Statistica

Variabili doppie

Domenico De Stefano

a.a. 2024/2025

Indice

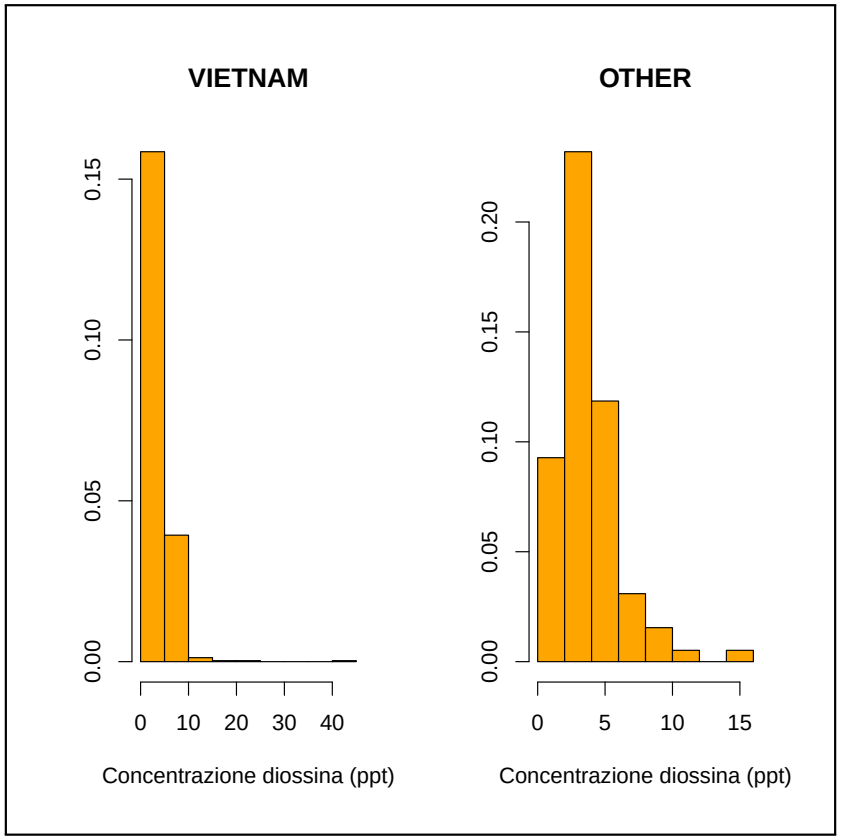
Introduzione

Le misure di posizione e scala che abbiamo introdotto fino ad ora si riferivano ad una sola variabile.

A volte, però, abbiamo in qualche modo dato un'occhiata a due variabili congiuntamente.

Esempio: dataset `vets`

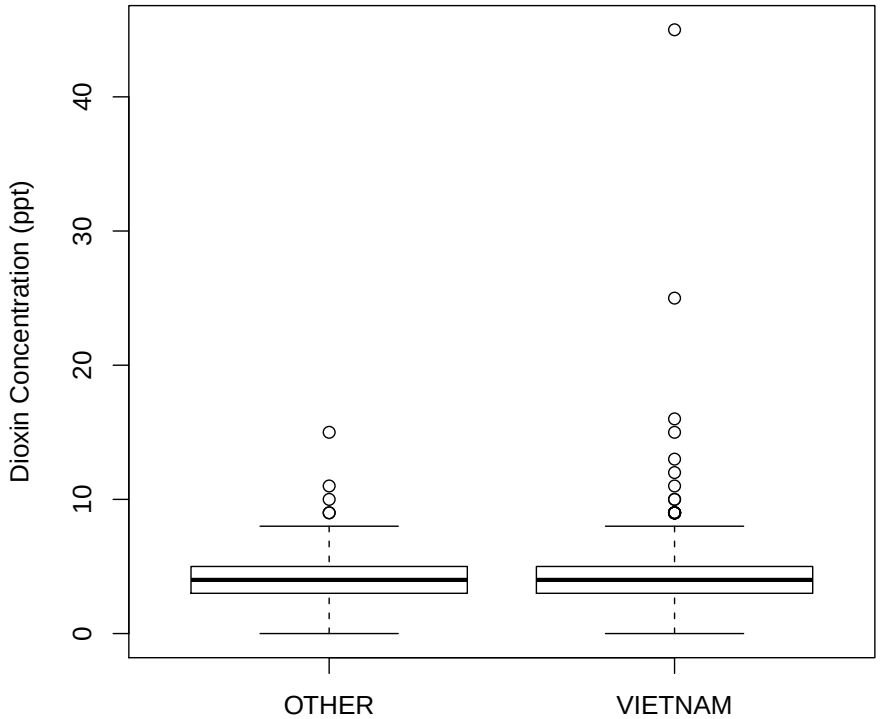
Qui abbiamo guardato le distribuzioni della concentrazione di diossina condizionate al luogo di servizio veterano



Quindi, abbiamo considerato 2 variabili.

Esempio: dataset `vets`

Anche qui



Come descrivereste queste distribuzioni condizionate?

In generale

Si ha una distribuzione doppia quando si esaminano congiuntamente due variabili sulle unità statistiche del nostro collettivo.

In generale

Si ha una distribuzione doppia quando si esaminano congiuntamente due variabili sulle unità statistiche del nostro collettivo.

Come nel caso di distribuzioni relative ad una singola variabile, si parlerà di distribuzioni doppie disaggregate quando si elencano le N coppie di modalità e di distribuzioni doppie di frequenze, quando le osservazioni sono aggregate per modalità o classi.

Esempio: dataset babies

Consideriamo la variabile doppia: (X, Y)=(durata della gravidanza, fumo). La distribuzione di frequenze assolute è data da

X	Y		totale
	$Y = N$	$Y = S$	
34	1	0	1
35	2	1	3
36	1	2	3
37	2	0	2
38	2	3	5
39	3	4	7
40	3	0	3
41	1	2	3
42	1	4	5
totale	16	16	32

Esempio: dataset `babies`

La distribuzione doppia appena vista "contiene" varie distribuzioni di frequenza. Infatti:

- Il "centro" della tabella (in questo caso le 9 righe e le 2 colonne centrali) mostra il numero di individui che presentano **una particolare modalità della coppia** (X, Y). Mostra cioè la distribuzione congiunta.

Esempio: dataset babies

La distribuzione doppia appena vista "contiene" varie distribuzioni di frequenza. Infatti:

- Il "centro" della tabella (in questo caso le 9 righe e le 2 colonne centrali) mostra il numero di individui che presentano **una particolare modalità della coppia** (X, Y). Mostra cioè la distribuzione congiunta.
- La 1^a colonna, per esempio, mostra la distribuzione della durata della gravidanza per le madri non fumatrici, cioè la distribuzione della variabile condizionata $(X|Y = N)$. Analogamente, la 2^a mostra la distribuzione della variabile condizionata $(X|Y = S)$. Quindi, le colonne riportano le distribuzioni della variabile condizionata $X|Y$.

Esempio: dataset babies

La distribuzione doppia appena vista "contiene" varie distribuzioni di frequenza. Infatti:

- Il "centro" della tabella (in questo caso le 9 righe e le 2 colonne centrali) mostra il numero di individui che presentano **una particolare modalità della coppia** (X, Y). Mostra cioè la distribuzione congiunta.
- La 1^a colonna, per esempio, mostra la distribuzione della durata della gravidanza per le madri non fumatrici, cioè la distribuzione della variabile condizionata $(X|Y = N)$. Analogamente, la 2^a mostra la distribuzione della variabile condizionata $(X|Y = S)$. Quindi, le colonne riportano le distribuzioni della variabile condizionata $X|Y$.
- La 1^a riga mostra, per tutte gravidanze durate 34 settimane, quante sono da riferirsi a madri non fumatrici e fumatrici, cioè la distribuzione della variabile condizionata $(Y|X = 34)$. Quindi, le righe riportano le distribuzioni della variabile condizionata $Y|X$.

Esempio: dataset `babies`

La distribuzione doppia appena vista "contiene" varie distribuzioni di frequenza. Infatti:

- Il "centro" della tabella (in questo caso le 9 righe e le 2 colonne centrali) mostra il numero di individui che presentano **una particolare modalità della coppia** (X, Y). Mostra cioè la distribuzione congiunta.
- La 1^a colonna, per esempio, mostra la distribuzione della durata della gravidanza per le madri non fumatrici, cioè la distribuzione della variabile condizionata $(X|Y = N)$. Analogamente, la 2^a mostra la distribuzione della variabile condizionata $(X|Y = S)$. Quindi, le colonne riportano le distribuzioni della variabile condizionata $X|Y$.
- La 1^a riga mostra, per tutte gravidanze durate 34 settimane, quante sono da riferirsi a madri non fumatrici e fumatrici, cioè la distribuzione della variabile condizionata $(Y|X = 34)$. Quindi, le righe riportano le distribuzioni della variabile condizionata $Y|X$.
- L'ultima colonna, mostra la distribuzione della durata della gravidanza a prescindere dalla condizione rispetto al fumo. L'ultima riga, invece, mostra la distribuzione della variabile `fumo` a prescindere dalla condizione rispetto alla durata della gravidanza. Sono cioè rappresentate le distribuzioni marginali.

Tabella a doppia entrata

Una distribuzione doppia di frequenze è normalmente chiamata tabella (di contigenza) a doppia entrata.
In generale, una tabella di contingenza (con due variabili) si presenta nella forma:

			Y			
X	y_1	$\cdots$	y_j	$\cdots$	y_t	totale
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	n_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{it}	n_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_s	n_{s1}	$\cdots$	n_{sj}	$\cdots$	n_{st}	n_{s0}
totale	n_{01}	$\cdots$	n_{0j}	$\cdots$	n_{0t}	N

Tabella a doppia entrata (cont)

Nella tabella

- X e Y sono le due variabili considerate

Tabella a doppia entrata (cont)

Nella tabella

- X e Y sono le due variabili considerate
- $\{x_1, \ldots, x_s\}$ sono le modalità di X

Tabella a doppia entrata (cont)

Nella tabella

- X e Y sono le due variabili considerate
- $\{x_1, \ldots, x_s\}$ sono le modalità di X
- $\{y_1, \ldots, y_t\}$ sono le modalità di Y

Tabella a doppia entrata (cont)

Nella tabella

- X e Y sono le due variabili considerate
- $\{x_1, \ldots, x_s\}$ sono le modalità di X
- $\{y_1, \ldots, y_t\}$ sono le modalità di Y
- n_{ij} è la *frequenza congiunta* assoluta per $X = x_i$ e $Y = y_j$

Tabella a doppia entrata (cont)

Nella tabella

- X e Y sono le due variabili considerate
- $\{x_1, \ldots, x_s\}$ sono le modalità di X
- $\{y_1, \ldots, y_t\}$ sono le modalità di Y
- n_{ij} è la *frequenza congiunta* assoluta per $X = x_i$ e $Y = y_j$
- n_{0j}, è il totale della colonna j, $n_{0j} = \sum_{i=1}^{s} n_{ij}$.

Tabella a doppia entrata (cont)

Nella tabella

- X e Y sono le due variabili considerate
- $\{x_1, \ldots, x_s\}$ sono le modalità di X
- $\{y_1, \ldots, y_t\}$ sono le modalità di Y
- n_{ij} è la *frequenza congiunta* assoluta per $X = x_i$ e $Y = y_j$
- n_{0j}, è il totale della colonna j, $n_{0j} = \sum_{i=1}^{s} n_{ij}$. Quindi è la frequenza assoluta marginale per la modalità y_j di Y.
- n_{i0}, è il totale della riga i: $n_{i0} = \sum_{j=1}^{t} n_{ij}$. Quindi è la frequenza assoluta marginale per la modalità x_i di X.

NB. La scelta di quale variabile (X o Y) mettere sulle righe/colonne e quale indice massimo (s o t) associare alle modalità delle variabili è libera.

Esempio: il disastro del Titanic

Tabella a doppia entrata per le variabili Classe di Viaggio del Passeggero (class) e Sopravvivenza (Survival)

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

Esempio: il disastro del Titanic

Tabella a doppia entrata per le variabili Classe di Viaggio del Passeggero (class) e Sopravvivenza (Survival)

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

- 118 passeggeri di seconda classe sopravvissero

Esempio: il disastro del Titanic

Tabella a doppia entrata per le variabili Classe di Viaggio del Passeggero (class) e Sopravvivenza (Survival)

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

- 118 passeggeri di seconda classe sopravvissero
- 178 passeggeri di terza classe sopravvissero

I passeggeri di terza classe avevano minori chance di sopravvivere?

Esempio: il disastro del Titanic

Alla domanda precedente si risponde meglio guardando alle frequenze relative (o, che poi è lo stesso, alle percentuali).

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201

Esempio: il disastro del Titanic

Alla domanda precedente si risponde meglio guardando alle frequenze relative (o, che poi è lo stesso, alle percentuali).

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201

- In seconda classe, sopravvisse il 41.4% (0.414 in termini di frequenza relativa) dei passeggeri

Esempio: il disastro del Titanic

Alla domanda precedente si risponde meglio guardando alle frequenze relative (o, che poi è lo stesso, alle percentuali).

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201

- In seconda classe, sopravvisse il 41.4% (0.414 in termini di frequenza relativa) dei passeggeri
- In terza, sopravvisse il 25.2% (0.252 in termini di frequenza relativa) dei passeggeri

Tabella a doppia entrata (cont)

Come l'esempio del Titanic dimostra, il calcolo delle frequenze relative in una tabella a doppia entrata è più delicato, perché la tabella contiene tante distribuzioni.

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201

Qui, abbiamo calcolato le frequenze percentuali della variabile condizionata `Sopravvivenza|Classe di Viaggio del Passeggero`. Si noti che, per ogni classe di viaggio del passeggero, le percentuali sommano a 100!

Rappresentazioni grafiche

Anche nel caso di variabili statistiche bivariate, le rappresentazioni grafiche aiutano molto (se ben fatte) ad interpretare i dati.

Rappresentazioni grafiche

Anche nel caso di variabili statistiche bivariate, le rappresentazioni grafiche aiutano molto (se ben fatte) ad interpretare i dati.

La rappresentazione dipende dalla natura delle variabili (qualitativi, quantitativi) e dalla forma in cui ci sono forniti i dati (aggregata/non aggregata).

Rappresentazioni grafiche

Anche nel caso di variabili statistiche bivariate, le rappresentazioni grafiche aiutano molto (se ben fatte) ad interpretare i dati.

La rappresentazione dipende dalla natura delle variabili (qualitativi, quantitativi) e dalla forma in cui ci sono forniti i dati (aggregata/non aggregata).

Abbiamo già visto alcune di queste rappresentazioni (verranno richiamate per dare loro un nome); altre sono nuove.

Rappresentazioni grafiche

Anche nel caso di variabili statistiche bivariate, le rappresentazioni grafiche aiutano molto (se ben fatte) ad interpretare i dati.

La rappresentazione dipende dalla natura delle variabili (qualitativi, quantitativi) e dalla forma in cui ci sono forniti i dati (aggregata/non aggregata).

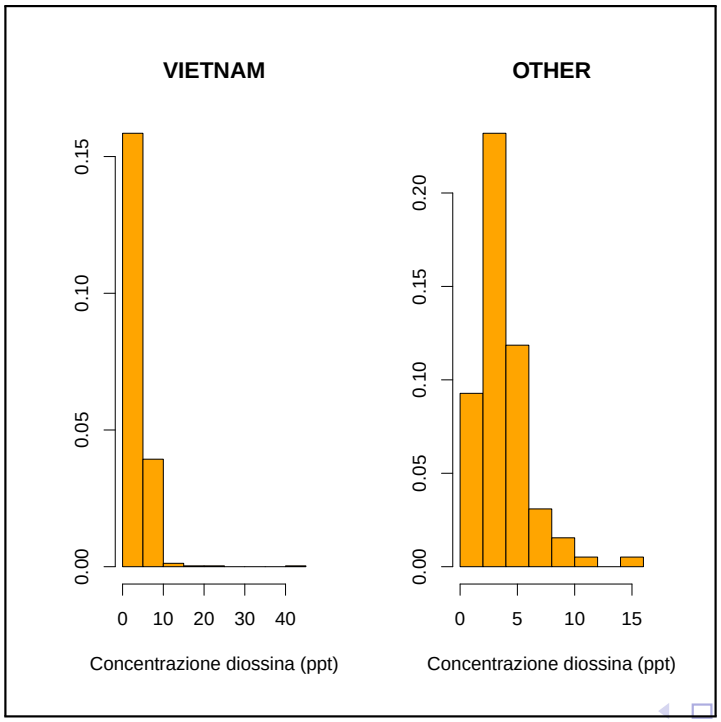
Abbiamo già visto alcune di queste rappresentazioni (verranno richiamate per dare loro un nome); altre sono nuove.

Per ogni grafico, si provi a fornire una lettura di quanto il grafico ci sta dicendo.

Istogrammi appaiati o affiancati (side-by-side histograms)

Esempio: dataset `vets`

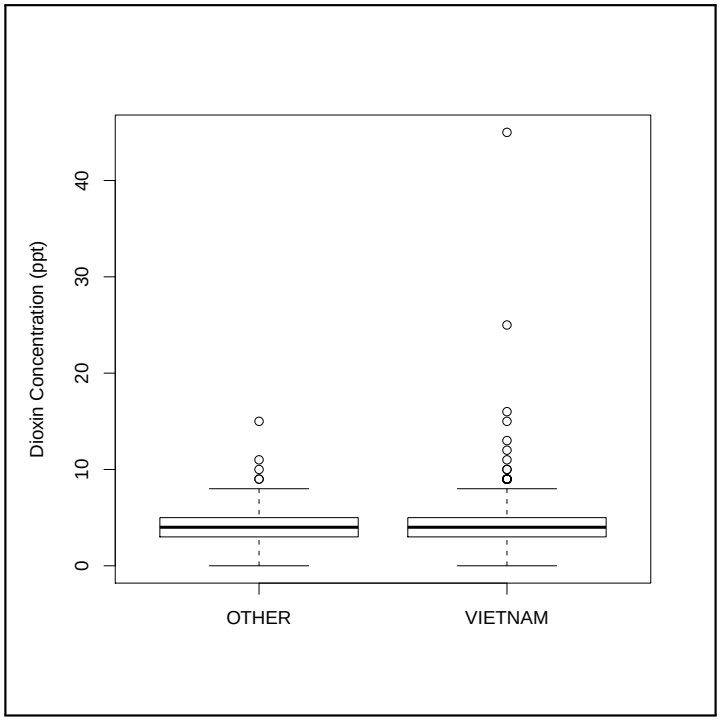
- $X \rightarrow$ Concentrazione di diossina (quantitativa continua)
- $Y \rightarrow$ Luogo di servizio (qualitativa)
- rappresentazione di $X|Y$.
- dati grezzi (distribuzione disaggregata) poi aggregati in tabella di frequenza in classi



Boxplot appaiati o affiancati (side-by-side boxplots)

Esempio: dataset `vets`

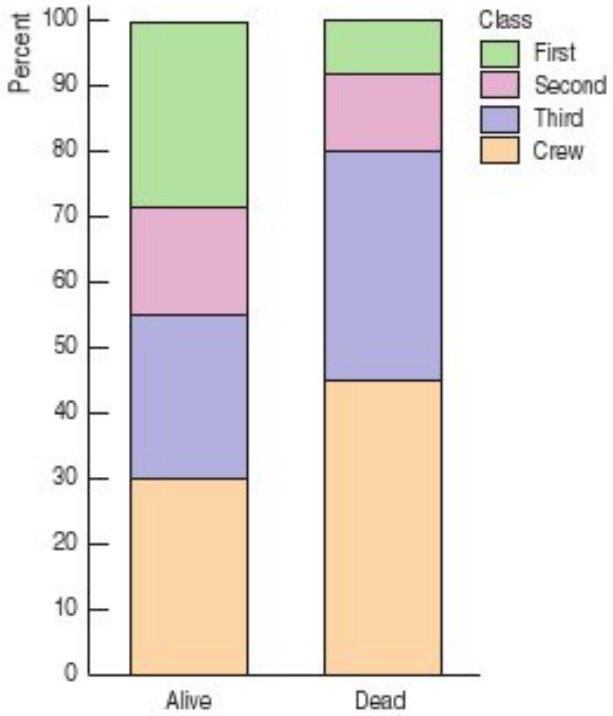
- $X \rightarrow$ Concentrazione di diossina (quantitativa continua)
- $Y \rightarrow$ Luogo di servizio (qualitativa)
- rappresentazione di $X|Y$.
- tendenzialmente dati grezzi (distribuzione disaggregata)



Diagrammi a barre condizionati (component bar charts)

Esempio: Titanic

- $X \rightarrow$ Classe di Viaggio del Passeggero (qualitativa)
- $Y \rightarrow$ Sopravvivenza (qualitativa)
- rappresentazione di $X|Y$
- dati aggregati in tabella di contingenza



Esercizio: caccia grossa

Esempio: chi tra maschie e femmine è alla caccia di un partner tra i compagni di studio?

		in cerca di un partner		
		No	Si	Total
sesso	F	86	51	137
	M	52	18	70
	Totale	138	69	207

Esercizio: alla caccia di un partner (cont)

Cosa rappresentano i due grafici?

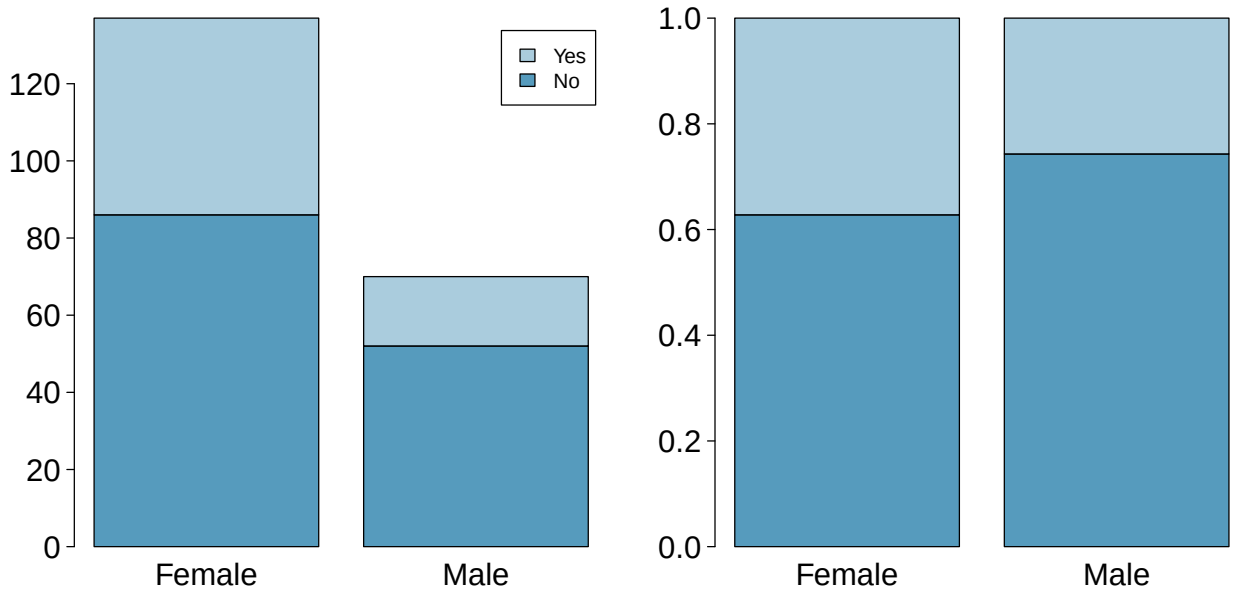


Diagramma a barre condizionate per la variabile “in cerca di partner” (qualitativa) condizionata alla variabile “sesso” (qualitativa); a destra usando le frequenze assolute; a sinistra quelle relative.

Indice

Relazioni tra variabili

A ben vedere, il commento più naturale che si può fare leggendo i grafici o le tabelle precedenti era del tipo: il comportamento di questa variabile cambia al cambiare dell'altra, oppure, questa variabile è influenzata da quest'altra.

Relazioni tra variabili

A ben vedere, il commento più naturale che si può fare leggendo i grafici o le tabelle precedenti era del tipo: il comportamento di questa variabile cambia al cambiare dell'altra, oppure, questa variabile è influenzata da quest'altra.

Quindi, quando guardiamo a più di una variabile, viene naturale esplorare se esiste una qualche associazione tra le stesse.

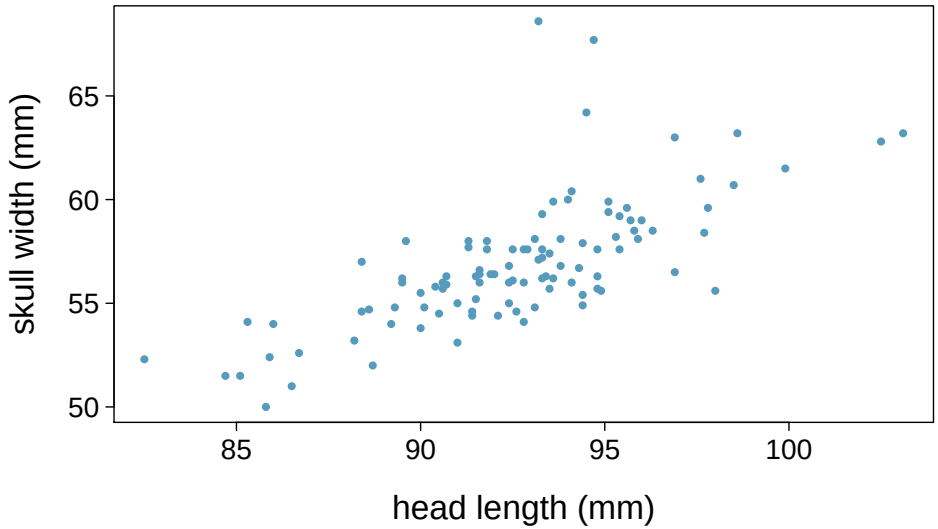
Associazione e indipendenza

- Quando due variabili mostrano qualche forma di relazione tra loro, si parla di associazione.
- Quando due variabili non mostrano alcuna forma di connessione tra loro, si parla di indipedenza.

Variabili associate sono anche dette dipendenti e viceversa.

Esercizio

Sulla base del diagramma a dispersione sulla destra, quale delle seguenti affermazioni è corretta?



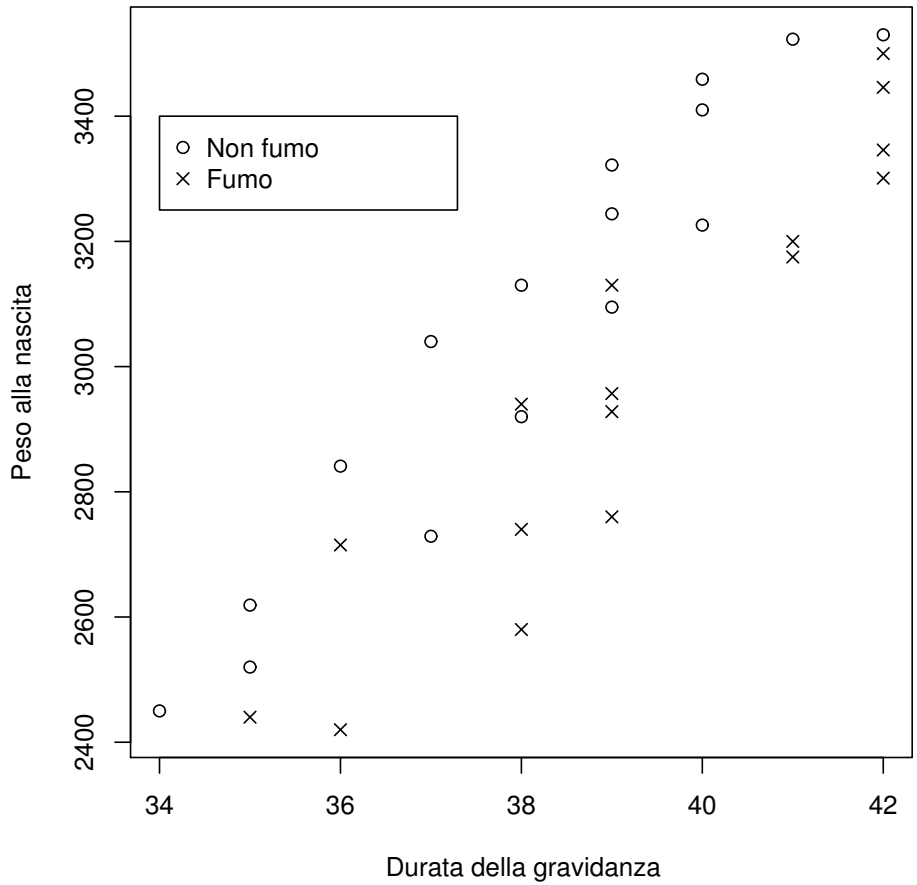
(a) Non c'è relazione tra lunghezza della testa (head length) e ampiezza del cranio (skull width) degli opossum.

(b) Head length e skull width sono associati (positivamente).

(c) Head length e skull width sono associati (negativamente).

Esercizio

Sulla base del diagramma a dispersione, sembra esserci associazione tra il peso alla nascita e la durata della gravidanza da madri fumatrici e non fumatrici?



Riprendiamo il Titanic

Riprendiamo la tabella che abbiamo analizzato in precedenza

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201

I passeggeri di terza classe avevano maggiori chance di sopravvivere?

Riprendiamo il Titanic (cont)

Per rispondere, abbiamo guardato alla variabile condizionata Sopravvivenza|Classe di Viaggio Passeggero. Sembrerebbe sensato affermare che la sopravvivenza dipende dalla classe di viaggio.

X (la sopravvivenza) *dipende* da Y (la classe in cui viaggiava il passeggero) poichè le distribuzioni di X condizionate ad Y sono diverse nel senso che hanno *frequenze relative diverse*

Indipendenza in distribuzione

Date due variabili X e Y, si dice che la X è *indipendente in distribuzione* da Y se, qualunque sia la modalità con cui si manifesta la Y, la distribuzione relativa condizionata di X rimane sempre la stessa (le modalità assunte dalla Y non modificano la distribuzione di X).
cioè le frequenze relative delle distribuzioni condizionate della X rispetto alla Y devono essere tutte fra loro uguali e uguali alla distribuzione marginale relativa della X.
Formalmente X è *indipendente in distribuzione* da Y se, per qualsivoglia i,

$$\frac{n_{i1}}{n_{01}} = \frac{n_{i2}}{n_{02}} = \cdots = \frac{n_{ij}}{n_{0j}} = \cdots = \frac{n_{it}}{n_{0t}}, i = 1, \ldots, s \qquad (1)$$

Indipendenza in distribuzione

Date due variabili X e Y, si dice che la X è *indipendente in distribuzione* da Y se, qualunque sia la modalità con cui si manifesta la Y, la distribuzione relativa condizionata di X rimane sempre la stessa (le modalità assunte dalla Y non modificano la distribuzione di X).
cioè le frequenze relative delle distribuzioni condizionate della X rispetto alla Y devono essere tutte fra loro uguali e uguali alla distribuzione marginale relativa della X.
Formalmente X è *indipendente in distribuzione* da Y se, per qualsivoglia i,

$$\frac{n_{i1}}{n_{01}} = \frac{n_{i2}}{n_{02}} = \cdots = \frac{n_{ij}}{n_{0j}} = \cdots = \frac{n_{it}}{n_{0t}}, i = 1, \ldots, s \qquad (1)$$

Se la (1) non è vera diremo che X *dipende in distribuzione* da Y.

Indipendente in distribuzione (cont)

Dalla (1) discende immediatamente che se le distribuzioni condizionate di X dato Y sono uguali tra di loro, allora sono anche uguali alla distribuzione marginale di X.

Indipendente in distribuzione (cont)

Dalla (1) discende immediatamente che se le distribuzioni condizionate di X dato Y sono uguali tra di loro, allora sono anche uguali alla distribuzione marginale di X.

L'uguaglianza, al solito, deve essere intesa nel senso delle frequenze relative.

Indipendenza in distribuzione (cont)

Per dimostrare la proposizione ci basta far vedere che la (1) implica

$$\frac{n_{i0}}{N} = \frac{n_{i1}}{n_{01}}, \quad i = 1, \ldots, s.$$

Ora, dalla (1) segue che $n_{ij} = (n_{i1} n_{0j})/n_{01}$.
Quindi,

$$\begin{aligned} \frac{n_{i0}}{N} &= \frac{\sum_{j=1}^{t} n_{ij}}{N} = \frac{\sum_{j=1}^{t} n_{i1} n_{0j}}{N n_{01}} = \\ &= \frac{n_{i1} \sum_{j=1}^{t} n_{0j}}{N n_{01}} = \frac{N n_{i1}}{N n_{01}} = \frac{n_{i1}}{n_{01}}. \end{aligned}$$

Esempio: indipendenza in distribuzione

Carattere X	Carattere Y				Totale
	$y1$	$y2$	$y3$	$y4$	
$x1$	5	30	15	10	60
$x2$	7	42	21	14	84
$x3$	3	18	9	6	36
$x4$	2	12	6	4	24
Totale	17	102	51	34	204

Esempio: indipendenza in distribuzione

Carattere X	Carattere Y				Totale
	y1	y2	y3	y4	
x1	0,083	0,5	0,25	0,167	1
x2	0,083	0,5	0,25	0,167	1
x3	0,083	0,5	0,25	0,167	1
x4	0,083	0,5	0,25	0,167	1
Totale	0,083	0,5	0,25	0,167	1

Esempio: indipendenza in distribuzione

Carattere X	Carattere Y				Totale
	y1	y2	y3	y4	
x1	0,2941176	0,2941176	0,2941176	0,2941176	0,2941176
x2	0,411764	0,411764	0,411764	0,411764	0,411764
x3	0,176470	0,176470	0,176470	0,176470	0,176470
x4	0,1176471	0,1176471	0,1176471	0,1176471	0,1176471
Totale	1	1	1	1	1

Frequenze attese

Poniamo

$$\hat{n}_{ij} = \frac{n_{i0} n_{0j}}{N}.$$

Se esiste indipendenza tra le due variabili, $n_{ij} = \hat{n}_{ij}$ per qualsivoglia i e per qualsivoglia j, ovvero, le $\hat{n}_{ij}$ sono le frequenze che ci aspettiamo di trovare quando esiste indipendenza.

Frequenze attese

Poniamo

$$\hat{n}_{ij} = \frac{n_{i0} n_{0j}}{N}.$$

Se esiste indipendenza tra le due variabili, $n_{ij} = \hat{n}_{ij}$ per qualsivoglia i e per qualsivoglia j, ovvero, le $\hat{n}_{ij}$ sono le frequenze che ci aspettiamo di trovare quando esiste indipendenza.

Per questo motivo, le $\hat{n}_{ij}$ sono chiamate le *frequenze attese* (sotto l'ipotesi di indipendenza in distribuzione).

Frequenze attese

Poniamo

$$\hat{n}_{ij} = \frac{n_{i0} n_{0j}}{N}.$$

Se esiste indipendenza tra le due variabili, $n_{ij} = \hat{n}_{ij}$ per qualsivoglia i e per qualsivoglia j, ovvero, le $\hat{n}_{ij}$ sono le frequenze che ci aspettiamo di trovare quando esiste indipendenza.

Per questo motivo, le $\hat{n}_{ij}$ sono chiamate le *frequenze attese* (sotto l'ipotesi di indipendenza in distribuzione).

Come è ovvio, le frequenze attese $\hat{n}_{ij}$ ci mostrano anche come le frequenze marginali si comporterebbero nel caso di indipendenza in distribuzione.

Esempio: indipendenza in distribuzione

Carattere X	Carattere Y				Totale
	y1	y2	y3	y4	
x1	5= 60*17/204	30	15	10	60
x2	7= 17*84/204	42	21	14	84
x3	3= 17*36/204	18	9	6	36
x4	2= 17*24/204	12	6	4	24
Totale	17	102	51	34	204

χ^2 (chi-quadro)

L'indice di uso più comune per *misurare* la dipendenza in distribuzione si basa sul confronto tra frequenze attese e frequenze osservate. Si tratta del cosidetto χ^2 di Pearson che è definito come

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}.$$

χ^2 è sempre maggiore o uguale a zero ed è uguale a 0 in caso di indipendenza ($n_{ij} = \hat{n}_{ij}$, per ogni i e per ogni j) e cresce man mano che le frequenze osservate si allontano da quelle attese.

χ^2 (cont)

Si può dimostrare che $\chi^2 \leq N \cdot \min(s-1, t-1)$.

χ^2 (cont)

Si può dimostrare che $\chi^2 \leq N \cdot \min(s-1, t-1)$.

Il massimo è raggiunto quando la distribuzione doppia assume una struttura particolare, quella di una tabella di *dipendenza perfetta*.

χ^2 (cont)

Si può dimostrare che $\chi^2 \leq N \cdot \min(s-1, t-1)$.

Il massimo è raggiunto quando la distribuzione doppia assume una struttura particolare, quella di una tabella di *dipendenza perfetta*.

Si chiama tabella di *dipendenza perfetta* la tabella tale che ad ogni modalità della variabile X corrisponde una sola modalità della variabile Y.

Tabella di dipendenza perfetta

Carattere X	Carattere Y				Totale
	y1	y2	y3	y4	
x1	45	0	0	0	45
x2	0	20	0	0	20
x3	0	0	0	92	92
x4	0	0	37	0	37
Totale	45	20	37	92	194

X^2 (cont)

Quindi, si può costruire un indice *normalizzato*, l'indice V di Cramer.

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(s-1, t-1)}} = \sqrt{\tilde{\chi}^2}$$

che assumerà valori tra 0 e 1: $0 \leq V \leq 1$.
(dove 0=indipendenza perfetta in distribuzione; 1=massima dipendenza)

Il caso del Titanic

Per il Titanic, le frequenza attese e X^2 valgono

Sopravvivenza	Passeggero I	II	III	equipaggio	
SI	104,9864	92,06497	228,0627	285,8860	711
NO	220,0136	192,93503	477,9373	599,1140	1490
totale	325	285	706	885	2201

$$X^2 = 133{,}05$$

$$\begin{aligned} X^2 &= \frac{(203-123,2)^2}{123,2} + \frac{(118-108,1)^2}{108,1} + \frac{(178-267,7)^2}{267,7} \\ &\quad + \frac{(122-201,8)^2}{201,8} + \frac{(167-176,9)^2}{176,9} + \frac{(528-438,3)^2}{438,3} \\ &= 133,05 \\ V &= \sqrt{\frac{133,05}{1316 \cdot \min(1,2)}} = 0,318 \end{aligned}$$

Quindi, c'è una qualche dipendenza (anche se non molto marcata) tra le due variabili.

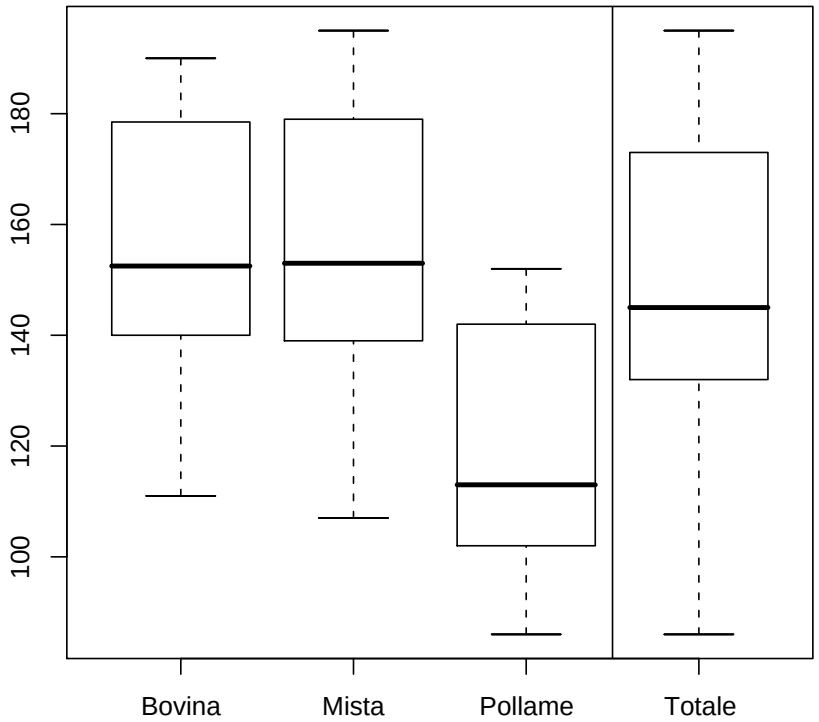
Esempio: hot dog

	Tipo di carne			
Calorie	Bovina	Mista	Pollame	Totale
86	0	0	1	1
87	0	0	1	1
94	0	0	1	1
99	0	0	1	1
102	0	0	2	2
106	0	0	1	1
107	0	1	1	2
111	1	0	0	1
113	0	0	1	1
129	0	0	1	1
131	1	0	0	1
132	1	0	1	2
135	1	1	1	3
136	0	1	0	1
138	0	1	0	1
139	1	1	0	2
140	0	1	0	1
141	1	0	0	1
142	0	0	1	1
143	0	0	1	1
144	0	0	1	1
146	0	1	1	2
147	0	1	0	1
148	1	0	0	1
149	2	0	0	2
152	1	0	1	2
153	1	1	0	2
157	1	0	0	1
158	1	0	0	1

Indipendenza/Dipendenza in media

Se una delle due variabili è quantitativa, ci possiamo spingere un po' oltre.

Esempio: hot dog



Esempio: hot dog

I boxplot appena visti ci dicono che non solo le tre distribuzioni delle "Calorie" condizionate al "Tipo di carne" $(Y|X)$ sono diverse ma suggeriscono anche, ad esempio, che le 3 medie sono diverse.

- Media di $X|Y =$bovina $= 156.85$
- Media di $X|Y =$mista $= 158.7059$
- Media di $X|Y =$pollame $= 118.7647$

Tra le due variabili X e Y esiste quindi dipendenza in media.

Indipendenza/Dipendenza in media

Una variabile, necessariamente numerica, X è *indipendente in media* da un'altra variabile Y, qualitativa o quantitativa, se le medie delle distribuzioni di X condizionate alle varie modalità della Y sono tutte uguali tra di loro.
In maniera analoga possiamo definire altri concetti di dipendenza/indipendenza (ad es. indipendenza in mediana, indipendenza in varianza,...).

- l'indipendenza in distribuzione implica l'indipendenza in media
- ma l'indipendenza in media non è sufficiente per concludere che esiste anche indipendenza in distribuzione.

Indice

Dipendenza in media

Quando esiste dipendenza in media, cerchiamo di stabilire se e in che misura le medie delle distribuzioni condizionate di una **variabile quantitativa**, diciamo X, variano al variare delle modalità dell'altra variabile, diciamo Y (variabile qualitativa).

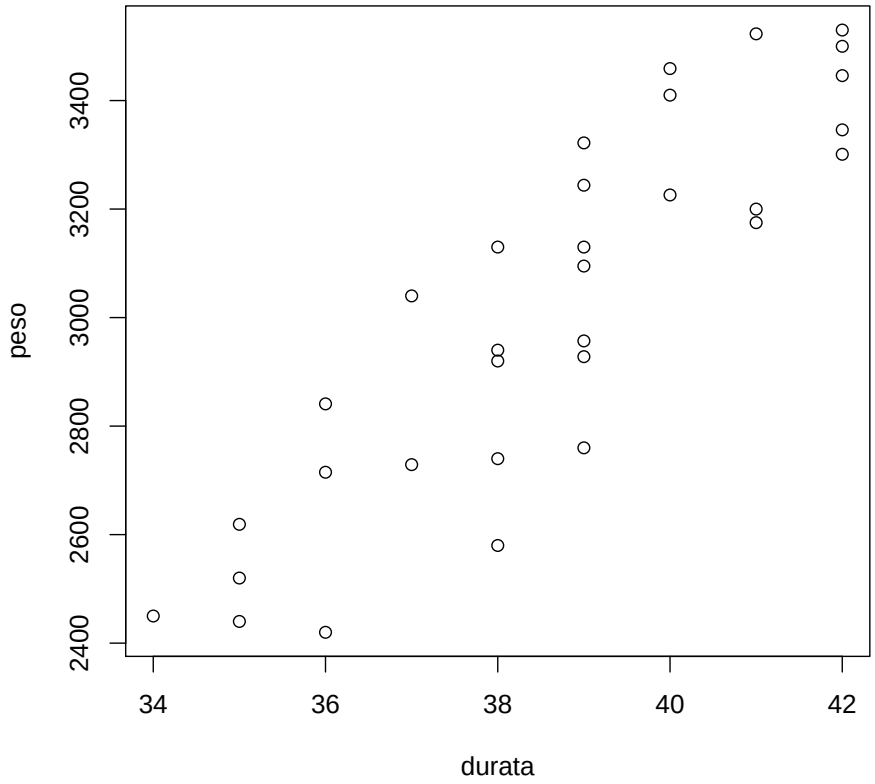
Dipendenza in media

Quando esiste dipendenza in media, cerchiamo di stabilire se e in che misura le medie delle distribuzioni condizionate di una **variabile quantitativa**, diciamo X, variano al variare delle modalità dell'altra variabile, diciamo Y (variabile qualitativa).

Se sia X che Y sono **entrambe variabili quantitative**, per studiare la dipendenza possiamo fare qualcosa di più.

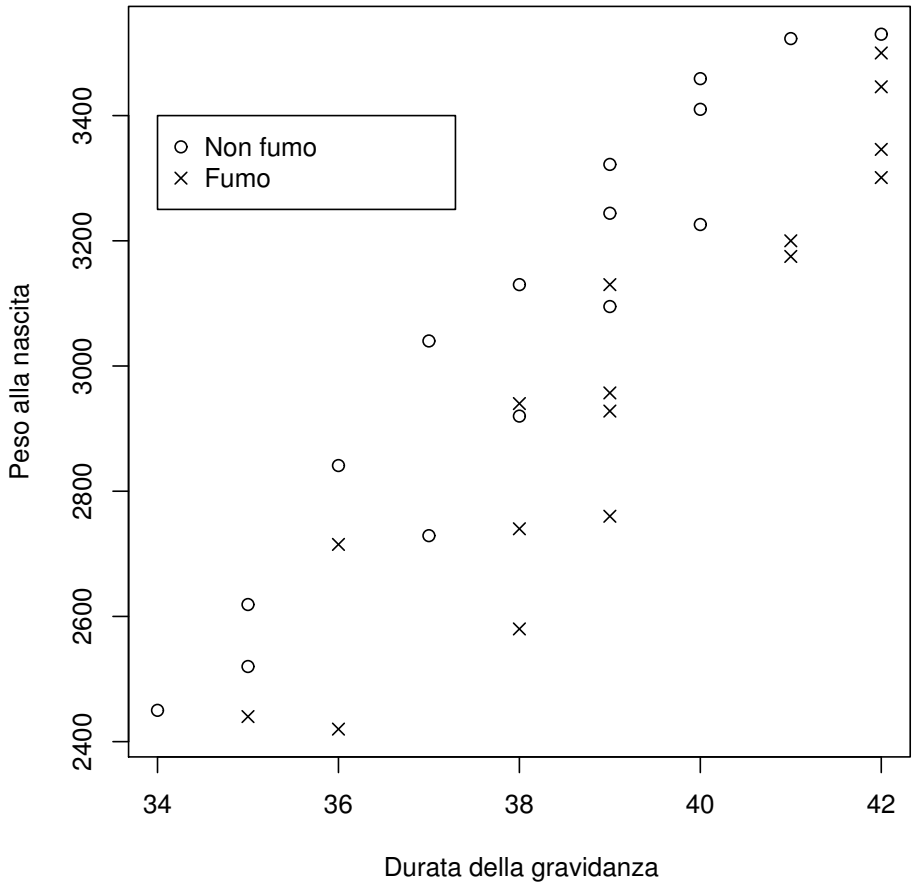
Descrivere la dipendenza

Sulla base del diagramma a dispersione, sembra esserci dipendenza in media tra il peso alla nascita e la durata della gravidanza da madri fumatrici e non fumatrici?



Descrivere la dipendenza (cont)

La dipendenza tra il peso alla nascita e la durata della gravidanza cambia tra madri fumatrici e non fumatrici?



Diagrammi di dispersione

Il diagramma di dispersione è la rappresentazione delle coppie

$$(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$$

ossia della distribuzione disaggregata della variabile doppia (X, Y).

Diagrammi di dispersione

Il diagramma di dispersione è la rappresentazione delle coppie

$$(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$$

ossia della distribuzione disaggregata della variabile doppia (X, Y).

Si dice che tra X e Y c'è associazione positiva quando essi tendono a crescere insieme.

Diagrammi di dispersione

Il diagramma di dispersione è la rappresentazione delle coppie

$$(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$$

ossia della distribuzione disaggregata della variabile doppia (X, Y).

Si dice che tra X e Y c'è associazione positiva quando essi tendono a crescere insieme.

Si dice che tra X e Y c'è associazione negativa quando essi tendono a decrescere insieme.

Diagrammi di dispersione

Il diagramma di dispersione è la rappresentazione delle coppie

$$(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$$

ossia della distribuzione disaggregata della variabile doppia (X, Y).

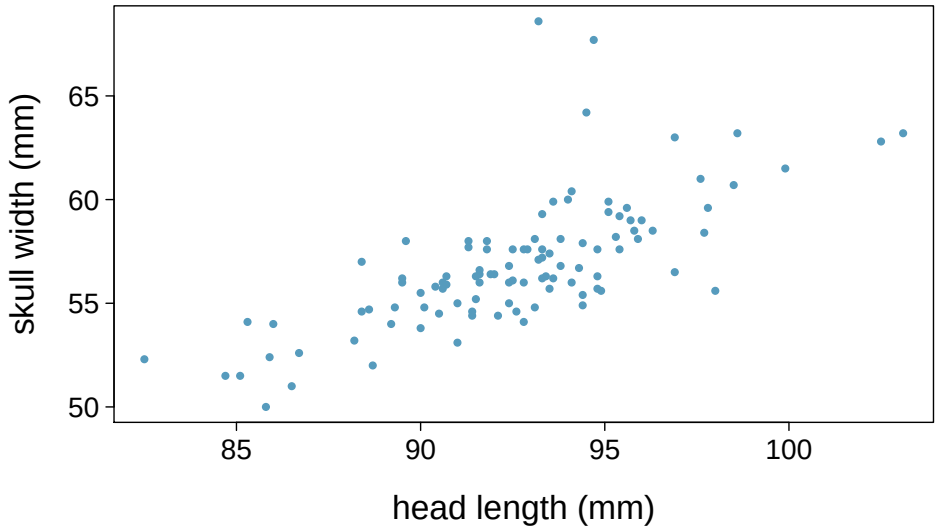
Si dice che tra X e Y c'è associazione positiva quando essi tendono a crescere insieme.

Si dice che tra X e Y c'è associazione negativa quando essi tendono a decrescere insieme.

È sempre utile cercare di stabilire, sulla base della sola osservazione del diagramma di dispersione, se esiste associazione positiva o negativa tra due caratteri

Esercizio

Sulla base del diagramma di dispersione sulla destra, quale delle seguenti affermazioni è corretta?



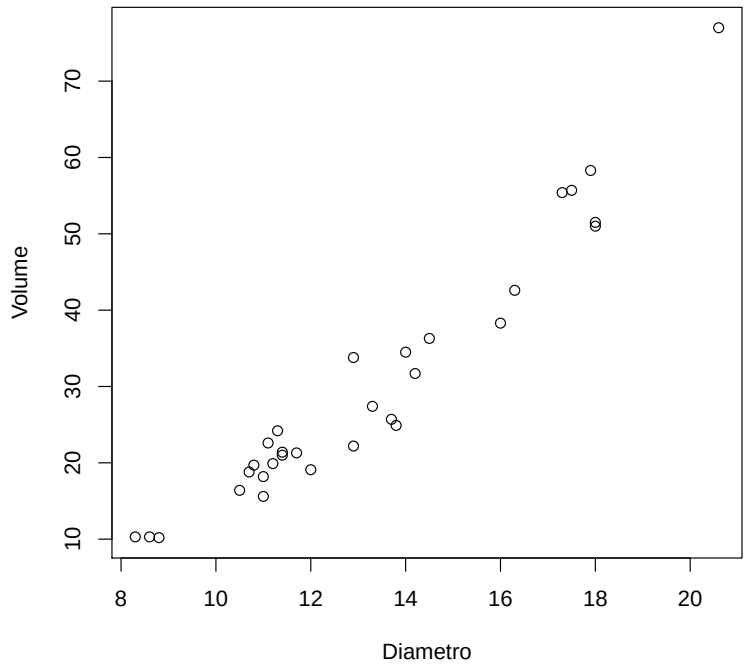
(a) Non c'è relazione tra lunghezza della testa (head length) e ampiezza del cranio (skull width) degli opossum.

(b) Head length e skull width sono associati (positivamente).

(c) Head length e skull width sono associati (negativamente).

Esempio: dataset `trees`

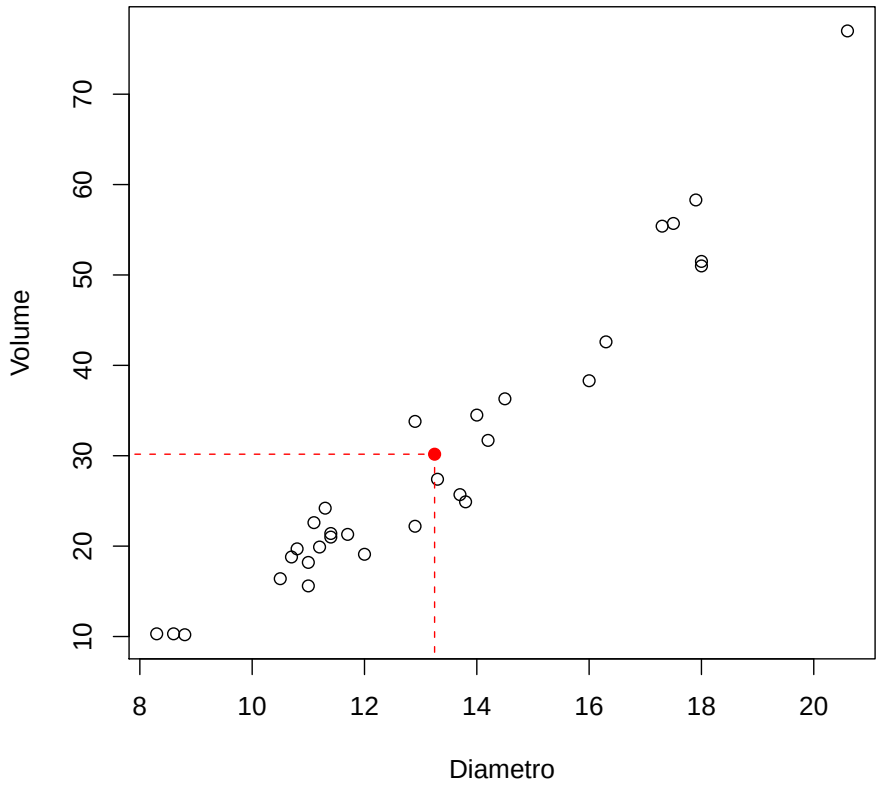
- $X \rightarrow$ Diametro del tronco (quantitativa continua)
- $Y \rightarrow$ Volume del tronco (quantitativa continua)
- rappresentazione di (X, Y)
- dati grezzi



Genesi di una misura di associazione

Esempio: dataset `trees`

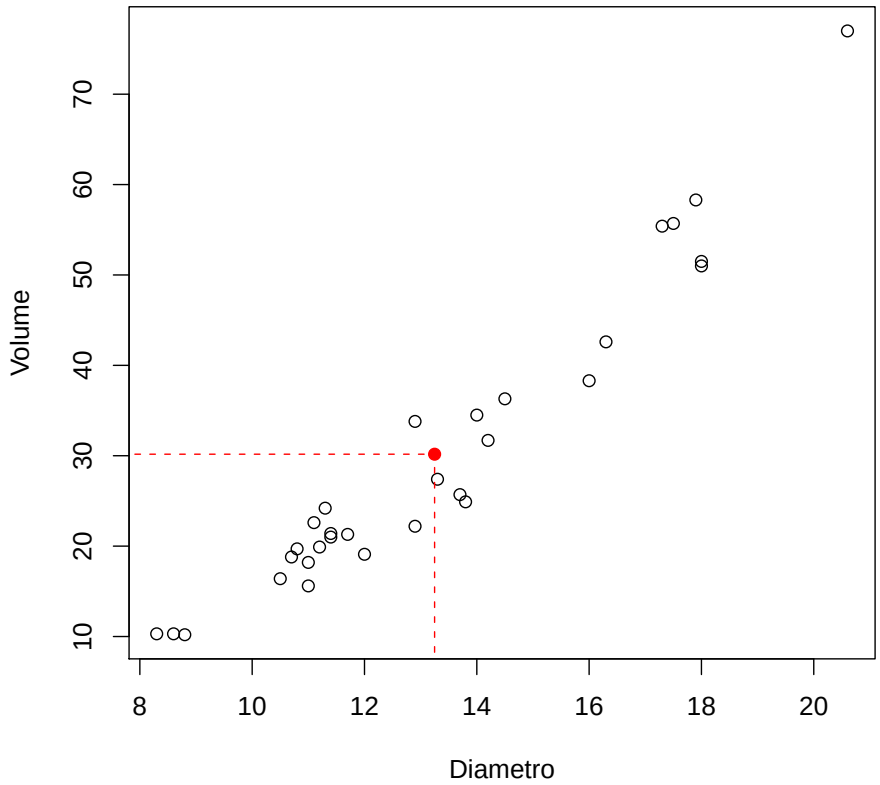
1. Il punto rosso è il punto di coordinate $(\bar{x}, \bar{y})$.



Genesi di una misura di associazione

Esempio: dataset `trees`

1. Il punto rosso è il punto di coordinate $(\bar{x}, \bar{y})$.
2. Valori maggiori della media di X corrispondono a valori maggiori della media per Y.
3. Valori inferiori alla media di X corrispondono a valori inferiori alla media per Y.



La covarianza

Questo suggerisce di partire dalla seguente quantità
$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$
dove (x_i, y_i), $i = 1, \ldots, N$, sono i dati disponibili su due variabili numeriche, mentre $\overline{x}$ e $\overline{y}$ indicano le due medie aritmetiche.

La covarianza

Questo suggerisce di partire dalla seguente quantità
$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$
dove (x_i, y_i), $i = 1, \ldots, N$, sono i dati disponibili su due variabili numeriche, mentre $\overline{x}$ e $\overline{y}$ indicano le due medie aritmetiche.

σ_{XY} è detta covarianza. Il suo numeratore è detto codevianza, indicata con C_{XY}.

La covarianza (cont)

1. In presenza di una qualche forma di relazione, più è forte la relazione tra le due variabili più ci aspettiamo che la covarianza diventi *grande in valore assoluto*. Infatti, più è forte la relazione, più grande dovrebbe essere il numero di addendi concordi nella somma. Inoltre, un certo numero di addendi sarà il prodotto di scarti dalle media grandi in valore assoluto.

2. In assenza di una qualche forma di relazione tra le due variabili, viceversa, gli addendi saranno in parte positivi ed in parte negativi. Quindi in questi casi ci aspettiamo che la covarianza *risulti nulla o comunque vicina allo zero*.

Calcolo della covarianza

Per il calcolo della covarianza è conveniente utilizzare la seguente relazione

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\,\overline{y}.$$

ovvero

$$(\text{covarianza}) = \left(\begin{array}{c}\text{media dei}\\ \text{prodotti}\end{array}\right) - \left(\begin{array}{c}\text{prodotto delle}\\ \text{medie}\end{array}\right).$$

Dataset `trees` (cont)

- $\sum_{i=1}^{N} x_i y_i = 13887,86$
- $\overline{x} = 13,24839$
- $\overline{y} = 30,17097$
- $\sigma_{XY} = 13887,86/31 - 13,24839 \times 30,17097 = 48,27871$

Grande quanto?

L'esempio sul dataset Ciliegi illustra uno dei problemi connessi con l'utilizzo della covarianza.

Grande quanto?

L'esempio sul dataset Ciliegi illustra uno dei problemi connessi con l'utilizzo della covarianza.

L'interpretazione del segno non pone nessuno problema. La covarianza indica una associazione tendenzialmente positiva tra diametro del tronco e volume del legno.

Grande quanto?

L'esempio sul dataset Ciliegi illustra uno dei problemi connessi con l'utilizzo della covarianza.

L'interpretazione del segno non pone nessuno problema. La covarianza indica una associazione tendenzialmente positiva tra diametro del tronco e volume del legno.

Ma quanto "forte" è questa dipendenza?

Grande quanto?

L'esempio sul dataset Ciliegi illustra uno dei problemi connessi con l'utilizzo della covarianza.

L'interpretazione del segno non pone nessuno problema. La covarianza indica una associazione tendenzialmente positiva tra diametro del tronco e volume del legno.

Ma quanto "forte" è questa dipendenza?

Per rispondere alla domanda avremmo bisogno di conoscere un estremo superiore, possibilmente con una chiara interpretazione, per il valore assoluto della covarianza.

Grande quanto? (cont)

Si dimostra che

$$-\sigma_Y \sigma_X \leq \sigma_{XY} \leq \sigma_Y \sigma_X.$$

Il coefficiente di correlazione (lineare)

I limiti per la covarianza suggeriscono che per affermare se la covarianza è "piccola" o è "grande" dobbiamo confrontarla con il prodotto delle deviazioni standard delle due variabili X e Y.

Il coefficiente di correlazione (lineare)

I limiti per la covarianza suggeriscono che per affermare se la covarianza è "piccola" o è "grande" dobbiamo confrontarla con il prodotto delle deviazioni standard delle due variabili X e Y.

In altre parole, dobbiamo costruire l'indice **normalizzato** (perchè compreso tra un minimo ed un massimo noti), chiamato coefficiente di correlazione (lineare)

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Il coefficiente di correlazione è spesso indicato con la lettera greca ρ.

Interpretazione di r

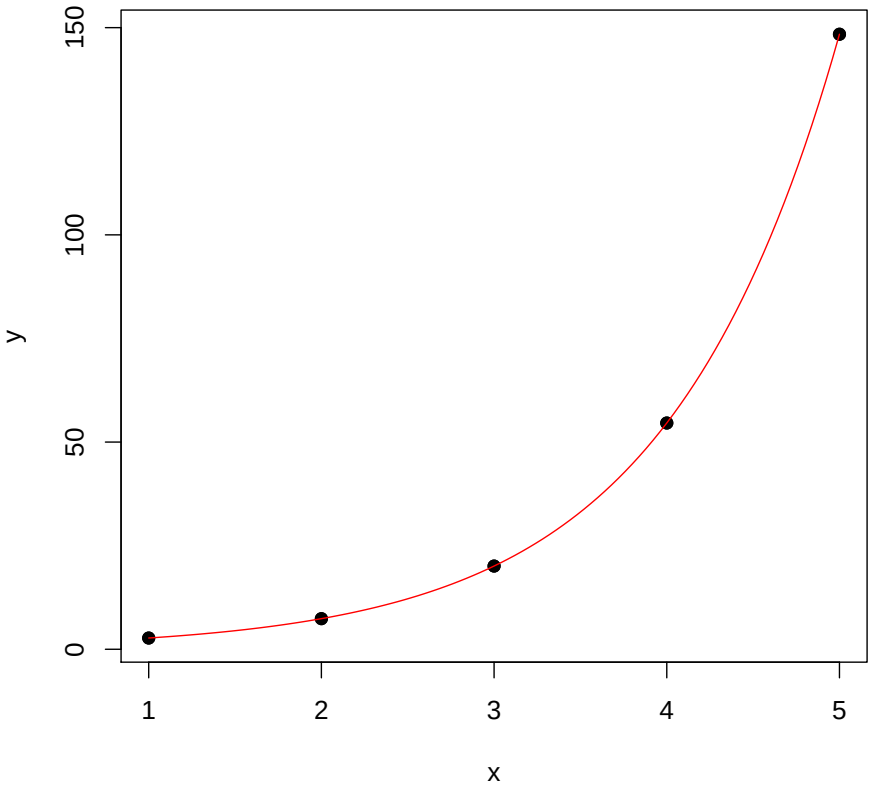
$$-1 \leq r \leq +1$$

- $r = -1$ perfetta dipendenza lineare negativa tra X e Y
- $r < 0$ associazione negativa tra X e Y
- $r = 0$ assenza di relazione monotona tra X e Y
- $r > 0$ associazione positiva tra X e Y
- $r = +1$ perfetta dipendenza lineare positiva tra X e Y

Ancora su $|r| = 1$

x	1	2	3	4	5
y	2.72	7.39	20.09	54.60	148.41

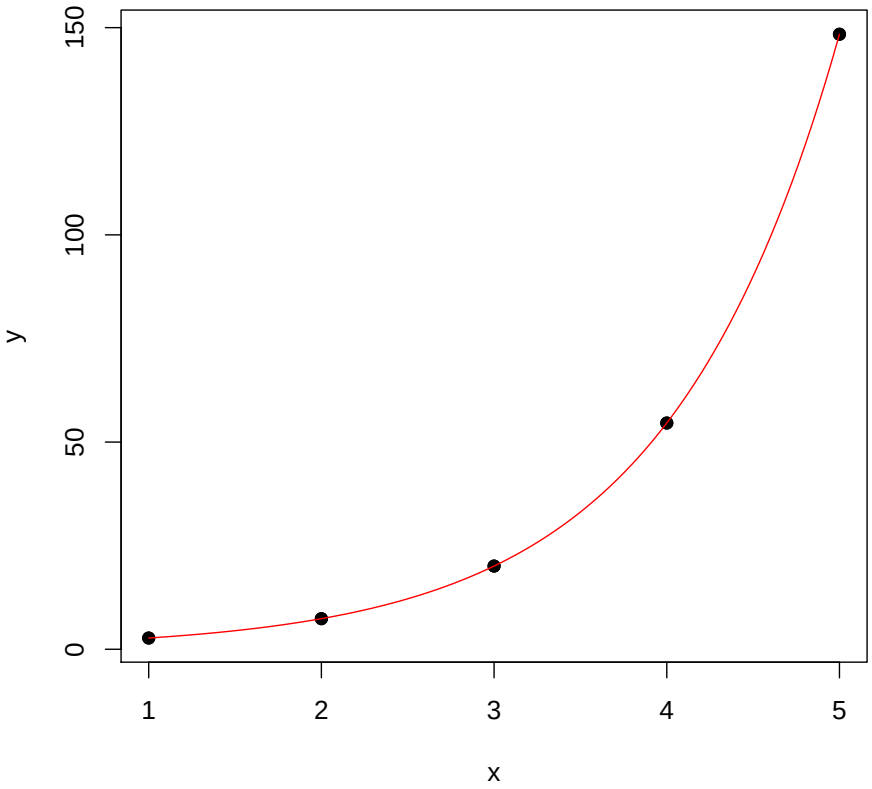
- I dati si dispongono sulla curva $Y = e^X$, quindi sono legati da una perfetta relazione monotona crescente, ma *non lineare*.



Ancora su $|r| = 1$

x	1	2	3	4	5
y	2.72	7.39	20.09	54.60	148.41

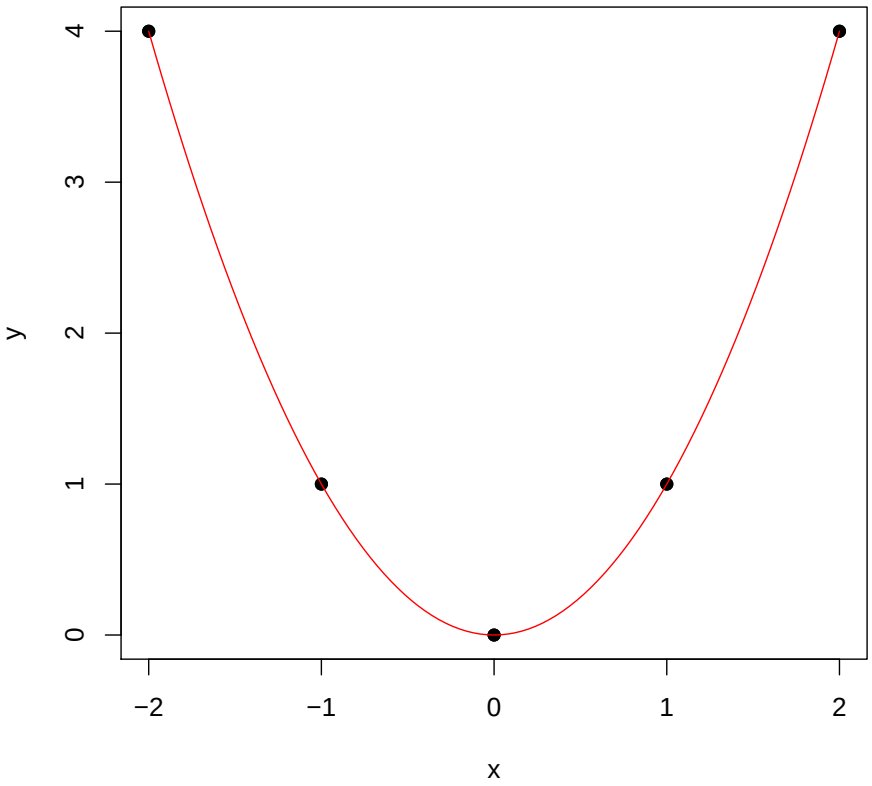
- I dati si dispongono sulla curva $Y = e^X$, quindi sono legati da una perfetta relazione monotona crescente, ma *non lineare*.
- Si ha $r = 0.886275$



Ancora su $r = 0$

x	-2	-1	0	1	2
y	4	1	0	1	4

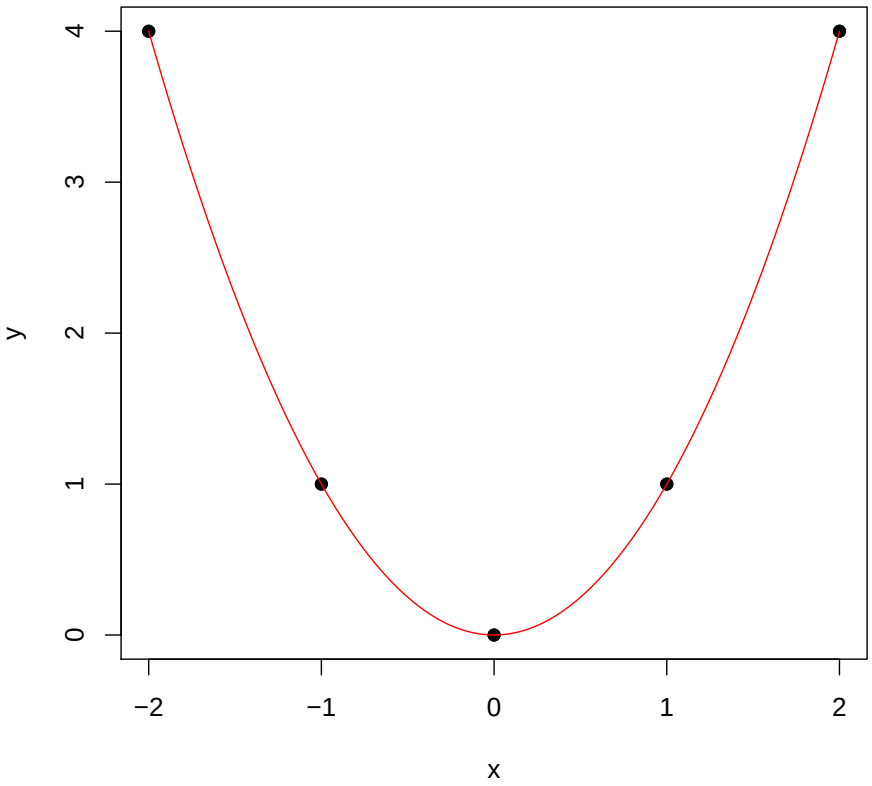
- I dati si dispongono sulla curva $Y = X^2$, quindi sono legati da una perfetta relazione, ma *non monotona*.



Ancora su $r = 0$

x	-2	-1	0	1	2
y	4	1	0	1	4

- I dati si dispongono sulla curva $Y = X^2$, quindi sono legati da una perfetta relazione, ma *non monotona*.
- Si ha $r = 0$.



Morale

- Un valore di r inferiore in valore assoluto a 1 non implica necessariamente assenza di un legame perfetto tra le variabili, ma assenza di un legame lineare perfetto.

- Un valore di r uguale a zero non implica necessariamente assenza di relazione tra le variabili, ma assenza di relazione lineare (più in generale, **monotona**).