UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

Statistica

Relazioni tra variabili

Francesco Pauli

A.A. 2017/2018

Premessa: dati

I dati sono aggiornati con le risposte al questionario di quest'anno (106 risposte), da cui le differenze rispetto ai lucidi del capitolo precedente.

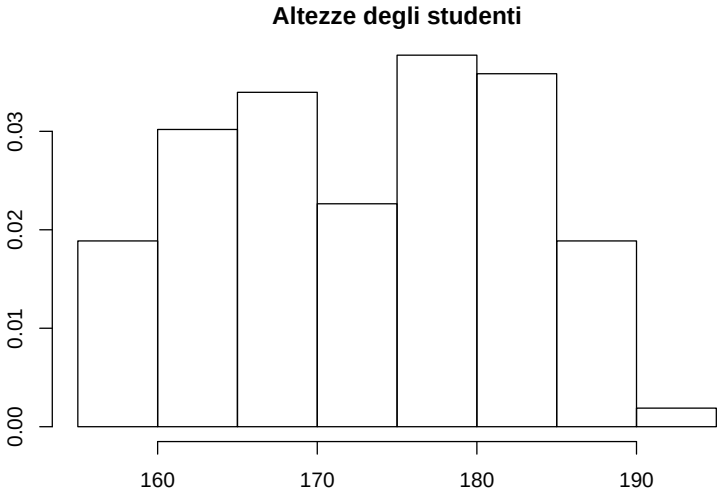
Inoltre, per alcuni esempi si utilizzano, oltre alle risposte date qal questionario queat'anno, quelle dell'anno precedente e quelle raccolte tra gli studenti di Scienze politiche e di Statistica gli anni scorsi (543 risposte in totale).

Solo una variabile...

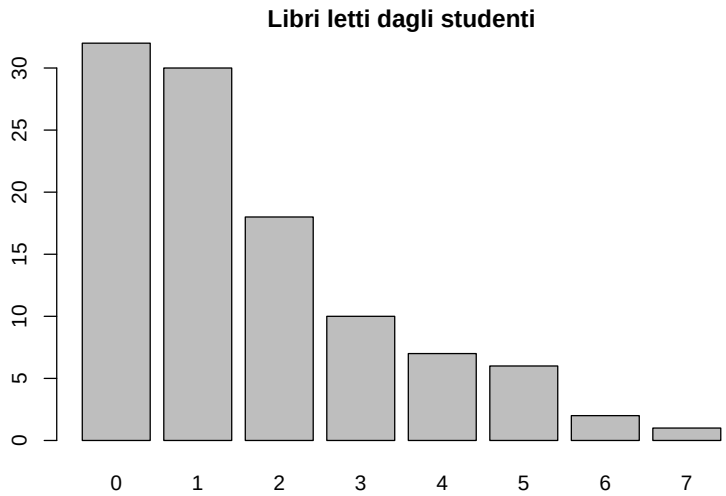
Finora abbiamo trattato di come

- rappresentare graficamente,
- sintetizzare numericamente (con medie, mediane, varianze, eccetera),

singole variabili, in modo da descrivere l'insieme delle unità statistiche rispetto a quel particolare carattere.



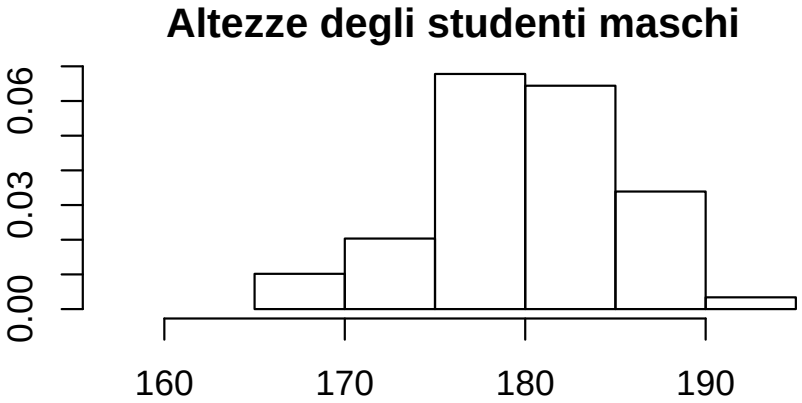
Altezza media=174
Altezza mediana=175
SD altezza=9.33

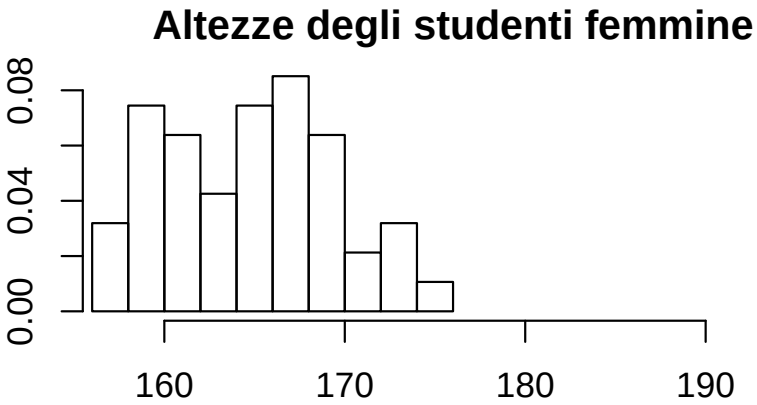


libri medio=1.63
libri mediano=1
SD # libri=1.67

... o quasi

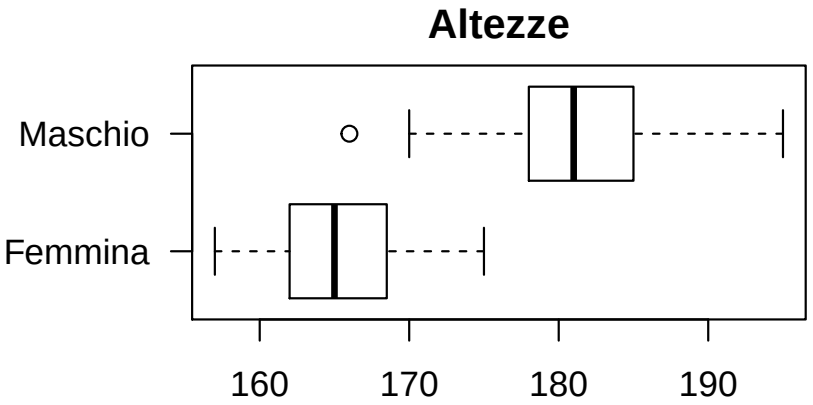
In molti casi abbiamo guardato congiuntamente a due variabili.





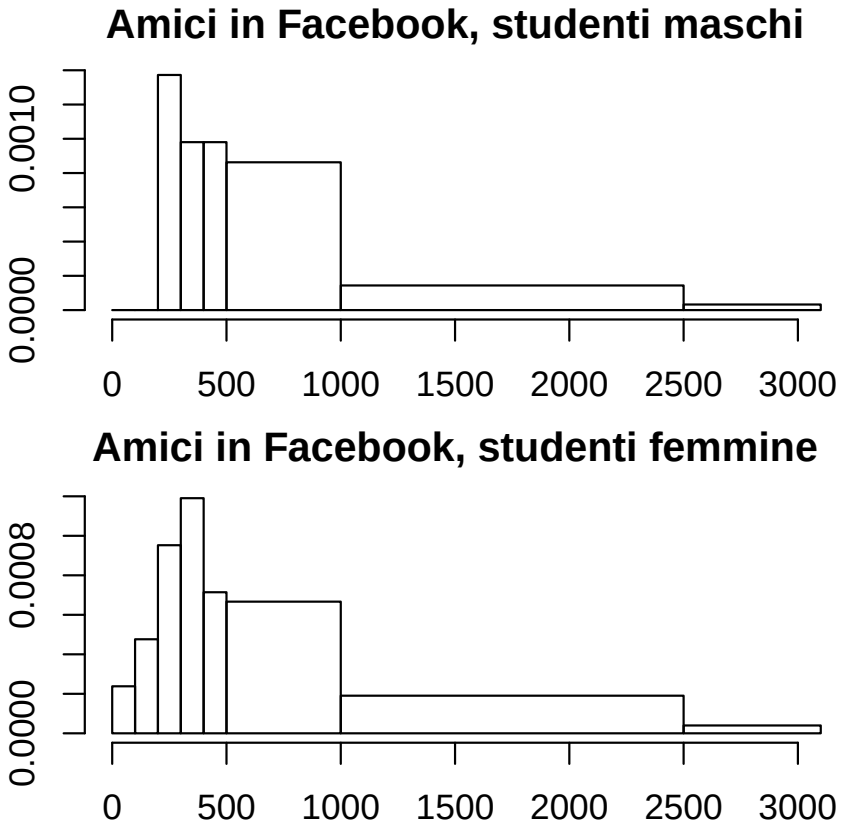
Altezza: media; mediana

- maschi: 181.1; 181;
- femmine: 165.4; 165



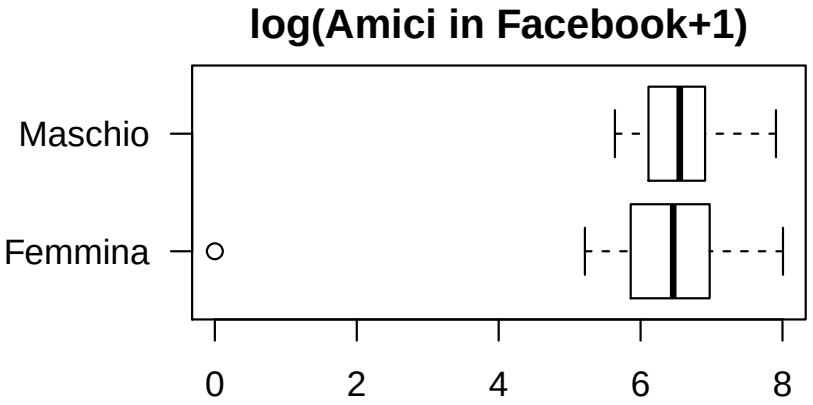
... o quasi

In molti casi abbiamo guardato congiuntamente a due variabili.



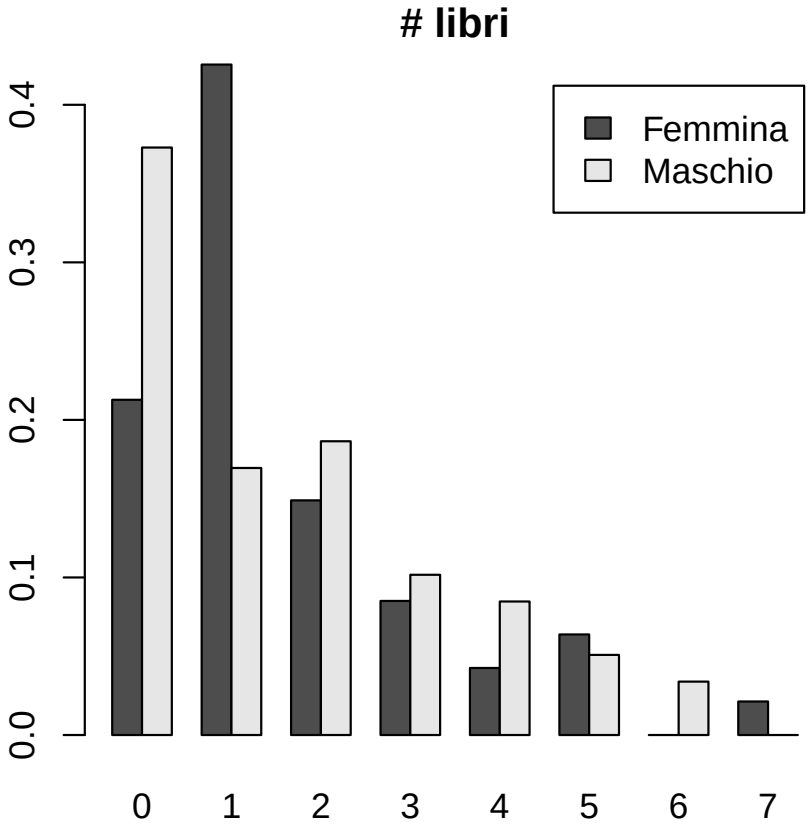
Amici in Facebook: media; mediana

- maschi: 811.5; 700;
- femmine: 854.9; 638



... o quasi

In molti casi abbiamo guardato congiuntamente a due variabili.



Libri medio e mediano

- maschi: 1.6; 1;
- femmine: 1.6; 1

... o quasi

In molti casi abbiamo guardato congiuntamente a due variabili.

Quando confrontiamo le distribuzioni di Y condizionate a diversi valori assunti da una seconda variabile X (usando gli stessi strumenti che usiamo per la distribuzione marginale di Y).

Si ha una distribuzione doppia quando si esaminano congiuntamente due caratteri nelle unità statistiche del collettivo.

Come nel caso di distribuzioni relative ad un singolo carattere, 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.

Indice

Variabili statistiche bivariate

Associazione tra variabili

Relazioni tra variabili

Esempio: ore di sonno e genere

Consideriamo la variabile doppia: (Y, X)=(ore di sonno, genere).
La distribuzione di frequenze assolute è data da

Y	X		totale
	$X = Fem$	$X = Mas$	
5	0	3	3
6	2	5	7
7	18	18	36
8	24	26	50
9	3	5	8
10	0	1	1
12	0	1	1
Totale	47	59	106

Esempio: ore di sonno e genere

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

- Il "centro" della distribuzione (in questo caso le 7 righe e le 2 colonne centrali) mostra il numero di individui che presentano una particolare modalità della coppia (Y, X): mostra cioè la **distribuzione congiunta**.

Esempio: ore di sonno e genere

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

- Il "centro" della distribuzione (in questo caso le 7 righe e le 2 colonne centrali) mostra il numero di individui che presentano una particolare modalità della coppia (Y, X): mostra cioè la **distribuzione congiunta**.
- La 1^a colonna, per esempio, mostra la distribuzione delle ore di sonno tra le femmine, cioè la distribuzione della variabile condizionata $(Y|X = Fem)$. Analogamente, la 2^a mostra la distribuzione della variabile condizionata $(Y|X = Mas)$. Quindi, le colonne riportano le **distribuzioni della variabile condizionata $Y|X$**.

Esempio: ore di sonno e genere

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

- Il "centro" della distribuzione (in questo caso le 7 righe e le 2 colonne centrali) mostra il numero di individui che presentano una particolare modalità della coppia (Y, X): mostra cioè la **distribuzione congiunta**.
- La 1^a colonna, per esempio, mostra la distribuzione delle ore di sonno tra le femmine, cioè la distribuzione della variabile condizionata $(Y|X = Fem)$. Analogamente, la 2^a mostra la distribuzione della variabile condizionata $(Y|X = Mas)$. Quindi, le colonne riportano le **distribuzioni della variabile condizionata $Y|X$**.
- La 3^a riga mostra, tra le persone che dormono 7 ore per notte, quante sono femmine e quanti maschi, cioè la distribuzione della variabile condizionata $(X|Y = 7)$. Quindi, le righe riportano le **distribuzioni della variabile condizionata $X|Y$**.

Esempio: ore di sonno e genere

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

- Il "centro" della distribuzione (in questo caso le 7 righe e le 2 colonne centrali) mostra il numero di individui che presentano una particolare modalità della coppia (Y, X): mostra cioè la **distribuzione congiunta**.
- La 1^a colonna, per esempio, mostra la distribuzione delle ore di sonno tra le femmine, cioè la distribuzione della variabile condizionata $(Y|X = Fem)$. Analogamente, la 2^a mostra la distribuzione della variabile condizionata $(Y|X = Mas)$. Quindi, le colonne riportano le **distribuzioni della variabile condizionata $Y|X$**.
- La 3^a riga mostra, tra le persone che dormono 7 ore per notte, quante sono femmine e quanti maschi, cioè la distribuzione della variabile condizionata $(X|Y = 7)$. Quindi, le righe riportano le **distribuzioni della variabile condizionata $X|Y$**.
- L'ultima colonna, mostra la distribuzione delle ore di sonno senza riguardo al genere. L'ultima riga, invece, mostra la distribuzione della variabile `Genere`. Sono cioè rappresentate le **distribuzioni marginali**.

Esempio: CdS e Sanremo

Consideriamo la variabile doppia: (Y, X)=(Corso di Studi, Sanremo).
La distribuzione di frequenze assolute è data da

Y	X			totale
	$X = NonV$	$X = VeNP$	$X = VeP$	
SP	167	22	48	244
EC	169	12	65	246
SIAFA	45	2	6	53
Totale	381	36	119	543

Tabella a doppia entrata

Una distribuzione doppia di frequenze è normalmente chiamata **tabella (di contingenza) a doppia entrata**.

In generale, una tabella di contingenza (con due variabili) si presenta nella forma:

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_t	totale
y_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	n_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{it}	n_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_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
- $\{x_1, \ldots, x_t\}$ sono le modalità di X

Tabella a doppia entrata (cont)

Nella tabella

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

Tabella a doppia entrata (cont)

Nella tabella

- X e Y sono le due variabili considerate
- $\{x_1, \ldots, x_t\}$ sono le modalità di X
- $\{y_1, \ldots, y_s\}$ sono le modalità di Y
- n_{ij} è la **frequenza congiunta** assoluta per $Y = y_i$ e $X = x_j$

Tabella a doppia entrata (cont)

Nella tabella

- X e Y sono le due variabili considerate
- $\{x_1, \ldots, x_t\}$ sono le modalità di X
- $\{y_1, \ldots, y_s\}$ sono le modalità di Y
- n_{ij} è la **frequenza congiunta** assoluta per $Y = y_i$ e $X = x_j$
- 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
- $\{x_1, \ldots, x_t\}$ sono le modalità di X
- $\{y_1, \ldots, y_s\}$ sono le modalità di Y
- n_{ij} è la **frequenza congiunta** assoluta per $Y = y_i$ e $X = x_j$
- 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 x_j di X.
- 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 y_i di Y.

NB. La scelta di quale variabile (X o Y) mettere sulle righe/colonne è libera.

Esempio: il disastro del Titanic

Tabella a doppia entrata per le variabili Passeggero (tipologia) e Sopravvivenza

	1st	2nd	3rd	Crew	Totale
Morto	122	167	528	673	1490
Sopravv.	203	118	178	212	711
Totale	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 percentuali) di Y condizionate a X.

		1st	2nd	3rd	Crew	Totale
Morto	freq. ass.	122	167	528	673	1490
	% di colonna	37.5%	58.6%	74.8%	76.0%	67.7%
Sopravv.	freq. ass.	203	118	178	212	711
	% di colonna	62.5%	41.4%	25.2%	24.0%	32.3%
Totale		325	285	706	885	2201

- In prima classe, sopravvisse il 62.5% dei passeggeri
- In seconda classe, sopravvisse il 41.4% dei passeggeri
- In terza, sopravvisse il 25.2% dei passeggeri
- Dell'equipaggio, sopravvisse il 24%

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.

		1st	2nd	3rd	Crew	Totale
Morto	freq. ass.	122	167	528	673	1490
	% di colonna	37.5%	58.6%	74.8%	76.0%	67.7%
Sopravv.	freq. ass.	203	118	178	212	711
	% di colonna	62.5%	41.4%	25.2%	24.0%	32.3%
Totale		325	285	706	885	2201

Qui, abbiamo calcolato le frequenze percentuali della variabile condizionata `Sopravvivenza|Passeggero`.
Si noti che, per ogni tipologia di passeggero, le percentuali sommano a 100.

Tabella a doppia entrata per un verso...

In generale, le frequenze relative per le distribuzioni di $Y|X$ si calcolano a partire dalle frequenze assolute così:

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_t	totale
y_1	n_{11}/n_{01}	$\cdots$	n_{1j}/n_{0j}	$\cdots$	n_{1t}/n_{0t}	n_{10}/N
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	n_{i1}/n_{01}	$\cdots$	n_{ij}/n_{0j}	$\cdots$	n_{it}/n_{0t}	n_{i0}/N
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	n_{s1}/n_{01}	$\cdots$	n_{sj}/n_{0j}	$\cdots$	n_{st}/n_{0t}	n_{s0}/N
totale	1	$\cdots$	1	$\cdots$	1	1

...e per un altro...

Viceversa, le frequenze relative per le distribuzioni di $X|Y$ si calcolano a partire dalle frequenze assolute così:

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_t	totale
y_1	n_{11}/n_{10}	$\cdots$	n_{1j}/n_{10}	$\cdots$	n_{1t}/n_{10}	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	n_{i1}/n_{i0}	$\cdots$	n_{ij}/n_{i0}	$\cdots$	n_{it}/n_{i0}	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	n_{s1}/n_{s0}	$\cdots$	n_{sj}/n_{s0}	$\cdots$	n_{st}/n_{s0}	1
totale	n_{01}/N	$\cdots$	n_{0j}/N	$\cdots$	n_{0t}/N	1

...o "per tutti e due"

Infine, possiamo costruire le frequenze relative per la distribuzione congiunta di (X, Y), che si calcolano a partire dalle frequenze assolute così:

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_t	totale
y_1	n_{11}/N	$\cdots$	n_{1j}/N	$\cdots$	n_{1t}/N	n_{10}/N
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	n_{i1}/N	$\cdots$	n_{ij}/N	$\cdots$	n_{it}/N	n_{i0}/N
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	n_{s1}/N	$\cdots$	n_{sj}/N	$\cdots$	n_{st}/N	n_{s0}/N
totale	n_{01}/N	$\cdots$	n_{0j}/N	$\cdots$	n_{0t}/N	1

Il disastro del Titanic

		1st	2nd	3rd	Crew	Totale
Morto	freq. ass.	122	167	528	673	1490
Sopravv.	freq. ass.	203	118	178	212	711
Totale	freq. ass.	325	285	706	885	2201

Il disastro del Titanic

		1st	2nd	3rd	Crew	Totale
Morto	freq. ass.	122	167	528	673	1490
	% di colonna	37.5%	58.6%	74.8%	76.0%	67.7%
Sopravv.	freq. ass.	203	118	178	212	711
	% di colonna	62.5%	41.4%	25.2%	24.0%	32.3%
Totale	freq. ass.	325	285	706	885	2201
	% di colonna	100%	100%	100%	100%	100%

Il disastro del Titanic

		1st	2nd	3rd	Crew	Totale
Morto	freq. ass.	122	167	528	673	1490
	% di riga	8.2%	11.2%	35.4%	45.2%	100%
Sopravv.	freq. ass.	203	118	178	212	711
	% di riga	28.6%	16.6%	25.0%	29.8%	100%
Totale	freq. ass.	325	285	706	885	2201
	% di riga	14.8%	12.9%	32.1%	40.2%	100%

Il disastro del Titanic

		1st	2nd	3rd	Crew	Totale
Morto	freq. ass.	122	167	528	673	1490
	% congiunta	5.6%	7.6%	24.0%	30.6%	67.7%
Sopravv.	freq. ass.	203	118	178	212	711
	% congiunta	9.2%	5.4%	8.1%	9.6%	32.3%
Totale	freq. ass.	325	285	706	885	2201
	% congiunta	14.8%	12.9%	32.1%	40.2%	100%

Il disastro del Titanic

		1st	2nd	3rd	Crew	Totale
Morto	freq. ass.	122	167	528	673	1490
	% di colonna	37.5%	58.6%	74.8%	76.0%	67.7%
	% di riga	8.2%	11.2%	35.4%	45.2%	100%
	% congiunta	5.6%	7.6%	24.0%	30.6%	67.7%
Sopravv.	freq. ass.	203	118	178	212	711
	% di colonna	62.5%	41.4%	25.2%	24.0%	32.3%
	% di riga	28.6%	16.6%	25.0%	29.8%	100%
	% congiunta	9.2%	5.4%	8.1%	9.6%	32.3%
Totale	freq. ass.	325	285	706	885	2201
	% di colonna	100%	100%	100%	100%	100%
	% di riga	14.8%	12.9%	32.1%	40.2%	100%
	% congiunta	14.8%	12.9%	32.1%	40.2%	100%

Esempio: CdS e Sanremo

Y	X			totale
	$X = NonV$	$X = VeNP$	$X = VeP$	
SP	167	22	48	244
EC	169	12	65	246
SIAFA	45	2	6	53
Totale	381	36	119	543

Per confrontare i diversi CdS calcolo le condizionate 'di riga'

Y	X		
	$X = NonV$	$X = VeNP$	$X = VeP$
SP	0.705	0.093	0.203
EC	0.687	0.049	0.264
SIAFA	0.849	0.038	0.113

Esempio: Genere e Sanremo

Y	X			totale
	$X = NonV$	$X = VeNP$	$X = VeP$	
Femmina	160	17	69	251
Maschio	221	19	50	292
Totale	381	36	119	543

Per confrontare i diversi generi calcolo le condizionate 'di riga'

Y	X		
	$X = NonV$	$X = VeNP$	$X = VeP$
Femmina	0.650	0.069	0.280
Maschio	0.762	0.066	0.172

Indice

Variabili statistiche bivariate
- Rappresentazioni grafiche
- Esempi
- Distribuzioni multiple

Associazione tra variabili

Relazioni tra variabili

Rappresentazioni grafiche

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

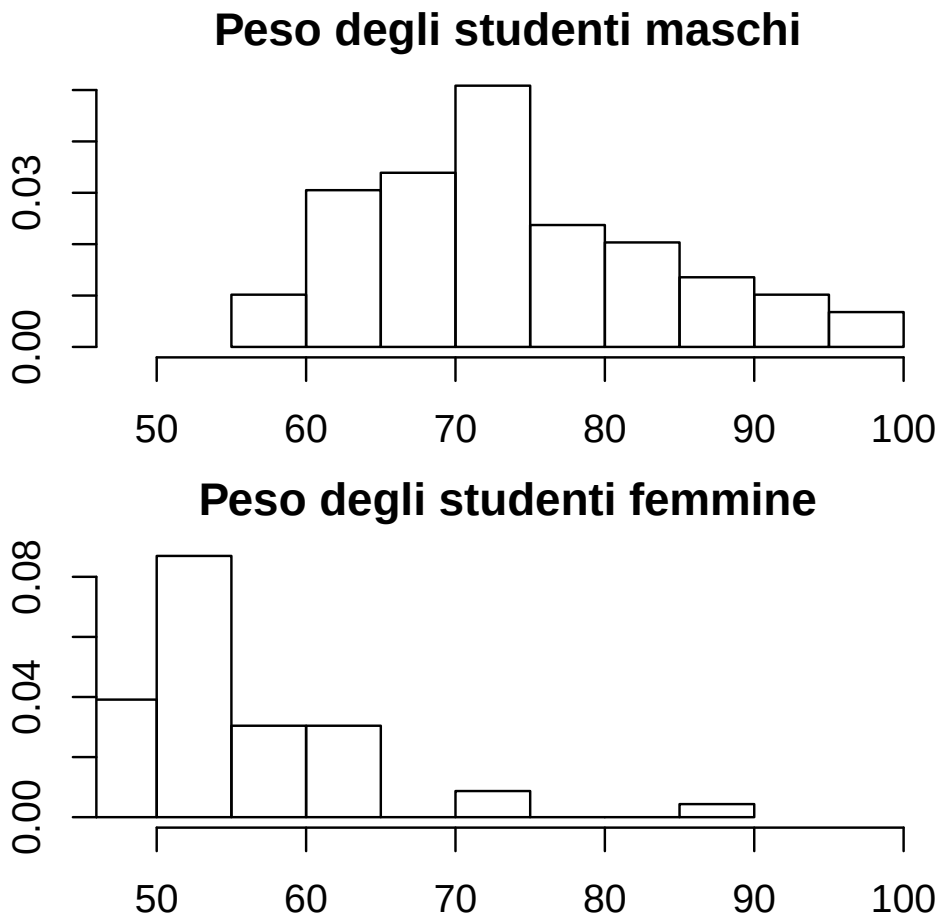
La rappresentazione dipende dalla natura delle variabili (qualitative, quantitative) e dalla forma in cui ci sono forniti i dati (aggregata/non aggregata).

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.

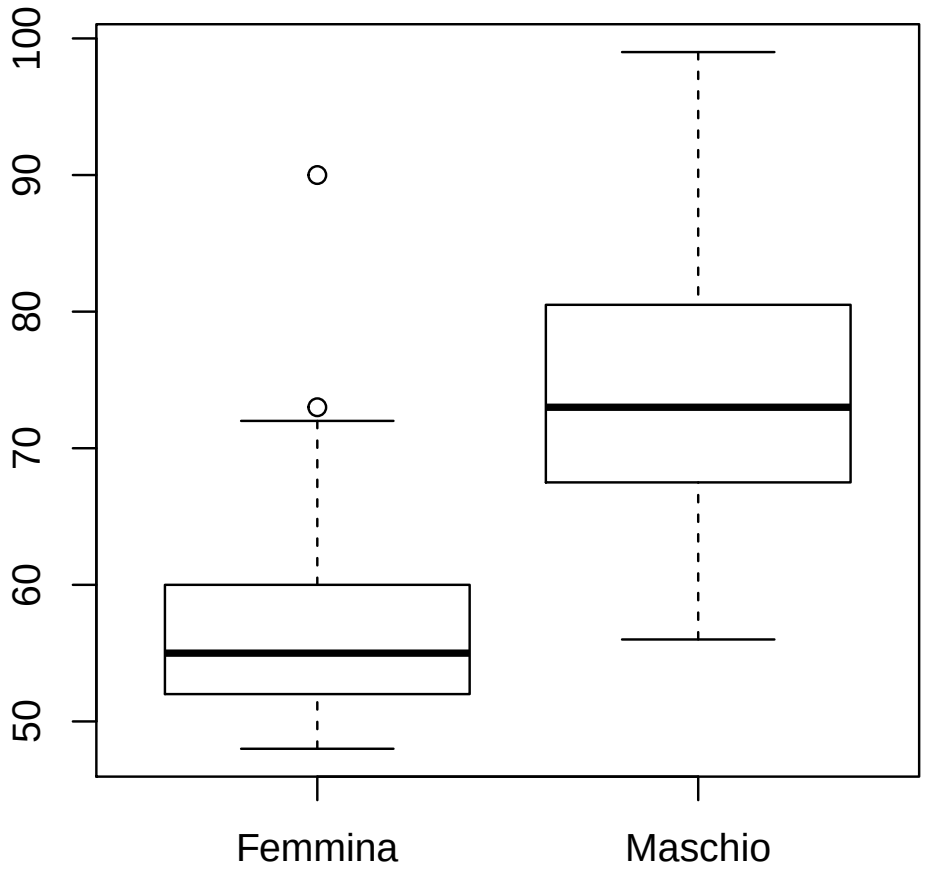
Istogrammi appaiati (side-by-side histograms)

- $Y \rightarrow$ Peso degli studenti (quantitativa continua)
- $X \rightarrow$ Genere (qualitativa)
- rappresentazione di $Y|X$.



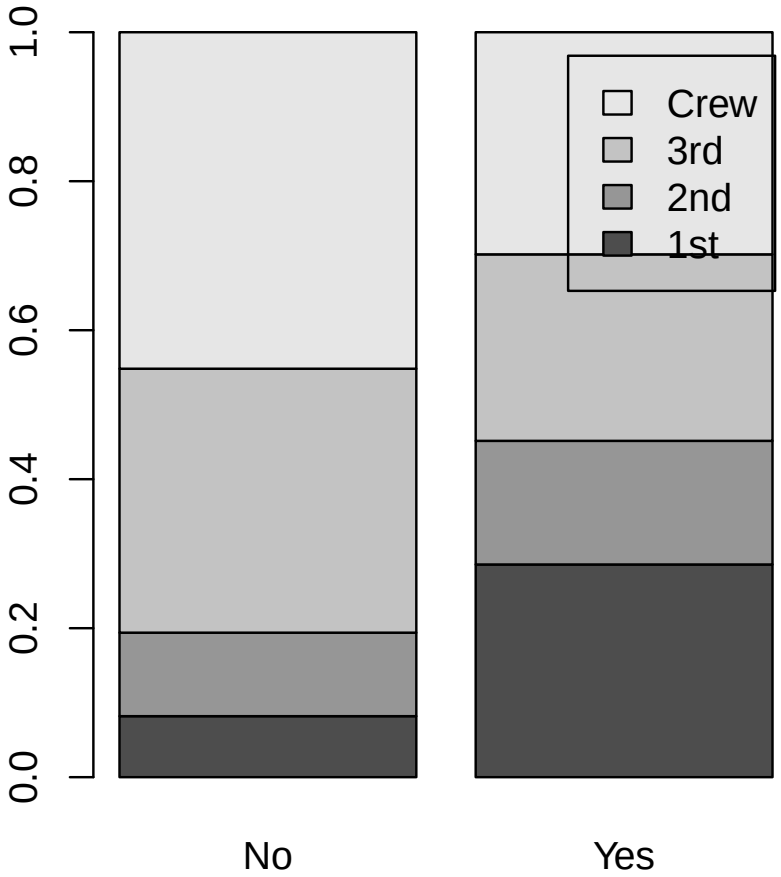
Diagrammi a scatola appaiati (side-by-side histograms)

- $Y \rightarrow$ Peso degli studenti (quantitativa continua)
- $X \rightarrow$ Genere (qualitativa)
- rappresentazione di $Y|X$.



Diagrammi a barre condizionati

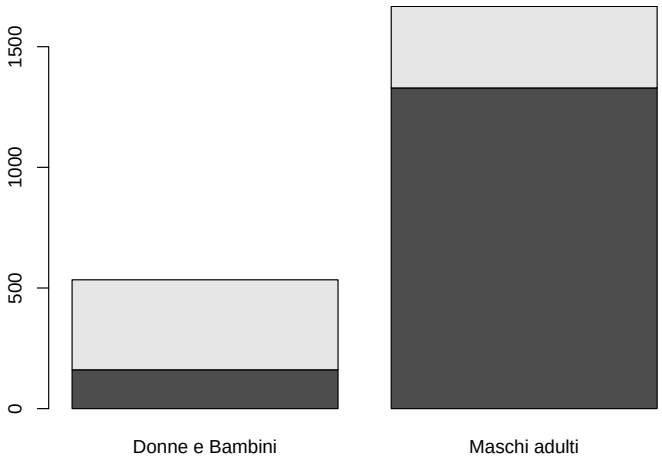
- $Y \rightarrow$ Classe / Equipaggio
- $X \rightarrow$ Sopravvivenza
- rappresentazione di $Y|X$.

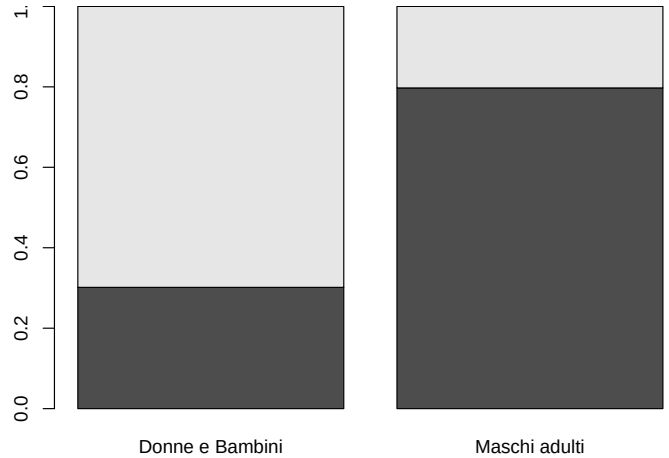


Prima le donne e i bambini?

	Donne e Bambini	Maschi adulti	Totale
Morto	161	1329	1490
Sopravv.	373	338	711
Totale	534	1667	2201

Cosa rappresentano i due grafici?





Indice

Variabili statistiche bivariate
- Rappresentazioni grafiche
- Esempi
- Distribuzioni multiple

Associazione tra variabili

Relazioni tra variabili

Esempio: internet ed età

Età	Frequenza nell'uso di internet				
	giornalmente	una o più volte la settimana	qualche volta al mese	qualche volta all'anno	mai
6-10	9	25,6	7,2	2,6	53,8
11-14	44,5	31,6	3,8	0,9	17,2
15-17	70,2	18,5	1,8	0,3	7,9
18-19	76,2	15,7	1,5	0,4	4
20-24	70,6	16,1	1,9	0,5	8,4
25-34	61	19,9	1,8	0,8	14,6
35-44	50,2	21,9	3	1	22,6
45-54	40,7	20,6	3,1	1,2	32,9
55-59	30,9	17,3	3,2	1,1	46,1
60-64	23,4	15,7	1,9	0,6	57,2
65-74	10,2	8,9	1,6	0,5	77,1
75 e più	1,9	1,8	0,4	0,2	93,9

Esempio: internet e condizione professionale

Condizione professionale	Frequenza nell'uso di internet				
	giornalmente	una o più volte la settimana	qualche volta al mese	qualche volta all'anno	mai
occupato	54,4	20,8	2,7	1	20
dirigenti, imprenditori, liberi professionisti	74,1	13	1,2	0,3	9,8
direttivo, quadro, impiegato	70,8	17,8	2	0,7	7,8
operaio, apprendista	34,5	26,5	3,7	1,4	32,6
lavoratore in proprio, coadiuvante familiare, co.co.co.	42,6	21,9	3,1	1,2	30
casalinga-o	10,4	10,9	2,2	0,7	73,8
studente	77,6	14,3	1,1	0,2	5
ritirato-a dal lavoro	8,9	8,5	1,4	0,4	79,3
in altra condizione	13,8	9,1	1,5	1,1	71
disoccupato alla ricerca di nuova occupazione	40,9	20	2,9	1,1	32,8
in cerca di prima occupazione	49,2	23,1	2,1	0,9	22
totale	38,1	16	2,2	0,8	41,4

Sport ed età

Età	continuativo	saltuario	qualche att.	nessuna
3-5	23,4	4,5	20	48,1
6-10	58,9	6,4	12,5	21,2
11-14	56,3	8,6	13,8	20,7
15-17	47,7	11,9	18,8	21
18-19	39,2	12,3	20	27,9
20-24	37,2	14,2	20,9	27,3
25-34	30	13,9	23,9	32
35-44	22,6	12	28,5	36,8
45-54	19,6	11,2	29,8	39,3
55-59	16,9	9,5	31,5	42
60-64	14,7	8,4	34,3	42,3
65-74	11,2	6	35	47,5
75 e pi	4,4	2,9	23,8	68,6

Sport e genere

Nella fascia d'età 20-24

	continuativo	saltuario	qualche att.	nessuna
Femmine	30,8	12,7	27	29,2
Maschi	44,6	19,1	12,5	23,3

In generale (3 anni o più)

	continuativo	saltuario	qualche att.	nessuna
Femmine	20,8	8,3	27,2	43,4
Maschi	29,7	11,1	24	34,8

Nel campione studenti

	Attività sportiva		
	No	Sì	Totale
Femmina	79	145	251
Maschio	38	216	292
Totale	117	361	543

Sport e genere

Nella fascia d'età 20-24

	continuativo	saltuario	qualche att.	nessuna
Femmine	30,8	12,7	27	29,2
Maschi	44,6	19,1	12,5	23,3

In generale (3 anni o più)

	continuativo	saltuario	qualche att.	nessuna
Femmine	20,8	8,3	27,2	43,4
Maschi	29,7	11,1	24	34,8

Nel campione studenti

	Attività sportiva		
	No	Sì	Totale
Femmina	79	145	251
Maschio	38	216	292
Totale	117	361	543

	Attività sportiva	
	No	Sì
Femmina	0.353	0.647
Maschio	0.150	0.850

Sport e genere

	Genere		
	Fem	Mas	Totale
Altri squadra	13	14	27
Altri non squadra	47	39	86
Basket	0	19	19
Calcio	2	61	63
Corsa	30	18	48
Nuoto	12	17	29
Palestra	60	61	121
Totale	164	229	393

Sport e genere

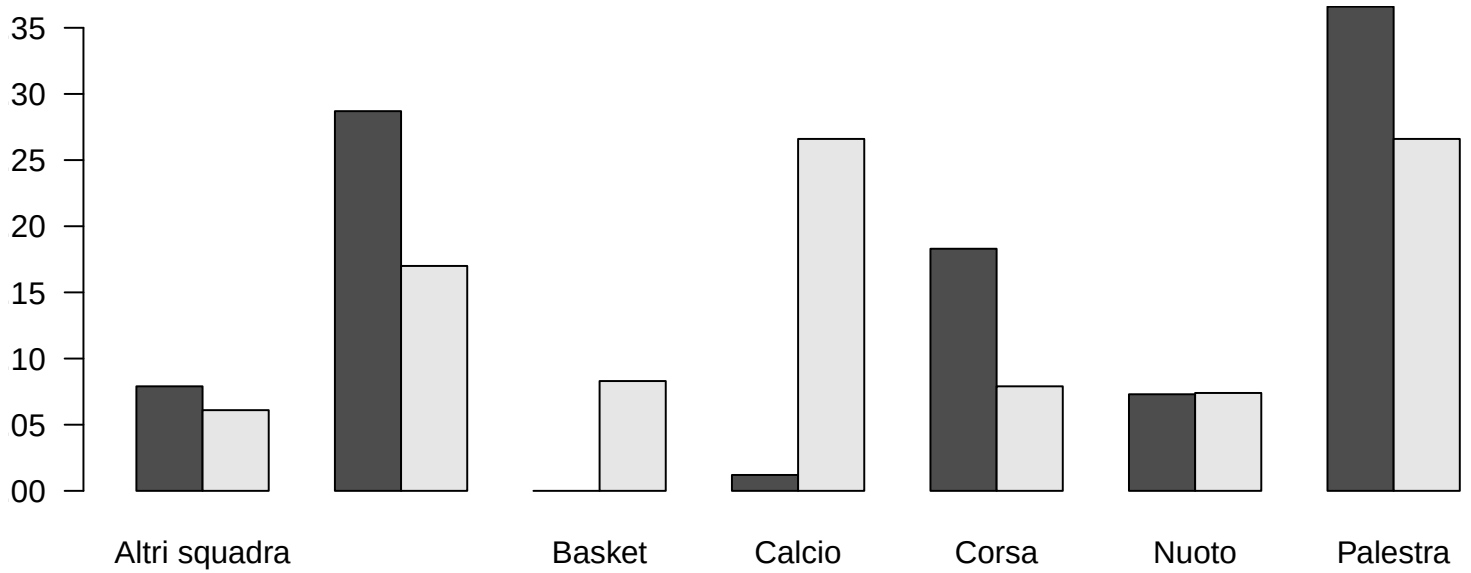
	Genere		
	Fem	Mas	Totale
Altri squadra	13	14	27
Altri non squadra	47	39	86
Basket	0	19	19
Calcio	2	61	63
Corsa	30	18	48
Nuoto	12	17	29
Palestra	60	61	121
Totale	164	229	393

	Genere	
	Fem	Mas
Altri squadra	0.079	0.061
Altri non squadra	0.287	0.170
Basket	0.000	0.083
Calcio	0.012	0.266
Corsa	0.183	0.079
Nuoto	0.073	0.074
Palestra	0.366	0.266

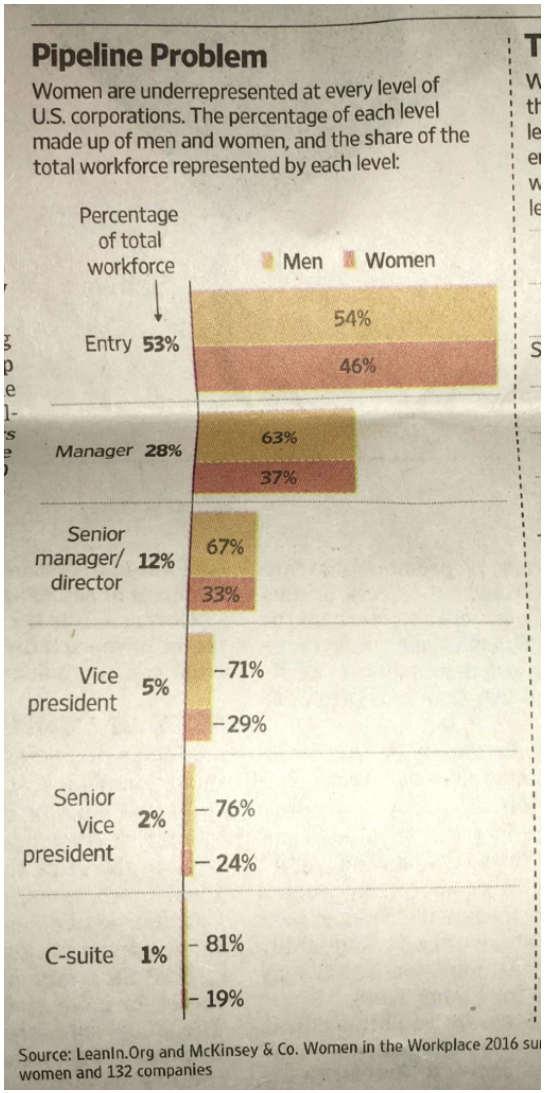
Sport e genere

	Genere		
	Fem	Mas	Totale
Altri squadra	13	14	27
Altri non squadra	47	39	86
Basket	0	19	19
Calcio	2	61	63
Corsa	30	18	48
Nuoto	12	17	29
Palestra	60	61	121
Totale	164	229	393

	Genere	
	Fem	Mas
Altri squadra	0.079	0.061
Altri non squadra	0.287	0.170
Basket	0.000	0.083
Calcio	0.012	0.266
Corsa	0.183	0.079
Nuoto	0.073	0.074
Palestra	0.366	0.266

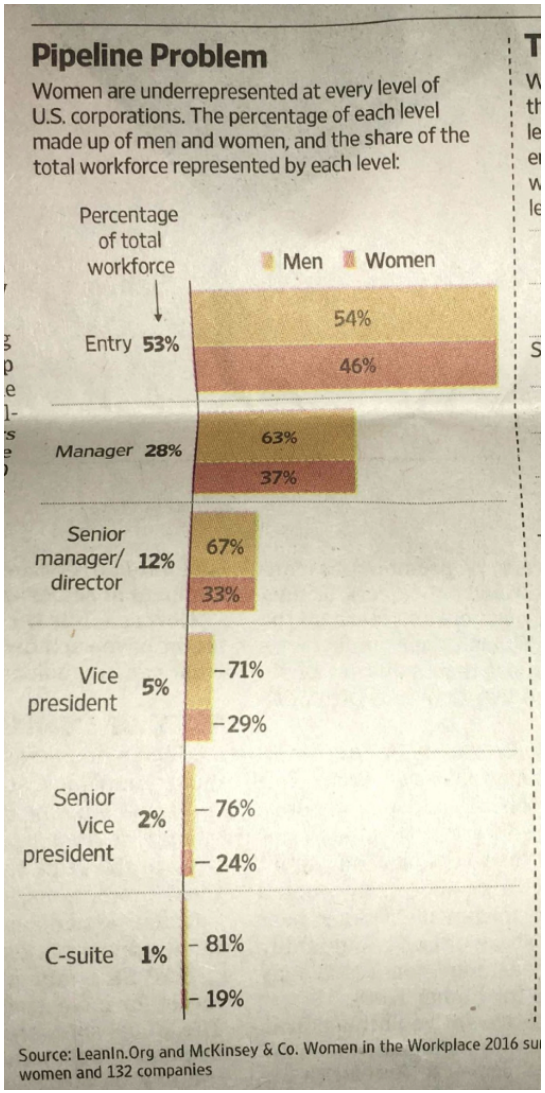


Disparità di genere nelle professioni

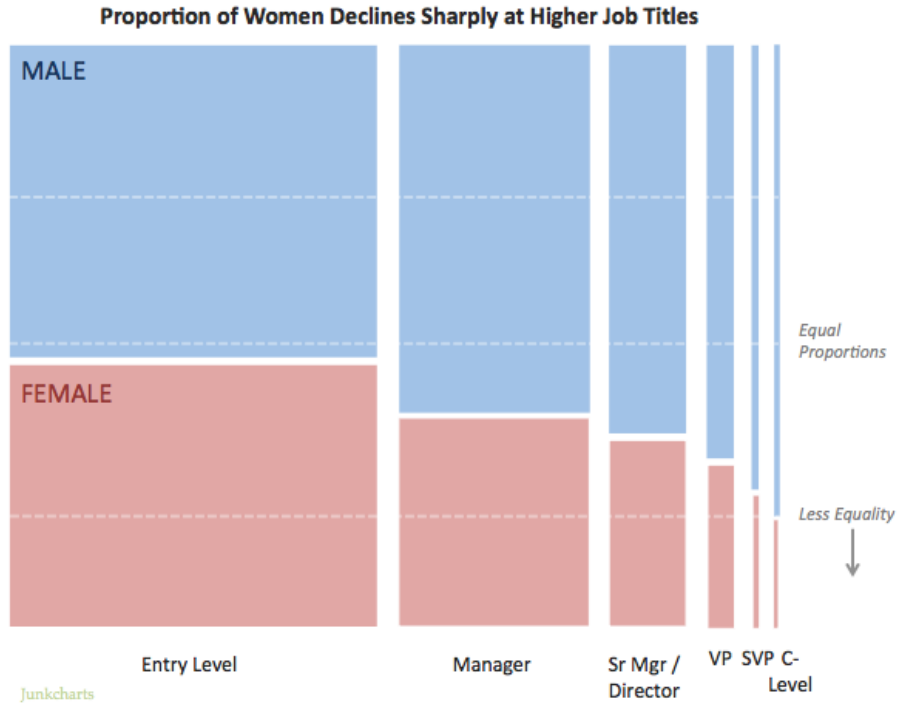


Il grafico a sinistra proviene dal Wall Street Journal, il messaggio che vuole comunicare è che le professioni più qualificate vedono una predominanza maschile.

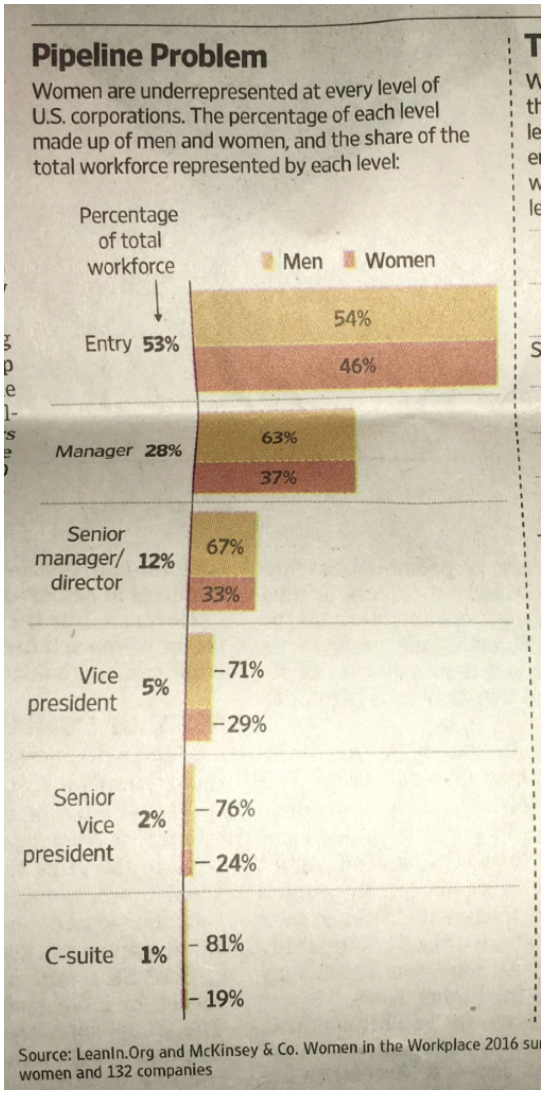
Disparità di genere nelle professioni



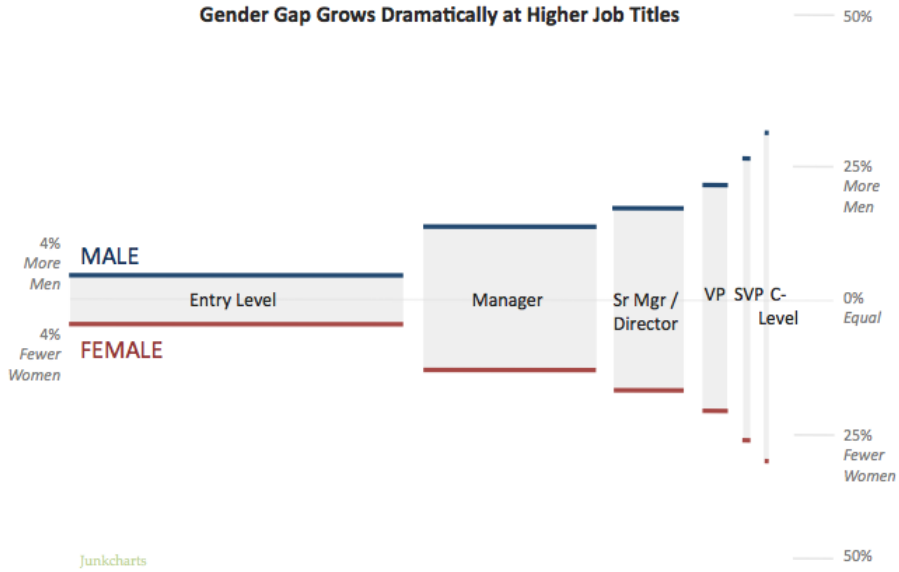
Il grafico a sinistra proviene dal Wall Street Journal, il messaggio che vuole comunicare è che le professioni più qualificate vedono una predominanza maschile.



Disparità di genere nelle professioni



Il grafico a sinistra proviene dal Wall Street Journal, il messaggio che vuole comunicare è che le professioni più qualificate vedono una predominanza maschile.



Indice

Variabili statistiche bivariate
- Rappresentazioni grafiche
- Esempi
- Distribuzioni multiple

Associazione tra variabili

Relazioni tra variabili

Variabili statistiche multivariate (cenno)

L'idea di variabile statistica bivariata può essere generalizzata senza difficoltà.

Variabili statistiche multivariate (cenno)

L'idea di variabile statistica bivariata può essere generalizzata senza difficoltà.

Si parla di variabile statistica

- trivariata, se si considerano congiuntamente tre caratteri;
- quadrivariata, se si considerano congiuntamente quattro caratteri;
- in generale, multivariata, se si considerano congiuntamente almeno due caratteri;

Distribuzioni di frequenza multiple (cenno)

Anche le distribuzioni di frequenza si generalizzano di conseguenza.

Distribuzioni di frequenza multiple (cenno)

Anche le distribuzioni di frequenza si generalizzano di conseguenza.

Si parla di distribuzione di frequenza

- tripla, se mostra la distribuzione di una variable statistica trivariata;
- quadrupla, se mostra la distribuzione di una variable statistica quadrivariata;
- in generale, multipla, se si considera una variable statistica multivariata;

Titanic, dati completi

Classe	Genere	Età	Sopravvissuto	
			No	Sì
1st	Male	Child	0	5
		Adult	118	57
	Female	Child	0	1
		Adult	4	140
2nd	Male	Child	0	11
		Adult	154	14
	Female	Child	0	13
		Adult	13	80
3rd	Male	Child	35	13
		Adult	387	75
	Female	Child	17	14
		Adult	89	76
Crew	Male	Child	0	0
		Adult	670	192
	Female	Child	0	0
		Adult	3	20

Esempio: esiti ammissione a Berkeley, 1973

I dati che abbiamo mostrato, anche se organizzati in forma non tabellare, rappresentano una distribuzione di frequenza tripla.

	Admit	Gender	Dept	Freq
1	Admitted	Male	A	512
2	Rejected	Male	A	313
3	Admitted	Female	A	89
4	Rejected	Female	A	19
5	Admitted	Male	B	353
6	Rejected	Male	B	207
7	Admitted	Female	B	17
8	Rejected	Female	B	8
9	Admitted	Male	C	120
10	Rejected	Male	C	205
11	Admitted	Female	C	202
12	Rejected	Female	C	391
13	Admitted	Male	D	138
14	Rejected	Male	D	279
15	Admitted	Female	D	131
16	Rejected	Female	D	244
17	Admitted	Male	E	53
18	Rejected	Male	E	138
19	Admitted	Female	E	94
20	Rejected	Female	E	299
21	Admitted	Male	F	22
22	Rejected	Male	F	351
23	Admitted	Female	F	24
24	Rejected	Female	F	317

Esempio: esiti ammissione a Berkeley, 1973

In forma tabellare:

Dept	Gender	Admit	
		Admitted	Rejected
A	Male	512	313
	Female	89	19
B	Male	353	207
	Female	17	8
C	Male	120	205
	Female	202	391
D	Male	138	279
	Female	131	244
E	Male	53	138
	Female	94	299
F	Male	22	351
	Female	24	317

Ammissioni a Berkely e genere

A partire dalla tabella a tre vie possiamo costruire delle tabelle a due vie (marginali), peresempio prendendo in considerazione solo l'esito (Admitted/Rejected) e il genere (Male/Female).

	Male	Female
Admitted	1198	557
Rejected	1493	1278

Guardando alle due variabili, e in particolare alle condizionate di riga (Admission|Gender)

	Male	Female
Admitted	0.45	0.30
Rejected	0.55	0.70

notiamo che le femmine sono ammesse in percentuale inferiore.
Possibili deduzioni

- i maschi sono più intelligenti
- Berkeley è sessista (a favore dei maschi)

Ammissioni a Berkely e genere, un'occhiata più accurata

Il 45% dei maschi viene ammesso contro il 30% delle femmine, ma come va per dipartimento?

Per ciascuno dei 6 dipartimenti (A,B,C,D,E,F) possiamo considerare la tabella congiunta ammissione-genere, ad esempio per il dipartimento A

	Male	Female
Admitted	512	89
Rejected	313	19
Totale	825	108

	Male	Female
Admitted	0.62	0.82
Rejected	0.38	0.18
Totale	1.00	1.00

Quindi nel dipartimento A la percentuale di ammessi è maggiore tra le femmine, contrariamente a quanto avviene sul totale.

Ripetiamo l'analisi sui vari dipartimenti.

Berkeley: analisi condizionata ai dipartimenti

Dept	Male			Female			Differenza
	Adm	Rej	Adm/Tot	Adm	Rej	Adm/Tot	
A	512	313	0.621	89	19	0.824	0.203
B	353	207	0.630	17	8	0.680	0.050
C	120	205	0.369	202	391	0.341	-0.029
D	138	279	0.331	131	244	0.349	0.018
E	53	138	0.277	94	299	0.239	-0.038
F	22	351	0.059	24	317	0.070	0.011

Le percentuali, dipartimento per dipartimento, sono abbastanza in linea, a parte che nel dipartimento A dove le femmine sono ammese in percentuale sensibilmente maggiore.

Notiamo però altre due cose

- il tasso di ammissione è parecchio diverso tra i vari dipartimenti
- la distribuzione di genere tra i dipartimenti è anche differente

Ammissioni condizionate e marginali

il tasso di ammissione è parecchio diverso tra i vari dipartimenti

	Admitted	Rejected	Totale
A	601	332	933
B	370	215	585
C	322	596	918
D	269	523	792
E	147	437	584
F	46	668	714

Ammissioni condizionate e marginali

il tasso di ammissione è parecchio diverso tra i vari dipartimenti

	Admitted	Rejected	Totale
A	0.64	0.36	1
B	0.63	0.37	1
C	0.35	0.65	1
D	0.34	0.66	1
E	0.25	0.75	1
F	0.06	0.94	1

Ammissioni condizionate e marginali

il tasso di ammissione è parecchio diverso tra i vari dipartimenti

	Admitted	Rejected	Totale
A	0.64	0.36	1
B	0.63	0.37	1
C	0.35	0.65	1
D	0.34	0.66	1
E	0.25	0.75	1
F	0.06	0.94	1

la distribuzione di genere tra i dipartimenti è anche differente

	Male	Female
A	825	108
B	560	25
C	325	593
D	417	375
E	191	393
F	373	341
Totale	2691	1835

Ammissioni condizionate e marginali

il tasso di ammissione è parecchio diverso tra i vari dipartimenti

	Admitted	Rejected	Totale
A	0.64	0.36	1
B	0.63	0.37	1
C	0.35	0.65	1
D	0.34	0.66	1
E	0.25	0.75	1
F	0.06	0.94	1

la distribuzione di genere tra i dipartimenti è anche differente

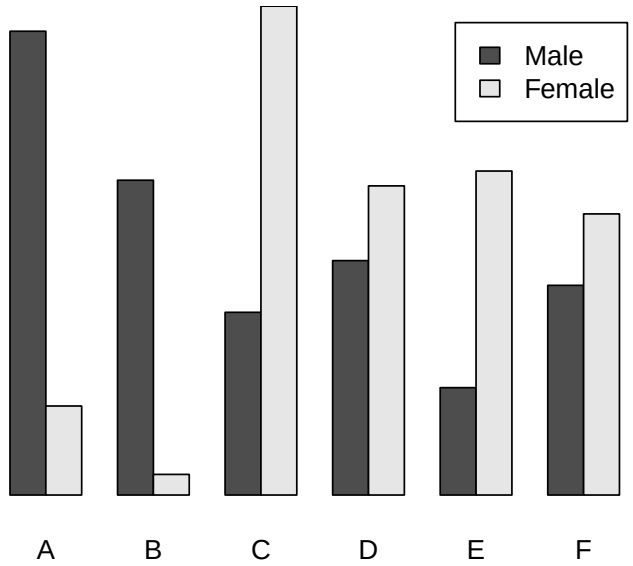
	Male	Female
A	0.31	0.06
B	0.21	0.01
C	0.12	0.32
D	0.15	0.20
E	0.07	0.21
F	0.14	0.19
Totale	1.00	1.00

Ammissioni condizionate e marginali

il tasso di ammissione è parecchio diverso tra i vari dipartimenti

	Admitted	Rejected	Totale
A	0.64	0.36	1
B	0.63	0.37	1
C	0.35	0.65	1
D	0.34	0.66	1
E	0.25	0.75	1
F	0.06	0.94	1

la distribuzione di genere tra i dipartimenti è anche differente

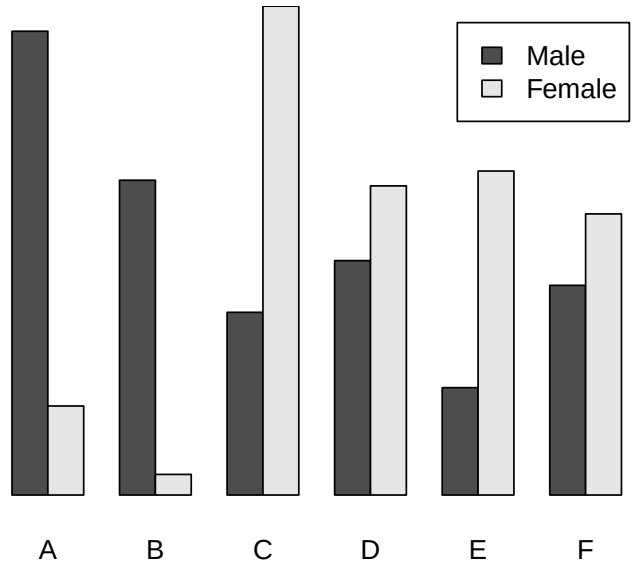


Ammissioni condizionate e marginali

il tasso di ammissione è parecchio diverso tra i vari dipartimenti

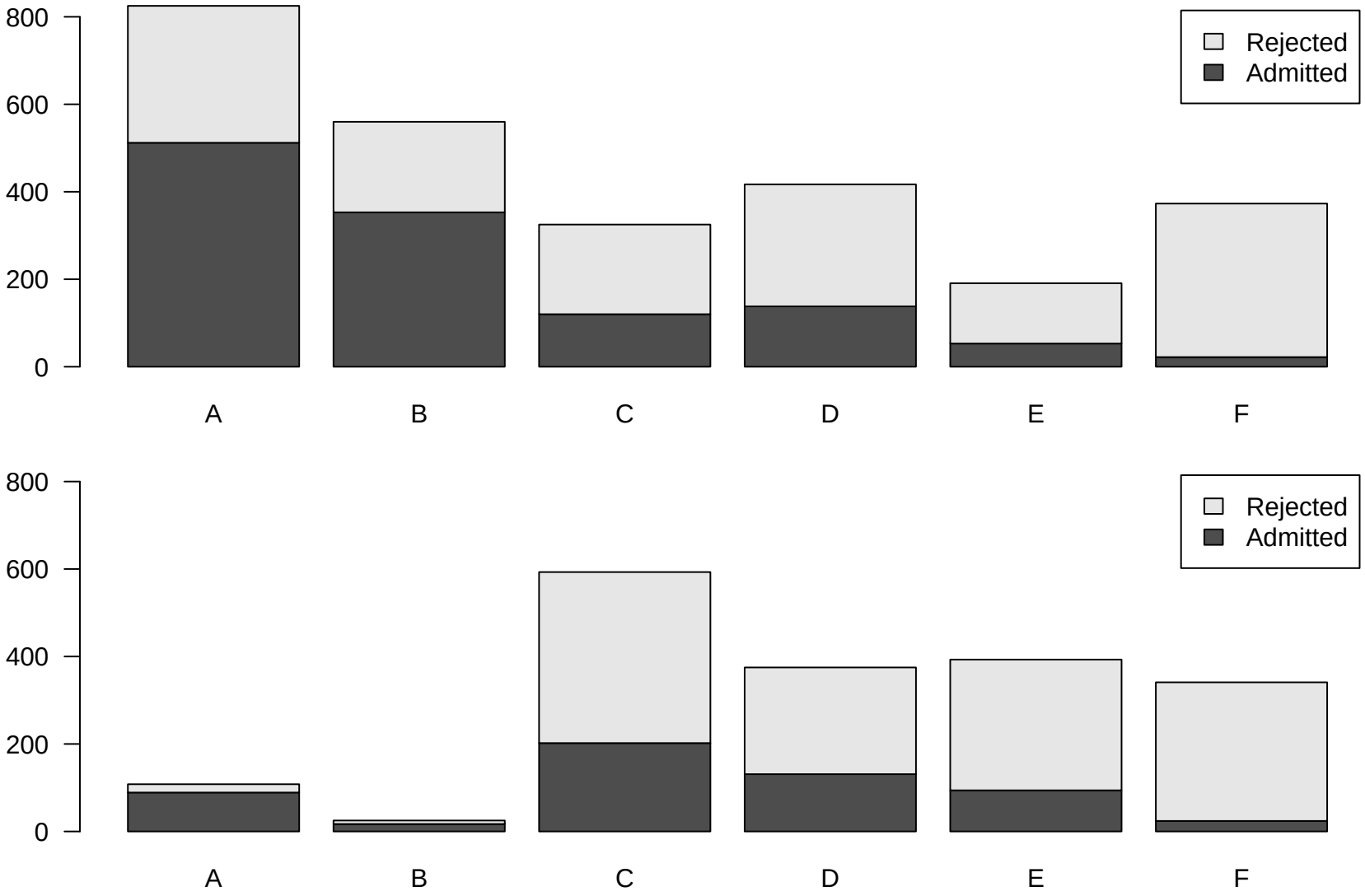
	Admitted	Rejected	Totale
A	0.64	0.36	1
B	0.63	0.37	1
C	0.35	0.65	1
D	0.34	0.66	1
E	0.25	0.75	1
F	0.06	0.94	1

la distribuzione di genere tra i dipartimenti è anche differente



Le femmine fanno domanda prevalentemente nei dipartimenti dove l'ammissione è più difficile, da cui l'apparente incoerenza tra le percentuali di ammisione condizionate e quella marginale (e le due spiegazioni avanzate all'inizio sono entrambe non supportate dai dati).

Visualizzazione



Paradosso di Simpson

Il fenomeno appena illustrato va sotto il nome di **paradosso di Simpson**.

Il riferimento è a Edward Simpson che lo descrisse nel 1951.

Si ha un paradosso di Simpson quando i dati mostrano un'associazione di un certo tipo tra due variabili se li si guarda condizionatamente a una terza, ma l'associazione opposta se si ignora la terza variabile.

Indice

Variabili statistiche bivariate

Associazione tra variabili

Relazioni tra variabili

Indice

Variabili statistiche bivariate

Associazione tra variabili

Dipendenza e indipendenza

Dipendenza in media, mediana,. . .

Relazioni tra variabili

Relazioni tra variabili

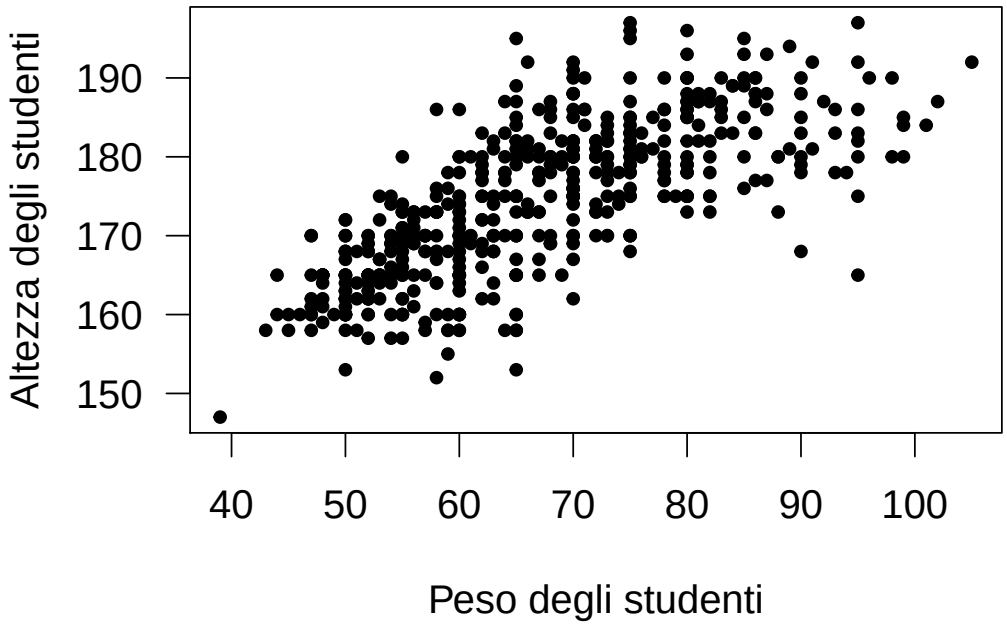
A ben vedere, il commento più naturale che abbiamo fatto leggendo i grafici precedenti era del tipo: il comportamento di questa variabile cambia al cambiare dell'altra, oppure, questa variabile è influenzata da quest'altra.

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

- Quando due variabili mostrano qualche forma di connessione tra loro, si parla di **associazione** o **dipendenza**.
- Quando due variabili non mostrano alcuna forma di connessione tra loro, si parla di **indipendenza**.

Esercizio

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



(a) Non c'è relazione tra altezza e peso;

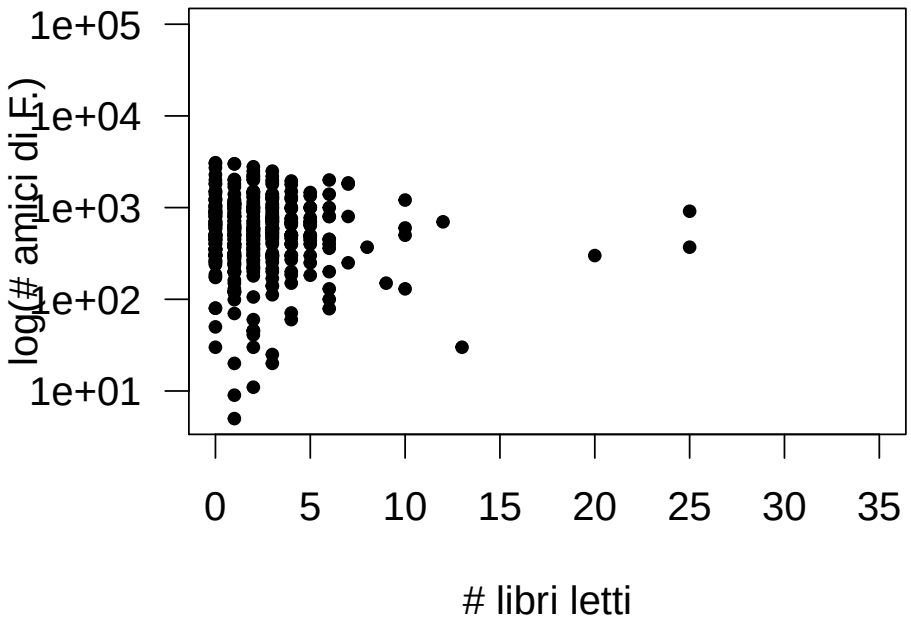
(b) altezza e peso sono associati (positivamente);

(c) altezza e peso sono associati (negativamente);

(d) un peso maggiore causa una maggiore altezza.

Esercizio

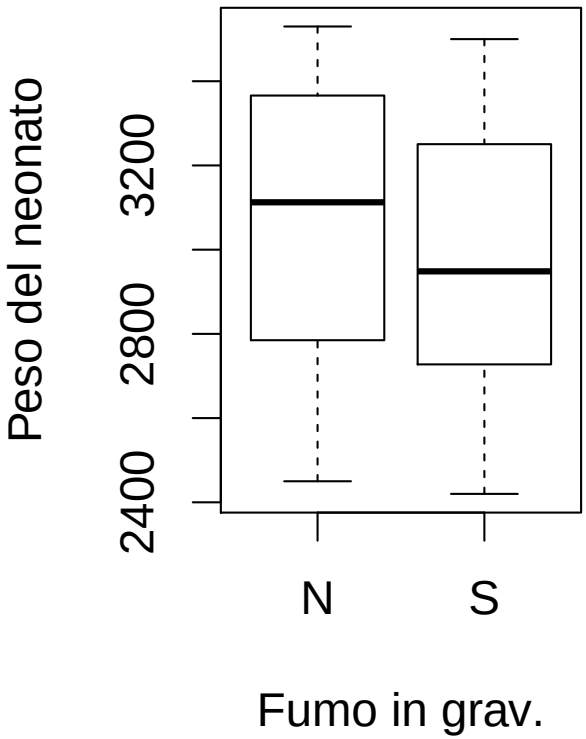
Sulla base del diagramma a dispersione sulla destra, quale delle seguenti affermazioni è corretta?



(a) Non c'è relazione tra numero di libri e numero di amici di Facebook;
(b) numero di libri e numero di amici di Facebook sono associati (positivamente);
(c) numero di libri e numero di amici di Facebook sono associati (negativamente);
(d) un maggior numero di libri letti causa un maggior/minor numero di amici su Facebook.

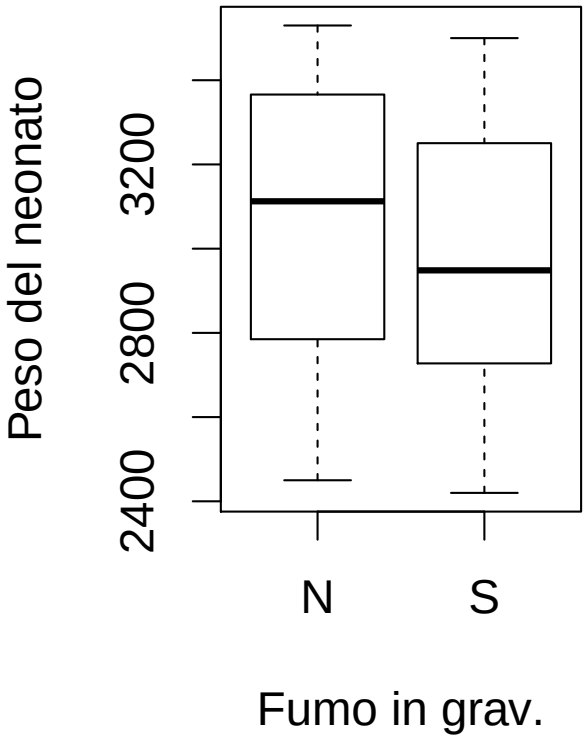
Esercizio

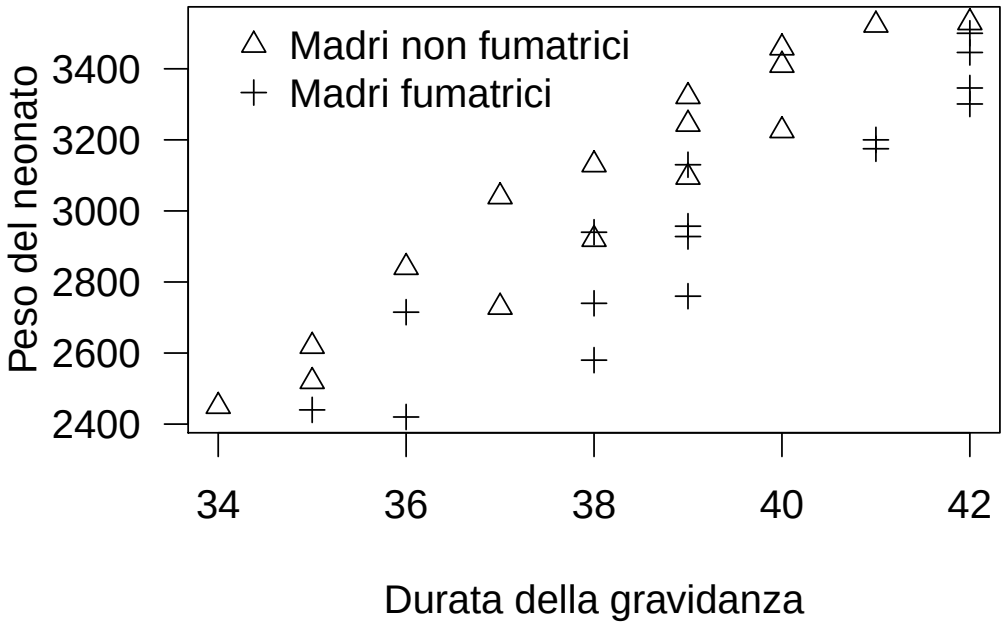
Per 32 neonati è noto il peso alla nascita, la durata della gravidanza, e se la madre fumasse durante la gravidanza. C'è associazione tra fumo e peso?



Esercizio

Per 32 neonati è noto il peso alla nascita, la durata della gravidanza, e se la madre fumasse durante la gravidanza. C'è associazione tra fumo e peso?





Riprendiamo il Titanic

Riprendiamo la tabella che abbiamo analizzato in precedenza

		1st	2nd	3rd	Crew	Totale
Morto	freq. ass.	122	167	528	673	1490
	% di colonna	37.5%	58.6%	74.8%	76.0%	67.7%
Sopravv.	freq. ass.	203	118	178	212	711
	% di colonna	62.5%	41.4%	25.2%	24.0%	32.3%
Totale		325	285	706	885	2201

I passeggeri di terza classe avevano minori chance di sopravvivere?

Riprendiamo il Titanic (cont)

Per rispondere, abbiamo guardato alla variabile condizionata Sopravvivenza|Tipologia. Sembrerebbe sensato affermare che l'esito dipende dalla classe.

	1st	2nd	3rd	Crew	Totale
Morto	37.5%	58.6%	74.8%	76.0%	67.7%
Sopravv.	62.5%	41.4%	25.2%	24.0%	32.3%
Totale	100%	100%	100%	100%	100%

Y (L'esito) dipende da X (la classe in cui viaggiava il passeggero) poiché le distribuzioni di Y condizionate ad X sono diverse nel senso che hanno frequenze relative diverse

Indipendenza in distribuzione

Diciamo che Y è **indipendente in distribuzione** da X se, per qualsivoglia $i = 1, \ldots, s$,

$$\frac{n_{i1}}{n_{01}} = \frac{n_{i2}}{n_{02}} = \ldots = \frac{n_{ij}}{n_{0j}} = \ldots = \frac{n_{it}}{n_{0t}}$$

Altrimenti, diremo che Y **dipende in distribuzione** da X.

Se le distribuzioni condizionate di Y dato X sono uguali tra di loro, allora sono anche uguali alla distribuzione marginale di Y.
L'uguaglianza, al solito, deve essere intesa nel senso delle frequenze relative.

Y	X					
	x_1	$\cdots$	x_j	$\cdots$	x_t	totale
y_1	$\frac{n_{11}}{n_{01}}$	$\cdots$	$\frac{n_{1j}}{n_{01}}$	$\cdots$	$\frac{n_{1t}}{n_{0t}}$	$\frac{n_{10}}{N}$
y_2	$\frac{n_{21}}{n_{01}}$	$\cdots$	$\frac{n_{2j}}{n_{01}}$	$\cdots$	$\frac{n_{2t}}{n_{0t}}$	$\frac{n_{20}}{N}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	$\frac{n_{i1}}{n_{01}}$	$\cdots$	$\frac{n_{ij}}{n_{01}}$	$\cdots$	$\frac{n_{it}}{n_{0t}}$	$\frac{n_{i0}}{N}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	$\frac{n_{s1}}{n_{01}}$	$\cdots$	$\frac{n_{sj}}{n_{01}}$	$\cdots$	$\frac{n_{st}}{n_{0t}}$	$\frac{n_{s0}}{N}$
totale	1	$\cdots$	1	$\cdots$	1	

Indipendenza in distribuzione (cont)

Per dimostrare la proposizione ci basta far vedere che

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

Ora, da $\frac{n_{ij}}{n_{0j}} = \frac{n_{i1}}{n_{01}}$ segue che $n_{ij} = \frac{n_{i1} n_{0j}}{n_{01}}$.
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

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

	x1	x2	x3	x4	marginale
y1	0.083	0.083	0.083	0.083	0.083
y2	0.500	0.500	0.500	0.500	0.500
y3	0.250	0.250	0.250	0.250	0.250
y4	0.167	0.167	0.167	0.167	0.167
totale	1.000	1.000	1.000	1.000	1.000

Simmetria dell'indipendenza in distribuzione

Se Y è indipendente da X allora X è indipendente da Y e viceversa.
Dimostrazione.
Se Y è indipendente da X allora

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i0}}{N},\ i = 1, \ldots, s;\ j = 1, \ldots, t. \tag{1}$$

Simmetria dell'indipendenza in distribuzione

Se Y è indipendente da X allora X è indipendente da Y e viceversa.
Dimostrazione.
Se Y è indipendente da X allora

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i0}}{N},\ i = 1, \ldots, s;\ j = 1, \ldots, t. \tag{1}$$

che può essere riscritta nella forma

$$\frac{n_{ij}}{n_{i0}} = \frac{n_{0j}}{N},\ i = 1, \ldots, r;\ j = 1, \ldots, c$$

ovvero, l'indipendenza in distribuzione di Y da X implica l'uguaglianza di tutte le distribuzioni condizionate di X dato Y alla distribuzione marginale di X.

Simmetria dell'indipendenza in distribuzione

Se Y è indipendente da X allora X è indipendente da Y e viceversa.
Dimostrazione.
Se Y è indipendente da X allora

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i0}}{N},\ i = 1, \ldots, s;\ j = 1, \ldots, t. \tag{1}$$

che può essere riscritta nella forma

$$\frac{n_{ij}}{n_{i0}} = \frac{n_{0j}}{N},\ i = 1, \ldots, r;\ j = 1, \ldots, c$$

ovvero, l'indipendenza in distribuzione di Y da X implica l'uguaglianza di tutte le distribuzioni condizionate di X dato Y alla distribuzione marginale di X.
Quindi, tutte le distribuzioni condizionate di X dato Y sono tra di loro uguali.

Esempio: indipendenza in distribuzione

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

	x1	x2	x3	x4	totale
y1	0.294	0.412	0.176	0.118	1.000
y2	0.294	0.412	0.176	0.118	1.000
y3	0.294	0.412	0.176	0.118	1.000
y4	0.294	0.412	0.176	0.118	1.000
marginale	0.294	0.412	0.176	0.118	1.000

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

	x1	x2	x3	x4	Sum
y1	$5 = \frac{60 \times 17}{204}$	$7 = \frac{84 \times 17}{204}$	3	2	17
y2	$30 = \frac{60 \times 12}{204}$	42	18	12	102
y3	15	21	9	6	51
y4	10	14	6	4	34
Sum	60	84	36	24	204

X^2

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 X^2 di Pearson

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

X^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.

X^2 (cont)

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

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 del carattere X corrisponde una sola modalità del carattere Y.

Quindi, si può costruire un indice **normalizzato**

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

che assumerà valori tra 0 e 1: $0 \leq \tilde{X}^2 \leq 1$.

Tabella di dipendenza perfetta

	x1	x2	x3	x4	Sum
y1	45	0	0	0	45
y2	0	20	0	0	20
y3	0	0	0	92	92
y4	0	0	37	0	37
Sum	45	20	37	92	194

Indice

Variabili statistiche bivariate

Associazione tra variabili

Dipendenza e indipendenza

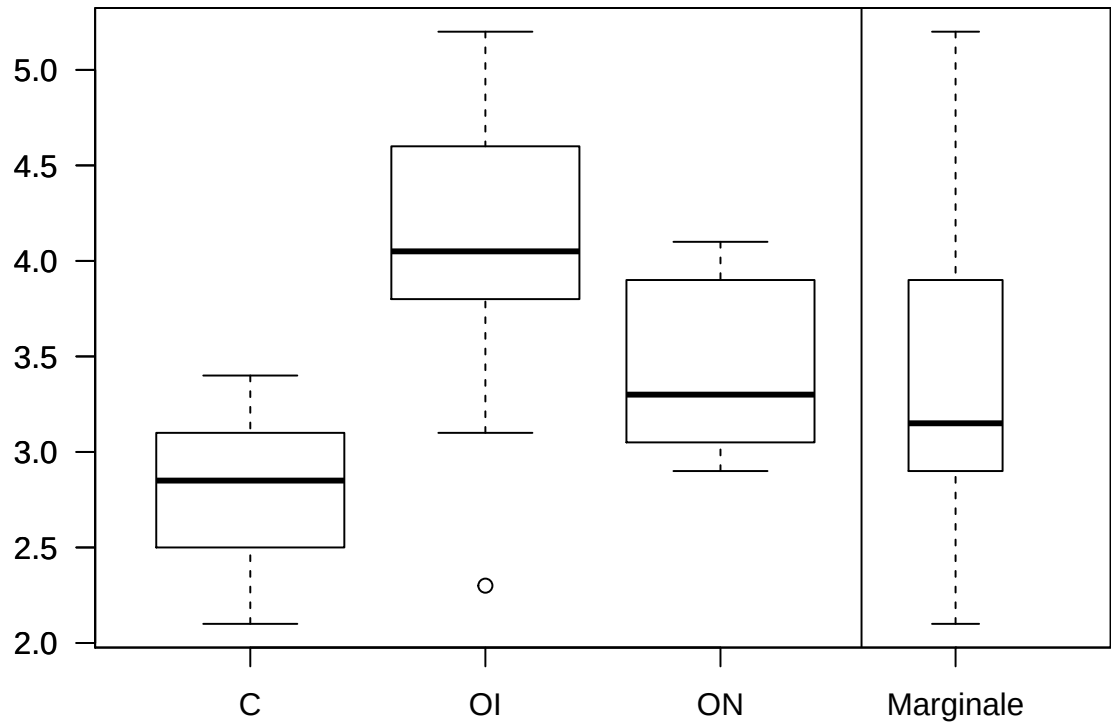
Dipendenza in media, mediana,. . .

Relazioni tra variabili

Indipendenza/Dipendenza in media

Se una delle due variabili è quantitativa, possiamo guardare alla cosa in temrini di indici di posizione

Esempio: colesterolo



Esempio: colesterolo

I boxplot appena visti ci dicono che

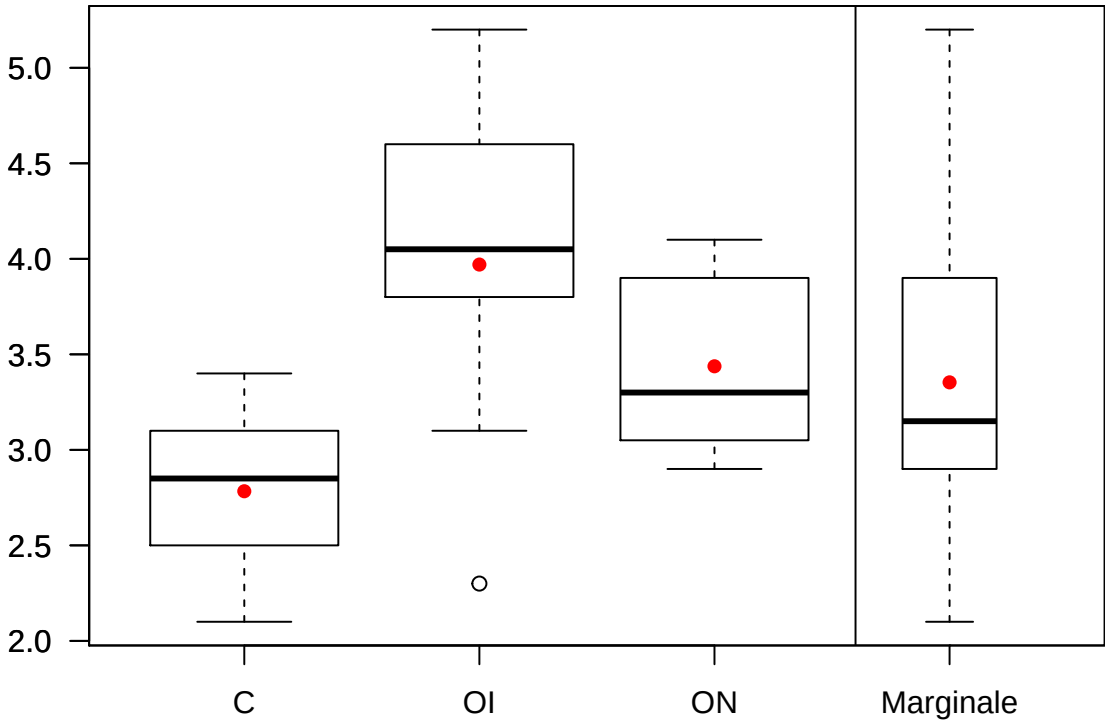
- le tre distribuzioni dei fosfati condizionate al Tipo di paziente" $(Y|X)$ sono diverse
- le 3 medie sono diverse.

Infatti

- Media di $Y|X = C = 2.783$
- Media di $Y|X = OI = 3.97$
- Media di $Y|X = ON = 3.438$
- Media marginale di $Y = 3.353$

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

Esempio: colesterolo



Indipendenza/Dipendenza in media

Una variabile, necessariamente numerica, Y è **indipendente in media** da un'altra variabile X, qualitativa o quantitativa, se le medie delle distribuzioni di Y condizionate alle varie modalità della X sono tutte uguali tra di loro.

Sempre in generale, l'applicazione $x_j \rightarrow \text{media}(Y|X = x_j)$ viene chiamata **funzione di regressione di Y su X**. Quindi, possiamo anche dire che abbiamo indipendenza in media se e solo se la funzione di regressione è costante.

In maniera analoga possiamo definire altri concetti di dipendenza/indipendenza (ad es. indipendenza in mediana, indipendenza in varianza,...).

Un'osservazione importante

Questi concetti di indipendenza sono **più deboli** di quello di indipendenza in distribuzione. Discutiamo questo punto con riferimento alla sola indipendenza in media.

Una tabella di contingenza generica è del tipo

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_t	totale
y_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	n_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{it}	n_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	n_{s1}	$\cdots$	n_{sj}	$\cdots$	n_{st}	n_{s0}
totale	n_{01}	$\cdots$	n_{0j}	$\cdots$	n_{0t}	N

Indip distribuzione $\Rightarrow$ Indip media

Supponiamo che Y sia una variabile numerica. Allora è immediato verificare che **se esiste indipendenza in distribuzione tra Y e X allora esiste anche indipendenza in media**.

Se esiste indipendenza in distribuzione, si ha, per esempio,

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i1}}{n_{01}}.$$

Ma essendo

$$\bar{y}_j = \frac{\sum_{i=1}^{s} y_i n_{ij}}{n_{0j}},$$

è immediato, sostituendo, vedere che

$$\bar{y}_j = \frac{\sum_{i=1}^{s} y_i n_{i1}}{n_{01}} = \bar{y}_1.$$

Indip distribuzione $\Rightarrow$ Indip media

Supponiamo che Y sia una variabile numerica. Allora è immediato verificare che **se esiste indipendenza in distribuzione tra Y e X allora esiste anche indipendenza in media**.

Se esiste indipendenza in distribuzione, si ha, per esempio,

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i1}}{n_{01}}.$$

Ma essendo

$$\bar{y}_j = \frac{\sum_{i=1}^{s} y_i n_{ij}}{n_{0j}},$$

è immediato, sostituendo, vedere che

$$\bar{y}_j = \frac{\sum_{i=1}^{s} y_i n_{i1}}{n_{01}} = \bar{y}_1.$$

Quindi la media di $Y|X = x_j$ è uguale alla media di $Y|X = x_1$, $\forall j$.

Indip media $\not\Rightarrow$ Indip distribuzione

Dall'altra parte è facile costruire tabelle in cui esiste indipendenza in media ma non indipendenza in distribuzione:

	X		
Y	x_1	x_2	totale
-2	0	1	1
-1	1	0	1
0	1	1	2
1	1	0	1
2	0	1	1
totale	3	3	6

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

Notazione

μ_Y : media di $Y \longrightarrow \frac{\sum_{i=1}^{s} y_i n_{i0}}{N}$;

σ_Y^2 : varianza di $Y \longrightarrow \frac{\sum_{i=1}^{s} (y_i - \mu_Y)^2 n_{i0}}{N}$;

D_Y : devianza di $Y \longrightarrow \sum_{i=1}^{s} (y_i - \mu_Y)^2 n_{i0} = N\sigma_Y^2$;

$\mu_Y(x_j)$ media di $Y|X = x_j \longrightarrow \frac{\sum_{i=1}^{s} y_i n_{ij}}{n_{0j}}$.

$\sigma_Y^2(x_j)$ varianza di $Y|X = x_j \longrightarrow \frac{\sum_{i=1}^{s} (y_i - \mu_Y(x_j))^2 n_{ij}}{n_{0j}}$.

$D_Y(x_j)$ devianza di $Y|X = x_j \longrightarrow \sum_{i=1}^{s} (y_i - \mu_Y(x_j))^2 n_{ij} = n_{0j}\sigma_Y^2(x_j)$.

σ_S^2 : varianza tra $\longrightarrow \frac{\sum_{j=1}^{t} (\mu_Y(x_j) - \mu_Y)^2 n_{0j}}{N}$;

D_S : devianza tra $\longrightarrow \sum_{j=1}^{t} (\mu_Y(x_j) - \mu_Y)^2 n_{0j}$;

σ_R^2 : varianza entro $\longrightarrow \frac{\sum_{j=1}^{t} \sigma_Y^2(x_j) n_{0j}}{N} = \frac{\sum_{j=1}^{t} \sum_{i=1}^{s} (y_i - \mu_Y(x_j))^2 n_{ij}}{N}$;

D_R : devianza entro $\longrightarrow \sum_{j=1}^{t} \sigma_Y^2(x_j) n_{0j} = \sum_{j=1}^{t} \sum_{i=1}^{s} (y_i - \mu_Y(x_j))^2 n_{ij}$;

Terminologia

D_S : devianza **tra i gruppi** $\longrightarrow$ devianza **spiegata**;

D_S/N : varianza **tra i gruppi** $\longrightarrow$ varianza **spiegata**;

D_R : devianza **entro i gruppi** $\longrightarrow$ devianza **residua**;

D_R/N : varianza **entro i gruppi** $\longrightarrow$ varianza **residua**.

$$D_Y = D_S + D_R$$

$$\sigma^2_Y = \sigma^2_S + \sigma^2_R$$

Misura della dipendenza in media

- Se la varianza tra i gruppi è nulla, le medie condizionate sono tutte uguali a $\overline{y}$ e quindi esiste indipendenza in media.
- Se la varianza tra i gruppi è molto grande rispetto alla varianza entro i gruppi, allora buona parte della variabilità totale mostrata dai dati diventa interpretabile in termini di differenze tra le medie condizionate. ($\rightarrow$ le differenze tra le medie "spiegano" una larga frazione delle differenze che osserviamo nei dati).

Rapporto di correlazione

Per misurare la forza della dipendenza in media, sembra allora ragionevole usare

$$\eta^2 = \eta^2_Y = \frac{D_S}{D_Y} = 1 - \frac{D_R}{D_Y}$$

In altre parole

$$\eta^2_Y = \frac{\text{varianza tra i gruppi}}{\text{varianza totale}} = 1 - \frac{\text{varianza entro i gruppi}}{\text{varianza totale}}$$

equivalente a

$$\eta^2_Y = \frac{\text{devianza tra i gruppi}}{\text{devianza totale}} = 1 - \frac{\text{devianza entro i gruppi}}{\text{devianza totale}}$$

η^2_Y

Per misurare la forza della dipendenza in media, sembra allora ragionevole usare

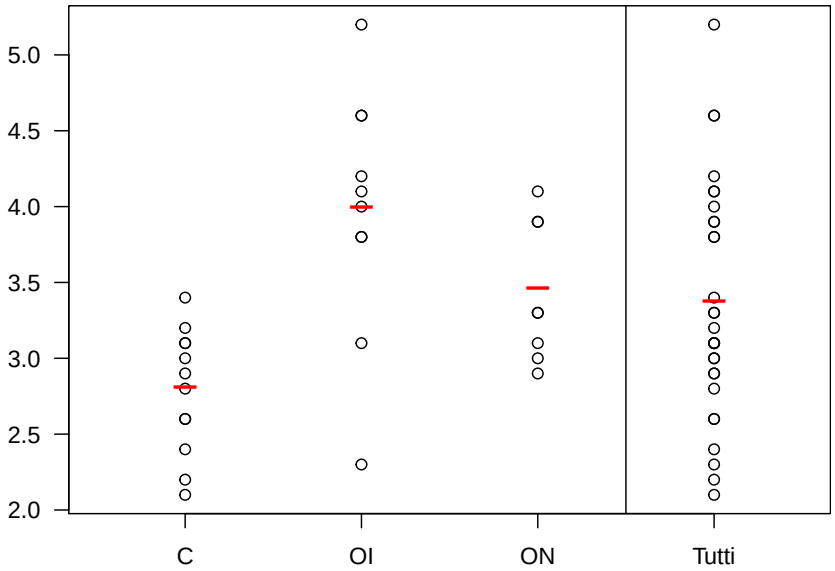
$$\eta^2 = \eta^2_Y = \frac{D_S}{D_Y} = 1 - \frac{D_R}{D_Y}$$

$$0 \leq \eta^2_Y \leq 1$$

- $\eta^2 = 0$: indipendenza in media
- $\eta^2 = 1$: dipendenza perfetta

η^2 non è ovviamente né definito né sensato quando $\sigma^2_Y = 0$.

η^2_Y, esempio colesterolo



x_j	n_{0j}	$\mu_Y(x_j)$	$\sigma^2_Y(x_j)$
C	12	2.78	0.1531
OI	10	3.97	0.5981
ON	8	3.44	0.1873

varianza entro i gruppi ≈ 0.31055

varianza tra i gruppi ≈ 0.25861

varianza totale ≈ 0.56916

e, quindi, $\eta^2 \approx 0.45437$. Il valore trovato indica che siamo in presenza di un legame di dipendenza in media di discreta entità.

Scomposizione della varianza: a cosa serve

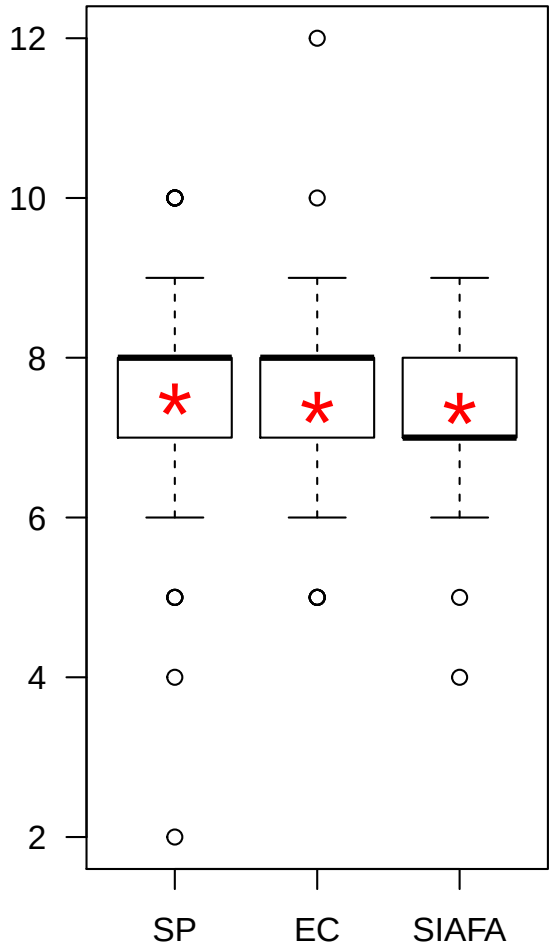
È uno strumento per studiare in che misura dei gruppi di unità differiscano in media rispetto a una variabile quantitativa.

Con riferimento ai dati raccolti sugli studenti, dove i gruppi sono i CdL (Economia, Scienze Politiche, Statistica), ci si può porre questo problema relativamente a diverse variabili

- Altezza/Peso
- Ore di Sonno
- Ore di Studio
- Voto di maturità

Cosa ci aspettiamo?

Scomposizione della varianza: Ore di sonno



	N_j	$\bar{y}_j$	σ_j^2
SP	236	7.55	1.15
EC	244	7.45	0.80
SIAFA	53	7.43	1.08

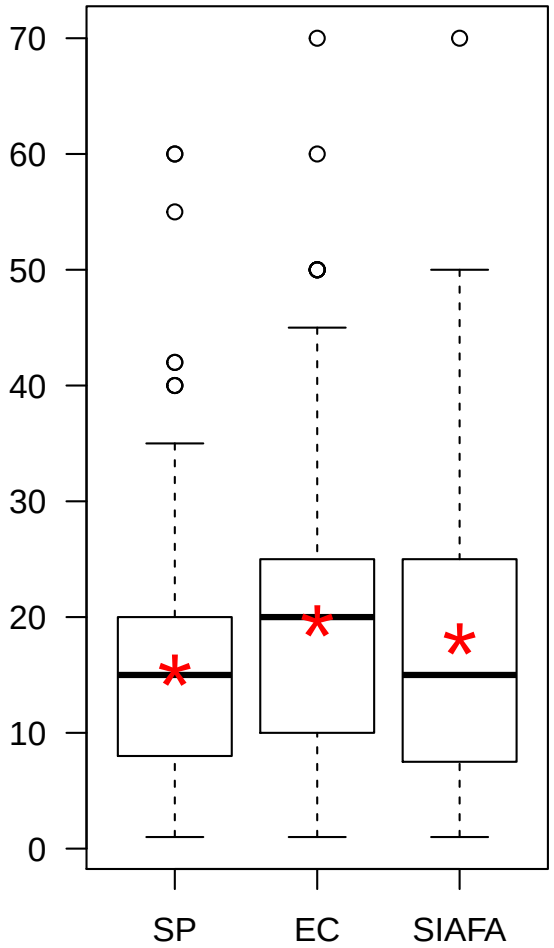
La scomposizione è come segue

$$\begin{aligned} \frac{1}{N}\sum_{j=1}^{L} N_j \sigma_j^2 &= 0.9826 \\ \frac{1}{N}\sum_{j=1}^{L} N_j (\bar{y}_j - \bar{y})^2 &= 0.002807 \\ \hline \sigma^2 &= 0.9854 \end{aligned}$$

Un indice di quanto diversi sono i gruppi è

$$\eta^2 = \frac{0.002807}{0.9854} = 0.002849$$

Scomposizione della varianza: Ore di studio



	N_j	$\bar{y}_j$	σ_j^2
SP	225	15.89	104.92
EC	238	20.15	132.93
SIAFA	51	18.59	183.10

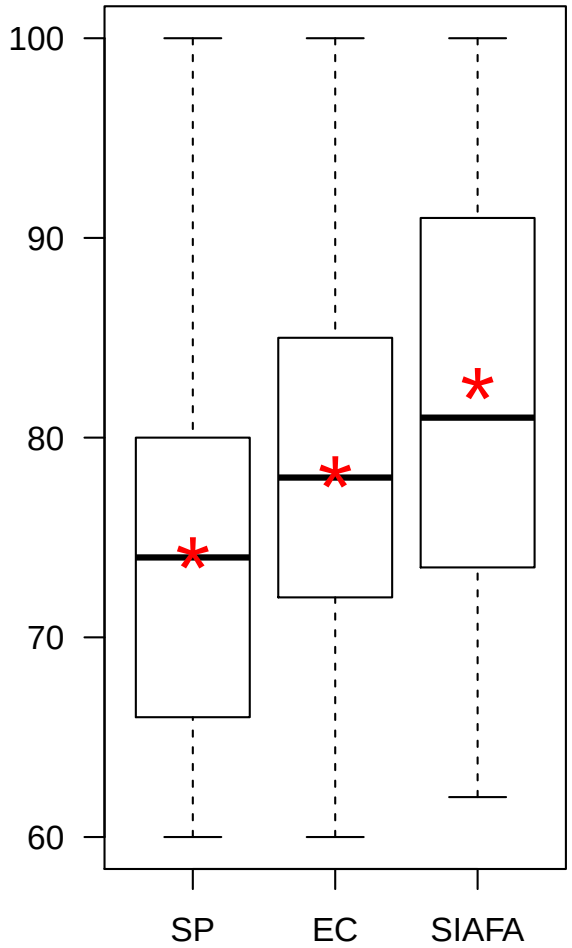
La scomposizione è come segue

$$\begin{array}{lcl} \frac{1}{N}\sum_{j=1}^{L} N_j\sigma_j^2 & = & 125.6 \\ \frac{1}{N}\sum_{j=1}^{L} N_j(\bar{y}_j - \bar{y})^2 & = & 4.103 \\ \hline \sigma^2 & = & 129.8 \end{array}$$

Un indice di quanto diversi sono i gruppi è

$$\eta^2 = \frac{4.103}{129.8} = 0.03161$$

Scomposizione della varianza: Voto di matura



	N_j	$\bar{y}_j$	σ_j^2
SP	230	74.50	108.28
EC	246	78.56	95.96
SIAFA	47	82.98	127.04

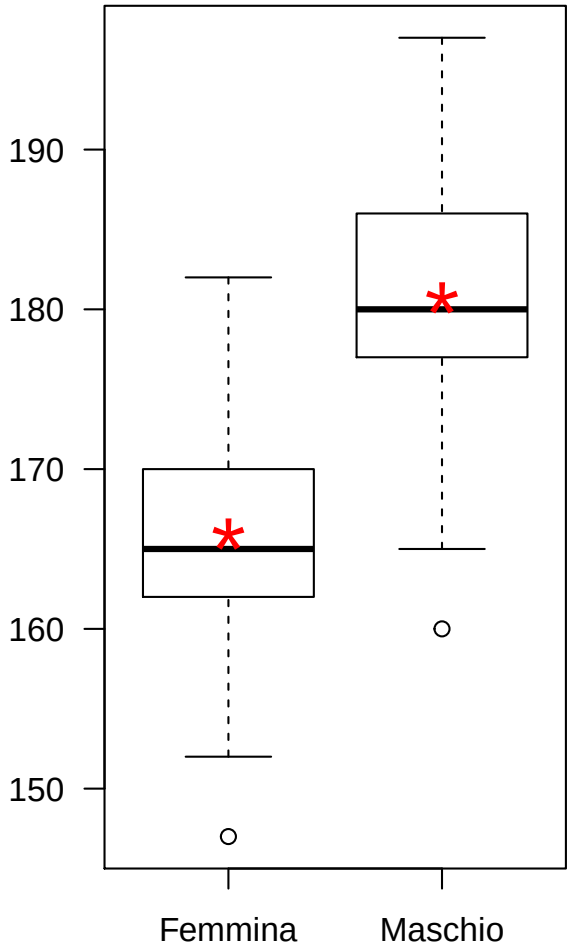
La scomposizione è come segue

$$\begin{array}{rcl} \frac{1}{N}\sum_{j=1}^{L} N_j \sigma_j^2 & = & 104.2 \\ \frac{1}{N}\sum_{j=1}^{L} N_j (\bar{y}_j - \bar{y})^2 & = & 7.067 \\ \hline \sigma^2 & = & 111.2 \end{array}$$

Un indice di quanto diversi sono i gruppi è

$$\eta^2 = \frac{7.067}{111.2} = 0.06355$$

Scomposizione della varianza: Altezza e Genere



	N_j	$\bar{y}_j$	σ_j^2
Femmina	244	166.28	32.91
Maschio	283	181.06	40.77

La scomposizione è come segue

$\frac{1}{N}\sum_{j=1}^{L} N_j\sigma_j^2$	$=$	37.13
$\frac{1}{N}\sum_{j=1}^{L} N_j(\bar{y}_j - \bar{y})^2$	$=$	54.32
σ^2	$=$	91.45

Un indice di quanto diversi sono i gruppi è

$$\eta^2 = \frac{54.32}{91.45} = 0.594$$

Indice

Variabili statistiche bivariate

Associazione tra variabili

Relazioni tra variabili

Dipendenza in media

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

Il rapporto di correlazione misura il grado di dipendenza in media per dati in forma di una distribuzione doppia di frequenza ed è tanto maggiore quanto più le medie differiscono tra loro.

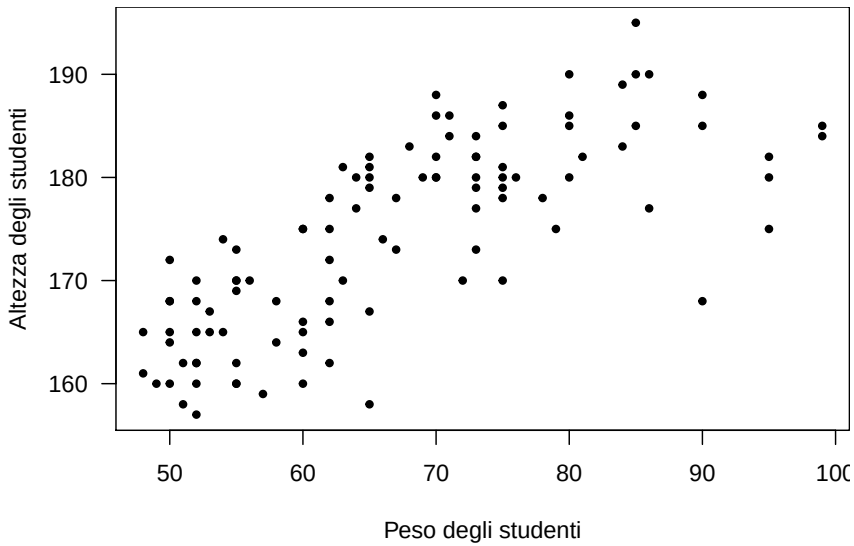
Se sia X che Y sono caratteri quantitativi, per studiare la dipendenza disponiamo di un ulteriore strumento.

Descrivere la dipendenza: diagramma 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 doppia disaggregata della variabile doppia (X, Y).

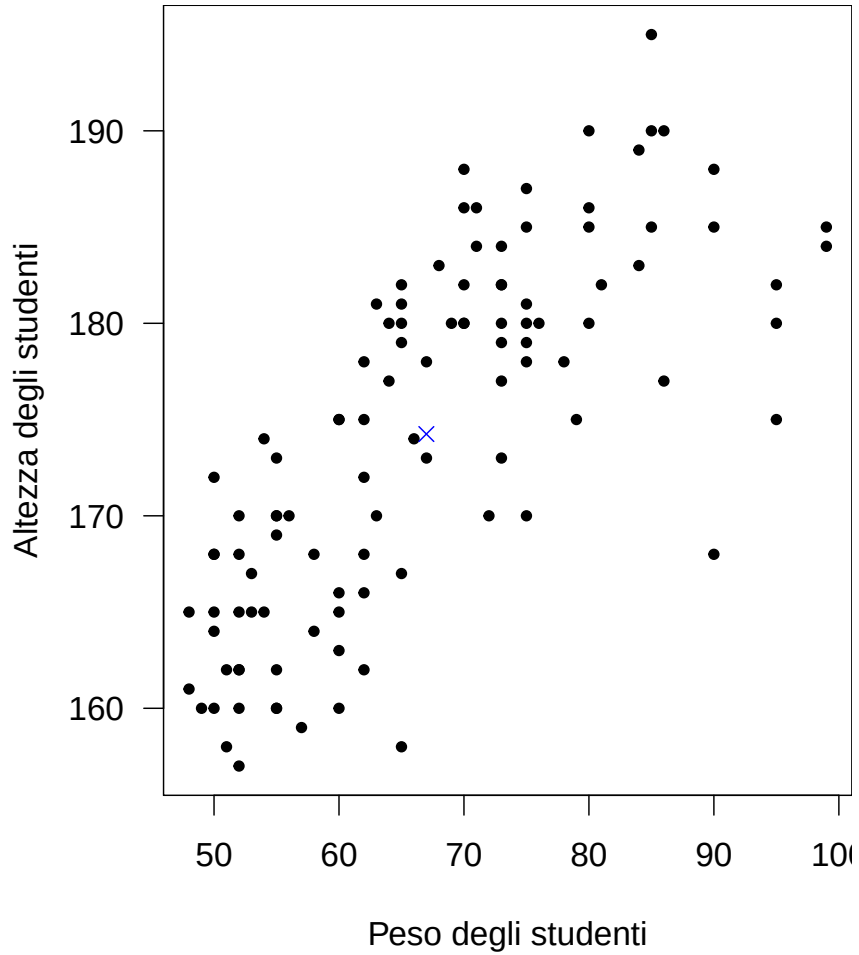


- Si dice che tra X e Y c'è associazione **positiva** quando essi tendono a crescere insieme.
- Si dice che tra X e Y c'è associazione **negativa** quando essi tendono decrescere insieme.

Il diagramma di dispersione è uno strumento per esplorare graficamente la presenza di associazione positiva o negativa tra due caratteri.

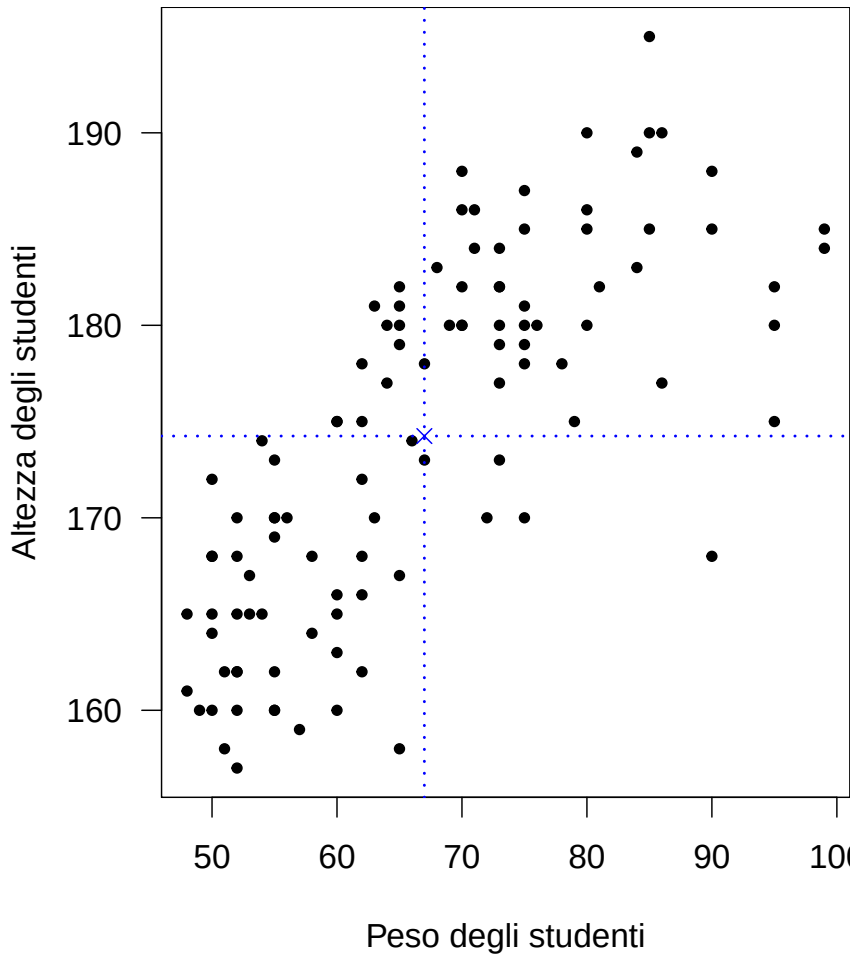
Misurare l'associazione

- La x blu è il punto di coordinate $(\bar{x}, \bar{y})$.



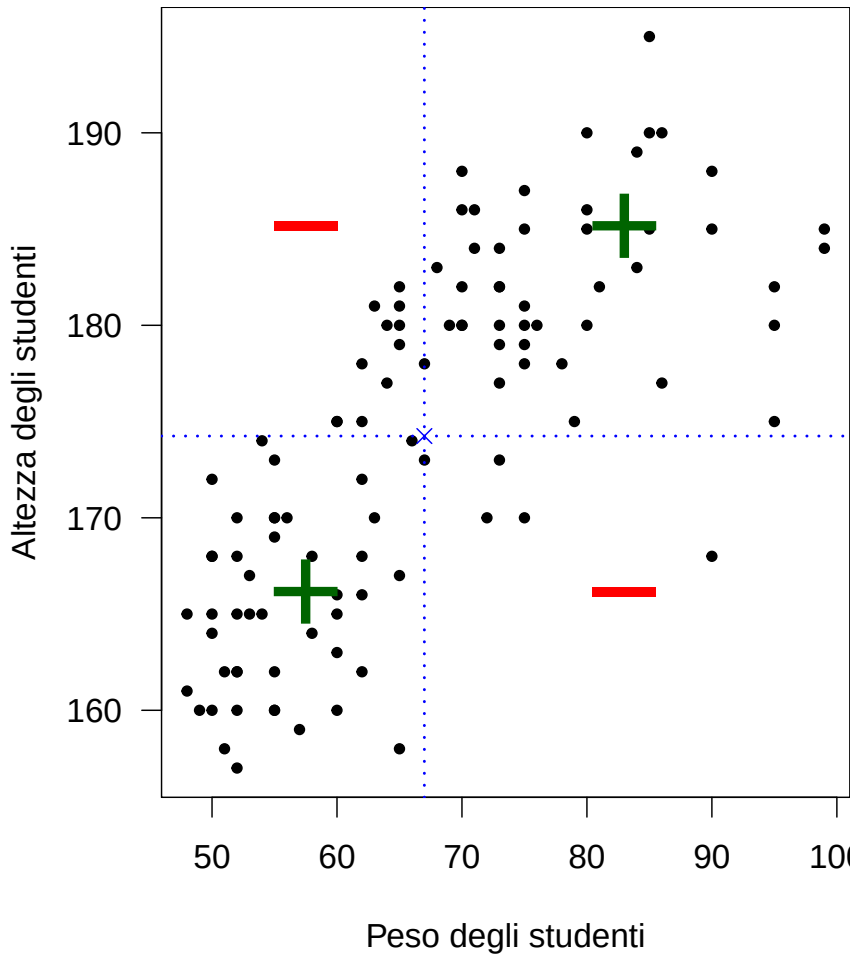
Misurare l'associazione

- La x blu è il punto di coordinate $(\bar{x}, \bar{y})$.
- Valori maggiori della media di X corrispondono a valori maggiori della media per Y.
- Valori inferiori alla media di X corrispondono a valori inferiori alla media per Y.



Misurare l'associazione

- La x blu è il punto di coordinate $(\bar{x}, \bar{y})$.
- Valori maggiori della media di X corrispondono a valori maggiori della media per Y.
- Valori inferiori alla media di X corrispondono a valori inferiori alla media per Y.
- più osservazioni cadono nelle regioni contrassegnate da un "+" rispetto a quante ne cadono nelle regioni contrassegnate da un "-" più è manifesta l'associazione.



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.

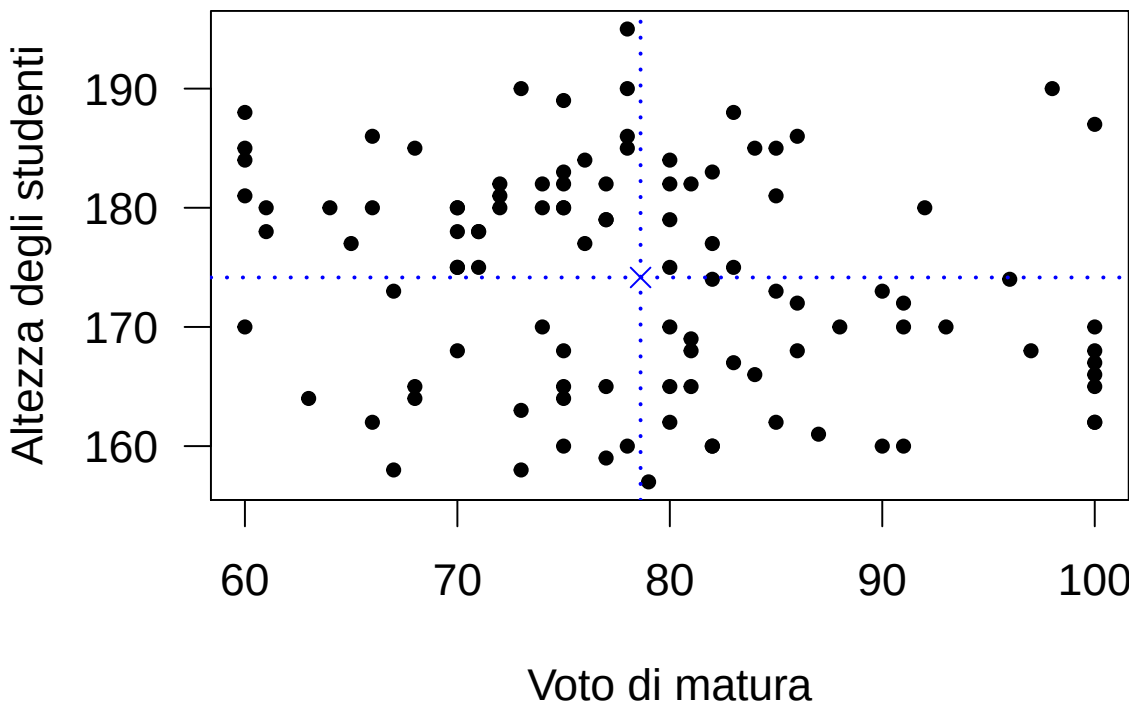
σ_{XY} è detta **covarianza**. Il suo numeratore è detto **codevianza**, indicata con C_{XY}.

La covarianza (cont)

1. In presenza di una qualche forma di relazione monotona, 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 monotona 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.

Covarianza: esempio



Calcolo della covarianza

Per il calcolo della covarianza si ha la seguente formula

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

ovvero

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

Calcolo della covarianza

Per il calcolo della covarianza si ha la seguente formula

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

Per dimostrarlo si scriva

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i(y_i - \overline{y}) - \frac{\overline{x}}{N}\sum_{i=1}^{N}(y_i - \overline{y})$$

Dove

$$\frac{\overline{x}}{N}\sum_{i=1}^{N}(y_i - \overline{y}) = 0$$

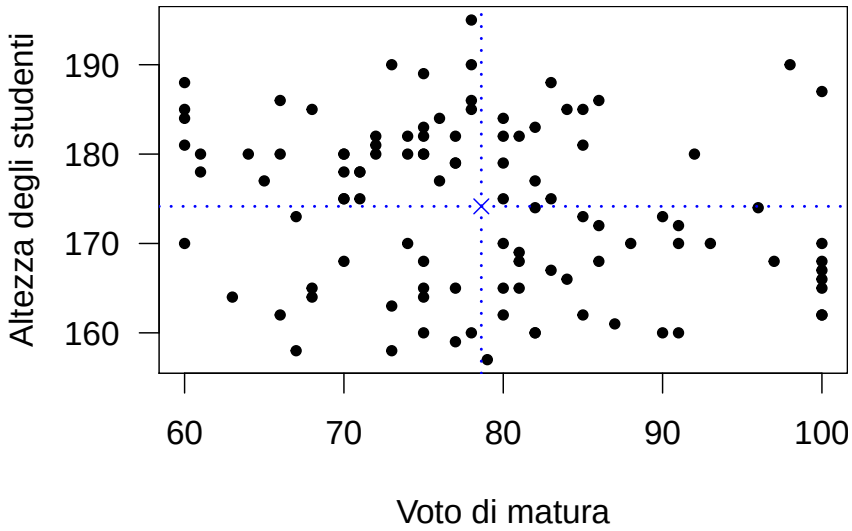
$$\frac{1}{N}\sum_{i=1}^{N} x_i(y_i - \overline{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{y}\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\,\overline{y}.$$

La varianza della somma di due variabili

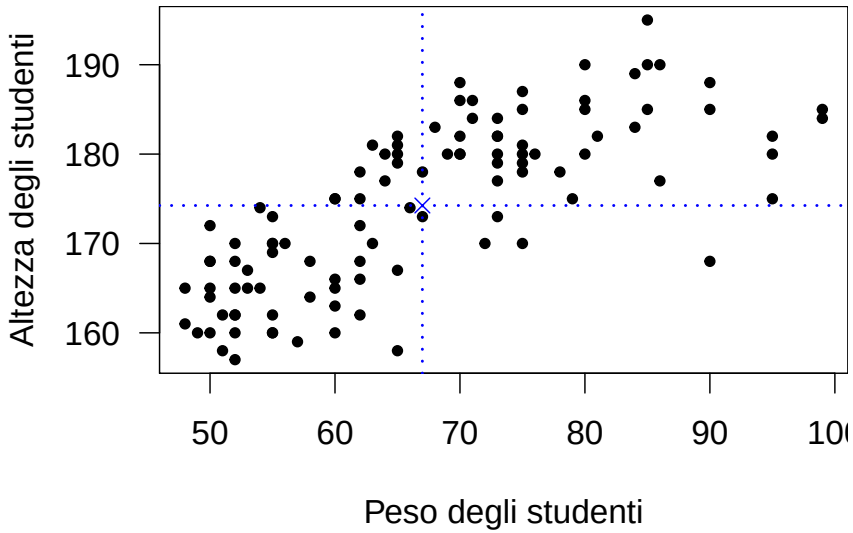
La covarianza risulta utile anche quando vogliamo ottenere la varianza di una nuova variabile che sia la somma di altre due variabili. Infatti

$$
\begin{aligned}
\text{var}(X+Y) &= \frac{1}{N}\sum_{i=1}^{n}((x_i+y_i)-\text{media}(X+Y))^2 = \\
&= \frac{1}{N}\sum_{i=1}^{N}((x_i+y_i)-(\bar{x}+\bar{y}))^2 = \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2+\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2+\frac{2}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \\
&= \sigma_X^2+\sigma_Y^2+2\sigma_{XY}
\end{aligned}
$$

Covarianza: esempio

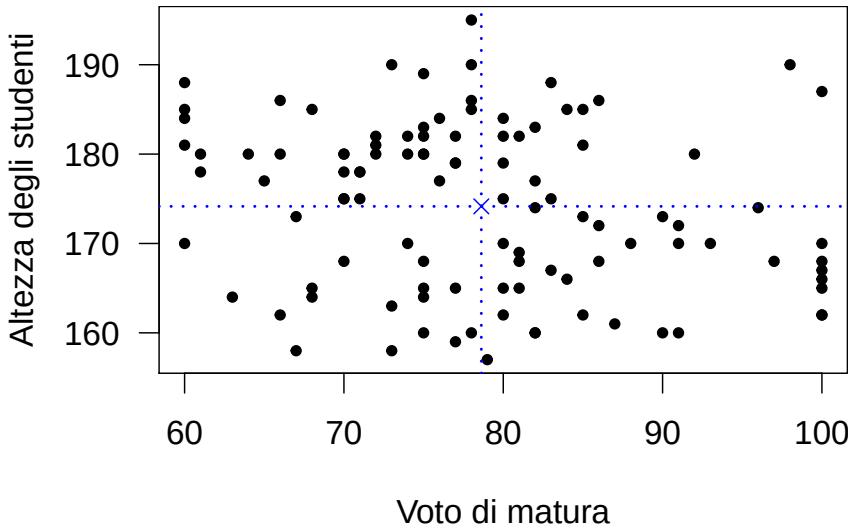


$$\frac{1}{N}\sum_i x_i y_i = 1.36705 \times 10^4$$
$$\bar{x} = 78.6226$$
$$\bar{y} = 174.151$$
$$\sigma_{XY} = -21.6789$$

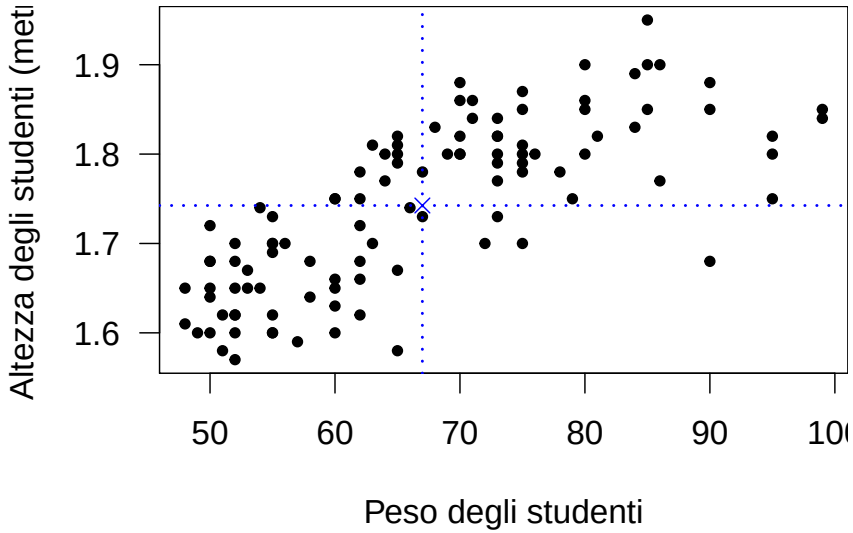


$$\frac{1}{N}\sum_i x_i y_i = 1.17631 \times 10^4$$
$$\bar{x} = 66.9905$$
$$\bar{y} = 174.248$$
$$\sigma_{XY} = 90.1928$$

Covarianza: esempio



$$\frac{1}{N}\sum_i x_i y_i = 1.36705 \times 10^4$$
$$\bar{x} = 78.6226$$
$$\bar{y} = 174.151$$
$$\sigma_{XY} = -21.6789$$



$$\frac{1}{N}\sum_i x_i y_i = 117.631$$
$$\bar{x} = 66.9905$$
$$\bar{y} = 1.74248$$
$$\sigma_{XY} = 0.901928$$

Grande quanto?

L'esempio su altezza e peso 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 le due grandezze

---

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

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

Per dimostrarlo, consideriamo la seguente trasformazione di scala per X e Y:

$$X^* = \frac{X}{\sigma_X} \quad \text{e} \quad Y^* = \frac{Y}{\sigma_Y}.$$

Sappiamo che $\sigma^2_{X^*} = \sigma^2_{Y^*} = 1$.
Calcoliamo la varianza della variabile $X^* + Y^*$.

Grande quanto? (cont)

Calcoliamo la varianza della variabile $X^* + Y^*$.

$$\begin{aligned}
\text{var}\left(X^* + Y^*\right) &= \sigma^2_{X^*} + \sigma^2_{Y^*} + 2\text{cov}\left(X^*, Y^*\right) = \\
&= 1 + 1 + 2\frac{1}{N}\sum_{i=1}^{N}\left[\left(x_i^* - \overline{x^*}\right)\left(y_i^* - \overline{y^*}\right)\right] = \\
&= 2 + \frac{2}{N}\sum_{i=1}^{N}\left[\left(\frac{x_i}{\sigma_X} - \frac{\bar{x}}{\sigma_X}\right)\left(\frac{y_i}{\sigma_Y} - \frac{\bar{y}}{\sigma_Y}\right)\right] = \\
&= 2 + \frac{2}{\sigma_X\sigma_Y}\left\{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x})(y_i - \bar{y})\right]\right\} = \\
&= 2\left(1 + \frac{1}{\sigma_X\sigma_Y}\sigma_{XY}\right)
\end{aligned}$$

Grande quanto? (cont)

Ora poiché quest'ultimo valore ottenuto è uguale a una varianza, non può essere negativo, cioè

$$2\left(1+\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right)\geq 0$$

Grande quanto? (cont)

Ora poiché quest'ultimo valore ottenuto è uguale a una varianza, non può essere negativo, cioè

$$2\left(1+\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right)\geq 0$$

$$\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\geq -1$$

Grande quanto? (cont)

Ora poiché quest'ultimo valore ottenuto è uguale a una varianza, non può essere negativo, cioè

$$2\left(1+\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right)\geq 0$$

$$\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\geq -1$$

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

Grande quanto? (cont)

Ora poiché quest'ultimo valore ottenuto è uguale a una varianza, non può essere negativo, cioè

$$2\left(1+\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right)\geq 0$$

$$\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\geq -1$$

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

Inoltre

$$\text{var}\left(X^*+Y^*\right)=0 \iff X^*+Y^*=\text{cost.}$$

Grande quanto? (cont)

Ora poiché quest'ultimo valore ottenuto è uguale a una varianza, non può essere negativo, cioè

$$2\left(1+\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right)\geq 0$$

$$\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\geq -1$$

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

Inoltre

$$\operatorname{var}(X^*+Y^*)=0 \iff X^*+Y^*=\text{cost.}$$

Quindi

$$\sigma_{XY}=-\sigma_X\sigma_Y \iff Y^*=\text{cost}-X^*$$

cioè Y è funzione lineare decrescente di X e viceversa.

Grande quanto? (cont)

Calcolando poi la varianza tra X^* e $-Y^*$ si ottiene

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

e

$$\sigma_{XY} = \sigma_X \sigma_Y \iff Y^* = \text{cost} + X^*$$

cioè se e solo se Y è funzione lineare crescente di X e viceversa.

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 degli scarti quadratici medi.

In altre parole, dobbiamo costruire l'indice normalizzato, chiamato **coefficiente di correlazione (lineare)**

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

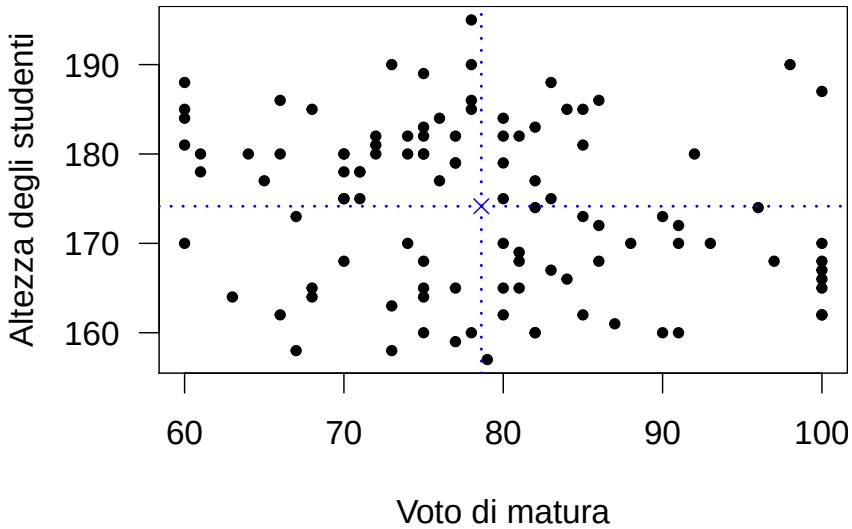
Il coefficiente di correlazione è spesso indicato con la lettera greca ρ.

Interpretazione di r

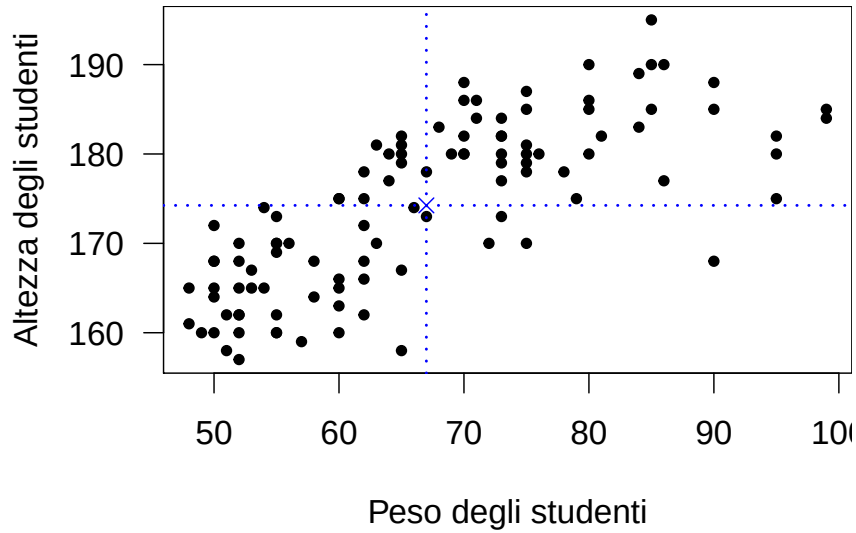
$$-1 \leq r \leq +1$$

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

Covarianza: esempio

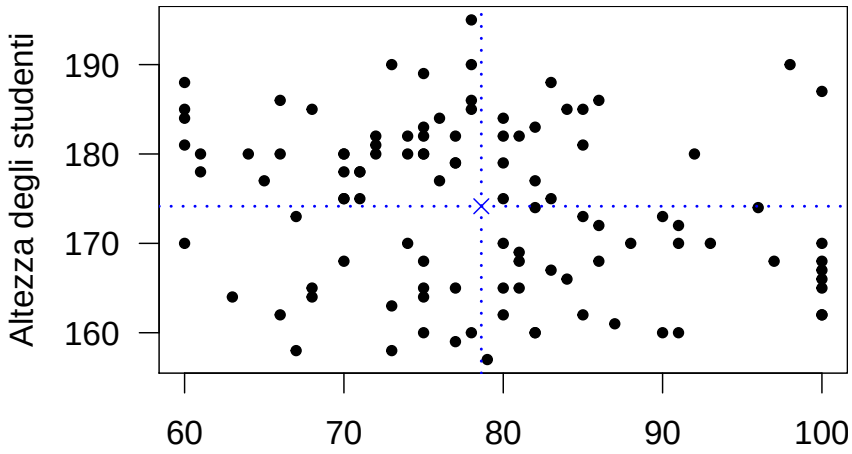


$$\frac{1}{N}\sum_i x_i y_i = 1.36705 \times 10^4$$
$$\bar{x} = 78.6226$$
$$M(x^2) = 6291.06$$
$$\sigma_X^2 = 109.537$$
$$\bar{y} = 174.151$$
$$M(y^2) = 3.04148 \times 10^4$$
$$\sigma_Y^2 = 86.2414$$
$$\sigma_{XY} = -21.6789$$
$$\rho_{XY} = \frac{-21.6789}{\sqrt{110 \times 86.2}} = -0.223$$

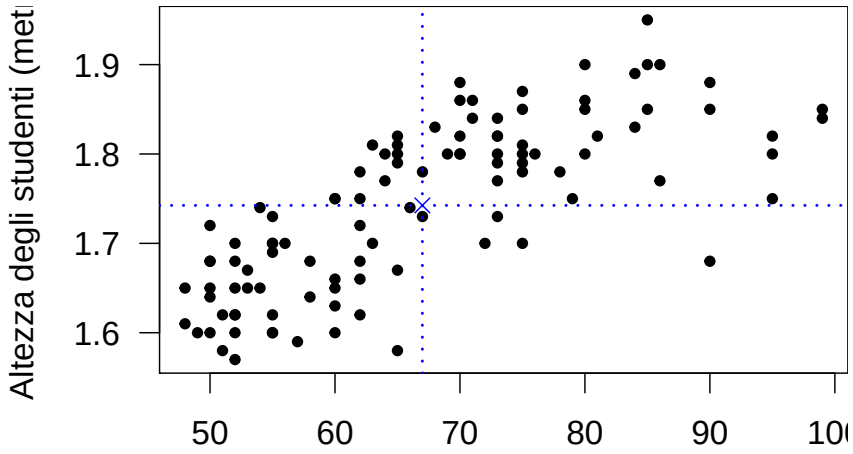


$$\frac{1}{N}\sum_i x_i y_i = 1.17631 \times 10^4$$
$$\bar{x} = 66.9905$$
$$M(x^2) = 4657.16$$
$$\sigma_X^2 = 169.438$$
$$\bar{y} = 174.248$$
$$M(y^2) = 3.04483 \times 10^4$$
$$\sigma_Y^2 = 86.072$$
$$\sigma_{XY} = 90.1928$$
$$\rho_{XY} = \frac{90.1928}{\sqrt{169 \times 86.1}} = 0.7469$$

Covarianza: esempio



$$\frac{1}{N}\sum_i x_i y_i = 1.36705 \times 10^4$$
$$\bar{x} = 78.6226$$
$$M(x^2) = 6291.06$$
$$\sigma_X^2 = 109.537$$
$$\bar{y} = 174.151$$
$$M(y^2) = 3.04148 \times 10^4$$
$$\sigma_Y^2 = 86.2414$$
$$\sigma_{XY} = -21.6789$$
$$\rho_{XY} = \frac{-21.6789}{\sqrt{110 \times 86.2}} = -0.223$$



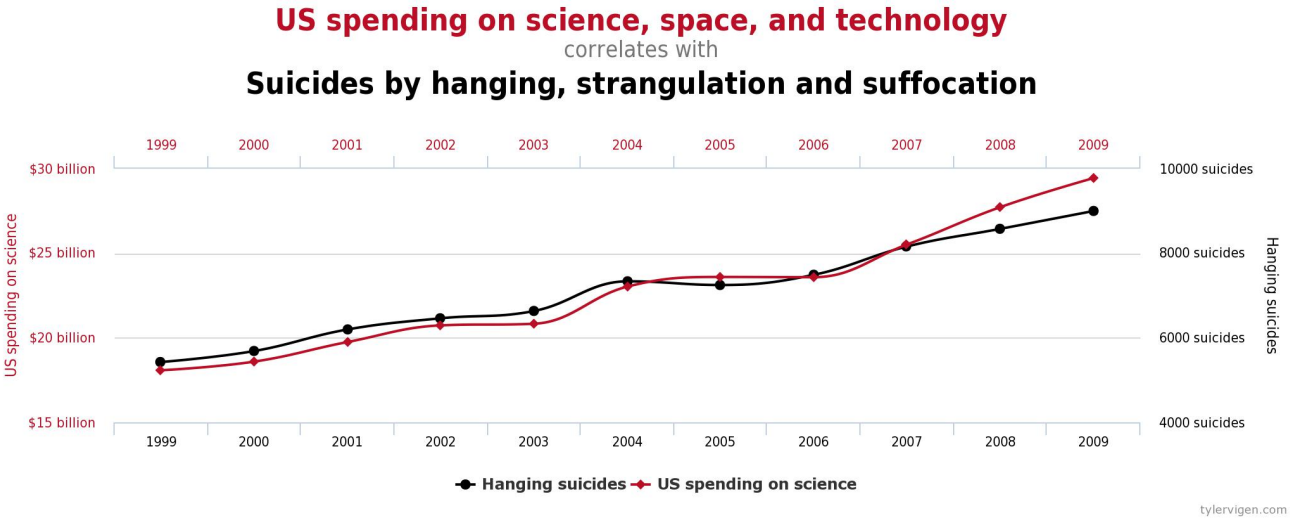
$$\frac{1}{N}\sum_i x_i y_i = 117.631$$
$$\bar{x} = 66.9905$$
$$M(x^2) = 4657.16$$
$$\sigma_X^2 = 169.438$$
$$\bar{y} = 1.74248$$
$$M(y^2) = 3.04483$$
$$\sigma_Y^2 = 0.0086072$$
$$\sigma_{XY} = 0.901928$$
$$\rho_{XY} = \frac{0.901928}{\sqrt{169 \times 0.00861}} = 0.7469$$

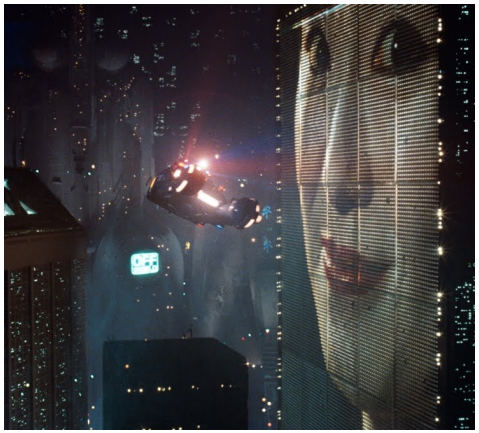
Relazione $\not\Rightarrow$ causa ed effetto

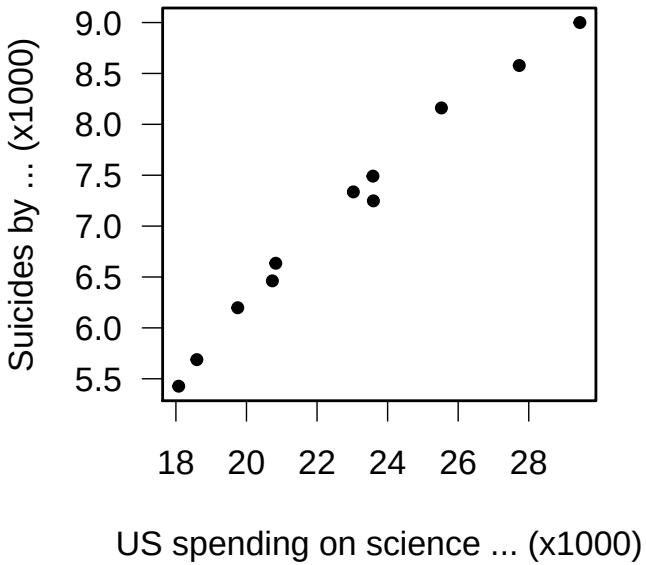
Attenzione all'interpretazione!

- Quando mettiamo in relazione due variabili e troviamo una forte associazione è forte la tentazione di interpretarlo come se x "causasse" y o viceversa.
- Una relazione statistica, anche stretta, tra y e x **non implica** una relazione causa effetto.
- Per esempio, entrambe potrebbero essere legate a una terza variabile, che le "causa" entrambe.
- Ci sono metodi statistici per l'inferenza su relazioni causa effetto ma richiedono una maggiore sofisticazione oppure un campione costruito in un certo modo.
- I prossimi esempi vengono dal sito www.tylervigen.com

La scienza rende il mondo arido e triste



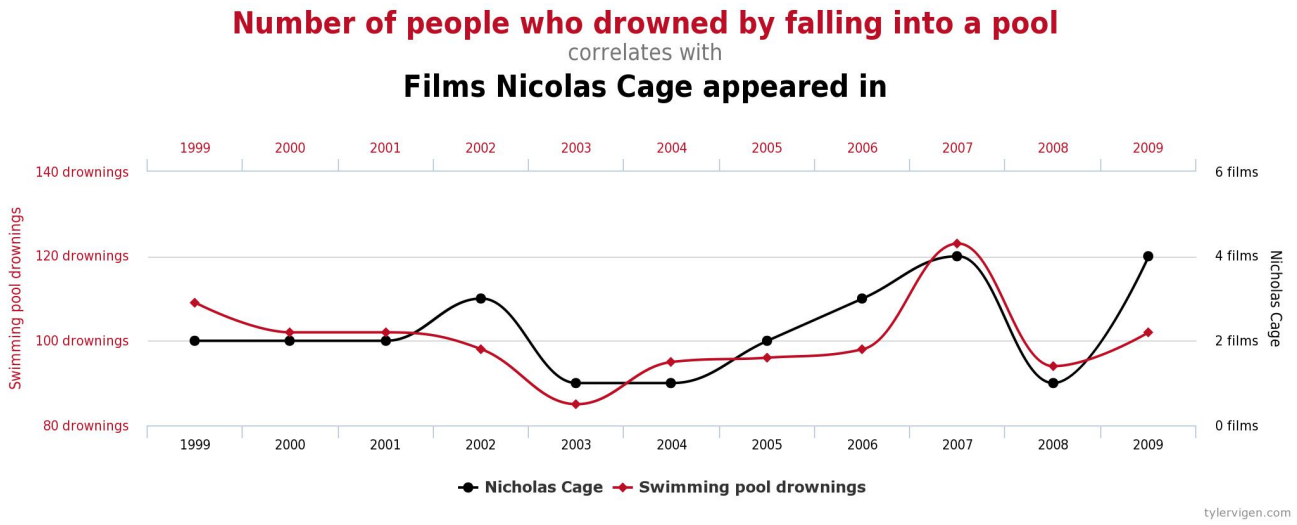


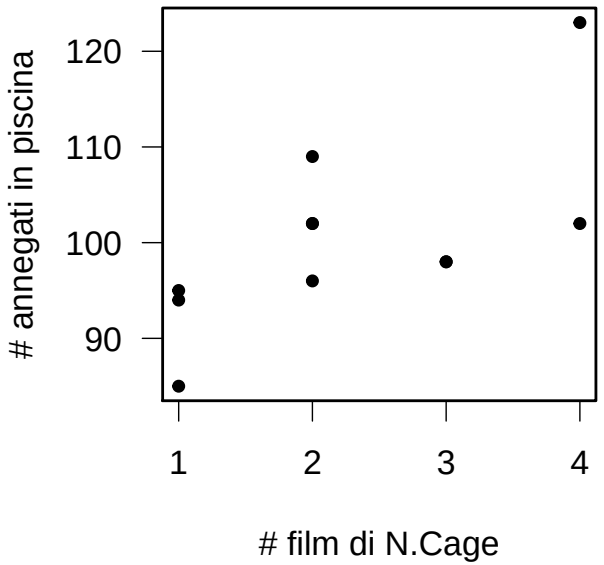


La correlazione è 0.992, ma diminuire la spesa per scienza e tecnologia non è una strategia per diminuire i suicidi.

Quale potrebbe essere una spiegazione per questa correlazione?

Nicolas Cage un pericolo per i nuotatori (in piscina)

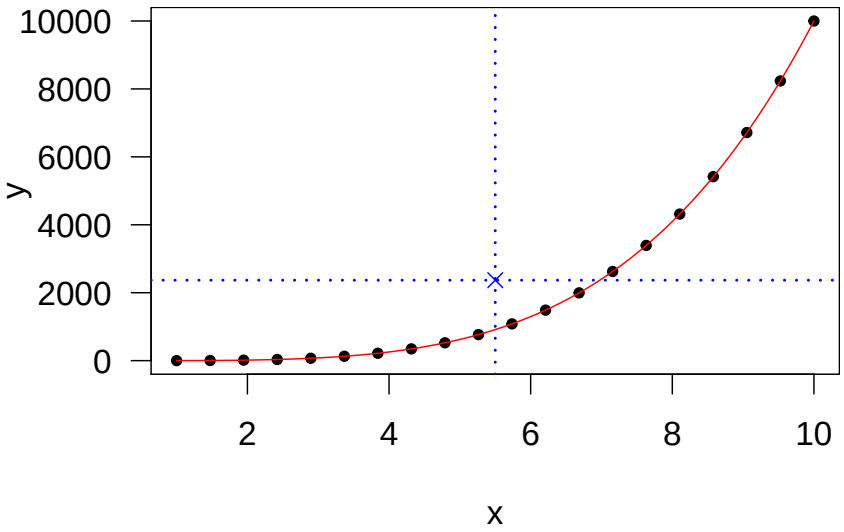




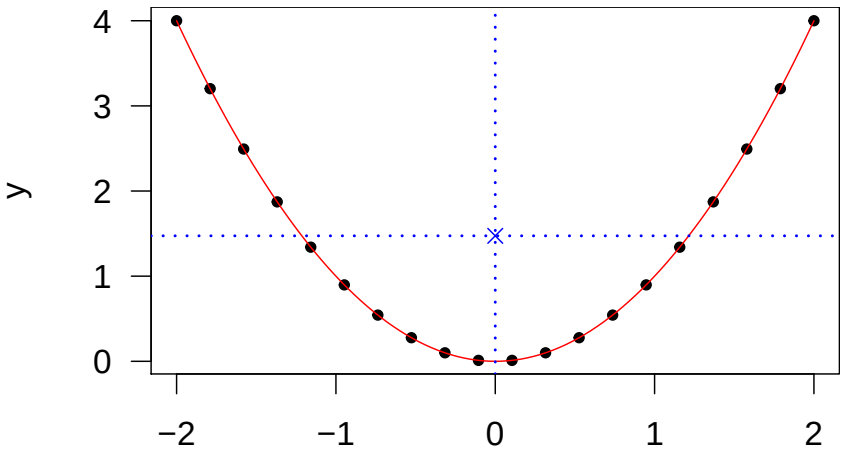
La correlazione è 0.666 ma ...

Notare come il grafico sopra suggerisca una relazione molto stretta, ridimensionata dal diagramma di dispersione: quello sopra **non è un buon modo di rappresentare due serie storiche.**

Attenzione a cosa misura r



- I dati si dispingono sulla curva $Y = X^4$.
- la relazione è perfetta ma non lineare e non monotona.
- $r = 0.8852$



- I dati si dispingono sulla curva $Y = X^2$.
- la relazione è perfetta ma non lineare e non monotona.
- $r = 0$

Morale

r misura la correlazione lineare tra le variabili.

- 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).