Statistica Sociale

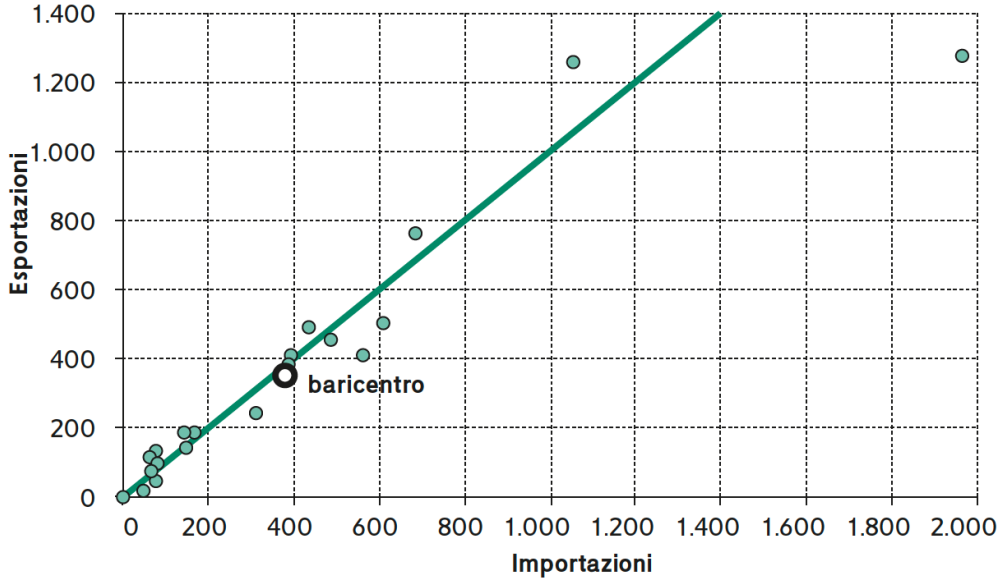
Rappresentazione grafica

- Se entrambi i caratteri sono **quantitativi** possiamo utilizzare il grafico di dispersione

Importazioni ed esportazioni dei paesi OCSE nel 2010 (fonte Nazioni Unite)

Paese	Import.	Esport.	Paese	Import.	Esport.	Paese	Import.	Esport.
Danimarca	84,7	96,8	Italia	487,0	446,9	Finlandia	68,8	69,5
Irlanda	60,7	118,3	Spagna	315,5	246,3	Svizzera	166,9	185,8
Inghilterra	561,5	410,2	Portogallo	75,6	48,7	Austria	150,3	144,6
Olanda	440,6	440,6	Grecia	50,7	20,9	Turchia	140,9	185,5
Belgio	393,5	409,3	Islanda	3,9	4,6	USA	1.968,8	1.277,6
Germania	1.056,2	1.261,6	Norvegia	77,3	131,4	Canada	390,5	386,0
Francia	605,6	515,3	Svezia	148,5	158,1	Giappone	692,4	769,8

- Grafico di dispersione delle importazioni e delle esportazioni dei paesi OCSE, 2010



Interpretazione

Il grafico di dispersione è la rappresentazione delle coppie (x1, y1), (x2, y2),...(xN,yN) ossia della distribuzione disaggregata della variabile doppia (X,Y).

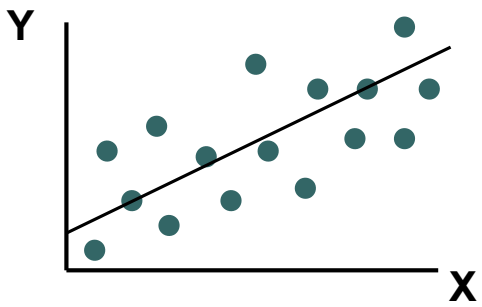
Si dice che tra X e Y c'è un'associazione positiva quando essi tendono a crescere insieme.

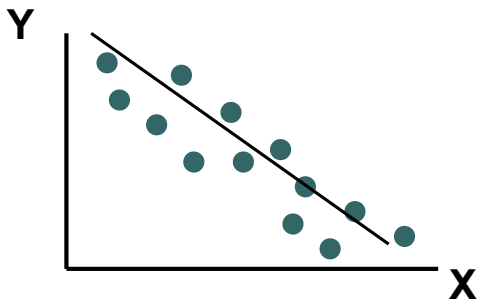
Si dice che tra X e Y c'è un'associazione negativa quando essi tendono a diminuire insieme.

E sempre utile cercare di stabilire, sulla base della sola osservazione del diagramma di dispersione, se esiste associazione positiva o negativa tra due caratteri

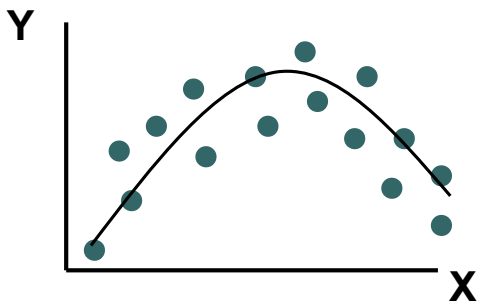
Grafico di dispersione (scatterplot): esempi di relazione/1

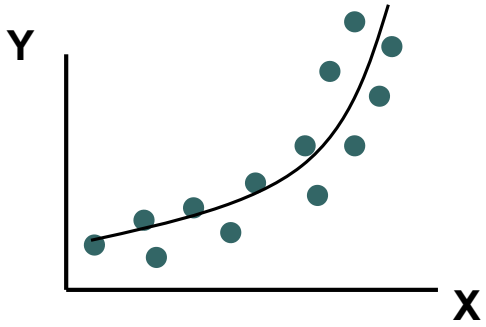
Relazione Lineare





Relazione non lineare





Il grafico di dispersione fornisce un riscontro sul fatto che la relazione sia approssimativamente lineare o non lineare (curvilinea)

Grafico di dispersione (scatterplot): esempi di relazione/2

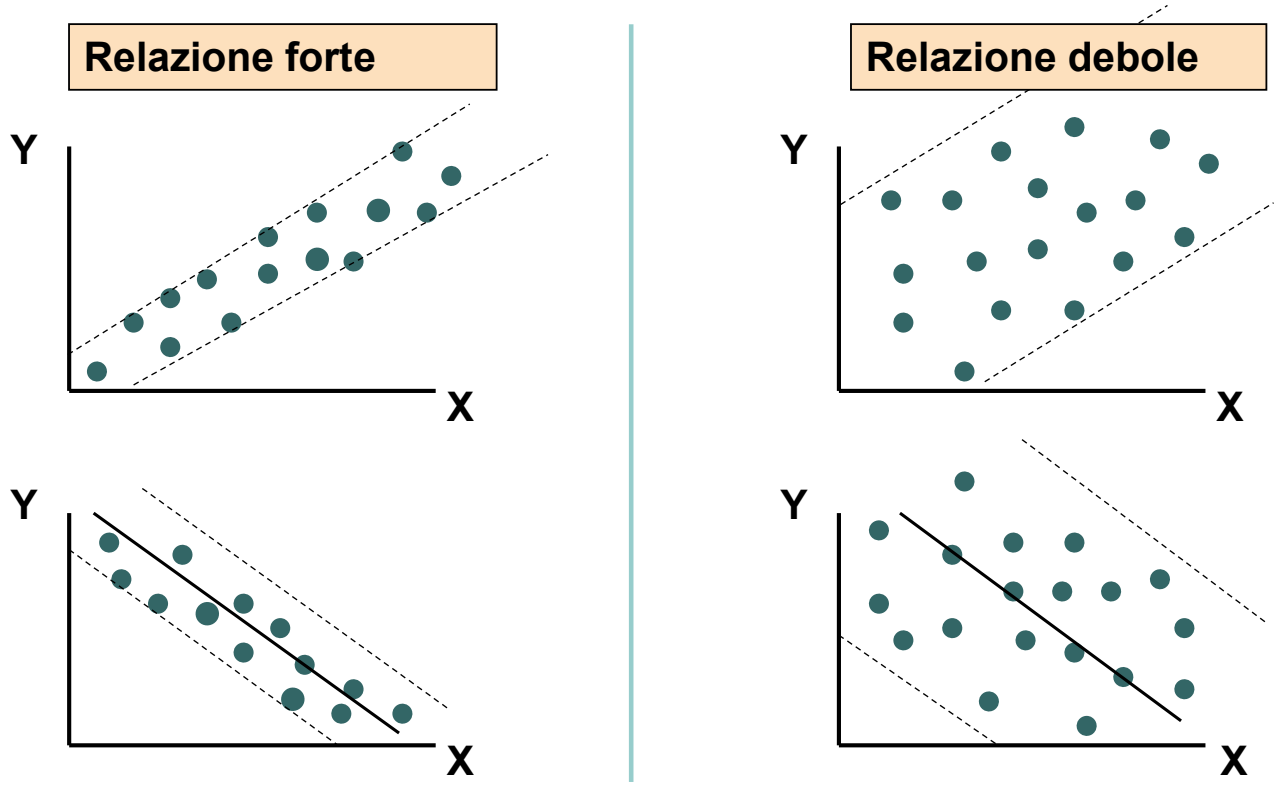
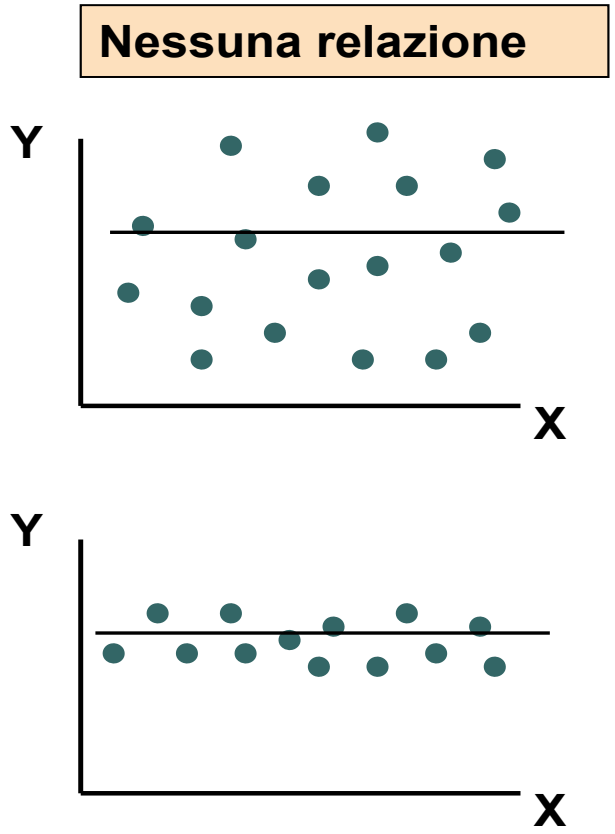


Grafico di dispersione (scatterplot): esempi di relazione/3

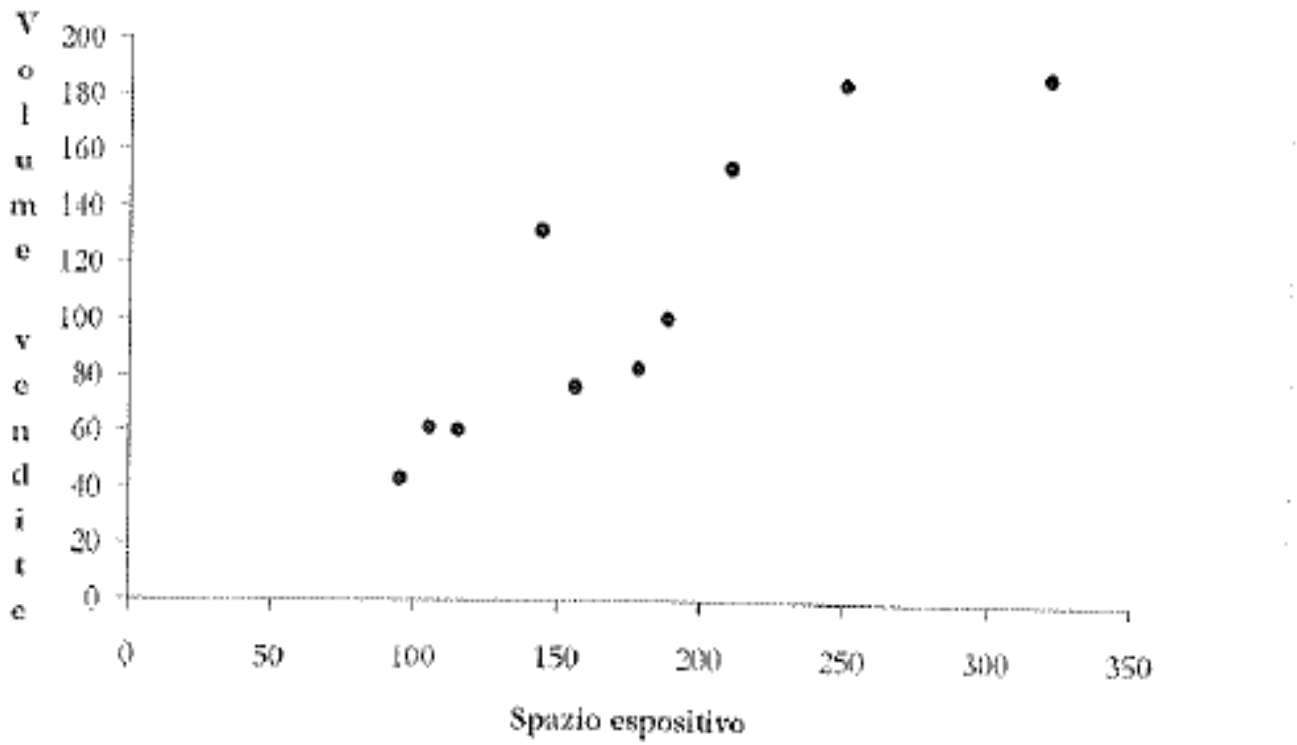


Esempio: vendite e superfice negozio

Volume vendite settimanali (*100 €)	Spazio espositivo (m^2)
43,2	95
132,0	144
155,0	210
76,0	156
100,9	188
187,4	321
185,0	250
60,7	115
82,9	178
61,3	105

Rilevazione delle variabili su 10 negozi

Esempio: vendite e superfice negozio



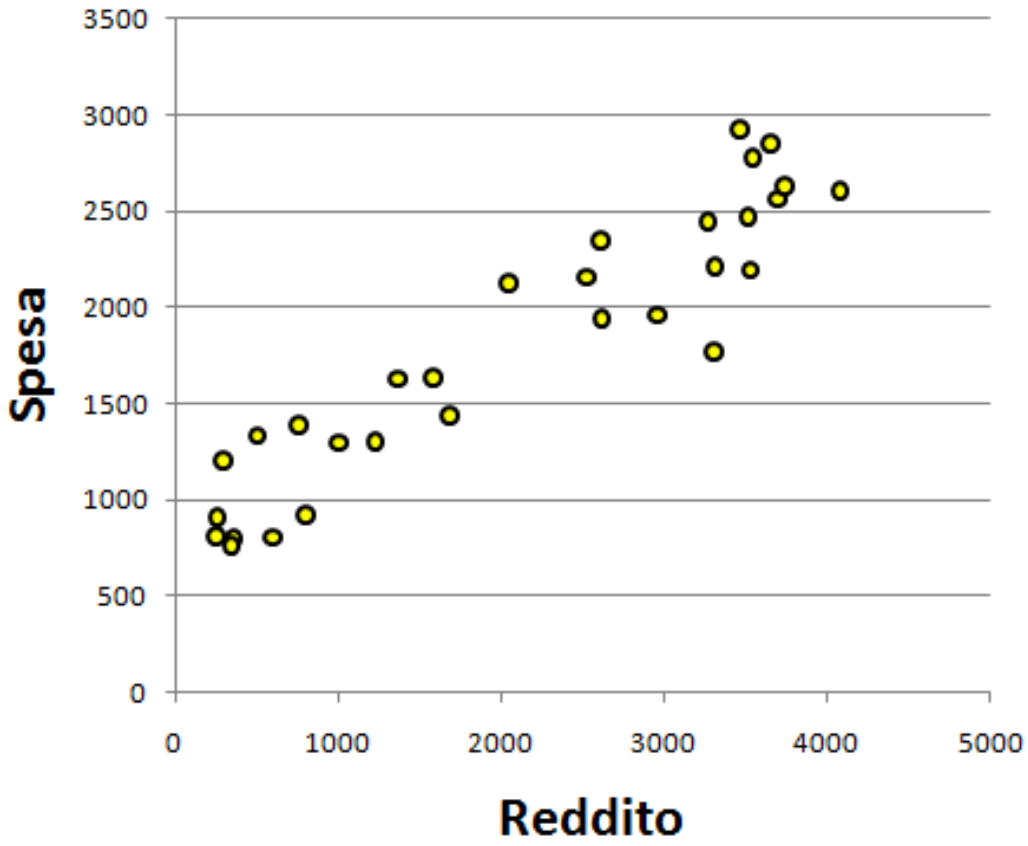
E’ ragionevole ipotizzare che

a maggiori spazi espositivi tendano a corrispondere valori più elevati delle vendite.

La disposizione dei punti sembra essere approssimata bene da una retta.

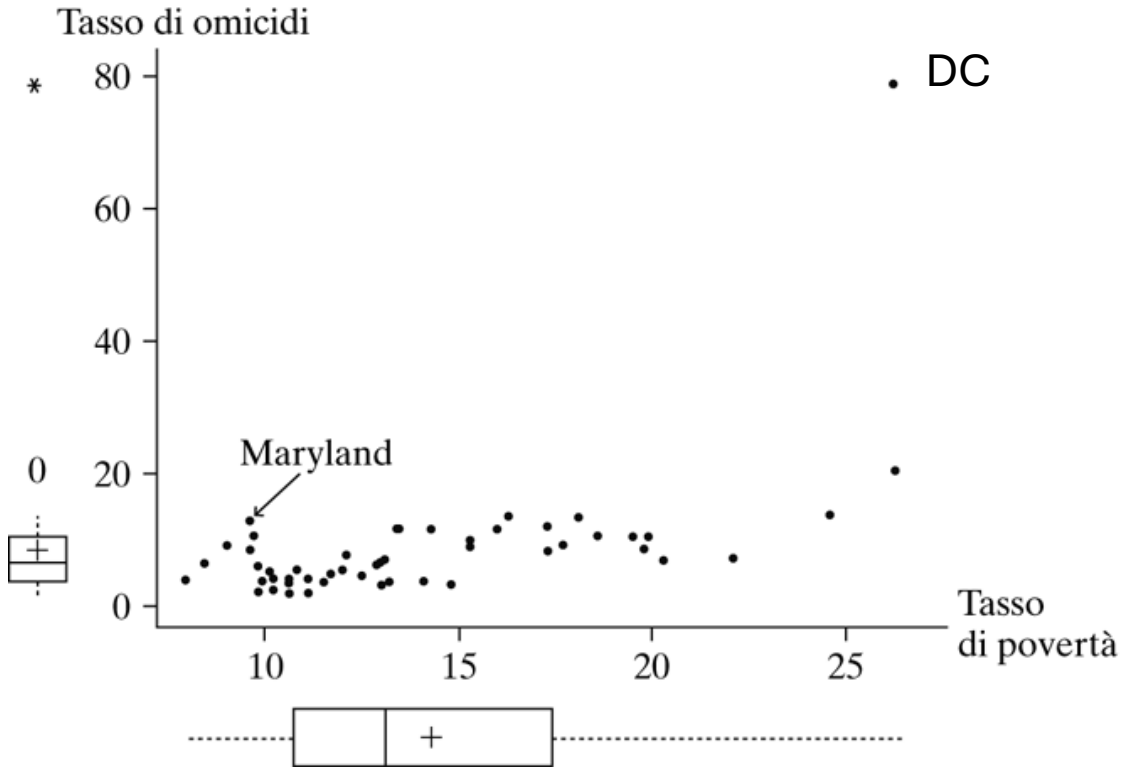
Diagramma di dispersione (scatter plot) per esame grafico della relazione tra le variabili

Esempio: reddito e spesa mensile



Esempio: tasso di omicidi e di povertà

(50 stati USA + Washington DC = 51 osservazioni)



La disposizione dei punti sembra essere approssimata bene da una retta (n.b. un punto –DC– è lontano dal resto delle osservazioni).

Indipendenza statistica

- Attraverso le tabelle di contingenza è possibile studiare l'eventuale dipendenza di una variabile dall'altra
- È possibile verificare ciò attraverso la presenza o meno di indipendenza. Se due variabili categoriali non sono indipendenti, allora sono associate
- Due variabili categoriali sono statisticamente indipendenti se nella popolazione le distribuzioni condizionate di una variabile rispetto a ciascuna categoria dell'altra sono identiche
- Due variabili categoriali sono statisticamente dipendenti se nella popolazione le distribuzioni condizionate di una variabile rispetto a ciascuna categoria dell'altra non sono identiche
- Molto utile è la trasformazione delle frequenze in percentuali, per agevolare l'analisi

	Orientamento Politico			
Gruppo Etnico	Dem	Indip	Repub	Totale
Bianchi	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Neri	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Ispanici	110 (44%)	35 (14%)	105 (42%)	250 (100%)

Che tipo di distribuzioni sono?

Nelle slide precendenti

Distribuzioni condizionate Y|X: come le interpreto?

Titolo di studio	Sinistra	C.-sin.	C.-des.	Destra	Totale
Licenza media	0,083	0,500	0,250	0,167	1,000
Diploma	0,083	0,500	0,250	0,167	1,000
Laurea triennale	0,083	0,500	0,250	0,167	1,000
Laurea magistrale	0,083	0,500	0,250	0,167	1,000
Marginale Y	**0,083**	**0,500**	**0,250**	**0,167**	1,000

Condiziono sulla riga

Ci dice se l'orientamento politico varia a seconda del titolo di studio. Nel nostro caso le righe sono tutte identiche e coincidono esattamente con la distribuzione marginale di Y.

Distribuzioni condizionate X|Y: come le interpreto?

Titolo di studio	Sinistra	C.-sin.	C.-des.	Destra	Marg. X
Licenza media	0,294	0,294	0,294	0,294	0,294
Diploma	0,412	0,412	0,412	0,412	0,412
Laurea triennale	0,176	0,176	0,176	0,176	0,176
Laurea magistrale	0,118	0,118	0,118	0,118	0,118
Totale	**1,000**	**1,000**	**1,000**	**1,000**	1,000

Condiziono sulla colonna

Ci dice se il titolo di studio cambia a seconda dell'orientamento politico. Le colonne sono tutte identiche e coincidono con la marginale di X. Sapere come vota una persona non cambia nulla sulla distribuzione del titolo di studio.

- Due variabili categoriali sono statisticamente indipendenti se nella popolazione le distribuzioni condizionate di una rispetto a ciascuna categoria dell'altra sono identiche

	Orientamento Politico				
Sesso	Dem	Indip	Repub	Totale	*n*
Femmine	38%	34%	28%	100%	1511
Maschi	31%	38%	32%	101%	1260

	Orientamento Politico			
Gruppo Etnico	Dem	Indip	Repub	Totale
Bianchi	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Neri	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Ispanici	110 (44%)	35 (14%)	105 (42%)	250 (100%)

L'indipendenza statistica è una proprietà simmetrica per le due variabili: se le distribuzioni condizionate per ogni riga sono identiche, sono identiche anche quelle per colonna

- Ad esempio per l'orientamento ed il gruppo etnico le distribuzioni condizionate del gruppo etnico rispetto alle modalità dell'orientamento politico sono:

- Dall'evidenza empirica, l'orientamento politico dipende dal sesso perché le distribuzioni condizionate percentuali sono diverse

- Nel campione osservato, l'orientamento politico non dipende dal gruppo etnico perché le distribuzioni condizionate percentuali sono uguali

Test chi-quadro di indipendenza

- Il concetto di indipendenza statistica è riferito alla popolazione, in genere noi disponiamo di dati campionari. Le distribuzioni condizionate campionarie possono essere diverse pur essendo le variabili indipendenti a livello di popolazione
- Per verificare statisticamente la reale esistenza di indipendenza tra due variabili categoriali (a livello di popolazione), possiamo applicare il test chi-quadro per l'indipendenza
- Le ipotesi saranno:
  - H_0: le variabili sono statisticamente indipendenti.
  - H_a: le variabili sono statisticamente dipendenti.

- Requisiti minimi per l'applicazione del test: campionamento casuale o esperimento randomizzato e campione sufficientemente grande.

Frequenze attese per l'indipendenza

Il test del chi-quadro si basa sul confronto tra frequenze osservate e frequenze attese

La frequenza attesa n'_{ij} è quella che potremmo osservare in presenza di indipendenza tra le due variabili, corrisponde cioè alla numerosità attesa in una cella se le due variabili sono indipendenti:

$$n'_{ij} = \frac{n_{i.} n_{.j}}{n} = \frac{totale\ di\ riga\ per\ la\ modalità\ i\ * totale\ di\ colonna\ per\ la\ modalità\ j}{numerosità\ campionaria\ totale}$$

Nel caso di indipendenza tra sesso e orientamento politico avremo:

$$n'_{11} = \frac{1511 * 959}{2771} = 522{,}9 \qquad n'_{21} = \frac{1260 * 959}{2771} = 436{,}1$$

$$n'_{12} = \frac{1511 * 991}{2771} = 540{,}4 \qquad n'_{22} = \frac{1260 * 991}{2771} = 450{,}6$$

$$n'_{13} = \frac{1511 * 821}{2771} = 447{,}7 \qquad n'_{23} = \frac{1260 * 821}{2771} = 373{,}3$$

	Orientamento Politico			
Sesso	Dem	Indip	Repub	Totale
Donne	573 (522.9)	516 (540.4)	422 (447.7)	1511
Uomini	386 (436.1)	475 (450.6)	399 (373.3)	1260
Totale	959	991	821	2771

- Se c'è indipendenza, ci aspettiamo che le frequenze osservate siano «vicine» a quelle attese
- Calcoliamo le differenze tra le frequenze osservate e quelle attese:

	Orientamento Politico			
Sesso	Dem	Indip	Repub	Totale
Donne	573 (522.9)	516 (540.4)	422 (447.7)	1511
Uomini	386 (436.1)	475 (450.6)	399 (373.3)	1260
Totale	959	991	821	2771

$$n_{11} - n'_{11} = 573 - 522{,}9 = 50{,}1$$
$$n_{12} - n'_{12} = 516 - 540{,}4 = -24{,}4$$
$$n_{13} - n'_{13} = 422 - 447{,}7 = -25{,}7$$

$$n_{21} - n'_{21} = 386 - 436{,}1 = -50{,}1$$
$$n_{22} - n'_{22} = 475 - 450{,}6 = 24{,}4$$
$$n_{23} - n'_{23} = 399 - 373{,}3 = 25{,}7$$

Riusciamo a valutare se le differenze sono grandi o piccole?

La statistica test del chi-quadro

Poichè in H_0 si è ipotizzata l'indipendenza tra le due variabili, il test statistico viene costruito con l'intenzione di evidenziare l'allontanamento da H_0

Si basa sulle differenze tra frequenze osservate e frequenze attese

La statistica test è la statistica chi-quadro χ^2 data da

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

Sommiamo per ogni cella, il rapporto tra differenza al quadrato tra frequenza attesa e osservata e la frequenza attesa

- Se $\chi^2 = 0$ le due variabili sono indipendenti (applicando il test a dati campionari, può bastare che sia sufficientemente piccolo)
- Al crescere del valore di χ^2 aumenta l'evidenza contro H_0
- χ^2 non può essere negativo

Questo valore viene confrontato con i valori della distribuzione teorica sotto l'ipotesi di indipendenza

Si ottiene il p-value, un valore che misura quanto i dati osservati sono compatibili con l'ipotesi di indipendenza

Valori soglia per il p-value sono in genere: 0,1; 0,05;0,01

$$n_{11} - n'_{11} = 573 - 522{,}9 = 50{,}1$$
$$n_{12} - n'_{12} = 516 - 540{,}4 = -24{,}4$$
$$n_{13} - n'_{13} = 422 - 447{,}7 = -25{,}7$$
$$n_{21} - n'_{21} = 386 - 436{,}1 = -50{,}1$$
$$n_{22} - n'_{22} = 475 - 450{,}6 = 24{,}4$$
$$n_{23} - n'_{23} = 399 - 373{,}3 = 25{,}7$$

$$\chi^2 = \left(\frac{50{,}1^2}{522{,}9} + \frac{(-24{,}4)^2}{540{,}4} + \cdots + \frac{24{,}4^2}{450{,}6} + \frac{25{,}7^2}{373{,}3}\right) = 4{,}8 + \cdots + 1{,}8 = 16{,}2$$

Il p-value in questo caso è pari a 0,0003

Concludiamo che c'è una forte evidenza empirica contro l'ipotesi di indipendenza H_0, quindi le due variabili sesso e orientamento politico sembrano essere associate nella popolazione

Se le variabili fossero indipendenti, dovrebbe essere davvero inusuale per un campione casuale avere un valore della statistica χ^2 così elevato

V di Cramer

Quindi, si può costruire un indice normalizzato, l'indice V di Cramér, che assumerà valori tra

0 e 1: $0 \leq V \leq 1$

(dove 0=indipendenza perfetta in distribuzione; 1=massima dipendenza)

$$V = \sqrt{\frac{\chi^2}{n * min(r-1,\ c-1)}}$$

r=righe, c=colonne

V	Intensità dell'associazione
0,00 – 0,10	Trascurabile
0,10 – 0,30	Debole
0,30 – 0,50	Moderata
0,50 – 1,00	Forte

Perchè??

Vantaggi rispetto al chi-quadro grezzo:

- Va sempre tra 0 e 1, indipendentemente dalla dimensione della tavola
- Non richiede tavole critiche
- Misura *quanto* sono associati due caratteri, non solo *se* lo sono

Misura della correlazione per variabili quantitative parte 1

- Indicato anche con: r, rho, ρ o anche R
- Il coefficiente è simmetrico: $r_{xy} = r_{yx}$ e misura l'*interdipendenza* tra X e Y
- Il valore di *r* non risente dell'unità di misura e dell'ordine di grandezza di X e Y
- Assume valori compresi tra -1 e 1 (estremi inclusi)

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

$$-1 \leq r \leq +1$$

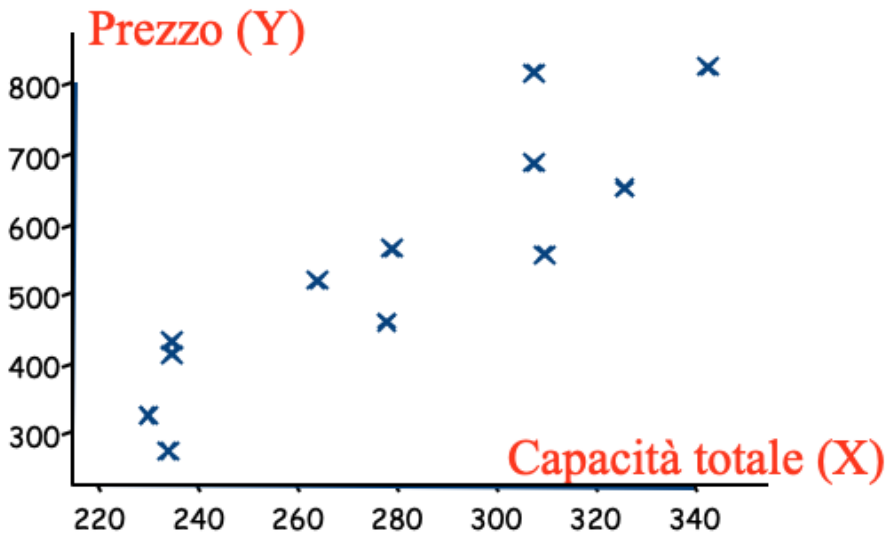
- r_{xy}= -1 perfetta relazione lineare inversa (negativa) tra X e Y
- r_{xy}= 1 perfetta relazione lineare concorde (positiva) tra X e Y
- r_{xy}= -0 X e Y sono incorrelati

Coefficiente di correlazione – Esempio

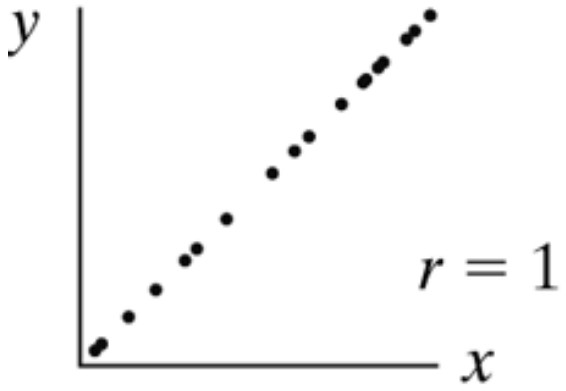
12 congelatori: capacità totale (in litri) e prezzo (in euro)

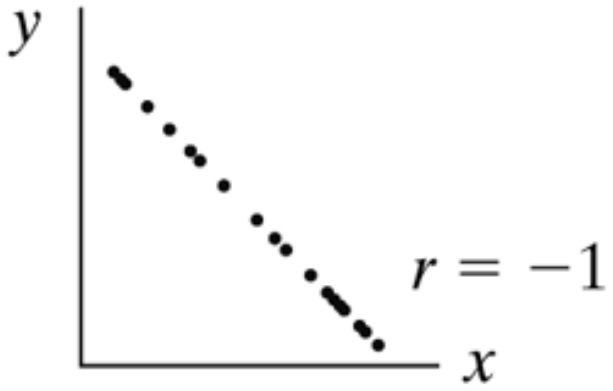
Capacità totale	Prezzo	Capacità totale	Prezzo
230	325	279	564
234	274	308	815
235	412	308	685
235	431	310	556
264	518	326	651
278	460	343	824

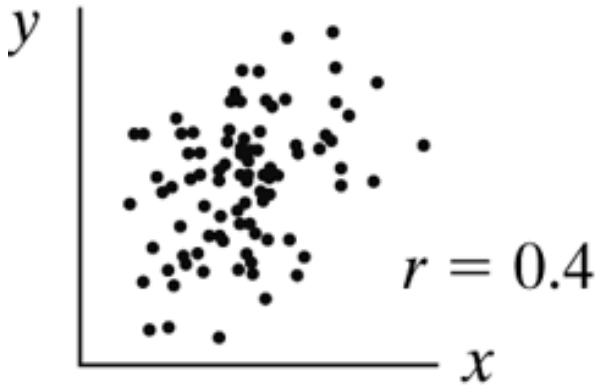
Grafico di dispersione

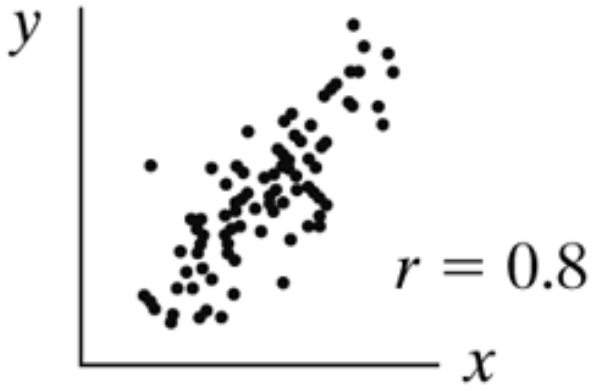


Grafici a dispersione e diversi valori di *r*









Importante:

Un valore di r inferiore in valore assoluto a 1 non implica necessariamente assenza di un legame perfetto tra le variabili, ma assenza di un legame lineare perfetto.

Un valore di r uguale a zero non implica necessariamente assenza di relazione tra le variabili, ma assenza di relazione lineare (più in generale, monotona).