STATISTICA 1 - Analisi Bivariate

Riccardo Corradin, Andrea Gilardi

Introduzione

- Le metodologie statistiche presentate finora rientrano nella cosiddetta **statistica descrittiva univariata**: p caratteri sono rilevati in una popolazione ed essi vengono studiati **separatamente uno alla volta**.
- E' però ragionevole supporre che possano esistere delle **relazioni** tra le caratteristiche degli individui in una popolazione. Ad esempio, l'altezza ed il genere di una persona sono tipicamente legati al peso.
- L'obiettivo della statistica (descrittiva) **multivariata** è quello di esplorare tali relazioni per capire i pattern ed i nessi presenti nei dati.
- In queste corso ci concentreremo sulla statistica descrittiva **bivariata**, cioè analizzeremo i caratteri **due alla volta**.

Introduzione - esempio

Example

La seguente tabella[a] riassume il *numero di interventi* e la *percentuale di sopravvivenza a 30 giorni* per gli interventi effetts sui bambini di età minore di un anno negli ospedali britannici che, nel periodo 1991-95, avevano un reparto di cardiochirurgia infantile:

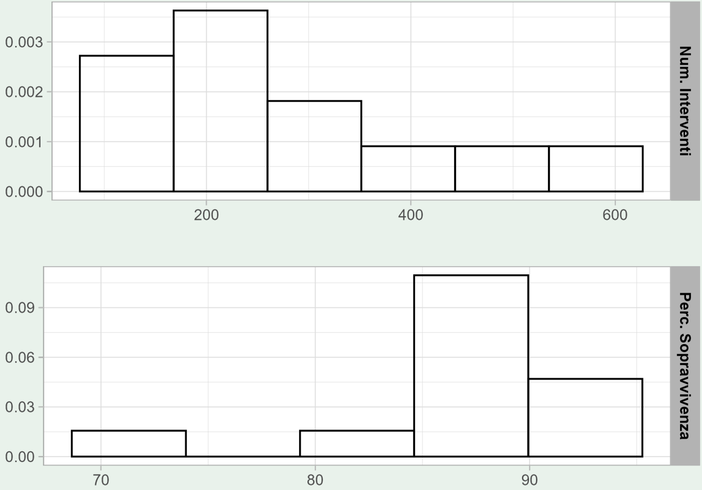
Ospedale	Num. Interventi	Perc. Sopravvivenza a 30 giorni
Birmingham	581	90%
Bristol	143	71.3
Brompton	301	89.7%
Great Ormond St.	482	89%
Guys	164	84.8%
Harefield	177	85.9%
...	...	...

[a] Fonte: *D.J. Spiegelhalter et al., Commissioned Analysis of Surgical Performance Using Routine Data: Lessons from the Bristol Inquiry.* Link.

Introduzione - esempio

Example

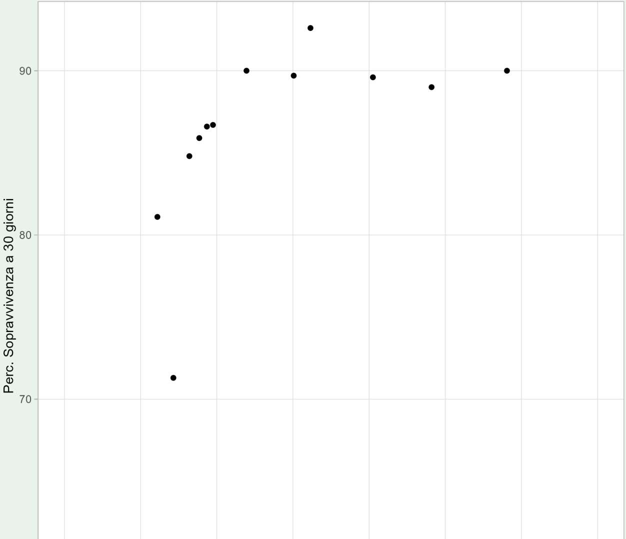
I seguenti istogrammi riassumono le due distribuzioni **singolarmente**:



Introduzione - esempio

Example

E' tuttavia interessante visualizzare il **legame**[a] tra queste due variabili:



Notazione

Tabelle a doppia entrata

- Supponiamo di aver osservato un **primo** carattere X che assume modalità $x_1, \ldots, x_r$ ed un **secondo** carattere Y che assume modalità $y_1, \ldots, y_c$.
- Tipicamente X e Y rappresenteranno due variabili *qualitative*.
- La **distribuzione congiunta** (i.e. la relazione) di X e Y può venire riassunta da una **tabella a doppia entrata** come segue:

$X \backslash Y$	y_1	$\ldots$	y_j	$\ldots$	y_c	Totale
x_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{ic}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_{r1}	$\ldots$	n_{rj}	$\ldots$	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$\ldots$	$n_{\cdot j}$	$\ldots$	$n_{\cdot c}$	N

Notazione

- Il simbolo n_{ij} indica le **frequenze assolute congiunte**: quante unità statistiche presentano **congiuntamente** le modalità x_i e y_j.
- $n_{i\cdot}$ indica la **somma** delle frequenze assolute congiunte poste sulla i-esima **riga**: $n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$.
- $n_{\cdot j}$ indica la **somma** delle frequenze assolute congiunte poste sulla j-esima **colonna**: $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$.
- Le frequenze $n_{i\cdot}$ e $n_{\cdot j}$ vengono denominate **frequenze assolute marginali**.
- Valgono le seguenti uguaglianze

$$\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j} = N$$

Notazione

- E' anche possibile definire le **frequenze relative congiunte** (f_{ij}) e le **frequenze relative marginali** ($f_{i\cdot}$ e $f_{\cdot j}$) come segue:

$$f_{ij} = \frac{n_{ij}}{N}; \qquad f_{i\cdot} = \frac{n_{i\cdot}}{N}; \qquad f_{\cdot j} = \frac{n_{\cdot j}}{N}$$

- Di conseguenza, una tabella a doppia entrata riassume **tre** distribuzioni: **una distribuzione congiunta** e **due distribuzioni marginali**.

$X \backslash Y$	y_1	$\ldots$	y_j	$\ldots$	y_c	Totale
x_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{ic}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_{r1}	$\ldots$	n_{rj}	$\ldots$	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$\ldots$	$n_{\cdot j}$	$\ldots$	$n_{\cdot c}$	N

Notazione

- Oltre a queste, è possibile studiare altre $r + c$ distribuzioni univariate ottenute esaminando **singolarmente** la i-esima riga o la j-esima colonna.

- Supponiamo di **restringere** le analisi alla prima riga della tabella. In tal caso, il totale è pari a $n_{1\cdot}$ diviso tra c gruppi: $\{n_{11}, \ldots, n_{1j}, \ldots, n_{1c}\}$:

$X \backslash Y$	y_1	$\ldots$	y_j	$\ldots$	y_c	Totale
x_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Tenendo fissa la prima riga, possiamo definire le **frequenze relative del carattere Y condizionate a $X = x_1$** come

$$f(y_j | x_1) = \frac{n_{1j}}{n_{1\cdot}}; \qquad j = 1, \ldots c.$$

Esse rappresentano la proporzione di unità che cadono nella classe (i, j) **condizionandoci** al fatto che il carattere X sia pari a x_1.

Notazione

- Possiamo generalizzare tale definizione a qualsiasi riga i della tabella:

$$f(y_j|x_i) = \frac{n_{ij}}{n_{i\cdot}}; \qquad j = 1, \ldots, c; i = 1, \ldots r.$$

- Analogamente possiamo studiare **singolarmente** le colonne della tabella, definendo le **frequenze relative del carattere X condizionate a $Y = y_j$**:

$$f(x_i|y_j) = \frac{n_{ij}}{n_{\cdot j}}; \qquad i = 1, \ldots, r; j = 1, \ldots, c$$

- Esistono quindi **r distribuzioni condizionate per il carattere Y** (una per ogni possibile valore X condizionante) e **c distribuzioni condizionate per il carattere X** (una per ogni possibile valore Y condizionante).

- Vale[1] inoltre che

$$\sum_{j=1}^{c} f(y_j|x_i) = 1 \qquad \sum_{i=1}^{r} f(x_i|y_j) = 1$$

[1] Si provi a dimostrare le due uguaglianze come esercizio.

Notazione

Example

Nel fantastico mondo di Flatlandia è tempo di eleggere il nuovo Presidente del Consiglio delle Forme. Gli abitanti sono divisi in: *Quadrati*, *Cerchi*, e *Rettangoli* e i tre candidati sono: *Archimede*, *Euclide*, e *Pitagora*. Da un campione casuale di $N = 120$ elettori si ricava che:

- tra i Quadrati, 15 voteranno Archimede, 10 Euclide e 15 Pitagora;
- tra i Cerchi, 20 voteranno Archimede, 5 Euclide e 5 Pitagora;
- tra i Rettangoli, 10 voteranno Archimede, 15 Euclide e 25 Pitagora.

Dopo aver costruito un'opportuna tabella a doppia entrata ed aver brevemente elencato tutte le distribuzioni riassunte in essa, si risponda alle seguenti:

1. Qual è la frequenza assoluta congiunta della coppia (*Cerchi*, *Archimede*)?
2. Qual è la frequenza relativa marginale dei voti per *Pitagora*?
3. Quanto vale la freq. relativa marginale dei voti per Archimede e la freq. relativa dei voti per Archimede condizionata alla popolazione dei Cerchi?

Si interpretino i risultati ottenuti.

Notazione - Esercizio

Example

Connessione

(Assenza di) Connessione

Un carattere X viene definito **indipendente in distribuzione** (o **non-connesso**) dal carattere Y se, per ogni $j = 1, ..., c$, tutte le frequenze relative condizionate di X dato $Y = y_j$ sono **identiche fra loro**:

$$f(x_i|y_1) = f(x_i|y_2) = \cdots = f(x_i|y_c) \quad \forall i = 1, \ldots, r$$

Example

La seguente tabella a doppia entrata

X \ Y	y_1	y_2	y_3	
x_1	5	10	15	30
x_2	3	6	9	18
x_3	2	4	6	12
	10	20	30	60

mostra un carattere X che è **indipendente in distribuzione** da Y. Infatti ...

(Assenza di) Connessione - Proprietà

- Se X è **indipendente in distribuzione** da Y allora le sue frequenze relative condizionate sono pari alla corrispondente frequenza relativa marginale:

$$f(x_i|y_1) = f(x_i|y_2) = \cdots = f(x_i|y_c) = f_{i\cdot} \quad \forall i = 1, \dots, r$$

- Se X è indipendente in distribuzione da Y allora **anche Y è indipendente in distribuzione da X**. In altre parole, se

$$f(x_i|y_1) = f(x_i|y_2) = \cdots = f(x_i|y_c) = f_{i\cdot} \quad \forall i = 1, \dots, r$$

  allora necessariamente vale anche che

$$f(y_j|x_1) = f(y_j|x_2) = \cdots = f(y_j|x_r) = f_{\cdot j} \quad \forall j = 1, \dots, c.$$

- Se X e Y sono indipendenti in distribuzione allora

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}$$

NB: Cosa implica la terza proprietà sulla distribuzione degli 0 in una tabella a doppia entrata?

(Assenza di) Connessione - Proprietà

Example

Data la seguente tabella a doppia entrata (che è la medesima della slide 13)

X \ Y	y_1	y_2	y_3	
x_1	5	10	15	30
x_2	3	6	9	18
x_3	2	4	6	12
	10	20	30	60

verifichiamo empiricamente le proprietà elencate in precedenza.

(Assenza di) Connessione - Proprietà

Example

Massima Connessione

- Si parla di **massima connessione unilaterale** di un carattere X da un carattere Y la situazione in cui se di una unità statistica è nota la modalità di Y allora è univocamente determinata anche la sua modalità di X:

X \ Y	y_1	y_2	y_3	y_4
x_1	0	3	0	0
x_2	5	0	0	2
x_3	0	0	4	0

- Si parla di **massima connessione bilaterale** la situazione in cui si ha **massima connessione unilaterale** di X da Y e, al contempo, **massima connessione unilaterale** di Y da X.

X \ Y	y_1	y_2	y_3
x_1	0	3	0
x_2	5	0	0
x_3	0	0	4

Massima Connessione - Esercizio

Example

Consideriamo due caratteri X e Y aventi le seguenti distribuzioni marginali:

X \ Y	y_1	y_2	y_3	
x_1				10
x_2				8
x_3				7
	8	7	10	25

Completare la tabella a doppia entrata nei due seguenti casi:

1. X e Y sono indipendenti in distribuzione;
2. X e Y sono massimamente connessi.

Contingenze

- Data una tabella di frequenze, è sempre possibile calcolare le quantità

$$\hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}.$$

  Esse rappresentano le **frequenze assolute teoriche in caso di indipendenza in distribuzione**.

- Una eventuale *discrepanza* tra n_{ij} e $\hat{n}_{ij}$ è **sintomo** di connessione tra X e Y. Definiamo quindi le **contingenze assolute** come

$$C_{ij} = n_{ij} - \hat{n}_{ij}$$

  e le **contingenze relative** come

$$\rho_{ij} = \frac{C_{ij}}{\hat{n}_{ij}} = \frac{n_{ij} - \hat{n}_{ij}}{\hat{n}_{ij}}.$$

- Interpretazione:
  - Se $C_{ij} \simeq 0$ allora X e Y sono *indipendenti in distribuzione*;
  - Se $C_{ij} > 0$ allora X e Y presentano una certa *attrazione*;
  - Se $C_{ij} < 0$ allora X e Y presentano una certa *repulsione*.

Contingenze

Example

Data la seguente tabella a doppia entrata

X \ Y	y_1	y_2	y_3	
x_1	15	10	15	40
x_2	20	5	5	30
x_3	10	15	25	50
	45	30	45	120

si calcolino le **contingenze relative** ρ_{ij}.

Indice quadratico di connessione di Pearson

- L'**indice quadratico di connessione di Pearson** sintetizza le contingenze relative tramite una media quadratica con pesi pari alle frequenze teoriche in caso di indipendenza, $\hat{n}_{ij}$:

$$M_2(\rho) = \sqrt{\frac{1}{N}\sum_{j=1}^{c}\sum_{i=1}^{r}\hat{n}_{ij}\rho_{ij}^2} = \sqrt{\frac{1}{N}\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}}$$

- E' possibile dimostrare che tale indice è anche esprimibile come $M_2(p) = \sqrt{\frac{1}{N}\chi^2}$ dove

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{\hat{n}_{ij}} - N.$$

  Questa formulazione evita il calcolo delle differenze $n_{ij} - \hat{n}_{ij}$.

- E' inoltre possibile dimostrare che $M_2(\rho) \leq \sqrt{\min(r-1, c-1)}$, il che ci permette di definire un **indice normalizzato**:

$$\tilde{M}_2(\rho) = \frac{M_2(\rho)}{\sqrt{\min(r-1, c-1)}}$$

Indice quadratico di connessione di Pearson

Example

Calcolare **l'indice quadratico di connessione di Pearson** (nella sua versione standard e normalizzata) per la seguente tabella a doppia entrata

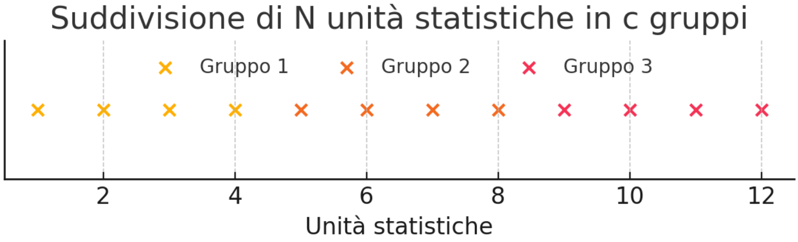
X \ Y	y_1	y_2	y_3	
x_1	15	10	15	40
x_2	20	5	5	30
x_3	10	15	25	50
	45	30	45	120

Dipendenza in media

Dipendenza in media

- Gli strumenti presentati finora vengono solitamente usati quando i caratteri di interesse sono di tipo *qualitativo* (nominale o ordinale). Se almeno uno dei due è di tipo *quantitativo* abbiamo a disposizione anche altre analisi.

- Sia X un carattere **quantitativo** e sia Y un carattere **qualitativo** che assume valori $y_1, \ldots, y_c$ con frequenze $n_{\cdot 1}, \ldots, n_{\cdot j}, \ldots, n_{\cdot c}$. Il carattere Y divide naturalmente le N osservazioni in c gruppi.



Dipendenza in media

- Se esiste una connessione tra X e Y (i.e. se non sono **indipendenti in distribuzione**), possiamo studiare come varia la *media* di X rispetto ai gruppi definiti da Y. Per ciascun gruppo j possiamo definire

$$M_1(X|y_j) = \bar{x}_j = \frac{1}{n_{\cdot j}} \sum_{i=1}^{r} x_i \cdot n_{ij} = \sum_{i=1}^{r} x_i f(x_i|y_j).$$

- Diciamo che X è **indipendente in media** da Y se

$$M_1(X|y_1) = M_1(X|y_2) = \cdots = M_1(X|y_c) = M_1(X)$$

dove $M_1(X) = \bar{x} = \frac{1}{N} \sum_{i=1}^{r} x_i n_{i\cdot}$. Se ciò non è verificato, diciamo che X **dipende in media** da Y.

Proprietà:

1 - La dipendenza in media non è l'unica forma di dipendenza tra variabili quantitative e qualitative. Possiamo anche parlare di *dipendenza in mediana*, *dipendenza in varianza* . . .

Dipendenza in media

2 - L'indipendenza in media **non** è una proprietà **simmetrica**. Infatti, data la loro natura, non ha senso chiedersi se Y dipende in media da X.

3 - L'indipendenza in distribuzione **implica** l'indipendenza in media:

$$\text{Indipendenza in Distribuzione} \Rightarrow \text{Indipendenza in media}$$

Come mostra il seguente esempio, il viceversa **non è vero**.

Example

Si dimostri che nella seguente tabella vi è **indipendenza in media** tra X e Y:

X \ Y	A	B	C	D
4	2	0	3	3
8	4	4	1	3
12	4	4	1	3
16	2	0	3	3

E' possibile che X e Y siano **indipendenti in distribuzione**?

Dipendenza in media

- Dopo aver stabilito che X e Y non sono indipendenti in media, potremmo anche essere interessati a misurare la *forza* di questa dipendenza.

- Il **Rapporto di correlazione di Pearson** raggiunge questo scopo:

$$\eta^2_{(X|Y)} = \frac{\text{Devianza fra i gruppi}}{\text{Devianza totale}}$$

dove

$$\text{Devianza fra i gruppi} = \sum_{j=1}^{c} n_{\cdot j}(\bar{x}_j - \bar{x})^2$$

e

$$\begin{aligned}\text{Devianza totale} &= \text{Devianza nei gruppi} + \text{Devianza fra i gruppi} \\ &= \sum_{j=1}^{c}\left(\sum_{i=1}^{r}(x_i - \bar{x}_j)^2 \cdot n_{ij}\right) + \sum_{j=1}^{c} n_{\cdot j}(\bar{x}_j - \bar{x})^2\end{aligned}$$

NB: Pensate alla *Devianza* come ad una *Varianza* che non viene divisa per N.

Dipendenza in media

L'indice $\eta^2_{(X|Y)}$ gode di alcune proprietà.

- L'indice è naturalmente **normalizzato**, cioè $0 \leq \eta^2_{(X|Y)} \leq 1$.

Analizziamo ora le due situazioni estreme.

- L'indice vale 0 se e solo se tutte le medie nei gruppi sono uguali a $\bar{x}$:

$$\sum_{j=1}^{c} n_{\cdot j}(\bar{x}_j - \bar{x})^2 = 0 \iff \bar{x}_j = \bar{x} \quad \forall j = 1, \ldots, c.$$

  Questa è esattamente la definizione di **indipendenza in media**. Quindi

$$\eta^2_{(X|Y)} = 0 \iff X \text{ e } Y \text{ sono indipendenti in media}$$

- L'indice vale 1 se e solo se la devianza nei gruppi è nulla, il che succede solamente quando tutte le distribuzioni condizionate di X assumono un solo valore che è forzatamente pari a $\bar{x}_j$. Di conseguenza

$$\eta^2_{(X|Y)} = 1 \iff X \text{ e } Y \text{sono } \textbf{massimamente connessi}$$

Dipendenza in media

- Inoltre, come già detto,

$$M_2(\rho) = 0 \Rightarrow \eta^2_{(X|Y)} = 0$$
$$\eta^2_{(X|Y)} = 0 \not\Rightarrow M_2(\rho) = 0$$

Example

In uno studio di mineralogia si vogliono studiare tre tipi di roccia differenti (Y) misurandone il numero di cristalli per unità di volume standard (X). I dati raccolti sono riassunti nella seguente tabella:

Roccia	Misurazioni
Granito	15, 18, 14, 17
Basalto	8, 7, 9, 10
Calcare	4, 5, 6, 5

Si calcoli il **rapporto di correlazione di Pearson**, $\eta^2_{(X|Y)}$, commentandone il valore ottenuto.

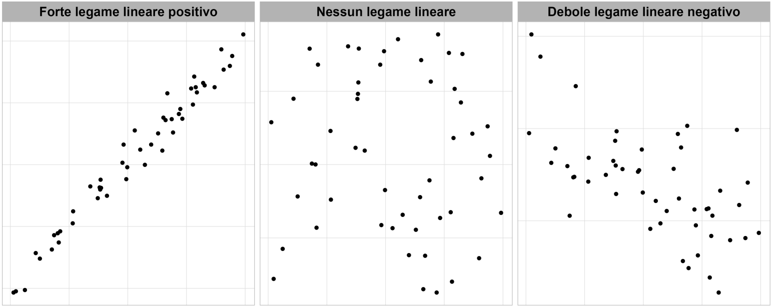
Covarianza e Correlazione

Covarianza

- Supponiamo ora che X e Y rappresentino due caratteri **quantitativi** rilevati **congiuntamente** su una popolazione di N unità. Siano

$$\{(x_1, y_1), \ldots, (x_N, y_N)\}$$

le coppie di osservazioni. Come già commentato, la relazione tra X e Y può essere esplorata graficamente tramite *diagrammi a dispersione*:



Covarianza

La **covarianza** tra due caratteri numerici X e Y è un indice che misura la forza della relazione **lineare** ed è definito come

$$\mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}).$$

Proprietà

- Se $\mathrm{cov}(x,y) = 0$, allora i due caratteri si dicono essere **incorrelati**.
- La covarianza è un indice **simmetrico**: $\mathrm{cov}(x,y) = \mathrm{cov}(y,x)$.
- La covarianza di un indice X con se stesso è pari alla sua **varianza**. La covarianza di X e $-X$ è pari alla varianza cambiata di segno:

$$\mathrm{cov}(x,x) = \mathrm{var}(x) \qquad \mathrm{cov}(x,-x) = -\mathrm{var}(x).$$

- La covarianza viene tipicamente **calcolata** sfruttando la seguente formula

$$\mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}.$$

Covarianza

- Se i dati di un problema sono espressi tramite una **tabella a doppia entrata**, allora possiamo calcolare $\mathrm{cov}(x, y)$ come

$$\mathrm{cov}(x,y) = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}(x_i - \bar{x})(y_j - \bar{y})$$

  o, sfruttando la formula precedente,

$$\mathrm{cov}(x,y) = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} x_i y_j - \bar{x}\bar{y}.$$

- Siano $V = \alpha_1 + \beta_1 X$ e $W = \alpha_2 + \beta_2 Y$ due trasformazioni lineari dei caratteri X e Y. Allora

$$\mathrm{cov}(v,w) = \beta_1 \beta_2 \mathrm{cov}(x,y)$$

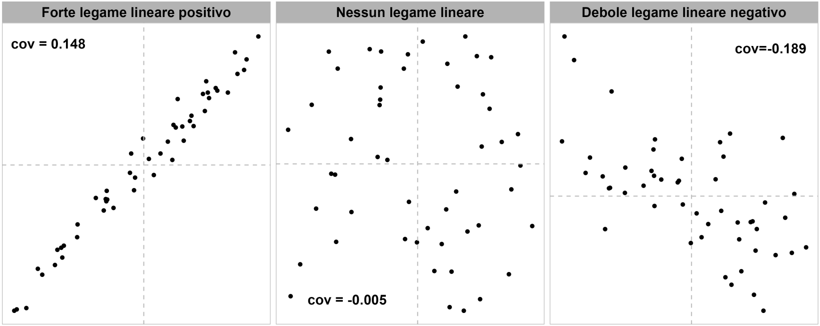
- Se $\eta^2_{(X|Y)} = 0$ oppure $\eta^2_{(Y|X)} = 0$ **allora** $\mathrm{cov}(x, y) = 0$. C

Domanda: Che relazione può sussistere tra indipendenza in distribuzione e incorrelazione?

Covarianza

La covarianza assume valori

- **positivi** se i termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono **concordi** (stesso segno);
- **prossimi a 0** se i termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono **in egual misura concordi e discordi** così da compensarsi;
- **negativi** se i termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono **discordi** (segni diversi).



Covarianza

Example

Supponiamo di aver intervistato $N = 5$ persone e di aver chiesto loro le seguenti informazioni: $X =$ 'Reddito Annuo' (in migliaia di euro) e $Y =$ 'Superficie della casa in mq2'. La seguente tabella riporta le informazioni rilevate

ID	1	2	3	4	5
Reddito (X)	28	42	33	85	66
Superficie (Y)	52	65	64	84	102

Rappresentare la relazione tra X e Y tramite un grafico opportuno e calcolare $\text{cov}(x, y)$ commentando il risultato ottenuto.

Covarianza

- Come si evince dal precedente esempio, è chiaro come interpretare il **segno** di una covarianza ma non il suo esatto **valore numerico**. . .

- Sarebbe utile poter definire una versione **normalizzata** di questo indice!! Siamo fortunati poichè è possibile dimostrare che

$$-\mathrm{sqm}(x)\mathrm{sqm}(y) \leq \mathrm{cov}(x, y) \leq \mathrm{sqm}(x)\mathrm{sqm}(y)$$

  dove $\mathrm{sqm}(x)$ e $\mathrm{sqm}(y)$ rappresentano, rispettivamente, lo **scarto quadratico medio** dei caratteri X e Y, indicati anche come σ_X e σ_Y.

Esercizio

Si dimostri empiricamente la proprietà di massimo e minimo della covarianza utilizzando i dati della slide precedente.

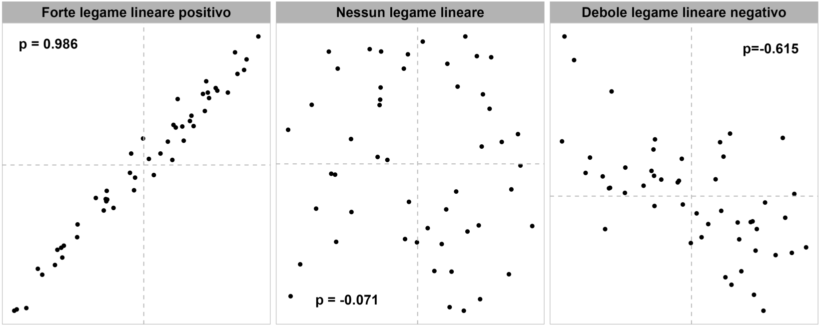
Correlazione

La covarianza nella sua versione **normalizzata** viene chiamata **correlazione** ed è definita come

$$\rho_{XY} = \text{cor}(x, y) = \frac{\text{cov}(x, y)}{\text{sqm}(x)\text{sqm}(y)}$$

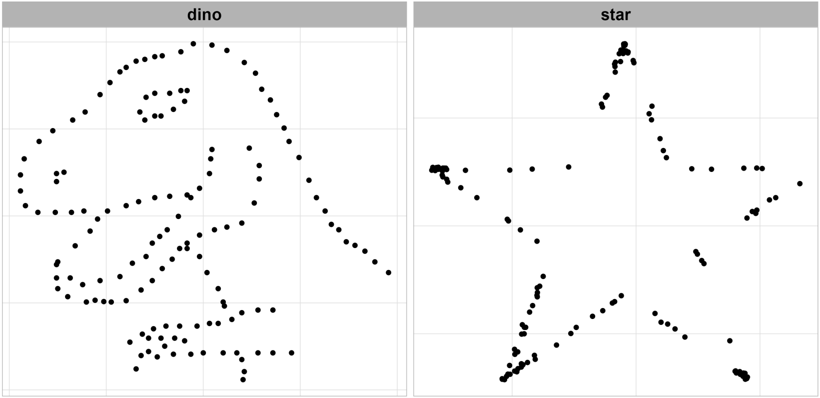
Proprietà:

- Per definizione, si ha che $-1 \leq \rho_{XY} \leq 1$.
- Inoltre, sempre per definizione, $\text{cor}(x, x) = 1$ e $\text{cor}(x, -x) = -1$. Quando si verifica quindi che $|\rho_{XY}| = 1$?



Correlazione

La **covarianza** e la **correlazione** misurano il legame **lineare** tra due variabili! In entrambi i seguenti casi si ha che $\rho_{XY} \simeq 0$ tuttavia. . .



Correlazione

Example

Nel 1886 Francis Galton ha pubblicato un articolo[a] in cui studiava come variassero le caratteristiche individuali da una generazione all'altra. In particolare egli era interessato a capire la relazione tra l'altezza dei genitori e quella dei loro figli (da adulti). La seguente tabella contiene un estratto dei dati pubblicati da Galton.

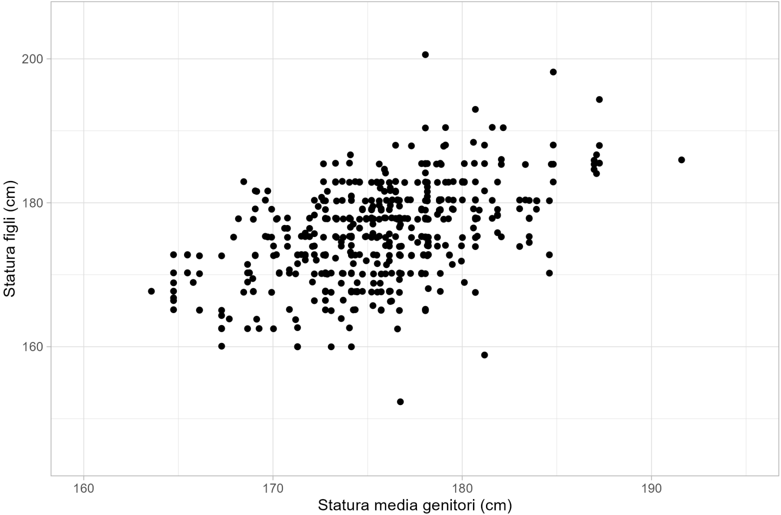
Altezza media genitori (cm)	176.5	168.9	171.5	168.9	174
Altezza figli (cm)	180.8	175.7	165.6	173.3	173

Dopo aver rappresentato i dati tramite un istogramma, si calcoli la correlazione tra le due misure interpretando il risultato ottenuto.

[a] Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature Journal of the Anthropological Institute, 15, 246-263

Regressione lineare semplice

La seguente figura mostra il grafico a dispersione della statura dei figli maschi rispetto ai loro padri usando i dati pubblicati da Galton.



Regressione lineare semplice

- I dati hanno un legame lineare positivo mediamente forte ($\rho_{XY} = 0.48$). Potremmo quindi essere interessati a rispondere alla seguente domanda:

  Come prevedere la statura dei figli conoscendo quella dei genitori?

- Adottiamo per il momento una ipotesi di **linearità**:

$$\text{AltFigli} = \alpha + \beta(\text{AltGenitori}) + (\text{Errore})$$

  La componente di **Errore** cattura la randomicità nell'altezza dei figli non spiegabile tramite le altezze dei genitori.

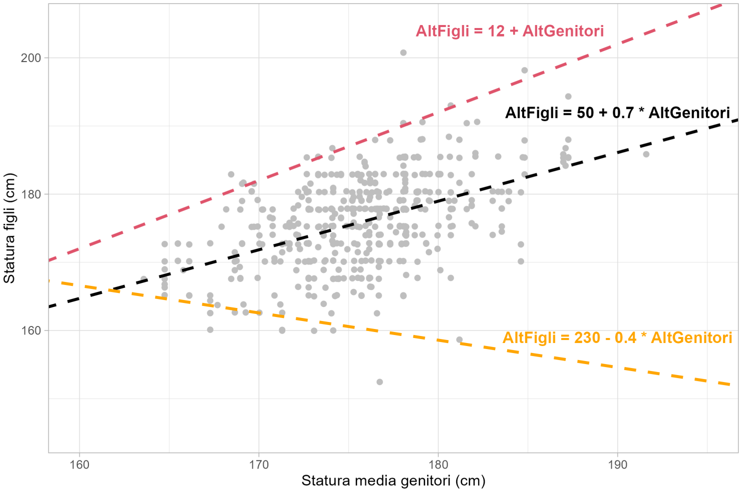
- I termini α e β rappresentano rispettivamente l'**intercetta** ed il **coefficiente angolare** della **retta di regressione**.

- Entrambi sono parametri ignoti che vanno **stimati** tramite i dati osservati. I valori stimati verranno indicati come $\hat{\alpha}$ e $\hat{\beta}$.

Regressione lineare semplice

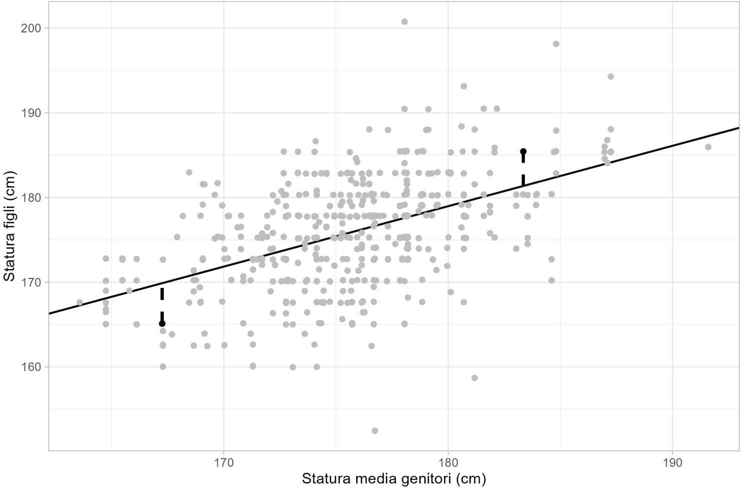
Siano $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ le altezze osservate sulle N unità statistiche (i.e. coppie padre/figlio). Esistono **infinite rette** del tipo

$$\text{AltFigli} = \hat{\alpha} + \hat{\beta}(\text{AltGenitori})$$



Regressione lineare semplice

Ogni retta che scegliamo creerà un **residuo** (le linee verticali tratteggiate): la **differenza** tra la statura di un figlio ed il valore risultante se usassimo **la retta per prevedere la statura a partire dall'altezza dei genitori**.



Regressione lineare semplice

- Per stimare i valori **ottimi** di α e β **minimizziamo** la somma dei residui al **quadrato**:

$$\arg\min_{\alpha,\beta} \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2.$$

- E' possibile dimostrare che la **soluzione ottima** a questo problema è

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}; \qquad \hat{\beta} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$$

- Nel caso dei dati di Galton si ha che

$$\bar{x} = 175.632; \qquad \bar{y} = 175.85; \qquad \mathrm{cov}(x,y) = 14.482; \qquad \mathrm{var}(x) = 20.305$$

così che

$$\hat{\beta} = \frac{14.482}{20.305} \simeq 0.7; \qquad \hat{\alpha} = 175.85 - 0.7 \cdot 175.632 \simeq 50.$$

Regressione lineare semplice

Ma perchè la **regressione lineare semplice** si chiama **regressione**? E' colpa di Francis Galton e del fenomeno denominato **regressione verso la media**.