POLO UNIVERSITARIO
FONDAZIONE
PORTOGRUARO CAMPUS

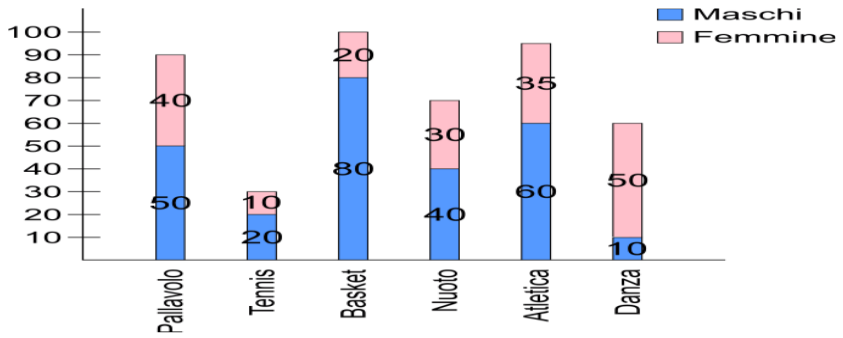
Corso di Statistica Sociale

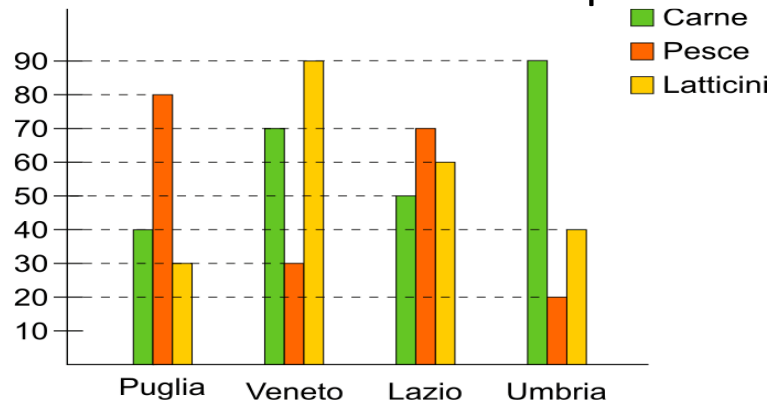
CORSO DI LAUREA: SCIENZE DELL'EDUCAZIONE

DOCENTE: FRANCESCO SANTELLI

Statistica Univariata

- Anche se non lo sapevamo, gran parte di ciò che abbiamo visto fino ad ora aveva a che fare con la **statistica univariata**
- E che significa? Significa che analizzavamo una variabile alla volta.
- Esempi? Istogramma delle età degli studenti (1 variabile: età), diagramma a torta della regione di provenienza (1 variabile: regione), varianza delle temperature a Portogruaro (1 variabile: temperatura)
- Eppure, fino ad ora qualche esempio di più variabili analizzate allo stesso tempo l'abbiamo visto…ricordate!?





Da 1 a 2 variabili: cambiano gli obiettivi

Obiettivi

Univariata

1 sola variabile. Vogliamo:

1) Sintetizzarla
2) Descriverla
3) Grafici maggiormente utilizzati: boxplot-istogrammi, grafici a barre semplici, grafici a torte

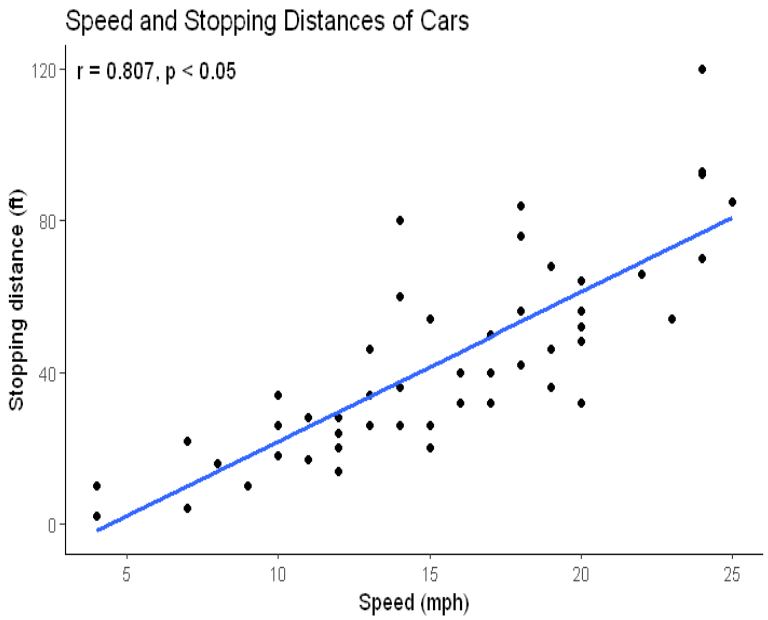
Bivariata

2 variabili. Vogliamo:

1) Capire se hanno una relazione
2) Capire se una causa l'altra
3) Grafici maggiormente utilizzati: grafici a barre sovrapposte o affiancate, boxplot condizionati, diagrammi a dispersione

Si passa da una idea puramente descrittiva di un fenomeno al cercare la sua spiegazione (introducendo una nuova variabile)

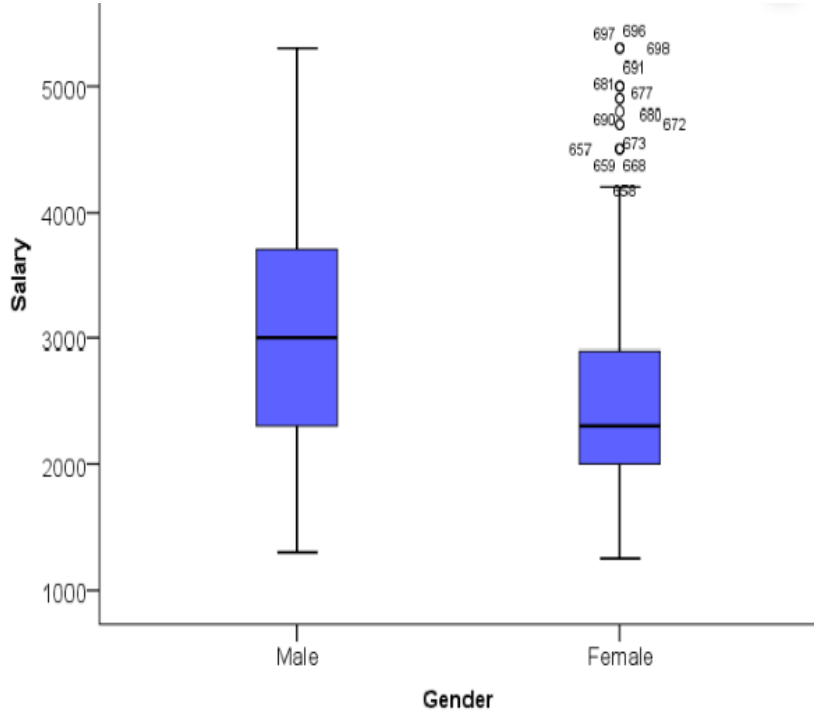
Che tipo di analisi bivariate possiamo fare?



Entrambe quantitative

Smoking in Public Places				
Gender	Favor	Oppose	No Opinion	Totals
Male	262	231	10	503
Female	302	205	5	512
Totals	564	436	15	1015

Entrambe qualitative (incroci di categorie)



Una numerica e una qualitativa

1) Caso: quantitative entrambe

	Spritz bevuti prima dell'esame	***Voto all'esame***	xi - media	yi - media	(xi - media)^2	(yi - media)^2
	0	28				
	0	28				
	1	27				
	2	24				
	7	18				
Media	***2***	***25***				

Completare questa tabella…e calcolare la varianza della 1° variabile (variabile X, spritz bevuti) e della 2° variabile (variabile Y, Voto)

$$\sigma_X^2 = ?$$

$$\sigma_Y^2 = ?$$

Obiettivo finale : capire che relazione c'è (statistica ma anche logica…) tra la variabile X e la variabile Y

Dobbiamo inserire nuova quantità…

	Spritz bevuti prima dell'esame	***Voto all'esame***	xi - media	yi - media	(xi - media)^2	(yi - media)^2	(xi-media)*(yi-media)	
	0	28	-2	3	4	9	-6	-2 per 3
	0	28	-2	3	4	9	-6	-2 per 3
	1	27	-1	2	1	4	-2	-1 per 2
	2	24	0	-1	0	1	0	0 per -1
	7	18	5	-7	25	49	-35	5 per -7
Media	***2***	***25***					***-49***	

Prendiamo gli scarti della 1 variabile (x) dalla propria media, gli scarti della 2 variabile (Y) dalla propria media, e li moltiplichiamo

E cosa ce ne facciamo!? Covarianza!

E' la quantità di prima (-49)

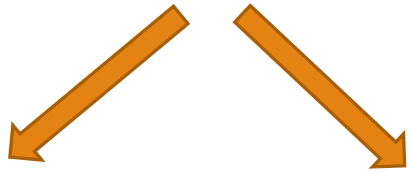
COVARIANZA = $\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{N} = -49 / 5 = 9,8$

E' la numerosità (cioè 5 studenti)

Significato della covarianza

Di per sé, la covarianza è molto difficile da interpretare (un po' come la varianza)

La cosa fondamentale è interpretare il segno

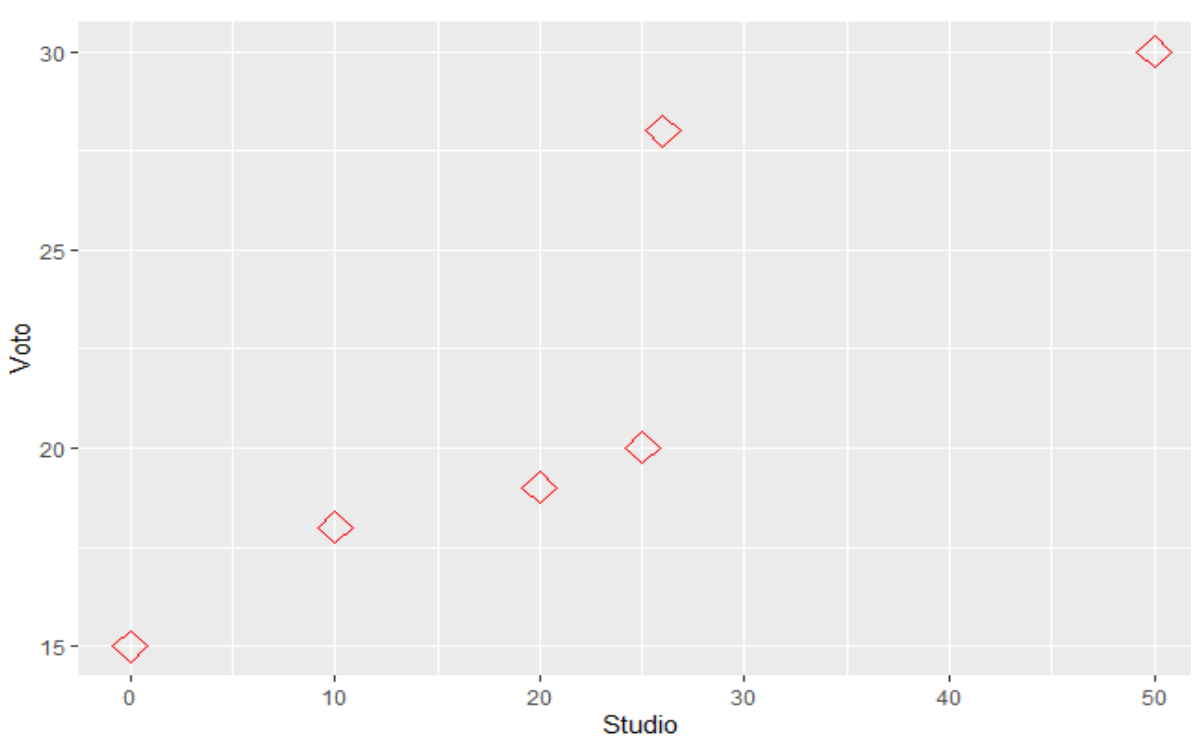


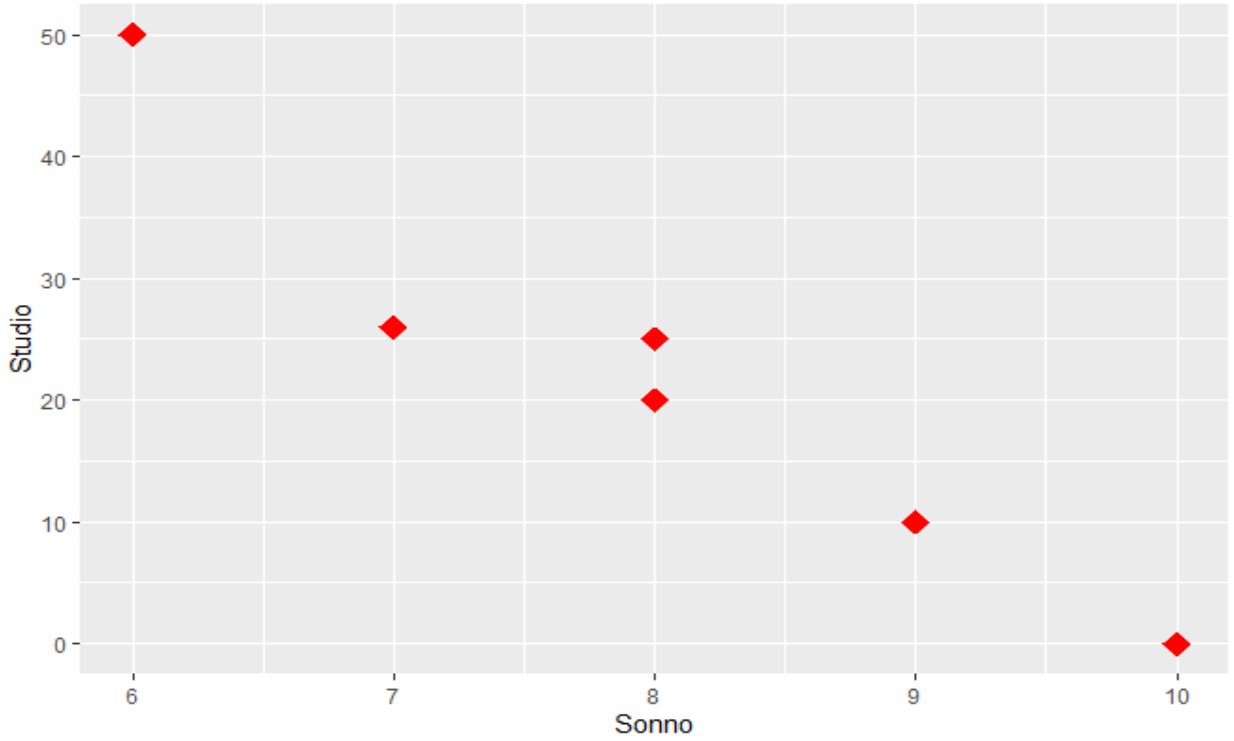
Segno negativo: le due variabili si dicono discordi: cioè all'aumentare di una, l'altra diminuisce e viceversa

Segno positivo: le due variabili si dicono concordi: cioè all'aumentare di una, l'altra aumenta ; al diminuire di una, l'altra diminuisce

Esempi: che segno vi aspettate per la covarianza di…n° ore di studio e voto all'esame? Temperature e n° di cappotti invernali venduti? N° di caffè venduti alla macchinetta del campus di Portogruaro e tonnellate di patate coltivate in Bielorussia?

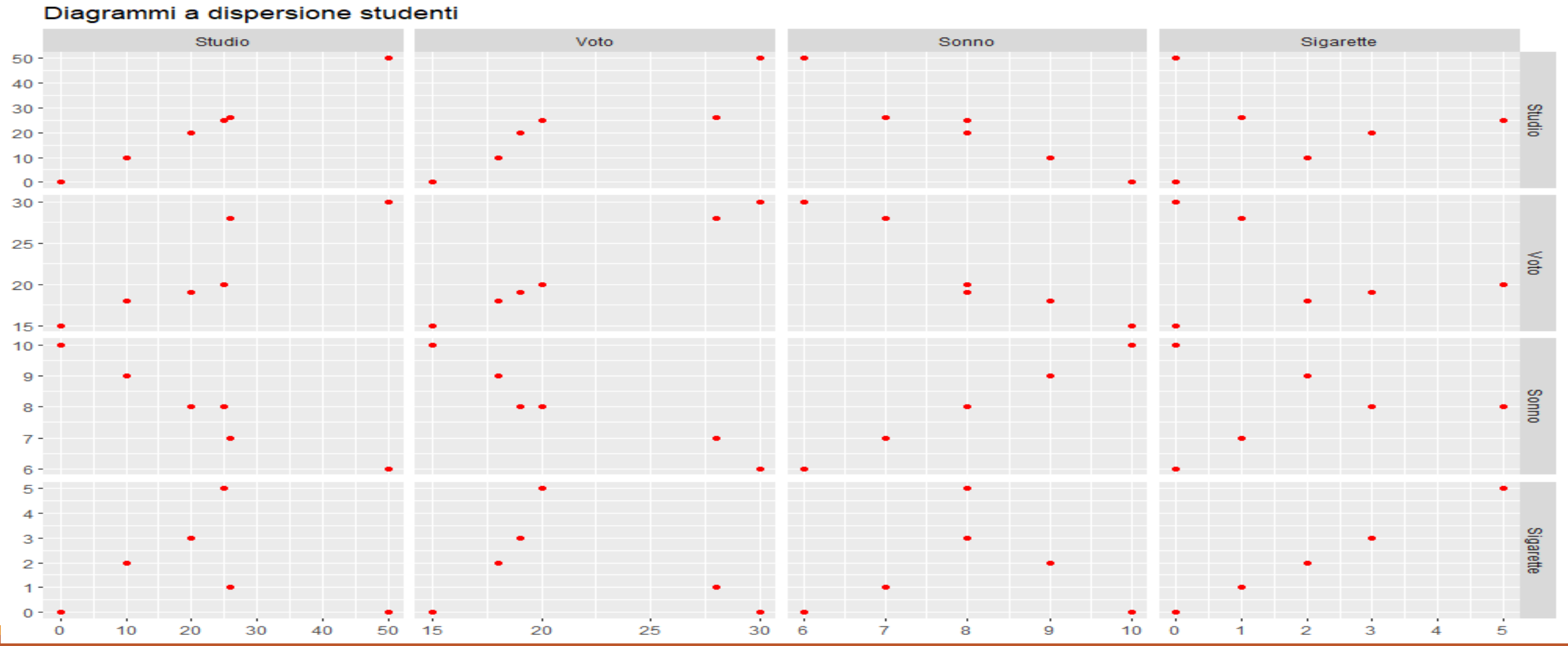
Il diagramma a dispersione (*scatterplot*)





Si costruisce mettendo una variabile quantitativa sulle X e una quantitativa sulle Y, è un diagramma cartesiano
Con coordinate (X, Y)

Altri Esempi di diagramma a dispersione



Correlazione

Come la varianza era di difficile interpretazione, così la covarianza… il diagramma a dispersione ci aiuta a capire il segno (positivo o negativo) della relazione ma non è così informativo sull'intensità. Con la correlazione abbiamo una lettura anche in termini di percentuale.

Si passa alla ***CORRELAZIONE***. Per il suo calcolo non serve nessuna quantità nuova. Si legge «rho»

Covarianza

$$\rho = \frac{\sigma_{XY}}{\sigma_X \bullet \sigma_Y}$$

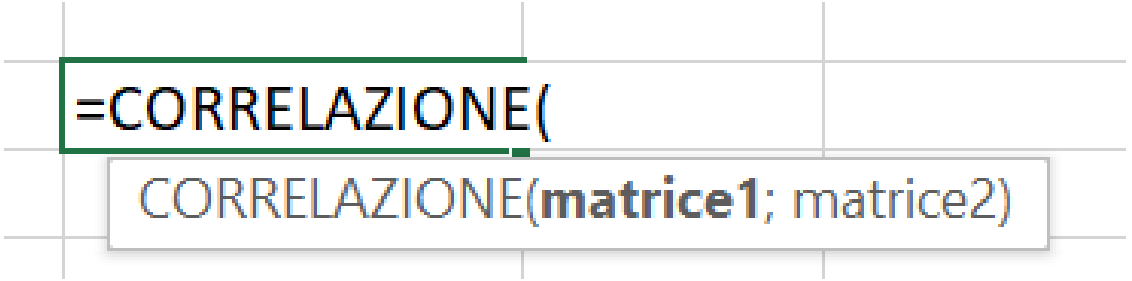
Prodotto deviazioni standard

Calcolo correlazione esempio di prima

	Spritz bevuti prima dell'esame	***Voto all'esame***	xi - media	yi - media	(xi - media)^2	(yi - media)^2	(xi-media)*(yi-media)	
	0	28	-2	3	4	9	-6	-2 per 3
	0	28	-2	3	4	9	-6	-2 per 3
	1	27	-1	2	1	4	-2	-1 per 2
	2	24	0	-1	0	1	0	0 per -1
	7	18	5	-7	25	49	-35	5 per -7
Media	***2***	***25***			6,8	14,4	***-49***	-9,8
					2,61	3,79		
				Varianze	6,8	14,4		
				Deviazioni standard	2,61	3,79		
				Correlazione	***-9,8/(2,61*3,79)***	***-99%***		

Su Excel…e interpretazione!

In Excel esiste una formuletta bellissima; inserendo i dati all'interno delle parentesi possiamo calcolare direttamente la correlazione tra due variabili senza fare tutti i calcoli di prima!



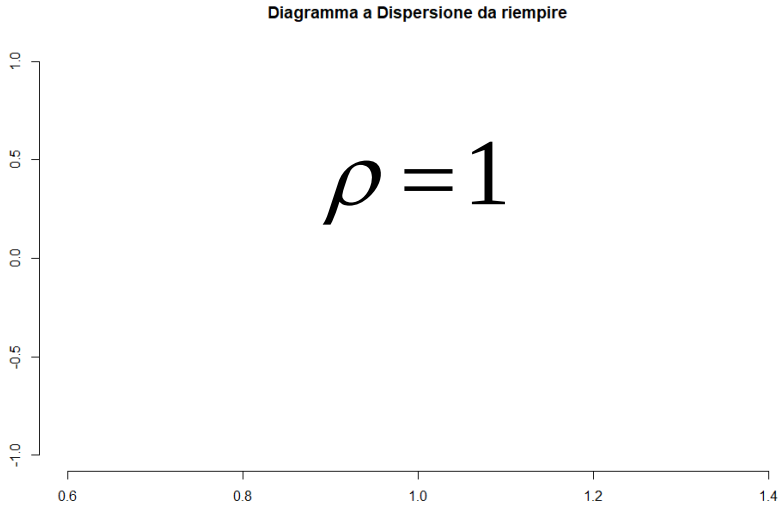
$$-1 \leq \rho = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \leq +1$$

Se è uguale a -1, massima correlazione inversa; le due variabili vanno sempre in direzione totalmente opposta!

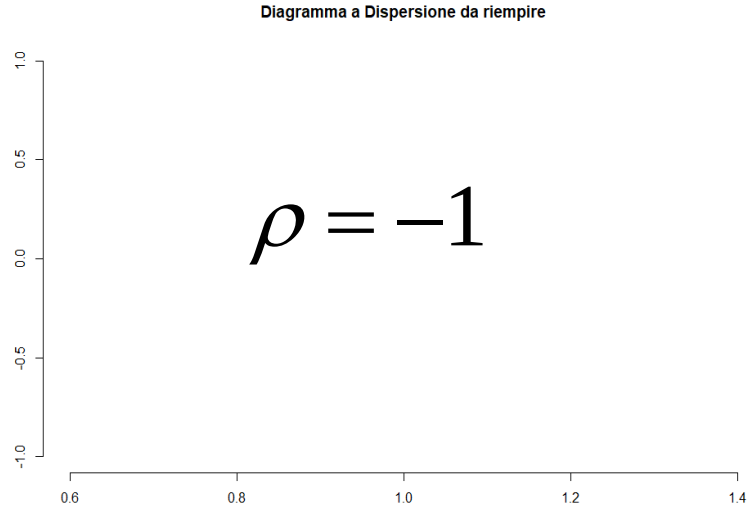
Se è uguale a 0, assenza di correlazione; le due variabili non hanno alcun legame

Se è uguale a +1, massima correlazione diretta; le due variabili vanno sempre in direzione perfettamente concorde!

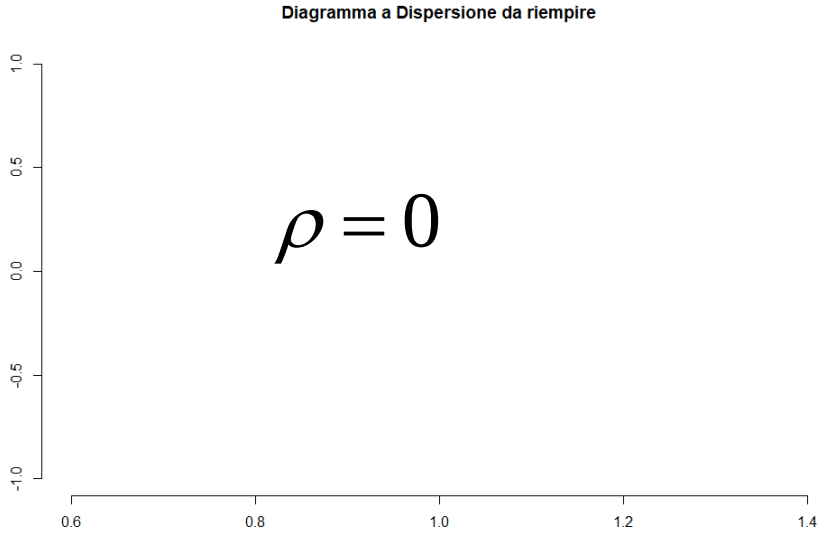
Esercizietto 1: grafici a dispersione



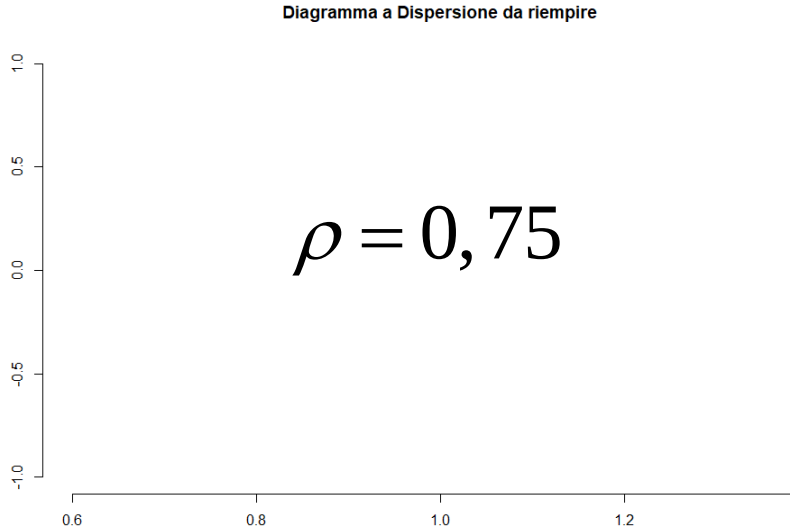
$$\rho = 1$$



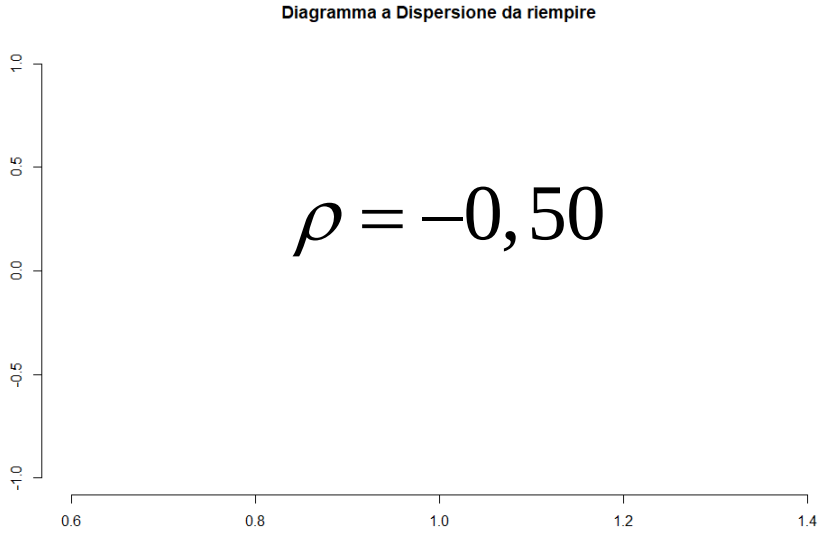
$$\rho = -1$$



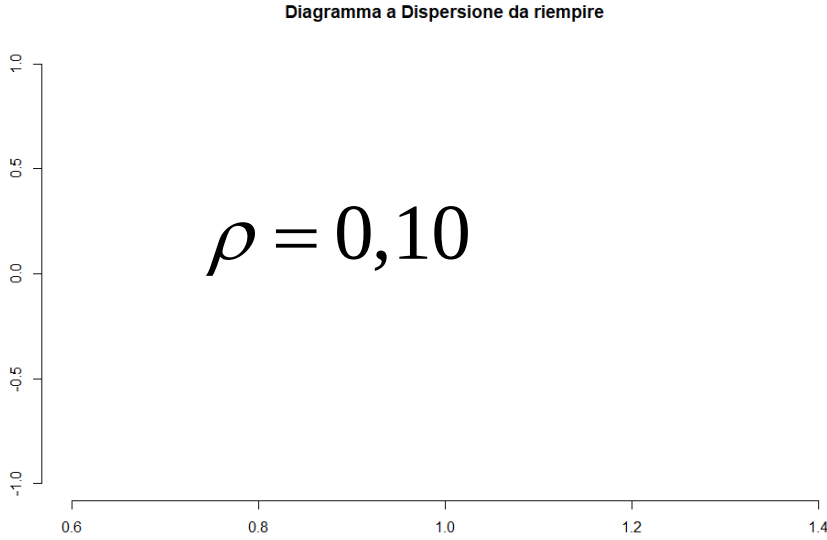
$$\rho = 0$$



$$\rho = 0,75$$



$$\rho = -0,50$$



$$\rho = 0,10$$

«Correlation is not causation»

Quando due variabili hanno una alta correlazione, vuol dire che in quei precisi dati c'è una relazione matematica; non sempre questo basta a dire che ci sia un nesso causa effetto!

Se cambiano i dati (i numeri) e cambia il contesto o si inseriscono altre variabili in esame…la correlazione potrebbe sparire o mutare valori

Per questo bisogna fare molta attenzione all'interpretazione; passare da una analisi di correlazione ad affermazioni del tipo «questo fenomeno A causa/influenza il fenomeno B» può essere un grave errore!

Un caso tipico è quello delle correlazioni spurie: «Quando C è presente, sia A *che* B sono osservati. (C è causa sia di A che di B.). Ma io osservo solo A e B, C non lo osservo»

Esempi carini qui:

https://www.fastcompany.com/3030529/hilarious-graphs-prove-that-correlation-isnt-causation

Da wikipedia..esempio di correlazione spuria

Rilevando anno dopo anno il numero di matrimoni e il numero di rondini in cielo, si può osservare ad esempio una forte correlazione tra i due fenomeni, il che non è dovuto al fatto che uno dei due influenza l'altro, ma semplicemente al fatto che in certi paesi le rondini compaiono durante le loro migrazioni in primavera ed autunno che sono pure i periodi preferiti dalle coppie nello scegliere il giorno delle nozze.

In altri termini se due fenomeni risultano statisticamente correlati tra loro, non vuol dire necessariamente che tra di essi sussista un legame diretto di causa-effetto, potendo essere tale correlazione del tutto casuale (cioè *spuria*) ***ovvero dipendente da una terza variabile in comune***, in assenza di meccanismo logico-causale plausibile che li metta in relazione tra loro.

È possibile rimediare a questo ordine di problemi mediante la misura e la comparazione della diversa strettezza delle correlazioni, se esistono sufficienti basi statistiche.

Per la prossima volta…

1) Calcolare le correlazioni associate a queste variabili:

Studio	***Voto***	***Sonno***	***Sigarette***
0	15	10	0
10	18	9	2
20	19	8	3
25	20	8	5
26	28	7	1
50	30	6	0

2) Ipotizzare per i vostri lavori di gruppo quali correlazioni interessanti ci siano da valutare

3) Provare a giocare a Guess The Correlation:

http://guessthecorrelation.com/