Statistica sociale - 11

Prof. Antonio Mussino

a. a. 2022-2023

SAPIENZA
UNIVERSITÀ DI ROMA

Relazioni bivariate - 1

- Quando mettiamo in relazione due (o più) variabili siamo entrati nella fase ***esplicativa*** dell'analisi dei dati:
  - vogliamo vedere se esista o meno un legame fra di esse, ovvero se il variare di una comporti, e in che modo, quello dell'altra.

- La relazione che vogliamo studiare è prettamente ***statistica***, ovvero legata al concetto di "media" in senso lato.

Relazioni bivariate - 2

- Se diciamo che c'è una relazione fra titolo di studio e frequenza della pratica sportiva, vuol dire che chi ha un titolo più alto in media pratica di più, ma ci possono essere laureati sedentari;
- se diciamo che c'è una relazione fra genere e disciplina praticata, vuol dire che tra gli uomini il calcio è lo sport più praticato, ma ci sono anche calciatrici, e tra le donne la ginnastica è la più praticata, ma ci sono molti uomini che vanno in palestra.
- In genere è più ***probabile*** che un laureato pratichi di più di un diplomato e tra gli uomini ci siano più calciatori.

Relazioni bivariate - 3

- Questo per dire che la relazione non implica un nesso di ***causa-effetto***:
  - capire se e quale sia questo nesso esula dai compiti della Statistica e rientra in quelli del ricercatore che analizza i dati.

Relazioni bivariate - 4

- Ad esempio, la relazione fra titolo di studio e pratica sportiva può essere spiegata considerando il fatto che:
  - chi ha un titolo più elevato ha un reddito più elevato e quindi più possibilità di spendere per praticare uno sport, quindi la relazione è indiretta e la causa della maggiore pratica è la maggiore capacità di spesa;
  - oppure si può considerare il fatto che l'attività sportiva rientra nella sfera culturale di un individuo e, in genere, più è alto il titolo di studio maggiore è il livello culturale.

Relazioni bivariate - 5

- In questi semplici esempi abbiamo già visto come, anche se la relazione è solo statistica, il ricercatore tende ad assegnare alle due variabili un ruolo diverso:
  - una delle due è la possibile causa e l'altra l'effetto, ovvero la prima influenza il variare dell'altra.
- Allora la prima è definita ***indipendente*** e l'altra ***dipendente***.
- La scelta di quale ruolo giochino le variabili è fatta soggettivamente dal ricercatore e ci sono anche casi in cui questa scelta non è possibile perché le due variabili giocano un ruolo simmetrico nell'analisi.

Relazioni bivariate - 6

- Semplificando al massimo la classificazione delle variabili, possiamo trovarci quindi di fronte a quattro situazioni:

  a) le due variabili sono entrambe (indipendente e dipendente) qualitative;

  b) le due variabili sono entrambe (indipendente e dipendente) quantitative:

  c) la variabile indipendente è qualitativa e quella dipendente quantitativa;

  d) la variabile indipendente è quantitativa e quella dipendente qualitativa.

Relazioni bivariate - 7

- Entreremo ora nel dettaglio dell'analisi delle relazioni a seconda del ruolo e della natura delle variabili:
  - i più rilevanti sono i casi **a)** e **b)** e sono disponibili anche strategie importanti per trattare il caso **c)**;
  - non entreremo nel merito del caso **d)**, che in effetti si verifica molto raramente.
- In quest'ultimo caso, ma vedremo accade anche per il **c)**, si preferisce accorpare i valori della variabile quantitativa in classi trasformandola in qualitativa e tornando così al caso **a)**.

Relazioni bivariate - 8

- La possibilità di ricondurre tutti gli altri al caso **a)**, ma soprattutto la netta prevalenza di variabili qualitative nell'area della Statistica sociale, ci spinge a iniziare e a trattare con maggiore accuratezza il caso della relazione fra variabili qualitative, che più precisamente definiremo ***associazione***.

Il caso di variabili qualitative - 1

- Per affrontare questo argomento dobbiamo definire e costruire un nuovo modo di rappresentare i dati: la ***tabella*** (o *tavola*) ***di contingenza****.

- È, di fatto, una trasformazione della matrice originaria e anch'essa si può considerare una matrice nella quale le righe e le colonne sono le modalità delle variabili studiate (due per volta) e nelle celle c'è la frequenza delle volte in cui le modalità si presentano associate nel collettivo.

* Anche detta tabella a doppia entrata, incrocio, tabulazione incrociata.

Il caso di variabili qualitative - 2

- Come esempio consideriamo
  - “età” (raggruppata in classi) e
  - “frequenza della pratica sportiva” (di fatto anch’essa raggruppata in classi),
- proprio per mostrare l’applicabilità di questa strategia anche con variabili quantitative.

ALLEGATO 8

Allegato 8

Tabella 3.8 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica sportiva” e “Età”

		Età in classi						
		15-19 anni	20-24 anni	25-34 anni	35-44 anni	45-54 anni	55-64 anni	Totale
Frequenza totale nell'anno della pratica sportiva	Mai	65	104	205	167	139	107	787
	Meno di 1 volta al mese	2	1	7	2	1	1	14
	1-3 volte al mese	3	5	10	4	0	0	22
	1 volta alla settimana	3	9	19	15	7	5	58
	2 volte alla settimana	23	15	24	10	10	3	85
	3 e più volte alla settimana	39	37	39	31	18	7	171
Totale		135	171	304	229	175	123	1137

Il caso di variabili qualitative - 3

- Logicamente l' "età" è la variabile ***indipendente*** e la "frequenza della pratica" la ***dipendente***, per cui disponiamo convenzionalmente la prima sulle colonne e la seconda sulle righe.
- La frequenza assoluta 65 indica che nel collettivo ci sono 65 intervistati che hanno meno di 19 anni e che non hanno mai praticato nell'anno precedente e così via.
- Nell'ultima riga e nell'ultima colonna troviamo i "***totali marginali***" (che in realtà non fanno parte della matrice), che corrispondono alle distribuzioni di frequenza uni-variate rispettivamente delle variabili per colonna e per riga.

Il caso di variabili qualitative - 4

- Il ricercatore utilizza questa presentazione perché vuole vedere se c'è ***dipendenza***, o meno, fra le due variabili e, in caso di dipendenza, che tipo e con quale intensità ci sia ***associazione*** fra di esse.

- Un altro caso, che presenta un numero minore di modalità e quindi è più immediato nel commento, è quello che mette in relazione la "frequenza della pratica" con il "genere" degli intervistati.

ALLEGATO 9

Allegato 9

Tabella 3.9 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

		Genere		Totale
		Maschi	Femmine	
Frequenza totale nell'anno della pratica sportiva	Mai	318	469	787
	Meno di 1 volta al mese	8	6	14
	1-3 volte al mese	16	6	22
	1 volta alla settimana	49	9	58
	2 volte alla settimana	48	37	85
	3 e più volte alla settimana	105	66	171
Totale		544	593	1137

Il caso di variabili qualitative - 5

- Ovviamente le frequenze proposte (quelle assolute) non sono utili per avere informazioni sulle eventuali associazioni, perché la numerosità dei gruppi di individui nelle diverse classi di età e nei diversi livelli di frequenza è differente:
- abbiamo bisogno di relativizzare l'informazione e, per far questo, calcoliamo le frequenze percentuali, che possono essere di tre tipi: ***percentuale*** *di* ***riga***, *di* ***colonna*** e *sul* ***totale***.

Il caso di variabili qualitative - 6

- Nella prima tabella sono riportate le percentuali di riga, ovvero le percentuali in relazione alle diverse modalità della variabile “frequenza della pratica”:
  - è un’informazione poco utile, ci dice quale genere è prevalente all’interno dei diversi livelli di impegno sportivo e di sedentarietà.

ALLEGATO 10

Allegato 10

Tabella 3.10 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

% di riga		Genere		Totale
		Maschi	Femmine	
Frequenza totale nell'anno della pratica sportiva	Mai	40,4%	59,6%	100,0%
	Meno di 1 volta al mese	57,1%	42,9%	100,0%
	1-3 volte al mese	72,7%	27,3%	100,0%
	1 volta alla settimana	84,5%	15,5%	100,0%
	2 volte alla settimana	56,5%	43,5%	100,0%
	3 e più volte alla settimana	61,4%	38,6%	100,0%
Totale		47,8%	52,2%	100,0%

Il caso di variabili qualitative - 7

- Per capire il motivo di tale affermazione consideriamo la "frequenza della pratica" solo per gli sportivi.
- In questo caso tutti gli intervistati hanno risposto alle due domande e non ci sono *mancate risposte*.
- Ma se riprendiamo in considerazione il questionario possiamo osservare come chi rispondeva di non praticare sport non doveva rispondere alle domande dalla numero 2 alla numero 11.

Il caso di variabili qualitative - 8

- Si tratta del caso di ***risposte non dovute***, e abbiamo visto come si trattano nelle distribuzioni di frequenza univariate:
  - nel caso bivariato la gestione è più semplice, basta che manchi l'informazione per una delle due variabili che l'unità statistica non è contata nelle celle.
- Si può, infatti, notare come il totale generale sia pari a 350, ovvero gli intervistati che praticano sport.

ALLEGATO 11

Allegato 11

Tabella 3.11 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

		Genere		Totale
		Maschi	Femmine	
Frequenza totale nell'anno della pratica sportiva	Meno di 1 volta al mese	8	6	14
	1-3 volte al mese	16	6	22
	1 volta alla settimana	49	9	58
	2 volte alla settimana	48	37	85
	3 e più volte alla settimana	105	66	171
Totale		226	124	350

Il caso di variabili qualitative - 9

- Se calcoliamo le percentuali di riga rispetto a questa tabella, vediamo che le percentuali dei maschi sono ***sempre*** più alte:
  - pertanto, per capire se vi è una differente modalità di prevalenza è necessario comparare le percentuali di ciascuna riga (possiamo definirle i ***profili***) con quelle della riga del totale (64,6% tra i maschi e 35,4% tra le femmine).

ALLEGATO 12 e 13

Allegati 12 e 13

Tabella 3.12 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

% di riga		Genere		
		Maschi	Femmine	Totale
Frequenza totale nell'anno della pratica sportiva	Meno di 1 volta al mese	57,1%	42,9%	100,0%
	1-3 volte al mese	72,7%	27,3%	100,0%
	1 volta alla settimana	84,5%	15,5%	100,0%
	2 volte alla settimana	56,5%	43,5%	100,0%
	3 e più volte alla settimana	61,4%	38,6%	100,0%
Totale		64,6%	35,4%	100,0%

Tabella 3.13 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

Inclusi i non praticanti **% di colonna**		Genere		
		Maschi	Femmine	Totale
Frequenza totale nell'anno della pratica sportiva	Mai	58,5%	79,1%	69,2%
	Meno di 1 volta al mese	1,5%	1,0%	1,2%
	1-3 volte al mese	2,9%	1,0%	1,9%
	1 volta alla settimana	9,0%	1,5%	5,1%
	2 volte alla settimana	8,8%	6,2%	7,5%
	3 e più volte alla settimana	19,3%	11,1%	15,0%
Totale		100,0%	100,0%	100,0%

Il caso di variabili qualitative - 10

- Più immediata è la lettura delle percentuali di colonna:
  - infatti i due sottoinsiemi che si mettono a confronto sono uniformati rispetto alla numerosità, in quanto si considera il risultato ogni 100 maschi e ogni 100 femmine.
- Vediamo così che le donne sono nettamente prevalenti se consideriamo la mancata pratica, ma quando si impegnano lo fanno con maggiore costanza e regolarità.
- Il confronto fra i ***profili colonna*** è estremamente efficace, rispetto al nostro obiettivo di scoprire le associazioni, in entrambe le tabelle.

Il caso di variabili qualitative - 11

- La scelta di privilegiare i profili colonna sui profili riga è dovuta al fatto che la variabile indipendente è posizionata sulle colonne;
- se fosse posta sulle righe, ovviamente, bisognerebbe invertire la scelta.

(ALLEGATO 14)

Allegato 14

Tabella 3.14 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

Esclusi i non praticanti	**% di colonna**	Genere		Totale
		Maschi	Femmine	
Frequenza totale nell'anno della pratica sportiva	Meno di 1 volta al mese	3,5%	4,8%	4,0%
	1-3 volte al mese	7,1%	4,8%	6,3%
	1 volta alla settimana	21,7%	7,3%	16,6%
	2 volte alla settimana	21,2%	29,8%	24,3%
	3 e più volte alla settimana	46,5%	53,2%	48,9%
Totale		100,0%	100,0%	100,0%

Il caso di variabili qualitative - 12

- L'ultima opportunità di calcolo di percentuali è relativa a quelle ***totali***;
  - come si può vedere dalla tabella seguente non vi è un'informazione aggiuntiva alla tabella originaria delle frequenze assolute:
  - queste percentuali (il totale 100% è relativo a tutta la tabella) non ci servono per studiare le eventuali associazioni.

Il caso di variabili qualitative - 13

- Pertanto questa modalità non si utilizza mai, a meno che non si voglia confrontare la situazione in questo collettivo (con numerosità 1137), con quella ottenuta in un altro collettivo di numerosità diversa:
  - è come se calcolassimo la distribuzione di frequenze percentuali di una variabile ricostruita associando in tutti i modi possibili le modalità delle due variabili studiate.

ALLEGATO 15

Allegato 15

Tabella 3.15 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

% sul totale		Genere		Totale
		Maschi	Femmine	
Frequenza totale nell'anno della pratica sportiva	Mai	28,0%	41,2%	69,2%
	Meno di 1 volta al mese	,7%	,5%	1,2%
	1-3 volte al mese	1,4%	,5%	1,9%
	1 volta alla settimana	4,3%	,8%	5,1%
	2 volte alla settimana	4,2%	3,3%	7,5%
	3 e più volte alla settimana	9,2%	5,8%	15,0%
Totale		47,8%	52,2%	100,0%

Il caso di variabili qualitative - 14

- Lo studio delle percentuali calcolate relativamente alle modalità della variabile indipendente è di per sé sufficiente per capire se e quali relazioni sono presenti nella tabella;
- si può, peraltro, cercare di sintetizzare con un ***indice*** il livello di associazione fra le due variabili.

Il caso di variabili qualitative - 15

- Questo indice dovrebbe essere pari a ***0*** in caso di ***assenza di associazione*** fra le due variabili e avere un suo ***massimo*** nel caso di ***massima associazione*** possibile;
- questo massimo potrebbe essere ***normalizzato***, costruendo così un indice compreso fra ***0*** e ***1*** (o fra ***0*** e ***100*** e così via):
- è facile e immediatamente evidente quale è la situazione per il livello 0, in quanto corrisponde alla situazione di assoluta ***indipendenza*** fra le due variabili;
- più complessa è la rappresentazione della ***massima dipendenza*** possibile, perché è legata alla struttura della tabella.

Il caso di variabili qualitative - 16

- Il caso di ***indipendenza*** si verifica quando tutti i profili per riga e per colonna sono uguali,
  - ad esempio se la composizione percentuale dei livelli di pratica è la stessa per maschi e femmine, indipendentemente dalla numerosità dei due gruppi (maschi e femmine)
  - e, simmetricamente, la percentuale di maschi e femmine è la stessa per ogni livello, uguale a quella presente nell'intero collettivo.

Il caso di variabili qualitative - 17

- Questo risultato teorico si ottiene, tabella per tabella, moltiplicando i totali di riga per i totali di colonna di ogni coppia di modalità e dividendo per il totale generale delle unità.
- Così nell'ALLEGATO 16 troviamo le frequenza che ci aspetteremmo se non ci fosse alcuna associazione fra il "genere" e la "frequenza di pratica" degli sportivi (cfr. ALLEGATO 11).
- Si definiscono ***frequenze attese***: sono valori teorici, come si nota dal fatto che ci siano valori decimali, non ammissibili per le frequenze.
- Le frequenze marginali di riga e di colonna sono le stesse della tabella d'origine.

Allegato 16

Tabella 3.16 - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

Frequenze attese		Genere		Totale
		Maschi	Femmine	
Frequenza totale nell'anno della pratica sportiva	Meno di 1 volta al mese	9,0	5,0	14,0
	1-3 volte al mese	14,2	7,8	22,0
	1 volta alla settimana	37,5	20,5	58,0
	2 volte alla settimana	54,9	30,1	85,0
	3 e più volte alla settimana	110,4	60,6	171,0
Totale		226,0	124,0	350,0

Il caso di variabili qualitative - 18

- Se le frequenze della tabella da noi ottenuta, che si definiscono ***frequenze osservate*** coincidessero con quelle *attese*, il valore dell'indice che stiamo cercando per sintetizzare la relazione fra le variabili sarebbe ovviamente uguale a 0!
- Più le due frequenze divergono più forte si può considerare l'associazione fra le variabili!

Il Chi-quadrato - 1

- Un indice che risponde a queste caratteristiche è il ***Chi-quadrato***, che si ottiene sommando gli scarti fra frequenze osservate e attese al quadrato, rapportati per normalizzarli alle frequenze attese.
- In realtà il Chi-quadrato ha un minimo teorico pari a 0, ma il suo massimo dipende dalle caratteristiche della tabella:
  - esso sarà infatti uguale al più piccolo dei seguenti valori:
    - dimensione del collettivo per numero delle righe meno 1 e
    - dimensione del collettivo per numero delle colonne meno 1.

Il Chi-quadrato - 2

- Pertanto per normalizzarlo ed avere un massimo pari a 1 dobbiamo dividerlo per questo massimo.
- Otteniamo così un nuovo indice che è la ***V di Cramer***.
- IL Chi-quadrato gioca un ruolo prioritario nel Test di ipotesi.

Il Chi-quadrato - 3

- La formula esatta della V di Cramer è la seguente:

$$\mathbf{V = Sqrt(\chi^2/N*(min\ (r,c) - 1))}$$

- Nella nostra tabella il Chi-quadrato è pari a **14,217**, N è **350**, il minimo fra r e c è **2** e quindi

V= .202.

Il Chi-quadrato - 4

- La formula esatta della V di Cramer è la seguente:

$$\mathbf{V = Sqrt(\chi^2/N*(min\ (r,c) - 1))}$$

- Nella nostra tabella il Chi-quadrato è pari a **14,217**, N è **350**, il minimo fra r e c è **2** e quindi

V= .202.

Il Chi-quadrato - 5

- Molti altri sono i coefficienti che possono essere calcolati per misurare l'associazione fra le variabili, a seconda se la relazione possa essere trattata
  - come **simmetrica** (***Phi quadro***, ***Phi***, ***Coefficiente di contingenza***) o
  - come **asimmetrica**, ovvero con lo studio solo dell'effetto di una variabile sull'altra;
- a seconda se le variabili abbiamo modalità
  - **ordinabili** o
  - **non ordinabili**.

Il Chi-quadrato - 6

- Qui si è voluto proporre solo i due coefficienti, il Chi quadrato e la V di Cramer, emblematici dell'approccio inferenziale e di quello descrittivo sintetico: lasciamo ai corsi di Statistica metodologica la descrizione delle caratteristiche e dell'utilizzabilità degli altri indici.

Il Chi-quadro nel test di ipotesi -1

- Come è possibile utilizzare il Chi quadrato in un'ottica inferenziale?
- Negli *output* dei principali *software*, accanto al valore vengono indicati i gradi di libertà* e il *p-value*, ovvero la probabilità di ottenere i valori osservati presenti nella tabella di contingenza, se fosse vera l'ipotesi di indipendenza.
- Se il *p-value* è pari a 0,007, vuol dire che, se fosse vera l'ipotesi di indipendenza, il risultato rappresentato dai valori osservati si verificherebbe in soli 7 campioni su 1000!

Il Chi-quadro nel test di ipotesi -2

- Si dice allora che il *p-value* è significativo: questo porta a rifiutare l'ipotesi nulla e ad affermare che fra le due variabili c'è una qualche associazione, con un rischio di sbagliare in tale affermazione pari proprio a 7 per 1000.
- Ovviamente, se volessimo dire qual è la misura di tale associazione, potremmo usare la già nota V di Cramer.
- In genere si accetta un errore al massimo dello 0,05. Se si vuole essere più sicuri dello 0,01.

I gradi di libertà

- I gradi di libertà («*df*» *degrees of freedom*) sono un parametro della distribuzione del Chi-quadrato, che si calcola sottraendo al numero delle osservazioni quello dei vincoli della tavola: le osservazioni sono le celle (righe per colonne), i vincoli sono le celle marginali (righe più colonne meno uno, il totale generale).
- In altre distribuzioni il calcolo è diverso, ma il significato è analogo.

La distribuzione campionaria del Chi-quadrato

- Tale distribuzione è nota ed è nelle appendici di ogni testo di Statistica, con df= (r-1)*(c-1) (con r = righe e c = colonne).
- Se non avessimo il *p-value*, potremmo consultare le tavole e confrontare il valore ottenuto con quello di riferimento per il livello di fiducia prescelto.
- Se quello ottenuto sui dati osservati è maggiore di quello tabulato si rifiuta l'ipotesi nulla, altrimenti non è possibile farlo.

Caratteristiche della tabella

- Un'importante cautela da considerare nell'uso del Chi-quadrato con fini inferenziali è quella di osservare la percentuale di celle che hanno una frequenza attesa inferiore a 5.
- Se questa percentuale supera il 20, il risultato del Chi-quadrato non è accettabile, perché il coefficiente è troppo condizionato dalle basse frequenze.
- In questi casi è necessario accorpare le modalità delle variabili finché non si raggiunga una numerosità adeguata.

Il caso c): distribuzioni quanti-qualitative

- Ricordiamo che, nel caso c), abbiamo considerato le situazioni nelle quali:
  - la variabile ***indipendente*** è ***qualitativa*** e
  - quella ***dipendente*** è ***quantitativa***.

- Ad esempio "genere" è "frequenza della pratica nell'ultima settimana", oppure "numero di sport praticati".

- Si parla, in questo caso, di ***confronto fra medie***.

Confronto fra medie - 1

- Ovviamente più le medie differiscono al variare delle modalità della variabile indipendente più forte sarà l'intensità della relazione, ovvero l'associazione.
- Per misurare la forza dell'associazione definiamo le seguenti quantità:
- **Devianza Totale**, ovvero la somma dei quadrati degli scarti dei singoli valori dalla media generale del collettivo;
- **Devianza Interna**, ovvero la somma dei quadrati degli scarti dei singoli valori dalla loro media parziale di modalità;
- **Devianza dei Gruppi**, ovvero la somma dei quadrati degli scarti delle singole medie dei gruppi dalla media generale del collettivo.

Confronto fra medie - 2

- Poiché vale la relazione:

Devianza Totale = Devianza Interna + Devianza Gruppi

- ovvero

$$1 = \frac{DT}{DT} = \frac{DI}{DT} + \frac{DG}{DT}$$

- il valore

$$\eta^2 = \frac{DG}{DT}$$

Dove η^2 ci dà la proporzione di devianza totale che è ***spiegata*** dalla variabile indipendente.

Confronto fra medie - 3

- Dove η^2 ci dà la proporzione di devianza totale che è ***spiegata*** dalla variabile indipendente.
- η^2 varia tra **0**, nessuna relazione fra le variabili, e **1**, relazione perfetta, ovvero tutta la devianza dipende dalla variabile indipendente (es. tutti gli uomini hanno dato una stessa risposta e così, diversa, tutte le donne!).
- In genere η^2 non è molto elevato: è difficile andare oltre il 25% - 30%, spesso si è sul 10%.
- *Aumenta all'aumentare del numero di modalità.*
- È assimilabile al coefficiente r^2.

ANOVA - 1

- Anche in questo caso possiamo vedere gli aspetti inferenziali.
- Nel nostro data set consideriamo le variabili BMI (come indicatore dello stile di vita attiva e dell'alimentazione) e colore della pelle: vogliamo testare l'ipotesi che gli stili di vita siano diversi nelle diverse etnie che vivono ad Aracaju (Sergipe).
- Calcoliamo le statistiche di sintesi per il BMI:

Report BMI

colore della pelle	Media	N	Deviazione std.
bianca	24,4733	237	3,63455
gialla	23,8859	113	3,39628
marrone	24,9505	564	4,25270
nera	25,0485	223	3,78158
Totale	24,7644	1137	3,97074

ANOVA - 2

- L' η^2 in questo caso è pari a 0,08, quindi c'è, anche se ridotta, un'associazione fra le due variabili.
- Ma questa associazione è valida per tutta la popolazione da cui è stato estratto il campione degli intervistati, ovvero per i cittadini di Aracaju?
- La media generale è 24,8 e la misura della dispersione dei risultati di tutti i cittadini rispetto ad essa è lo s.q.m (la radice quadrata della ***varianza***).
- Ma quanta parte di questa varianza è dovuta alle differenze fra i risultati dei singoli cittadini appartenenti a un'etnia dalla media dell'etnia?
- E quanta alle differenze fra le (medie delle) etnie?

ANOVA - 3

- Utilizziamo la relazione fra le ***devianze*** (numeratore della varianza) precedentemente introdotta.
- La **Devianza tra i gruppi** è quella ***spiegata*** dai diversi stili di vita delle etnie.
- La **Devianza all'interno dei gruppi** è quella ***non spiegata***.
- Come già visto la **DG** è nulla in caso di assenza di relazione fra le variabili studiate.
- Nella **Tabella ANOVA** (***Analisi della Varianza***), che segue, sono riportate le informazioni utili per testare questa relazione.

ANOVA - 4

Tabella ANOVA BMI * colore della pelle

	Devianza	df	Media dei quadrati	F	Sig.
Fra gruppi	144,813	3	48,271	3,078	,027
Entro gruppi	17766,234	1133	15,681		
Totale	17911,047	1136			

- df ovvero "degrees of freedom" (gradi di libertà) sono i denominatori delle rispettive varianze, utilizzate come stime campionarie: (n-1) per la totale; (k-1) fra i gruppi; (n-k) interna ai gruppi*;
- le Medie dei quadrati sono le ***varianze*** stimate; se i due valori fossero simili (rapporto circa 1), allora non vi sarebbe l'effetto etnia e non potremmo rifiutare l'ipotesi nulla.
- Quindi calcoliamo questo rapporto che chiamiamo **F**, la cui distribuzione campionaria è nota ed è nelle appendici di ogni testo di Statistica (con df= (k-1) e (n-k)).

---

* qui n=1137; k= 4.

ANOVA - 5

- Oppure si può utilizzare il valore di significatività (**Sig.**) o ***p-value***, che rappresenta la probabilità di avere un rapporto **F** di questa dimensione se fosse vera l'ipotesi nulla.

- Ovvero, poiché in questo caso **p = .027**, questo risultato si verificherebbe 2,7 volte ogni 100 campioni estratti dalla popolazione di riferimento se fosse vera l'ipotesi nulla.

- Se quindi testiamo l'ipotesi nulla con un livello di errore del 5% potremmo rifiutarla (2,7<5).

- Al contrario, se ci fossimo posti un livello di errore dell'1%, i risultati non ci avrebbero permesso di rifiutarla.

Il caso b): variabili quantitative

- Chiudiamo con l'analisi del caso b), quello in cui si studia l'associazione fra due variabili quantitative, che potremo chiamare correttamente "relazione".
- Si parla, infatti, di ***correlazione*** e di ***regressione*** fra le variabili studiate.
- Le possibilità di analisi sono, in questo caso, molto ampie, cominciando dalla possibilità di partire dallo studio della rappresentazione grafica.

La correlazione - 1

- Prendiamo in considerazione una nuova matrice dei dati e consideriamo tre delle variabili quantitative: CORSA, FLESS e SALTO.

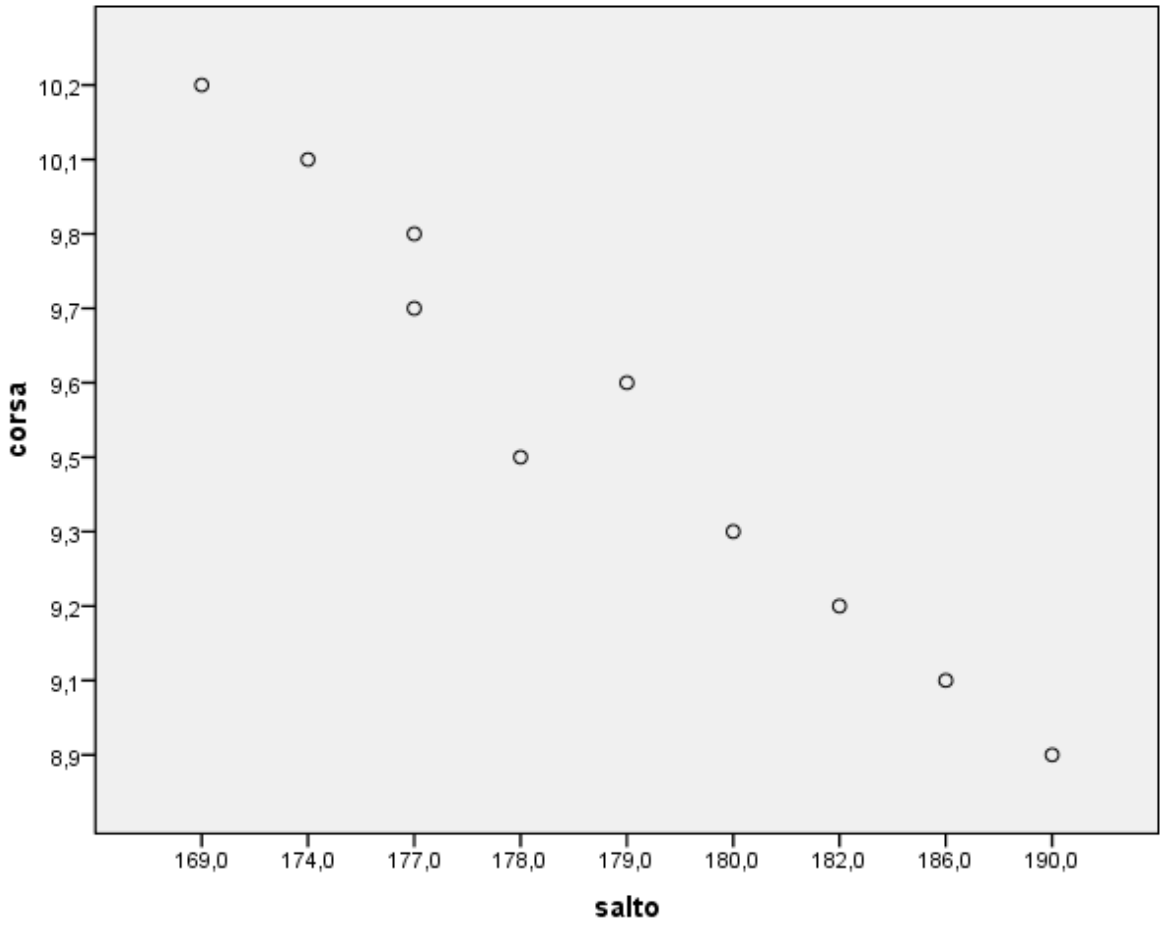
La matrice dei dati

Nome	Sport	corsa	salto	fless	spola	tapp	later	rank
Andrea	Calcio	9,8	177	6,2	17,9	38,6	Dx	6°
Carlo	Volley	10,2	169	10,2	18,2	38,4	Dx	22°
Enrico	Volley	9,5	178	11,9	17,6	38,1	Sn	3°
Gianni	Calcio	9,6	179	9,6	17,2	37,4	Dx	5°
Mario	Volley	9,2	182	6,4	16,8	36,2	Dx	10°
Mauro	Volley	9,1	186	10,1	16,4	37,4	Dx	7°
Nicola	Calcio	8,9	190	8,4	16,5	39,2	Sn	1°
Sandro	Atletica	9,3	180	10,4	17,0	39,6	Dx	12°
Silvano	Atletica	10,1	174	8,2	18,6	39,2	Dx	20°
Ugo	Atletica	9,7	177	8,4	17,9	38,1	Dx	14°

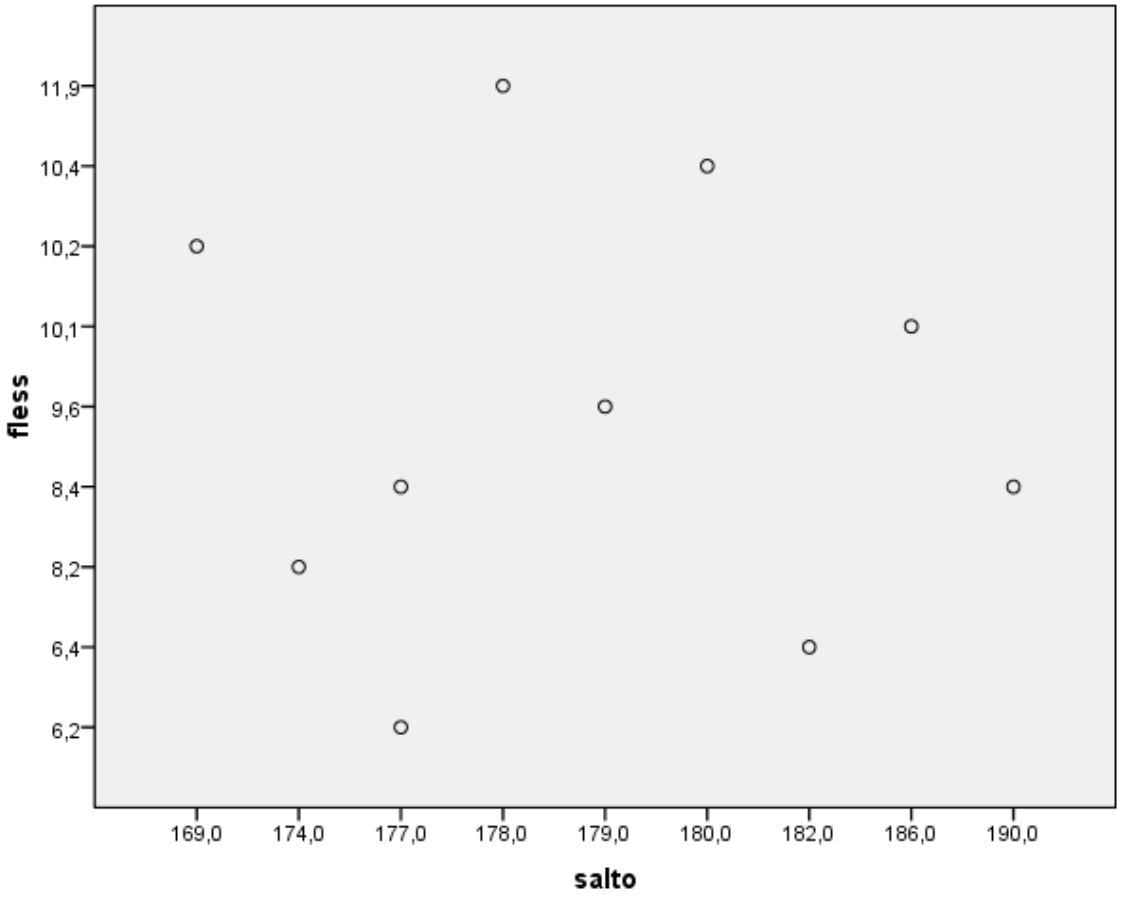
La correlazione - 2

- Poiché le variabili sono continue, si possono rappresentare su una retta di riferimento:
  - ✓CORSA sulla retta X_1
  - ✓SALTO sulla retta X_2
  - ✓FLESS sulla retta X_3
- Quindi le rette possono essere messe in relazione fra di loro su piani in coordinate cartesiane ortogonali (***diagrammi di dispersione***), ad esempio X_1 vs. X_2 e X_1 vs. X_3.

Plot: CORSA vs. SALTO



Plot: FLESS vs. SALTO



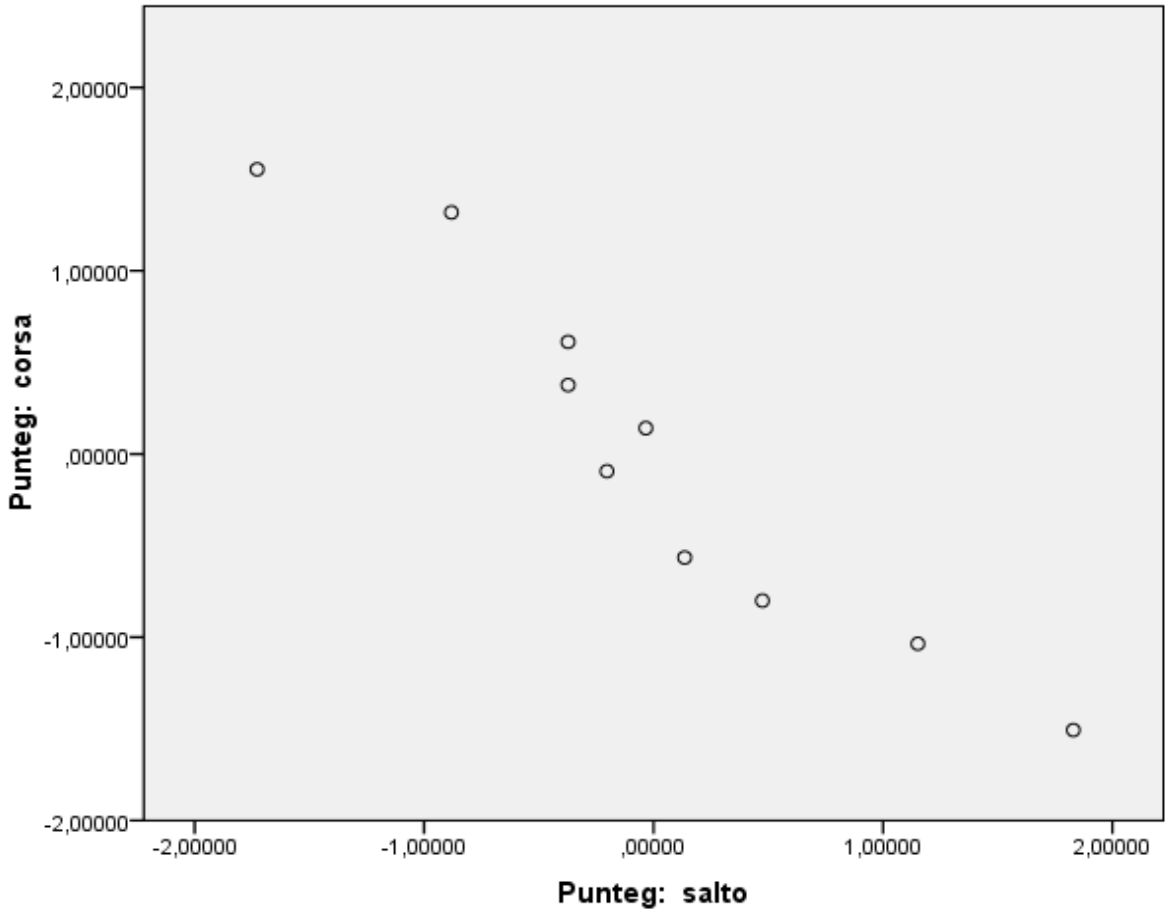
La correlazione - 3

- In questo modo su ogni dimensione unitaria (la retta è uno *spazio a una dimensione*) i risultati sono messi in ordine crescente e ogni allievo è rappresentato su un punto (coordinata).
- Mettendo in relazione due rette ogni allievo è rappresentato da un punto sul piano (*spazio a due dimensioni*), che si individua tramite le coordinate sulle rette.
- *Qualora le prove considerate fossero più di 2 (ad es. "p") lo spazio di riferimento sarebbe a "p" dimensioni.*

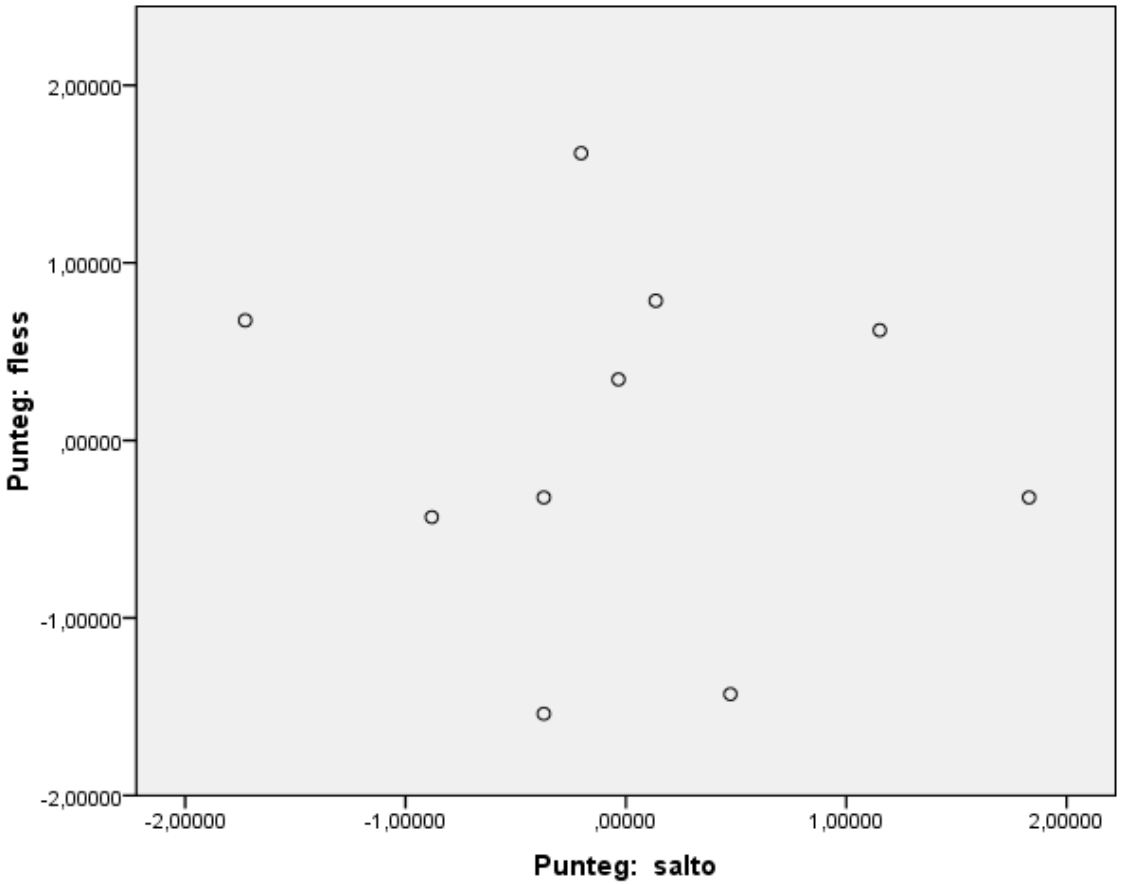
La correlazione - 4

- Possiamo vedere come le ***nuvole*** dei punti che rappresentano gli allievi si disperdono nel piano in maniera diversa, seguendo una certa regolarità nel primo caso e in maniera casuale nel secondo.
- Ma i due grafici possono essere fuorvianti, perché le unità di misura e/o la variabilità sono diverse.
- *Standardizziamo pertanto le tre variabili e vediamo il diverso risultato grafico.*

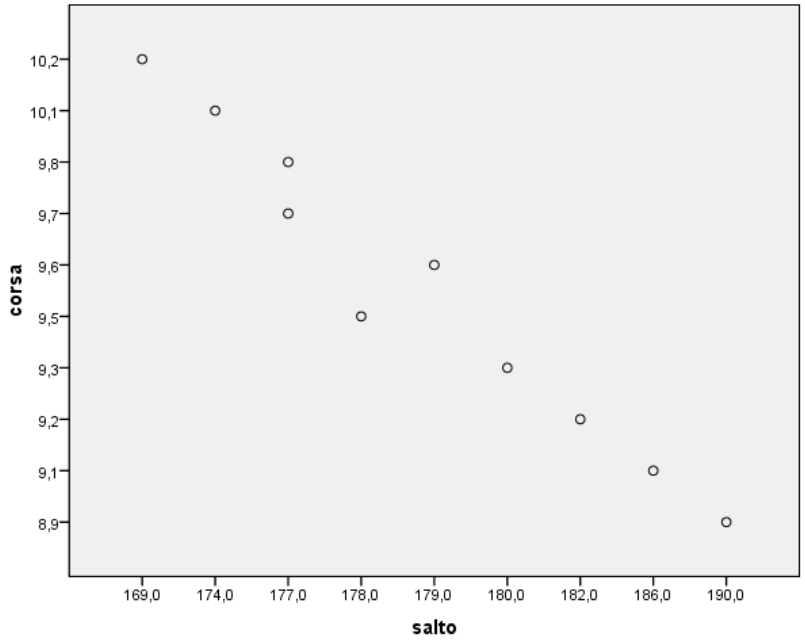
Plot ZCORSA vs. ZSALTO



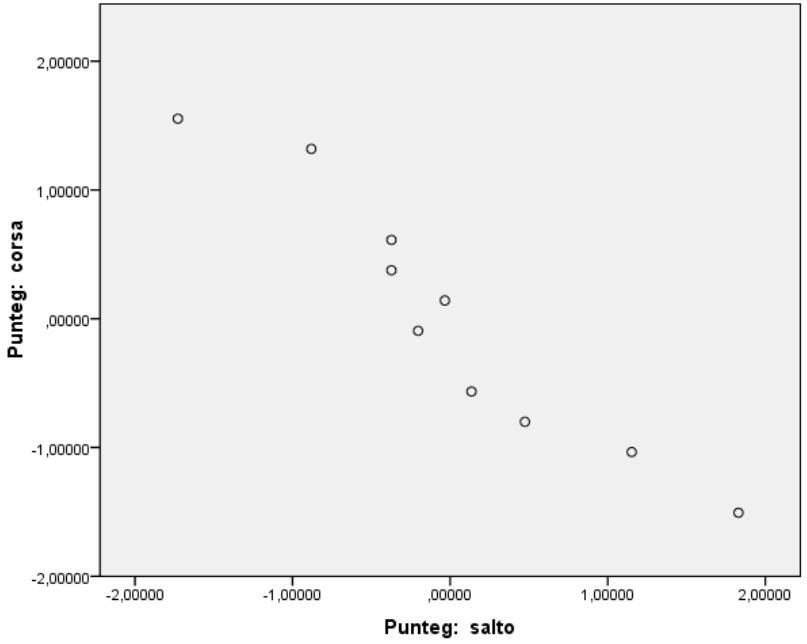
Plot: ZFLESS vs. ZSALTO



Plot CORSA vs. SALTO



Plot ZCORSA vs. ZSALTO



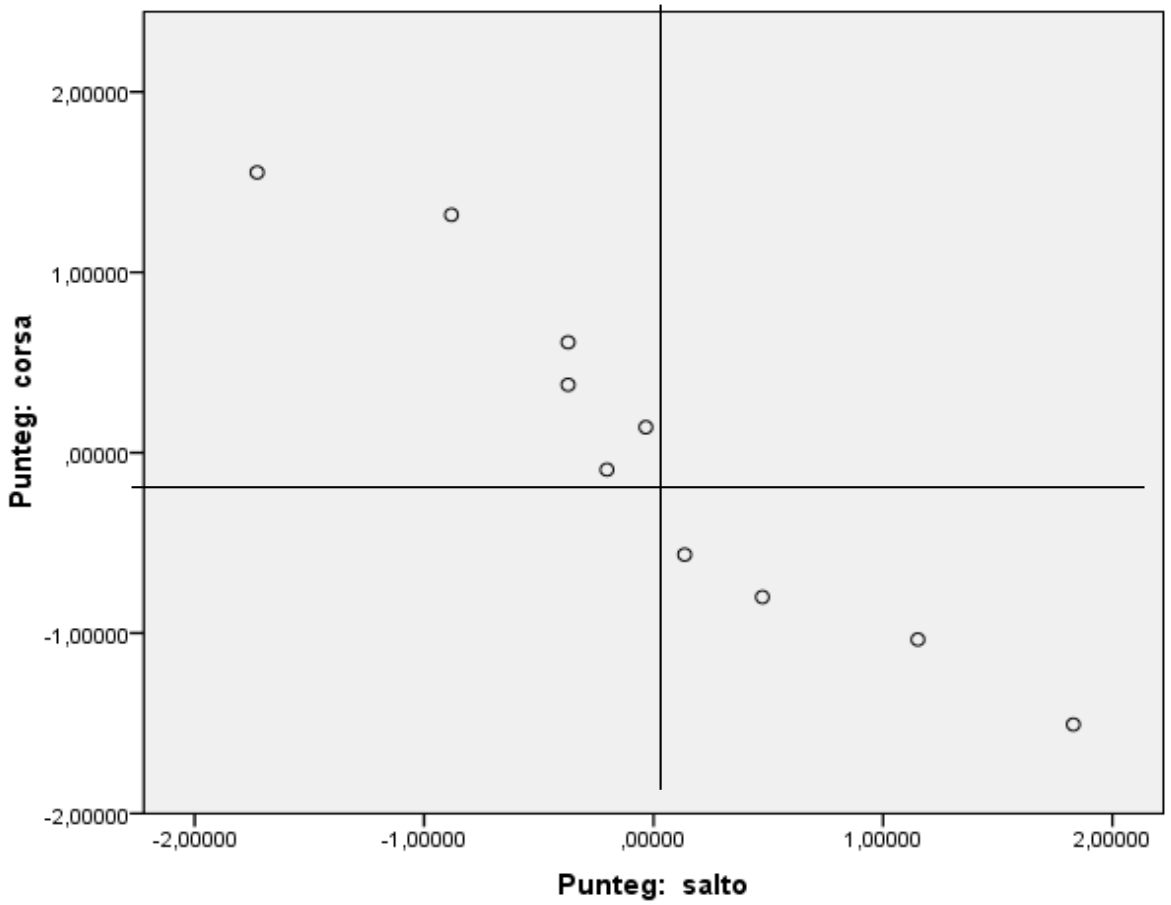
La correlazione - 5

- L’operazione di standardizzazione ha permesso di *centrare* le variabili, portando il centro del sistema di coordinate (0,0) a coincidere con i valori medi delle due variabili.
- Ha anche permesso di omogeneizzare la dispersione, per cui il peso delle due variabili nel determinare la forma della nuvola è uguale.

La correlazione - 6

- Tornando al grafico ZCORSA vs. ZSALTO, si può osservare come i ragazzi con risultati inferiori alla media nella corsa li abbiano ottenuti superiori alla media nel salto e viceversa: la nuvola dei punti si distribuisce unicamente nel II e IV quadrante del piano.

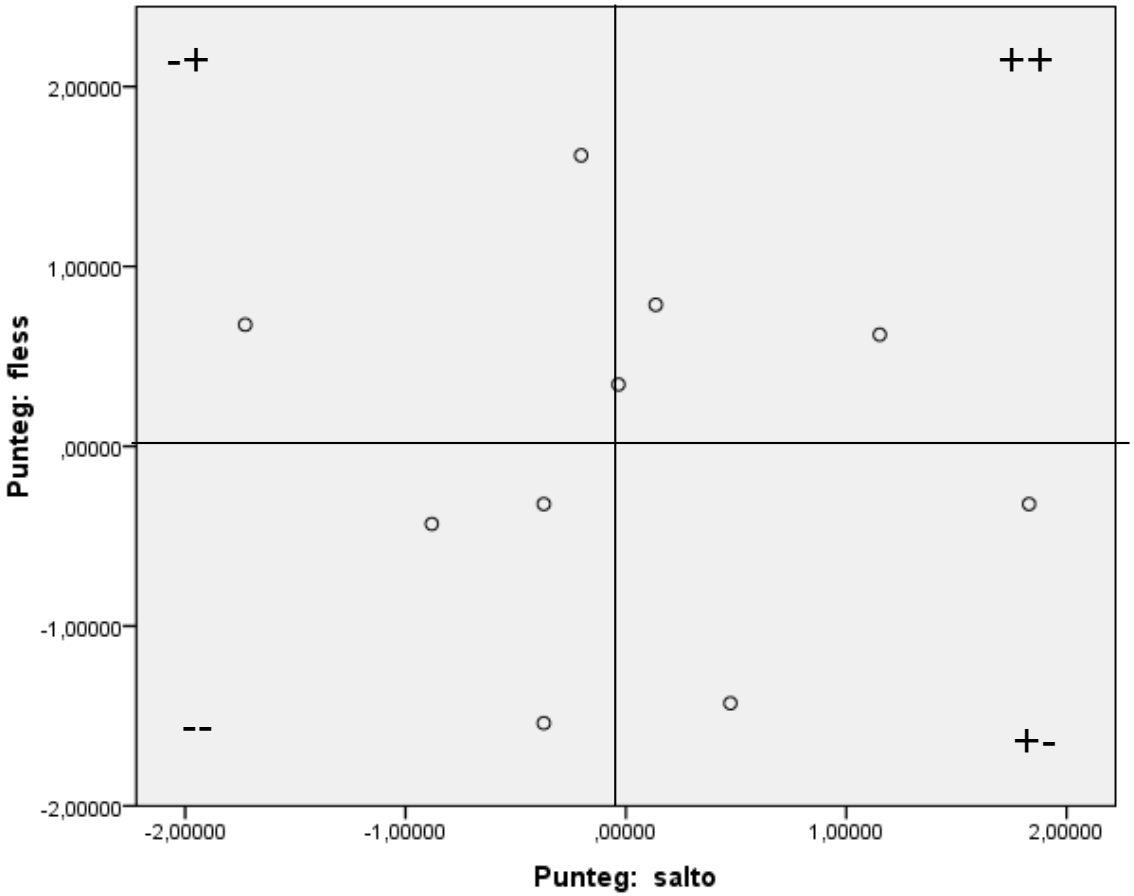
Plot ZCORSA vs. ZSALTO



La correlazione - 7

- Ovviamente è un fatto tecnico: le graduatorie dei risultati sono inverse (tempi e lunghezze). Quindi fra le due prove c'è discordanza tecnicamente e concordanza logicamente: gli allievi più bravi nella prima lo sono anche nella seconda.
- Invece nel grafico ZFLESS vs. ZSALTO non si individua nessuna relazione tendenziale fra i risultati nelle due prove, osservabile dalla dispersione della nuvola dei punti: questi si collocano in maniera uniforme nei quattro quadranti.

Plot: ZFLESS vs. ZSALTO



La correlazione - 8

- Siamo pronti a definire una misura sintetica delle relazioni che abbiamo presentato: il ***coefficiente di correlazione lineare di Pearson***.

$$-1 \leq \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^{n} (y_i - \mu_y)^2}} \leq +1$$

La correlazione - 9

- Questo indice è la media dei prodotti degli scarti standardizzati fra due variabili e, proprio per la standardizzazione, può assumere un *range* di valori compreso fra:
  - **-1** tra le due variabili vi è perfetta correlazione negativa, quindi discordanza, e
  - **+1** tra le due variabili vi è perfetta correlazione positiva, quindi concordanza, i punti sono allineati su una retta, passando per
  - **0** non vi è correlazione di tipo lineare, il che non vuol dire che non ci sia alcuna relazione.

La correlazione - 10

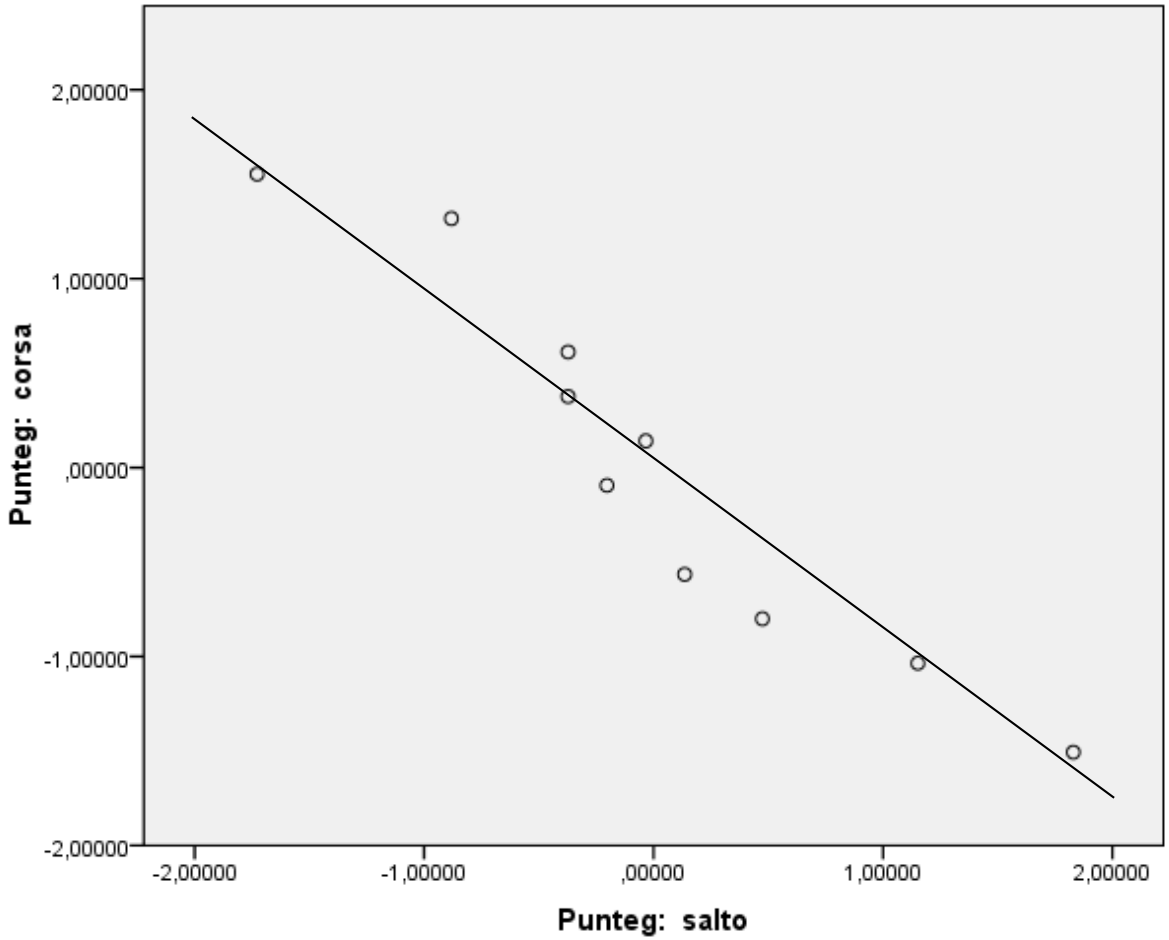
- Tra **-1** e **0** e, simmetricamente, tra **0** e **+1** vi sono infiniti valori, che indicano correlazione
  - bassa fra **0** e **-0,29** o fra **0** e **+0,29**;
  - media fra **-0,30** e **-0,59** o fra **+0,30** e **+0,59**;
  - alta fra **-0,60** e **-0,99** o fra **+0,60** e **+0,99**.

- Altra strategia va usata se si utilizza un coefficiente di correlazione campionaria per testare un ipotesi sulla correlazione in una popolazione (inferenza).

La correlazione -11

- Nel caso di correlazione alta, ossia quando le variabili hanno in comune qualcosa, possiamo cercare il significato di questa "comunalità": ad esempio cosa c'è in comune tra CORSA e SALTO?
- Questa "comunalità" può essere l'espressione di una variabile ***latente***, ossia non direttamente osservabile, che corrisponde a una retta (in quanto la correlazione misurata è quella lineare): questa retta è quella che passa (interpola) per la nuvola dei punti.
- La variabile latente è anche interpretabile, secondo la "Teoria dell'allenamento", come una capacità dell'atleta: la ***forza rapida***.

Plot ZCORSA vs. ZSALTO



La correlazione - 12

- Le ultime considerazioni aumentano di importanza quando si aumentano le dimensioni di riferimento, quando cioè le variabili poste a confronto simultaneamente sono più di tre e non è possibile rappresentare la nuvola dei punti in uno spazio per noi “leggibile”.

- Diviene così necessario cercare un ***punto di vista*** ottimale di dimensioni ridotte per leggere e sintetizzare i dati.

La correlazione - 13

- Pensiamo a una batteria di 30 prove: sarà necessario sintetizzare i risultati in un numero ridotto di prove per ***ordinare*** e ***classificare*** gli allievi, per vedere quali prove siano simili, quali attendibili, quali valide e quali dimensioni latenti soggiacciano alle loro prestazioni.

- Questo è il compito delle tecniche di **Analisi Multivariata**, sempre in un'ottica esplorativa (Analisi delle Componenti Principali, Analisi delle Corrispondenze Multiple, Cluster Analysis).

La regressione - 1

- Va tenuto presente che, finora, si sono studiate le relazioni fra due variabili seguendo un modello ***simmetrico***:
  - non si è considerata tecnicamente l'eventuale gerarchia tra le variabili stesse, ovvero non si è considerata una delle due indipendente e l'altra dipendente dai risultati della prima.
- Questo non toglie la possibilità di una gerarchia logica fra le due, ma nella correlazione le due variabili giocano un ruolo simmetrico.

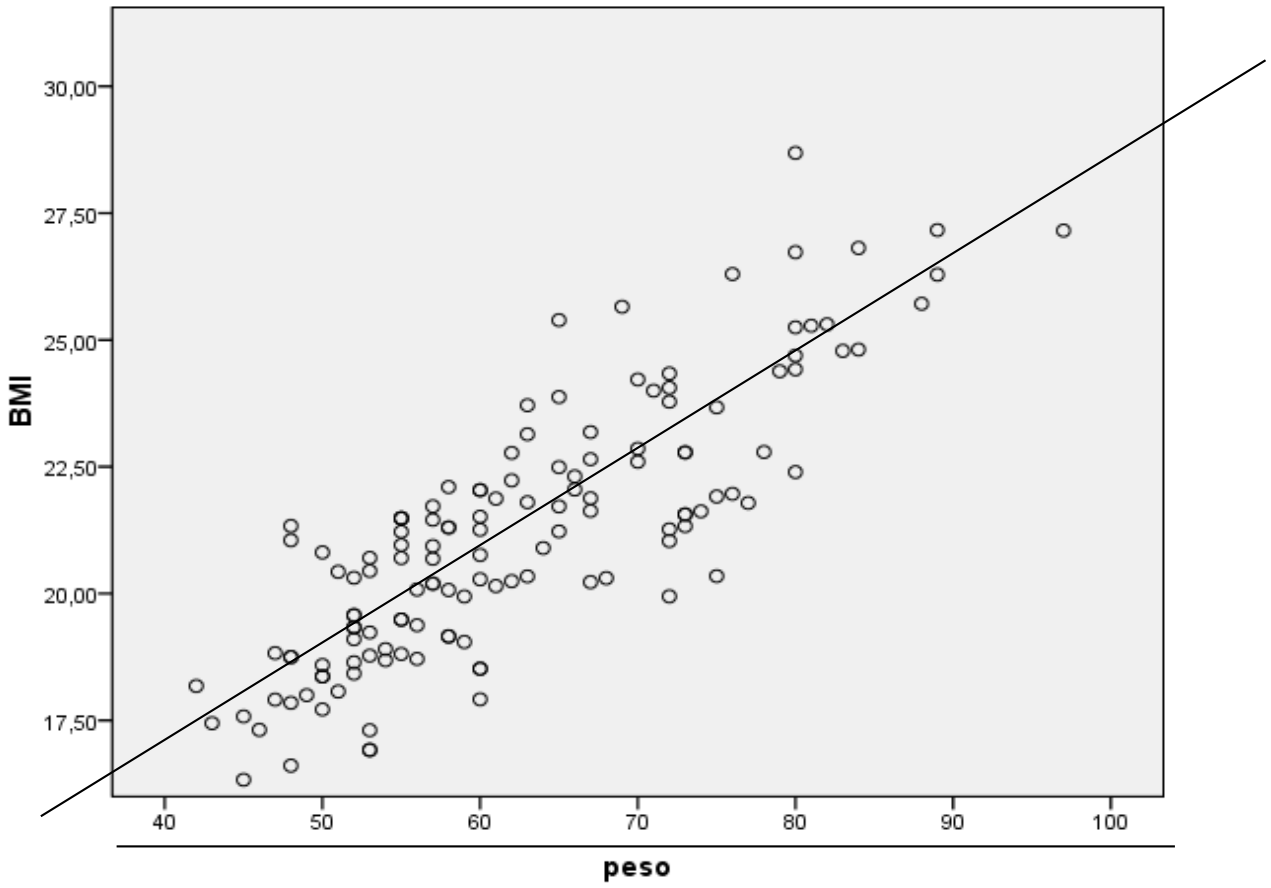
La regressione - 2

- Se c'è una correlazione, almeno "media" (oppure "significativa" in un'ottica inferenziale) e i risultati di una delle due variabili possono essere considerati dipendenti da quelli dell'altra, allora si può utilizzare il modello ***asimmetrico*** della **regressione**.

$$Y = a + b\,X$$

- Trovando i valori di **a** (intercetta sull'asse Y) e **b** (coefficiente angolare che indica l'inclinazione della retta) si possono prevedere i valori della **Y** al variare di quelli della **X**, anche in casi non osservati.

Plot: BMI vs. PESO



Ulteriori approfondimenti

Anomalie della Correlazione

- Abbiamo detto che **r** è il coefficiente di correlazione ***lineare*** (di Bravais Pearson): se però la relazione fra le variabili c'è, ma non è lineare (ad esempio è parabolica), **r=0**!

- Nel caso siano presenti valori estremi (***outlier***) la correlazione risulterà ***fittizia*** o ***soppressa***!!!!

- È necessario, quindi, studiare sempre il diagramma disispersione, per evitare questi errori!!!!

La fallacia ecologica

- Nel caso in cui le unità di analisi, su cui siano rilevate le due variabili messe in relazione, siano ***aggregati*** di individui (comune, municipio, regione e così via), si parla di ***correlazione ecologica***.
- In realtà il ricercatore vorrebbe conoscere la ***correlazione individuale*** (tra gli individui), ma questo non è possibile perché si è nel campo delle ***analisi secondarie*** (su dati rilevati da altri).
- **Pertanto non si dovrebbe mai interpretare una correlazione ecologica come correlazione individuale, da cui il termine di *fallacia ecologica.***
- Vedremo alcuni esempi nella seconda parte del corso.

La relazione spuria

- È il caso di "***presenza di covariazione, pur in assenza di causazione***"
- Esempio: correlazione fra numero di autopompe antincendio intervenute ed entità dei danni (in realtà dovuta alla dimensione dell'incendio.

- Anche questo caso sarà più facilmente trattabile con l'introduzione di più variabili!