$n_{.c}$	n

Supponiamo che la variabile X e la variabile Y (entrambe quantitative) siano rappresentate come distribuzione doppia di frequenze con una tabella a doppia entrata

Come si calcola in questo caso la covarianza?

$$COV(X,Y)=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}[x_i-M(X)][y_j-M(Y)]\cdot n_{ij}}{n}$$

Bisogna ricordare di moltiplicare ogni scarto misto per la frequenza congiunta corrispondente

In questo caso per semplificare il calcolo della covarianza è opportuno ricorrere ad una trasformazione della formula del tipo:

$$COV(X,Y)=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\left(x_i y_j n_{ij}\right)}{n}-M(X)\cdot M(Y)$$

28 – Esempio

Sono stati registrati per 100 possessori di automobile i **cavalli fiscali** e la **spesa settimanale per carburante** (in €): c'è concordanza tra le due variabili?

CV Fiscali	Consumi carburante (€/per sett.) 0 -| 25	25 -| 50	50 -| 100	
1 - 15	20	28	15	63
16 - 30	6	5	10	21
31 - 50	2	5	9	16
	28	38	34	100

Calcoliamo i cv fiscali medi e la spesa media

$$M(X) = \frac{8 \cdot 63 + 23 \cdot 21 + 40{,}5 \cdot 16}{100} = 16{,}35$$ (CV fiscali -> ≅1505 cc)

$$M(Y) = \frac{12{,}5 \cdot 28 + 37{,}5 \cdot 38 + 75 \cdot 34}{100} = 43{,}25$$ (€/per sett.)

Calcoliamo la covarianza

$$COV(X,Y) = 786{,}3 - 16{,}35 \cdot 43{,}25 = 79{,}16$$

Tra i due caratteri sussiste una relazione, e sono concordi

29 – Una misura normalizzata della covarianza

È possibile dimostrare che la covarianza assume valori nell'intervallo $[-\sigma_x\sigma_y ; +\sigma_x\sigma_y]$, dove σ_x e σ_y rappresentano le deviazioni standard di X e Y

Avendo a disposizione il limite inferiore e superiore della covarianza possiamo allora utilizzare

$$r(X,Y) = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}[x_i - M(X)][y_j - M(Y)]}{\sqrt{\sum_{i=1}^{r}[x_i - M(X)]^2 \times \sum_{j=1}^{c}[y_j - M(Y)]^2}}$$

Coefficiente di correlazione lineare di Bravais - Pearson

Il segno del coefficiente di correlazione corrisponde al segno della covarianza, poiché le quantità al denominatore sono sempre positive

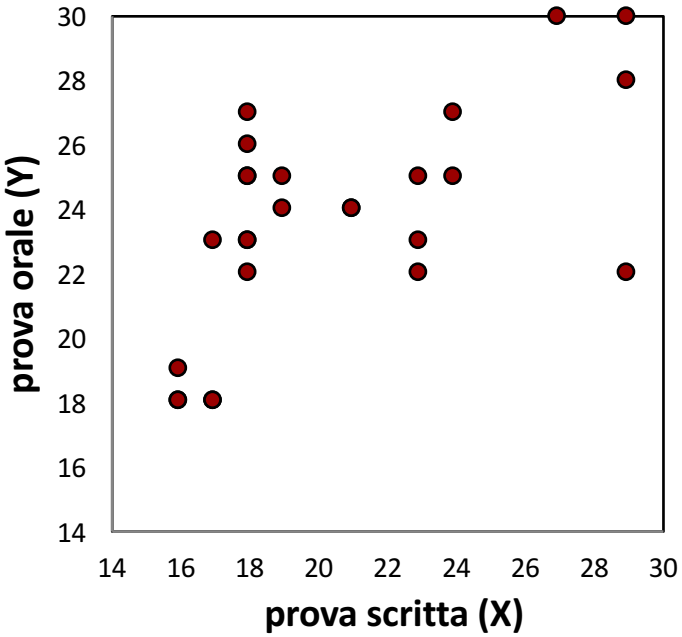
È un indice che varia tra **-1** e **+1**

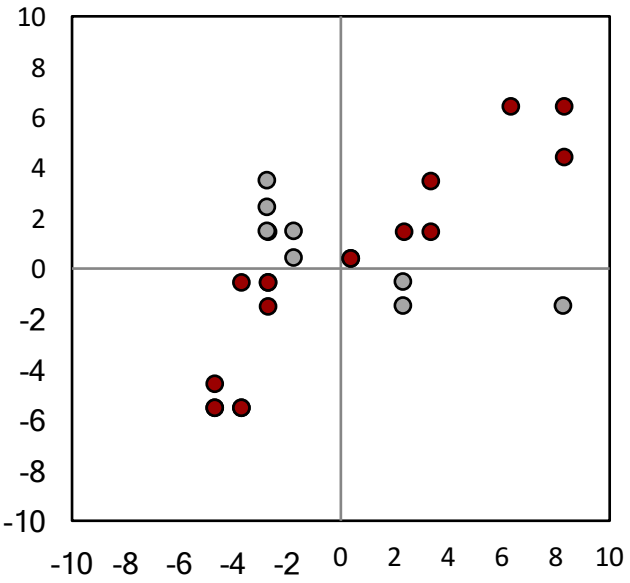
- **-1 se tra i due caratteri c'è perfetta dipendenza lineare e sono discordi**
- **0 se i due caratteri sono indipendenti (o la relazione non è lineare)**
- **+1 se tra i due caratteri c'è perfetta dipendenza lineare e sono concordi**

30 – Esempio

Su un collettivo di 26 studenti sono stati rilevati i voti ottenuti alla prova scritta (X) e orale (Y) dell'esame di statistica: verificare se i due caratteri sono concordi e l'intensità della relazione

Studente	X	Y	Studente	X	Y
1	29	22	**14**	18	23
2	29	28	**15**	18	27
3	27	30	**16**	18	26
4	24	25	**17**	18	22
5	24	27	**18**	18	25
6	23	23	**19**	18	23
7	23	22	**20**	17	18
8	23	25	**21**	17	23
9	21	24	**22**	17	18
10	21	24	**23**	16	18
11	19	25	**24**	16	18
12	19	24	**25**	29	30
13	18	25	**26**	16	19





Somma	536	614
Media	20.61538	23.61538
Varianza	17.23669	11.39053
Dev. Stand.	4.15	3.37

→ **COV(X,Y) = 8.8520**

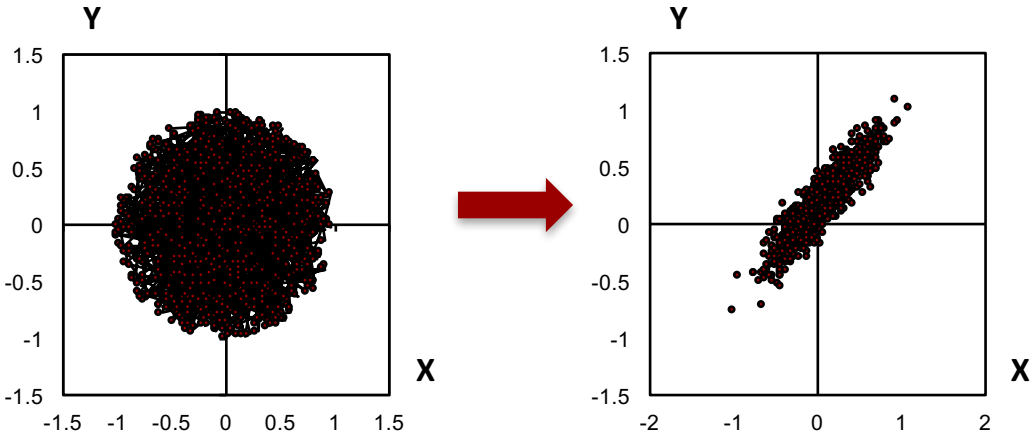
La covarianza è diversa da 0 e positiva, quindi tra i due caratteri sono concordi

$$r(X,Y) = \frac{8.8520}{4.15 \times 3.37} = 0.632$$

L'intensità della dipendenza è del 63.2%, quindi c'è un forte legame tra il voto allo scritto e il voto all'orale

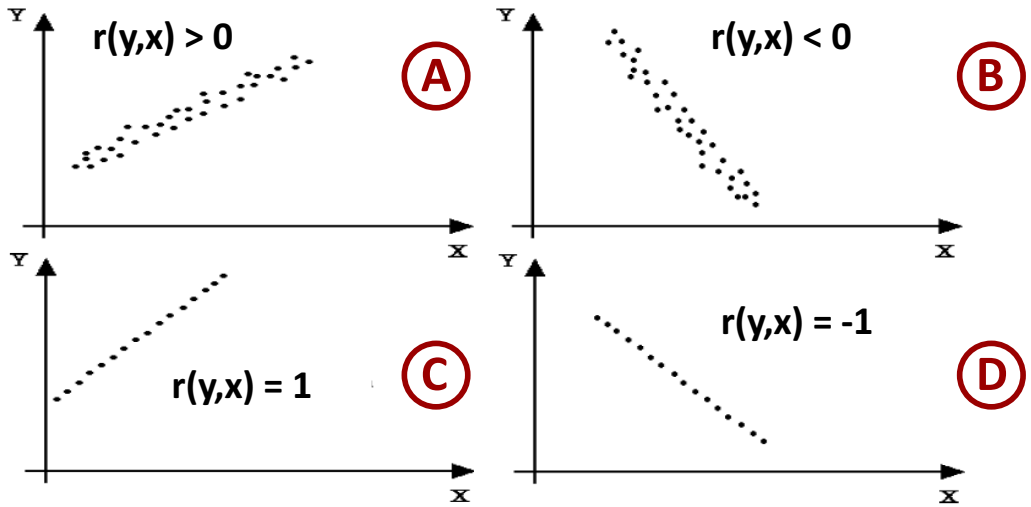
31 – Analisi grafica con il diagramma di dispersione — Unità n° 07

➡ Il diagramma di dispersione fornisce una idea immediata sia sulla direzione della relazione sia sulla sua intensità



La relazione tra i due caratteri tende a divenire più forte man mano che la nube di punti passa dalla forma circolare, all'ellisse ed alla retta

Si tratta comunque di una relazione <u>lineare</u>: i due caratteri possono essere dipendenti ma secondo una diversa relazione funzionale



Se il coefficiente è maggiore di 0 allora c'è un legame diretto tra i due caratteri **(A)**

Se invece il coefficiente è minore di 0 c'è un legame indiretto (inverso) tra i caratteri **(B)**

Se il coefficiente è uguale a 1 allora c'è una relazione di perfetta dipendenza lineare **(C)**

Se invece il coefficiente è uguale a -1 c'è una relazione di perfetta dipendenza lineare, ma di tipo inverso **(D)**

32 – Relazioni cause/effetto

Unità n° 07

L'esistenza di correlazione, per quanto intensa, non implica una relazione di causa/effetto

La correlazione indica soltanto che l'insieme dei valori assunti da una variabile tende a disporsi secondo una retta se rappresentato congiuntamente all'altro insieme. I *perché* vanno cercati al di fuori della statistica

ESEMPIO

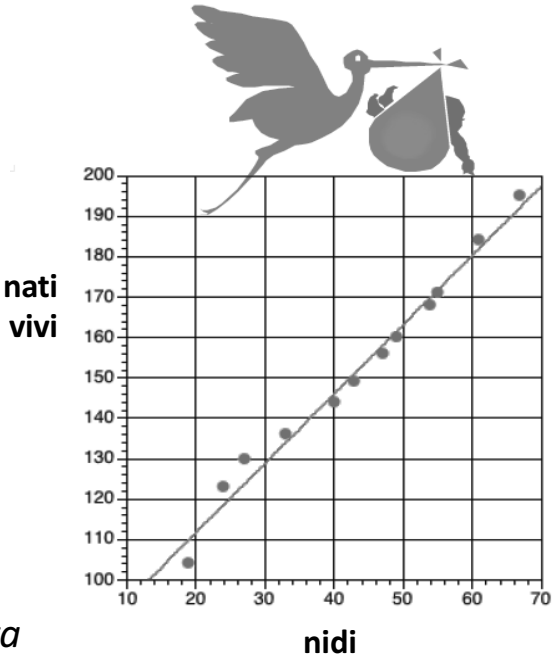
In una zona del Nord Europa è stato monitorato il numero di nidi costruiti dalle cicogne e il numero di nati vivi nel periodo della loro permanenza

	X	Y	X-M(X)	Y-M(Y)	s_{ij}
1972	19	104	-24	-48	1155,917
1973	24	123	-19	-29	551,8333
1974	27	130	-16	-22	352,0833
1975	33	136	-10	-16	160,5833
1976	40	144	-3	-8	24,91667
1977	43	149	0	-3	0,666667
1978	47	156	4	4	16,25
1979	49	160	6	8	47,91667
1980	54	168	11	16	175,5833
1981	55	171	12	19	227,1667
1982	61	184	18	32	573,9167
1983	67	195	24	43	1029,167

	X	Y		
SOMMA	519	1820		
MEDIA	43	152		
VAR	209,8542	627,2222	**COVAR**	359,6667
DEV.ST.	14,48634	25,04441		

$$r(X,Y) = \frac{359.67}{14.5 \times 25.05} = 0.991$$

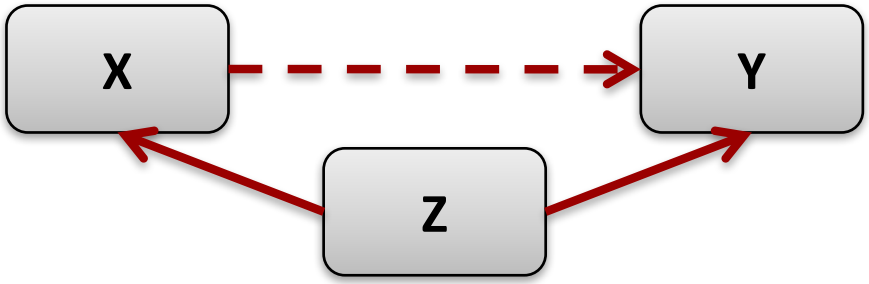
*Si parla in questi casi di **correlazione spuria** (c'è probabilmente una 3a variabile nascosta legata alle altre due)*



33 – La terza variabile nascosta… Unità n° 07

Spesso il valore di r(X,Y) altro non è che l'apparenza di un legame la cui sostanza è invece dovuta a fenomeni esterni

Tale legame non è distinguibile dal legame spurio che tra di esse si pone a causa della comune dipendenza da una terza variabile Z



Questo si verifica spesso a causa dell'esistenza di fenomeni tendenziali di lungo periodo che incidono allo stesso modo su variabili diverse

L'apprendimento di nuove parole non rende i piedi più grandi, come avere piedi più grandi non aiuta a conoscere nuove parole. C'è un terzo fattore nascosto dietro la correlazione: l'età

34 – Esercizi — Unità n° 07

1) *100 famiglie (coppie con o senza figli) sono state classificate secondo il numero di autovetture possedute (Y) e il reddito annuo lordo, in migliaia di Euro (X), ottenendo la seguente tabella*

Autovetture possedute	Reddito annuo lordo (in migliaia di Euro)				
	20-40	*40-60*	*60-100*	*100-120*	**Totale**
0	10	0	0	0	**10**
1	10	20	30	0	**60**
2	0	20	0	10	**30**
Totale	**20**	**40**	**30**	**10**	**100**

Studiare la relazione X|Y tra i due caratteri e valutarne l'intensità con un indice opportuno

---

2) *Vogliamo studiare la mobilità del voto degli elettori di una certa circoscrizione. Da un sondaggio telefonico risulta che:*

	Voterà			
Ha votato	Centro	Destra	Sinistra	Totale
Centro	8	11	2	21
Destra	2	9	2	13
Sinistra	4	2	10	16
Totale	14	22	14	50

C'è connessione tra il voto espresso in passato e le intenzioni di voto alle prossime elezioni?

35 – Relazione funzionale e relazione statistica

Negli studi empirici, la relazione tra Y e X non è mai funzionale (a un valore X corrispondono più valori di Y). Una **relazione statistica** tra la Y e la X può essere descritta da:

$$Y = f(X) + \varepsilon$$

La prima parte **f(X)** - la componente **DETERMINISTICA** - rappresenta il contributo della X, la seconda parte **ε** invece - la componente **STOCASTICA** - rappresenta il contributo di tutto ciò che non è stato osservato, ed è perciò considerato un termine di *errore*

Come visto quando si assume **f(X) = a+bx** la relazione è di tipo lineare e quindi il modello che viene utilizzato per studiare la variabile Y in funzione della variabile X è la cosiddetta

REGRESSIONE LINEARE SEMPLICE

36 – Regressione lineare semplice

Date due variabili X (indipendente) e Y (dipendente) si assume che:

$$y_i = a + bx_i + e_i$$

dove:

$a + bx_i$ *rappresenta una retta*

a *= ordinata all'origine* $\rightarrow$ *intercetta*

b *= coefficiente angolare* $\rightarrow$ *coefficiente di regressione*

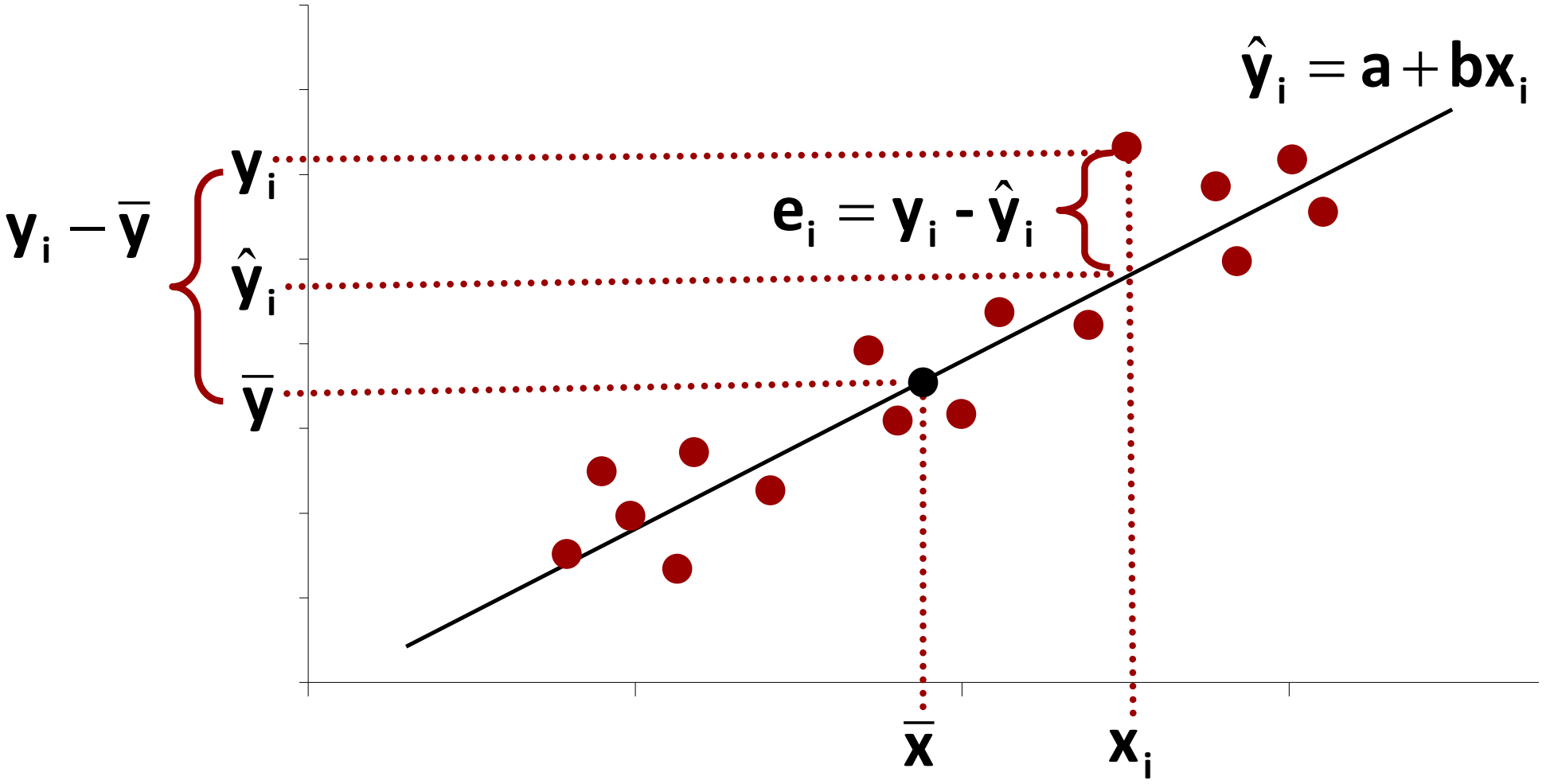
e_i *è un termine di errore (accidentale)*

Affinchè si possa parlare di regressione lineare semplice si devono considerare alcune assunzioni fondamentali:

1) la relazione tra X e Y deve essere lineare
2) i valori di X devono essere noti e senza errori
3) gli errori devono avere media nulla e varianza costante

37 – Retta di regressione

La retta di regressione è il modello attraverso il quale si rappresenta il fenomeno oggetto di studio e la relazione tra le variabili considerate:



38 – Calcolo di a e b

Per poter calcolare l'intercetta e il coefficiente di regressione è necessario minimizzare gli errori. Tale procedura è chiamata dei **minimi quadrati**:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min$$

Il risultato di tale operazione consente quindi di ottenere:

$$b = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{COV(X,Y)}{\sigma_X^2}$$

indica di quanto varia in media Y per una variazione unitaria di X

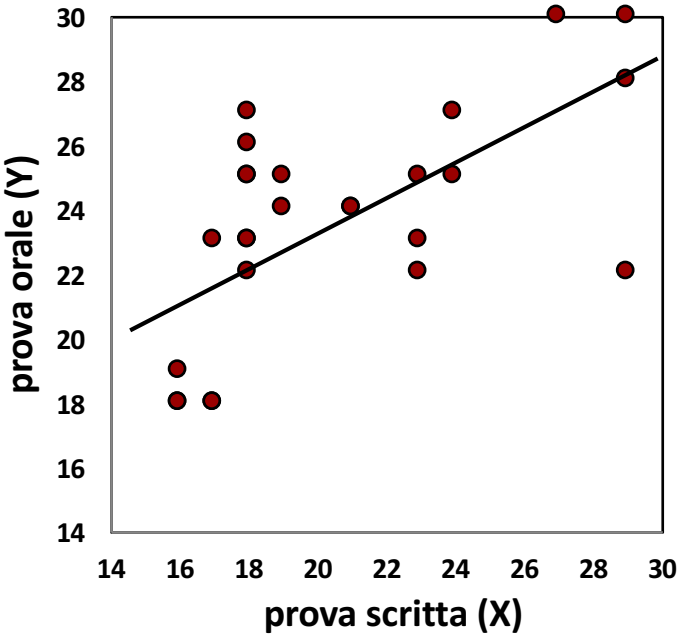
$$a = \bar{y} - b\bar{x}$$

indica il valore di Y per valori nulli di X

39 – Esempio

Studente	X	Y	Studente	X	Y
1	29	22	**14**	18	23
2	29	28	**15**	18	27
3	27	30	**16**	18	26
4	24	25	**17**	18	22
5	24	27	**18**	18	25
6	23	23	**19**	18	23
7	23	22	**20**	17	18
8	23	25	**21**	17	23
9	21	24	**22**	17	18
10	21	24	**23**	16	18
11	19	25	**24**	16	18
12	19	24	**25**	29	30
13	18	25	**26**	16	19

Somma	536	614
Media	20.61538	23.61538
Varianza	17.23669	11.39053
Dev. Stand.	4.15	3.37



$$y = 13.08 + 0.51x$$

il coefficiente di regressione indica che in media per ogni punto ottenuto in più alla prova scritta si otterrà mezzo punto in più alla prova orale

$$b = \frac{8.8520}{17.23669} = 0.5136$$

$$a = 23.6154 - 0.5136 \times 20.6154 = 13.0823$$

40 – Decomposizione della devianza

La variabilità della variabile Y può essere studiata attraverso una particolare decomposizione

$$DEV(Y) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$DEV(Y) = DEV(\hat{Y}) + DEV(E)$$

DEVIANZA DI REGRESSIONE

$$DEV(\hat{Y}) = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

è la parte di variabilità che il modello riesce a spiegare

DEVIANZA RESIDUA

$$DEV(E) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

è la parte di variabilità che rimane ignota

41 – Bontà di adattamento

Dato un insieme di osservazioni è sempre possibile costruire un modello di regressione lineare, ma è necessario valutare la bontà dell'approssimazione

In generale si ha un miglior adattamento quando DEV(E) è piccola, poiché minori sono gli scostamenti tra valori osservati e teorici della Y. È possibile costruire un indice che misura la bontà di adattamento al modello lineare:

COEFFICIENTE DI DETERMINAZIONE

$$R^2 = \frac{DEV(\hat{Y})}{DEV(Y)} = 1 - \frac{DEV(E)}{DEV(Y)}$$

- se $\sum (y_i - \hat{y}_i)^2 = 0$ $R^2=1$ — i valori osservati sono sempre uguali ai valori teorici (perfetto adattamento)
- se $\sum (\hat{y}_i - \bar{y})^2 = 0$ $R^2=0$ — I valori osservati sono sempre uguali alla media di Y (incorrelazione)

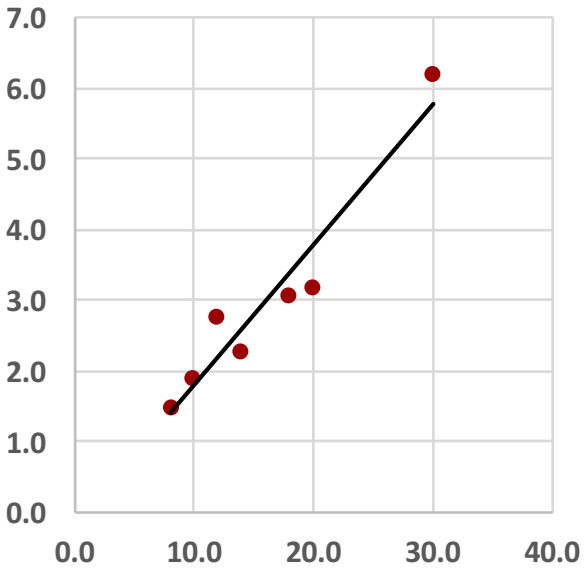
42 – Esempio

Consideriamo i 7 ipermercati operanti in un grande comune della Lombardia: esiste una relazione tra fatturato e numero di dipendenti?

	N. dipendenti (X)	Fatturato in milioni di € (Y)
A	10	1,9
B	18	3,1
C	20	3,2
D	8	1,5
E	30	6,2
F	12	2,8
G	14	2,3

media	16,0	3,00
varianza	48,0	2,04
dev. standard	6,93	1,43

$y = -0.1714 + 0.1982x$



$\hat{y}_i$	e_i^2	$(\hat{y}_i - \bar{y})^2$
1,81	0,01	1,42
3,40	0,09	0,16
3,79	0,35	0,62
1,41	0,01	2,52
5,78	0,18	7,67
2,21	0,35	0,63
2,60	0,09	0,16
TOT	1,08	13,20

$DEV(Y) = 14,28 \quad COV(X,Y) = 9,5143$

$$R^2 = \frac{13,20}{14,28} = 0,9244$$

Il 93% circa della variabilità del fatturato è spiegata dal modello di regressione

43 – Bontà di adattamento e correlazione

Si può dimostrare che il coefficiente di determinazione è legato al coefficiente di correlazione lineare dalla seguente relazione:

$$R^2 = r(X,Y)^2 = \left(\frac{\sigma_{XY}}{\sigma_X \sigma_Y} \right)^2$$

Se conosciamo il livello di correlazione tra X e Y è possibile calcolare la bontà di adattamento: il modello è tanto più valido quanto più la variabile risposta Y e la sua approssimazione lineare tramite X, indicata con $\hat{Y}$, hanno una correlazione vicina a 1

Riprendendo l'esempio dei 7 ipermercati avremo:

$$r(X,Y) = \frac{9{,}5143}{6{,}93 \times 1{,}43} = 0{,}9615 \quad \Rightarrow \quad R^2 = r(X,Y)^2 = (0{,}9615)^2 = 0{,}9244$$

44 – Estrapolazione

Una volta costruito il modello è possibile tentare di valutare in maniera attendibile il valore che assumerà la variabile di risposta in corrispondenza di un valore noto della variabile esplicativa

$$\hat{y}_{N+1} = a + bx_{N+1}$$

CONDIZIONI

1) validità della retta di regressione (R^2 prossimo ad 1)

2) valore noto della variabile esplicativa non lontano dai valori utilizzati nel calcolo della retta

ESEMPIO

Dato il modello che spiega la relazione tra fatturato e numero di dipendenti

$$y = -0.1714 + 0.1982x$$

A quanto ammonterebbe il fatturato di un supermercato che assumesse 32 dipendenti?

$$\hat{y} = -0.1714 + 0.1982 \cdot 32 = 6.171 \text{ (ML di €)}$$