

STATISTICA 1 - Analisi Bivariate

Riccardo Corradin, Andrea Gilardi

- 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 questo corso ci concentreremo sulla statistica descrittiva **bivariata**, cioè analizzeremo i caratteri **due alla volta**.

Example

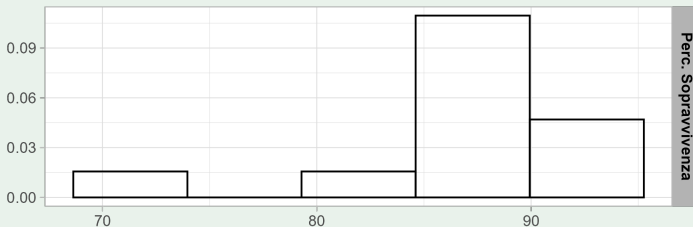
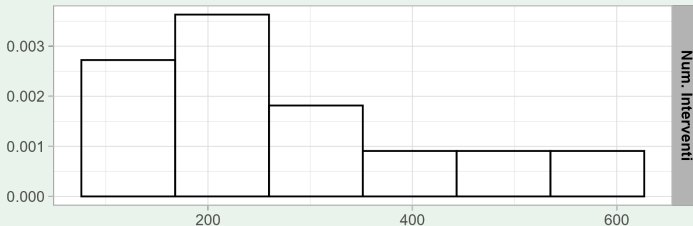
La seguente tabella^a riassume il *numero di interventi* e la *percentuale di sopravvivenza a 30 giorni* per gli interventi effettuati sui bambini di età minore di un anno negli ospedali britannici che, nel periodo 1991-95, avevano un reparto di cardiocirurgia infantile:

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%
...

^aFonte: D.J. Spiegelhalter et al., *Commissioned Analysis of Surgical Performance Using Routine Data: Lessons from the Bristol Inquiry*. [Link](#).

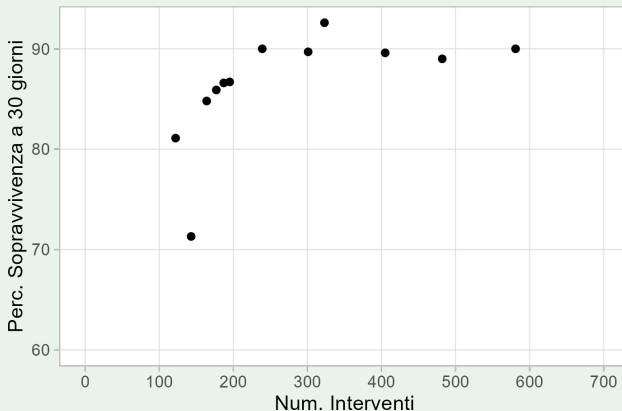
Example

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



Example

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



Cosa possiamo dedurre studiando le distribuzioni **univariate**? E quella **bivariata**?

^aQuesta rappresentazione grafica viene chiamata **grafico a dispersione**.

Notazione

Tabelle a doppia entrata

- Supponiamo di aver osservato un **primo** carattere X che assume modalità x_1, \dots, x_r ed un **secondo** carattere Y che assume modalità y_1, \dots, 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	\dots	y_j	\dots	y_c	Totale
x_1	n_{11}	\dots	n_{1j}	\dots	n_{1c}	$n_{1\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_{i1}	\dots	n_{ij}	\dots	n_{ic}	$n_{i\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_r	n_{r1}	\dots	n_{rj}	\dots	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	\dots	$n_{\cdot j}$	\dots	$n_{\cdot c}$	N

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

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

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

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

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

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

- 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: $\{n_{11}, \dots, n_{1j}, \dots, n_{1c}\}$:

$X \backslash Y$	y_1	\dots	y_j	\dots	y_c	Totale
x_1	n_{11}	\dots	n_{1j}	\dots	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

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

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

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

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

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

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

- 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).
- Vale¹ inoltre che

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

¹Si provi a dimostrare le due uguaglianze come esercizio.

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.

Example

Connessione

(Assenza di) Connessione

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

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

Example

La seguente tabella a doppia entrata

$X \setminus Y$	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:

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

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

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

allora necessariamente vale anche che

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

- Se X e Y sono indipendenti in distribuzione allora

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

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

Example

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

$X \setminus Y$	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.

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 :

$X \setminus Y$	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 .

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

Example

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

$X \setminus Y$	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.

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

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

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*.

Example

Data la seguente tabella a doppia entrata

$X \setminus Y$	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

$$M_2(\rho) = \sqrt{\frac{1}{N} \chi^2} \text{ dove}$$

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

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**:

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

Example

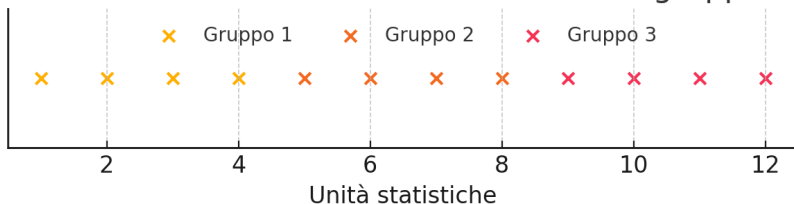
Calcolare l'**indice quadratico di connessione di Pearson** (nella sua versione standard e normalizzata) per la seguente tabella a doppia entrata

$X \setminus Y$	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

- 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, \dots, y_c con frequenze $n_1, \dots, n_j, \dots, n_c$. Il carattere Y divide naturalmente le N osservazioni in c gruppi.

Suddivisione di N unità statistiche in c gruppi



- 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

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

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:

- 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:

Indipendenza in Distribuzione \Rightarrow 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 :

$X \setminus Y$	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**?

- 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

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

e

Devianza totale = Devianza nei gruppi + Devianza fra i gruppi

$$= \sum_{j=1}^c \left(\sum_{i=1}^r (x_i - \bar{x}_j)^2 \cdot n_{ij} \right) + \sum_{j=1}^c n_{.j} (\bar{x}_j - \bar{x})^2$$

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

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} :

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

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

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

- Inoltre, come già detto,

$$M_2(\rho) = 0 \Rightarrow \eta_{(X|Y)}^2 = 0$$

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

$$\eta_{(X|Y)}^2 > 0 \Rightarrow M_2(\rho) > 0$$

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 coefficiente di correlazione parziale $\eta_{(X|Y)}^2$, commentandone il valore.

Correlazione

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

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

le coppie di osservazioni. Ad esempio:

- La **covarianza** tra X e Y è un indice che misura la forza della relazione tra X e Y ed è definita come

$$\text{cov}(x, y)$$