Statistica Sociale

10/04/2025

Analisi dell'associazione tra due variabili

Secondo te la tua famiglia è: (valori %)	I GRADO	II GRADO	Famiglie con solo 1 figlio	Famiglie con più figli	Famiglie monogenitoriali	Famiglie con più nuclei o ricostituite	TOTALE
Ricca	12,1	9,3	12,8	10,0	7,2	10,4	10,4
Né ricca né povera	80,4	80,1	79,4	80,7	79,6	80,7	80,2
Povera	2,1	4,6	2,5	3,2	7,0	3,2	3,7
Non risponde	5,4	6,0	5,3	5,7	6,2	5,7	5,7

Fonte: Istat, elaborazione del SPPS RAFVG

Indagine Istat 2021 "Bambini e ragazzi: comportamenti, atteggiamenti e progetti futuri", campione casuale nazionale di circa 41 mila alunni scuole secondarie di I e II grado. a.s. 2020/2021
FVG: oltre 2 mila bambini e ragazzi, (39,8% secondarie di I grado e 60,2% secondarie II grado; 16,1% con cittadinanza straniera)

SPPS: Servizio programmazione, pianificazione strategica, controllo di gestione e statistica della Direzione generale della Regione Autonoma Friuli Venezia Giulia (report VITA QUOTIDIANA A DISTANZA, 21/12/2023)

Analisi dell'associazione tra due variabili

L'obiettivo delle indagini statistiche che coinvolgono più variabili è quello di studiare l'associazione tra le variabili in esame

Associazione: quando una variabile cambia il suo valore, l'altra variabile tende ad assumere certi valori

Un'analisi tra due variabili è detta bivariata

C'è associazione tra:

- tipologia familiare e situazione economica?
- sesso e retribuzione?
- numero di ore passate all'aria aperta e l'età?

Utilizzando le tabelle di contingenza è possibile osservare la distribuzione dei soggetti secondo tutte le possibili combinazioni tra le modalità di due variabili

	Orientamento Politico			
Sesso	Democratici	Indipendenti	Repubblicani	Totale
Femmine	573	516	422	1511
Maschi	386	475	399	1260
Totale	959	991	821	2771

La tabella mostra la distribuzione degli intervistati al GSS 2004, secondo il sesso e l'orientamento politico

Tabelle di contingenza

Utilizzando le tabelle di contingenza è possibile osservare la distribuzione dei soggetti secondo tutte le possibili combinazioni tra le modalità di due variabili

Unità	Sesso	Età	Statura	Colore occhi
1	F	24	163	Marrone
2	F	21	165	Azzurri
3	M	34	185	Azzurri
4	F	22	164	Marroni
5	F	21	167	Marroni
6	F	22	175	Verdi
7	M	24	178	Verdi
8	F	21	155	Marroni

Sesso	Numerosità
F	6
M	2
	8

Colore occhi	Numerosità
Azzurri	2
Marroni	4
Verdi	2

Sesso/Colore occhi	Azzurri	Marroni	Verdi	
F	1	4	1	6
M	1	0	1	2
	2	4	2	8

Dati due caratteri X e Y si definisce distribuzione doppia di frequenze l'insieme delle frequenze congiunte n_{ij}, ovvero le frequenze assolute delle unità che presentano congiuntamente la modalità *i*-esima della variabile X e la modalità j-esima della variabile Y

- Si possono costruire per tutti i tipi di variabili ma per migliorare la leggibilità della tabella è opportuno ricodificare la variabile in categorie o classi (soprattutto nel caso di caratteri quantitativi continui)

Possiamo identificare le distribuzioni marginali e le distribuzioni condizionate

•Una distribuzione marginale di una variabile corrisponde alla distribuzione di frequenza della singola variabile (totali di riga/totali di colonna)
•Una distribuzione condizionata di una variabile corrisponde alla distribuzione di frequenza di una variabile condizionata rispetto ad una o più modalità dell'altra variabile

Distribuzione condizionata della *X* data *Y=y*

Distribuzione marginale della *X*

		Y					
		y_1	…	y_j	…	y_H	**Totale**
	x_1	n_{11}	…	n_{1j}	…	n_{1H}	$n_{1.}$
	…	…	…	…	…	…	…
X	x_i	n_{i1}	…	n_{ij}	…	n_{iH}	$n_{i.}$
	…	…	…		…	…	…
	x_K	n_{K1}	…	n_{Kj}	…	n_{KH}	$n_{K.}$
	Totale	$n_{.1}$	…	$n_{.j}$	…	$n_{.H}$	n

Distribuzione condizionata della *Y* data *X=x*

Distribuzione marginale della *Y*

Distribuzione doppia, marginale e condizionata

	Orientamento Politico			
Sesso	Democratici	Indipendenti	Repubblicani	Totale
Femmine	573	516	422	1511
Maschi	386	475	399	1260
Totale	959	991	821	2771

Con riferimento alla tabella, individuare

- la distribuzione congiunta della variabile Sesso e Orientamento Politico:
- la distribuzione marginale della variabile Sesso:
- la distribuzione condizionata della variabile Orientamento Politico rispetto alla modalità Femmine:
- la distribuzione condizionata della variabile Sesso rispetto alla modalità Repubblicani:

Distribuzioni doppie

Distribuzione marginale della X

$$n_{i.} = \sum_{j=1}^{H} n_{ij} = n_{i1} + n_{i2} + \cdots + n_{iH}, per\ i = 1, \dots, K$$

Distribuzione marginale della Y

$$n_{.j} = \sum_{i=1}^{K} n_{ij} = n_{1j} + n_{2j} + \cdots + n_{Kj}, per\ j = 1, \dots, H$$

Numerosità campionaria

$$n = \sum_{i=1}^{H} \sum_{j=1}^{K} n_{ij} = \sum_{i=1}^{K} n_{i.} = \sum_{j=1}^{H} n_{.j}$$

Distribuzione marginale del Sesso

$n_{1.} = 6$

$n_{2.} = 2$

Distribuzione marginale del Colore degli occhi

$n_{.1} = 2$

Sesso/Colore occhi	Azzurri	Marroni	Verdi	
F	1	4	1	6
M	1	0	1	2
	2	4	2	8

Dividendo per il corrispondente totale si ottiene:

- la distribuzione di frequenze doppie relative f_{ij} (dividendo le frequenze congiunte per il totale n)

Sesso/Colore occhi	Azzurri	Marroni	Verdi	
F	1	4	1	
M	1	0	1	
				8

Sesso/Colore occhi	Azzurri	Marroni	Verdi	
F	1/8	4/8	1/8	
M	1/8	0/8	1/8	
				1

- le distribuzioni marginali relative (dividendo le frequenze marginali per il totale n)

Sesso/Colore occhi	Azzurri	Marroni	Verdi	
F				6
M				2
	2	4	2	8

Sesso/Colore occhi	Azzurri	Marroni	Verdi	
F				6/8
M				2/8
	2/8	4/8	2/8	1

- le distribuzioni relative condizionate (dividendo le frequenze condizionate per il corrispondente totale di riga/colonna)

Sesso/Colore occhi	Azzurri
F	1
M	1
	2

Sesso/Colore occhi	Azzurri
F	1/2
M	1/2
	1

Sesso/Colore occhi	Azzurri	Marroni	Verdi	
F	1	4	1	6

Sesso/Colore occhi	Azzurri	Marroni	Verdi	
F	1/6	4/6	1/6	1

Moltiplicando le frequenze relative per 100 otteniamo le corrispondenti frequenze percentuali

Tabella 6.2.3 Laureati del 2011 in lauree magistrali per condizione occupazionale nel 2015 (fonte ISTAT)

Condizione occupazionale	Gruppo di corsi di laurea			
	Medico	Economico-statistico	Letterario	Totale
Occupati	9.090	14.787	7.361	31.238
In cerca di lavoro	126	1.534	2.146	3.806
Non cercano lavoro	202	350	522	1.074
Totale	9.418	16.671	10.029	36.118

Tabella 6.2.4 Distribuzioni doppie percentuali e distribuzioni percentuali condizionate

Condizione occupazionale		Gruppo di corsi di laurea			Totale (%)
		Medico	Economico-statistico	Letterario	
Occupati	(% totale)	25,2	40,9	20,4	86,5
	(% riga)	29,1	47,3	23,6	
	(% colonna)	96,5	88,7	73,4	
In cerca di lavoro	(% totale)	0,3	4,2	5,9	10,5
	(% riga)	3,3	40,3	56,4	
	(% colonna)	1,3	9,2	21,4	
Non cercano lavoro	(% totale)	0,6	1,0	1,4	3,0
	(% riga)	18,8	32,6	48,6	
	(% colonna)	2,1	2,1	5,2	
Totale		26,1	46,2	27,8	100,00

Distribuzione percentuale congiunta

Distribuzione condizionata della condizione occupazionale rispetto al gruppo di laurea

Distribuzione condizionata del gruppo di laurea rispetto alla condizione occupazionale

Distribuzione marginale percentuale

Medie condizionate

Per ogni distribuzione condizionata di un carattere quantitativo si può calcolare la media aritmetica condizionata

$$\bar{x}_{Y=y_j} = \frac{1}{n_{.j}} \sum_{i=1}^{K} x_i n_{ij} \qquad \bar{y}_{X=x_i} = \frac{1}{n_{i.}} \sum_{j=1}^{H} y_j n_{ij}$$

Il numero di medie condizionate aumenta all'aumentare delle modalità del carattere

Nel caso di carattere quantitativo suddiviso in classi si può calcolare la media aritmetica approssimata utilizzando il valore centrale di ogni classe

Distribuzione condizionata della X data $Y=y_j$

Distribuzione marginale della X

Y

X

	y_1	…	y_j	…	y_H	**Totale**
x_1	n_{11}	…	n_{1j}	…	n_{1H}	$n_{1.}$
…	…	…	…	…	…	…
x_i	n_{i1}	…	n_{ij}	…	n_{iH}	$n_{i.}$
…	…	…		…	…	…
x_K	n_{K1}	…	n_{Kj}	…	n_{KH}	$n_{K.}$
Totale	$n_{.1}$	…	$n_{.j}$	…	$n_{.H}$	n

Distribuzione condizionata della Y data $X=x_i$

Distribuzione marginale della Y

$$\bar{y}_{X=x_i} = \frac{1}{n_{i.}} \sum_{j=1}^{H} y_j n_{ij}$$

		Y	Numero di case	
		1	**2**	**3**
X Numero di auto	**1**	21	8	0	$29 = n_{1.}$
	2	12	11	1	24
	3	7	6	2	15
		40	25	3	68

$$\bar{y} = \frac{1}{68}(1 * 40 + 2 * 25 + 3 * 3) = 1{,}45$$

$\bar{y}_{X=1} = 1{,}28$
$\bar{y}_{X=2} = 1{,}54$
$\bar{y}_{X=3} = 1{,}67$

$\bar{x}_{Y=1} = 1{,}65$
$\bar{x}_{Y=2} = 1{,}92$
$\bar{x}_{Y=3} = 2{,}67$

$$\bar{x} = \frac{1}{18}(1 * 29 + 2 * 24 + 3 * 15) = 1{,}79$$

$$\bar{y}_{X=1} = \frac{1}{29}(1 * 21 + 2 * 8 + 3 * 0) = 1{,}28$$

$$\bar{y}_{X=2} = \frac{1}{24}(1 * 12 + 2 * 11 + 3 * 1) = 1{,}54$$

$$\bar{x}_{Y=1} = \frac{1}{40}(1 * 21 + 2 * 12 + 3 * 7) = 1{,}65$$

Calcola la media condizionata della variabile numero di case rispetto alle modalità 3 della variabile numero di auto:

Calcola le medie condizionate della variabile numero di auto rispetto alle modalità 2 e 3 della variabile numero di case:

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

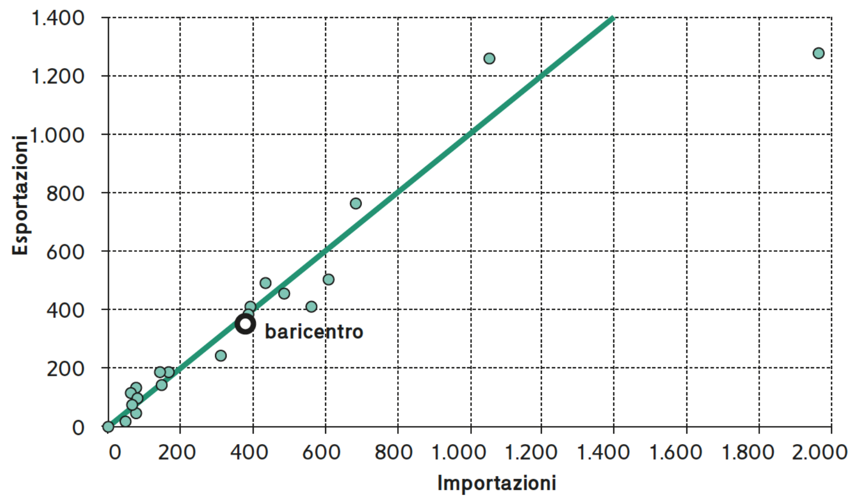
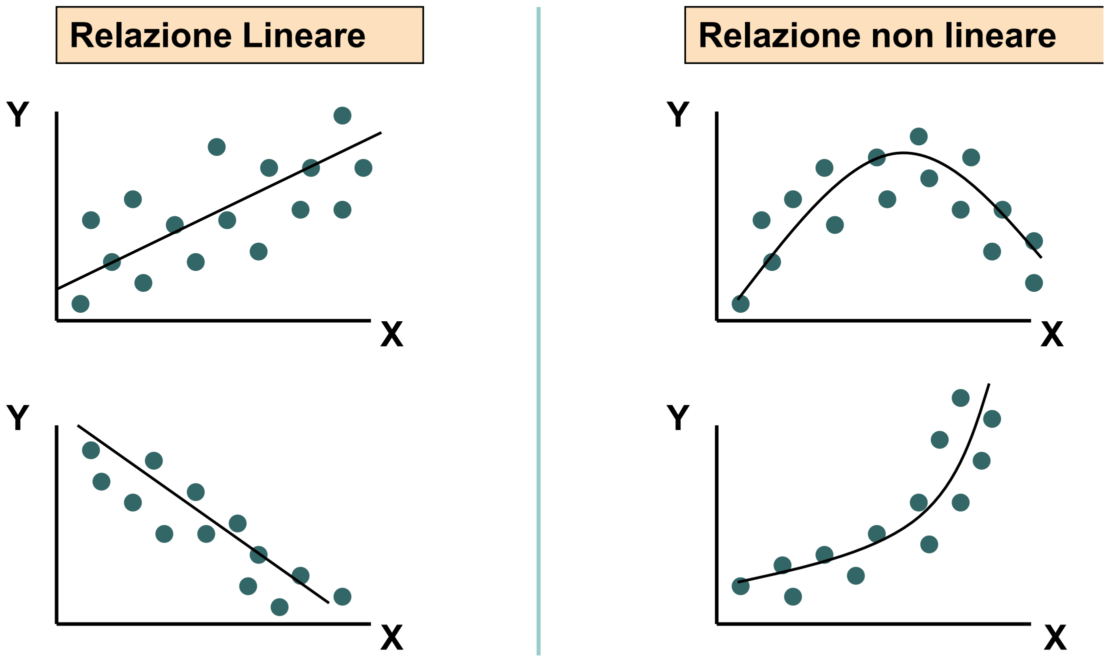


Grafico di dispersione (scatterplot): esempi di relazione/1



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

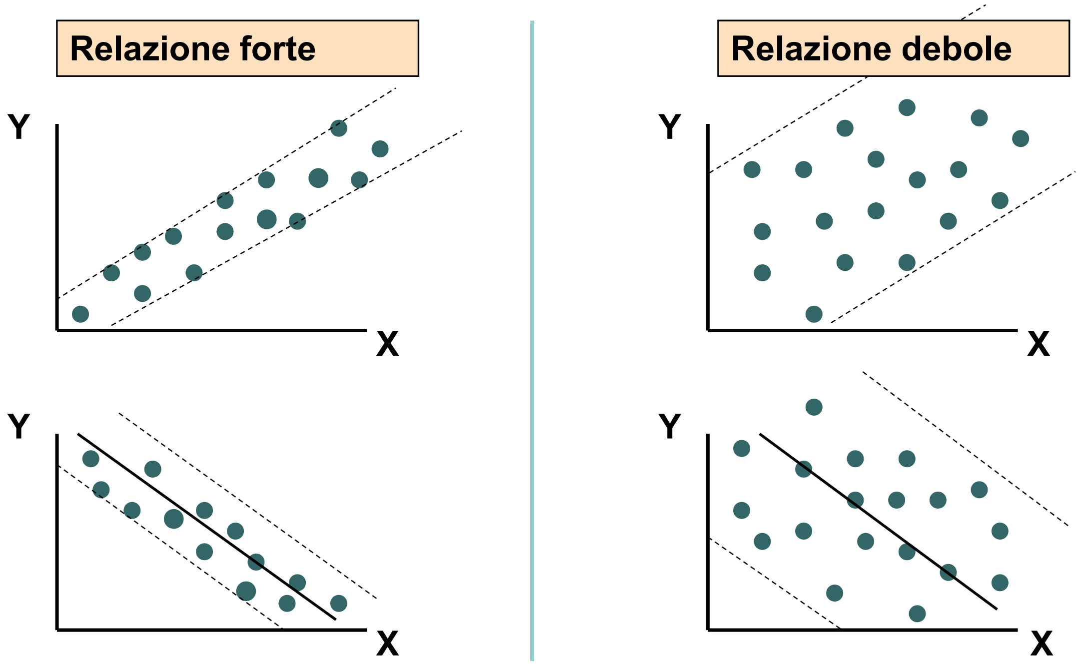
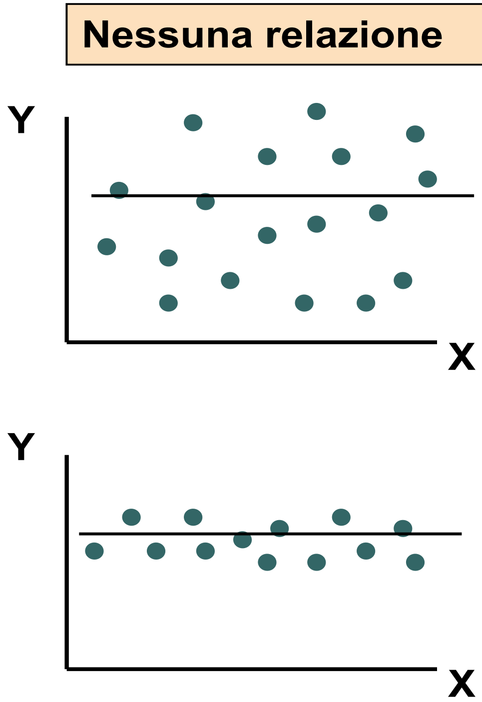


Grafico di dispersione (scatterplot): esempi di relazione/3

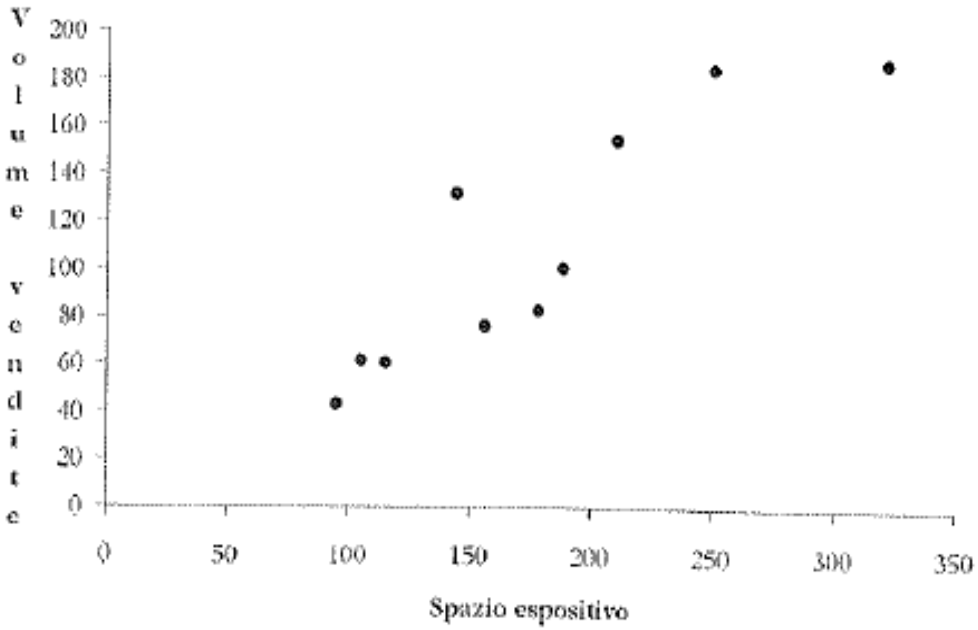


Esempio: vendite e superfice negozio

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

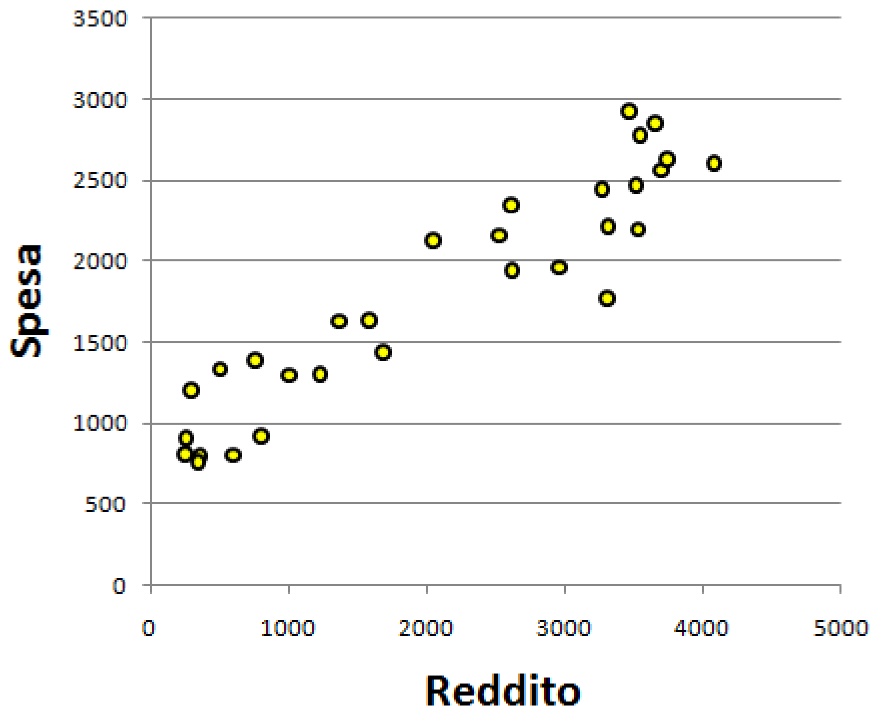


E' ragionevole ipotizzare che a maggiori spazi espositivi tendano a corrispondere valori più elevati delle vendite.

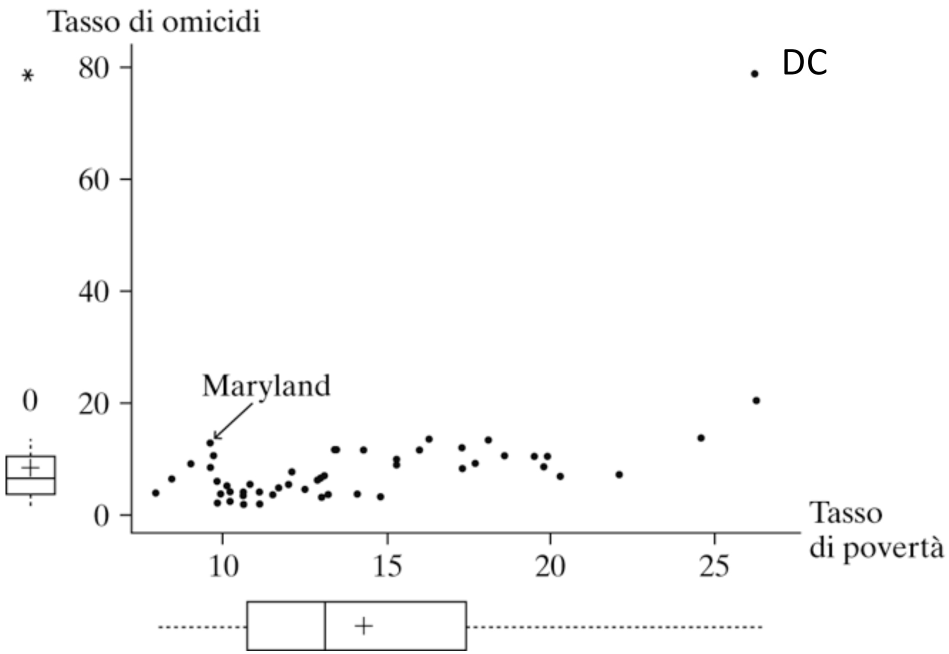
La disposizione dei punti sembra essere approssimata bene da una retta.

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 la comprensione

	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?

Due variabili categoriali sono statisticamente indipendenti se nella <u>popolazione</u> 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

Chi-quadro e associazione

Il test chi-quadro risponde alla domanda «C'è associazione?»

Esistono misure di associazione che sintetizzano la forza di dipendenza tra due variabili

Nel caso A della tabella vediamo un caso di indipendenza

Nel caso B vediamo un caso di dipendenza forte

Caso A	Razze	Opinione		Totale	Caso B	Opinione		Totale
		Favorevole	Contrario			Favorevole	Contrario	
	Bianchi	360	240	600		600	0	600
	Neri	240	160	400		0	400	400
	Totale	600	400	1000		600	400	1000

La statistica chi-quadro indica quanta evidenza c'è a favore della dipendenza, non ne misura la forza (valori più grandi si verificano quando la numerosità è grande)

Chi-quadro e associazione

	A			B			C		
	Sì	No	Totale	Sì	No	Totale	Sì	No	Totale
Bianchi	49	51	100	98	102	200	4900	5100	10 000
Neri	51	49	100	102	98	200	5100	4900	10 000
	100	100	200	200	200	400	10 000	10 000	20 000
	$\chi^2 = 0.08$ P-valore = 0.78			$\chi^2 = 0.16$ P-valore = 0.69			$\chi^2 = 8.0$ P-valore = 0.005		

Il valore di χ^2 non fornisce una misura della forza dell'associazione, in quanto fortemente dipendente dalla dimensione campionaria.

Misura della correlazione per variabili quantitative

Coefficiente di correlazione

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

numeratore della deviazione standard di X

numeratore della deviazione standard di Y

$$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \quad \text{e} \quad s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$$

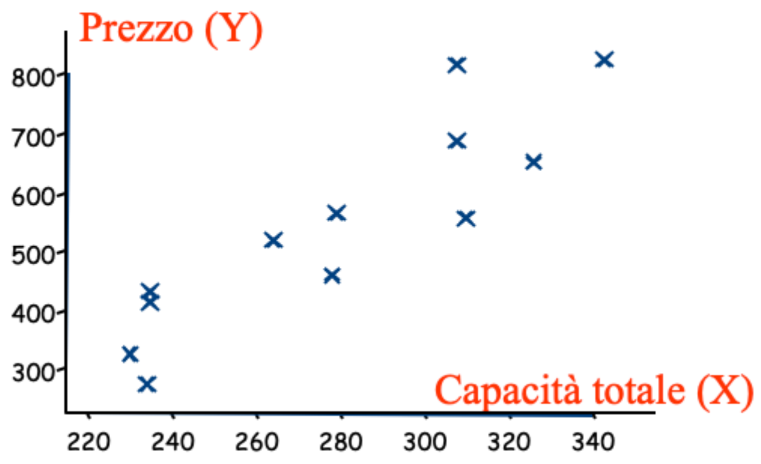
- Indicato anche con: r, rho, ρ o anche R
- Il coefficiente è simmetrico: $r_{xy} = r_{yx}$ e misura l'*interdipendenza* tra X e Y
- Il valore di *r* non risente dell'unità di misura e dell'ordine di grandezza di X e Y
- Assume valori compresi tra -1 e 1 (estremi inclusi)
- r_{xy}= -1 perfetta relazione 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 (anche se non si può escludere una relazione non lineare - ad es. parabolica – tra X e Y

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



Esempio calcolo coefficiente di correlazione

i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	230	325	-49.2	-217.9	2'420.64	47'480.41	10'720.68
2	234	274	-45.2	-268.9	2'043.04	72'307.21	12'154.28
3	235	412	-44.2	-130.9	1'953.64	17'134.81	5'785.78
4	235	431	-44.2	-111.9	1'953.64	12'521.61	4'945.98
5	264	518	-15.2	-24.9	231.04	620.01	378.48
6	278	460	-1.2	-82.9	1.44	6'872.41	99.48
7	279	564	-0.2	21.1	0.04	445.21	-4.22
8	308	815	28.8	272.1	829.44	74'038.41	7'836.48
9	308	685	28.8	142.1	829.44	20'192.41	4'092.48
10	310	556	30.8	13.1	948.64	171.61	403.48
11	326	651	46.8	108.1	2'190.24	11'685.61	5'059.08
12	343	824	63.8	281.1	4'070.44	79'017.21	17'934.18
Totale					**17'471.68**	**342'486.92**	**69'406.16**

$$\sum_{i=1}^{12}(x_i-\bar{x})^2 = 17'471.68$$

$$\sum_{i=1}^{12}(y_i-\bar{y})^2 = 342'468.92$$

$$\bar{x} = 279.2$$

$$\bar{y} = 542.9$$

$$\sum_{i=1}^{12}(x_i-\bar{x})(y_i-\bar{y}) = 69'406.16$$

$$\boxed{r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{69'406.16}{\sqrt{17'471.68\times 342'468.92}} = 0.897}$$

Grafici a dispersione e diversi valori di *r*