STATISTICA DESCRITTIVA – modulo 1

Corso di Laurea SMID

Elda Guala e Ivano Repetto
Dipartimento di Matematica - Università degli Studi di Genova

1. Variabili statistiche

I dati riportati sotto si riferiscono a 20 studenti universitari che hanno frequentato un corso di Statistica e sono stati raccolti facendo compilare ad ogni studente il seguente modulo:

SESSO	ALTEZZA	PESO	CORSO LAUREA	NUMERO SCARPA	COLORE OCCHI	COLORE CAPELLI	ATT SPORTIVA	DIPLOMA

CORSO LAUREA :	COLORE OCCHI	COLORE CAPELLI	ATT.SPORTIVA	DIPLOMA :
M= matematica B = biologia I = informatica	1 = scuri 2 = verdi 3 = azzurri	1 = scuri 2 = castani 3 = biondi	1 = nulla 2 = media 3 = alta	1 = liceo scientifico 2 = liceo classico 3 = ist. tecnico 4 = ist. magistrale 5 = altro

Tabella dei dati :

N	SESSO	ALTEZZA (cm)	PESO (kg)	CORSO LAUREA	NUMERO SCARPA	COLORE OCCHI	COLORE CAPELLI	ATT SPORTIVA	DIPLOMA
1	M	179	65	I	43	1	1	2	3
2	M	180	62	I	43	1	1	2	3
3	F	165	50	I	39	1	2	1	1
4	F	160	49	I	37	2	2	2	3
5	F	160	47	M	37	3	3	2	5
6	F	160	48	M	36	1	3	2	1
7	F	164	56	M	38	2	2	2	2
8	F	170	57	M	38	1	2	2	1
9	M	180	73	M	43	2	2	1	1
10	M	186	86	M	45	3	1	1	3
11	M	170	66	M	42	1	2	2	1
12	M	180	68	I	41	1	1	3	3
13	M	180	85	I	43	1	2	2	3
14	F	176	56	I	37	1	3	2	1
15	M	170	72	I	42	2	2	2	5
16	M	180	65	I	42	1	1	1	3
17	M	170	75	I	41	1	1	2	3
18	M	172	70	I	40	1	2	1	1
19	M	178	80	I	45	1	2	1	3
20	F	162	49	B	37	1	1	2	2

VARIABILI QUALITATIVE :

SESSO	CORSO LAUREA	COLORE OCCHI	COLORE CAPELLI	ATT SPORTIVA	DIPLOMA

VARIABILI QUANTITATIVE :

ALTEZZA (cm)	PESO (kg)	NUMERO SCARPA

2. Statistica descrittiva per variabili qualitative

Genere	*conteggi*	*percentuale*
M	12	60
F	8	40
totale	20	100

CLaurea	*conteggi*	*percentuale*
B	1	5
I	12	60
M	7	35
totale	20	100

Diploma	*conteggi*	*percentuale*
1	7	35
2	2	10
3	9	45
4	0	0
5	2	10
totale	20	100

Att.Sport.	*conteggi*	*percentuale*
1	6	30
2	13	65
3	1	5
totale	20	100

Occhi	*conteggi*	*percentuale*
1	14	70
2	4	20
3	2	10
totale	20	100

Capelli	*conteggi*	*percentuale*
1	7	35
2	10	50
3	3	15
totale	20	100

2.1 Relazioni fra variabili qualitative : tabelle

- Tabella di contingenza (variabili sesso e corso di laurea)

SESSO/C.LAUREA	*B*	*I*	*M*	*Totale*
F	1	3	4	8
M	0	9	3	12
Totale	1	12	7	20

- Tabella percentuale

SESSO/C.LAUREA	*B*	*I*	*M*	*Totale*
F	5	15	20	40
M	0	45	15	60
Totale	5	60	35	100

- Tabella profili riga

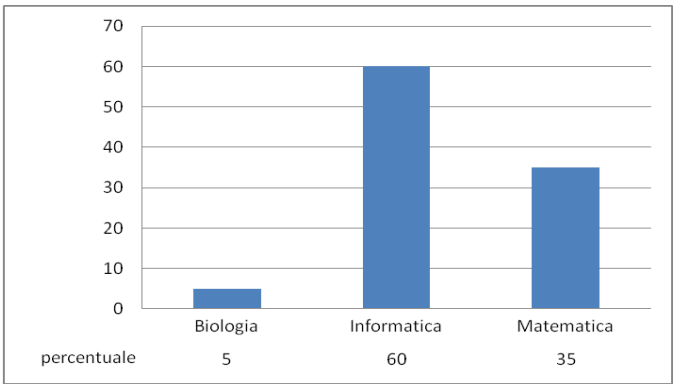
GENERE/C.LAUREA	*B*	*I*	*M*	***Totale***
F	12.5	37.5	50	100
M	0	75	25	100

- Tabella profili colonna

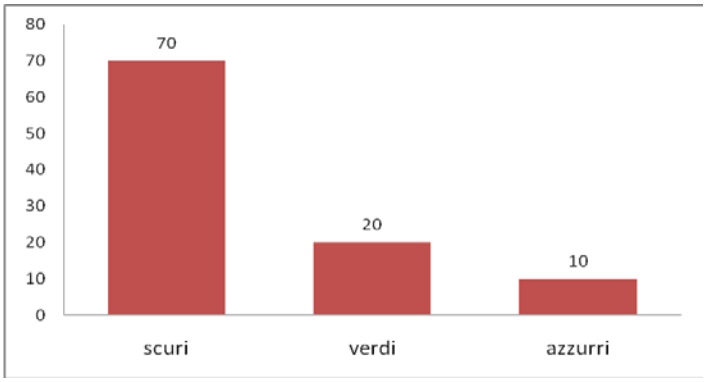
SESSO/C.LAUREA	*B*	*I*	*M*
F	100	25	57
M	0	75	43
Totale	100	100	100

2.2 Rappresentazione grafica di variabili qualitative : grafici a barra

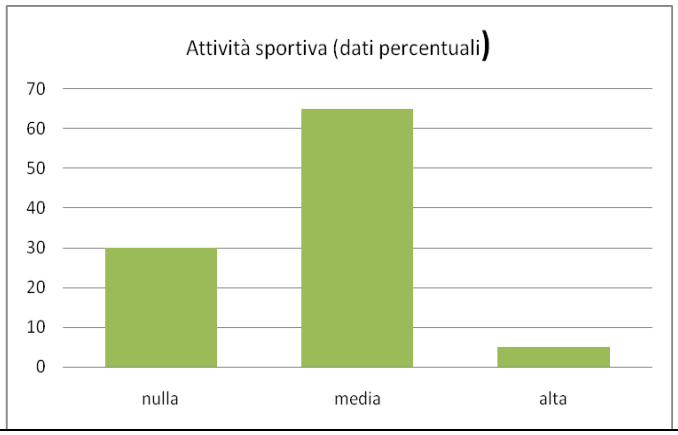
CLaurea	*conteggi*	*percentuale*
Biologia	1	5
Informatica	12	60
Matematica	7	35
totale	20	100



Occhi	*conteggi*	*percentuale*
scuri	14	70
verdi	4	20
azzurri	2	10
totale	20	100



Attività Sportiva	*conteggi*	*percentuale*
nulla	6	30
media	13	65
alta	1	5
totale	20	100



Esercizio 2.1

La tabella seguente riporta i dati di n=20 individui. La variabile Reddito e' suddivisa nelle tre classi, a reddito crescente, 1,2 e 3; la variabile Auto indica il numero di auto possedute dal nucleo familiare di cui la persona fa parte.

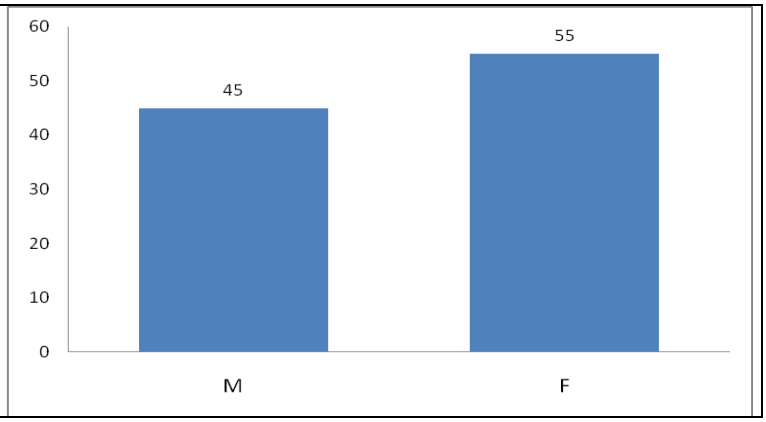
N	GENERE	REDDITO	AUTO	FUMATORE
1	M	1	1	NO
2	M	3	1	NO
3	F	3	2	NO
4	F	2	0	SI
5	F	3	3	NO
6	F	1	1	NO
7	M	2	1	NO
8	F	1	2	SI
9	F	2	1	SI
10	F	1	0	NO
11	M	2	1	SI
12	M	1	2	NO
13	M	1	2	SI
14	M	2	0	NO
15	F	3	3	NO
16	M	1	1	NO
17	M	3	2	NO
18	F	1	2	SI
19	F	2	0	SI
20	F	3	3	NO

1. Costruire le tabelle di contingenza e percentuali per ognuna delle variabili.
2. Costruire un grafico a barre dei conteggi per ognuna delle variabili.
3. Costruire un grafico a barre percentuale per ognuna delle variabili.
4. Costruire la tabella congiunta (conteggi e percentuale) per le variabili Sesso e Fumo.
5. Costruire la tabella congiunta (conteggi e percentuale) per le variabili reddito e Auto.
6. Costruire un grafico a barre della variabile reddito ripartito per la variabile Auto.

Risoluzione di alcune parti dell'esercizio 1

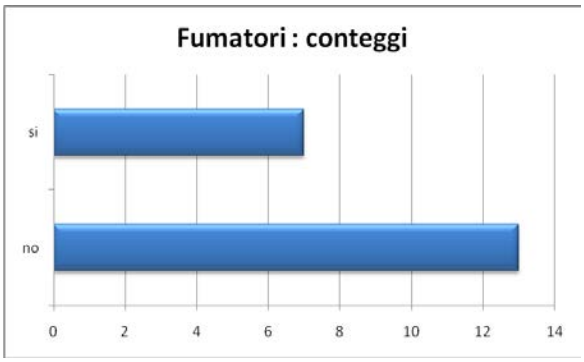
- Punti 1 e 2

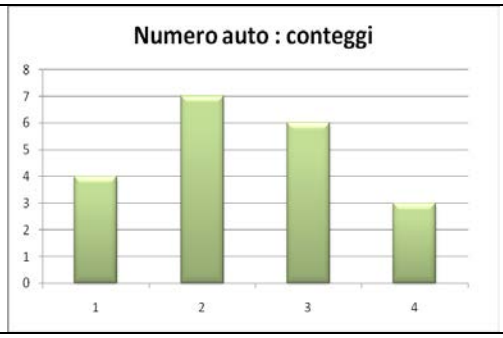
Genere	*conteggi*	*percentuale*
M	9	45
F	11	55
totale	20	100



Reddito	*conteggi*	*percentuale*
1	8	40
2	6	30
3	6	30
totale	20	100

fumatori	*conteggi*	*percentuale*
no	13	65
si	7	35
totale	20	100





Auto	*conteggi*	*percentuale*
0	4	20
1	7	35
2	6	30
3	3	15
totale	20	100

Punto 4

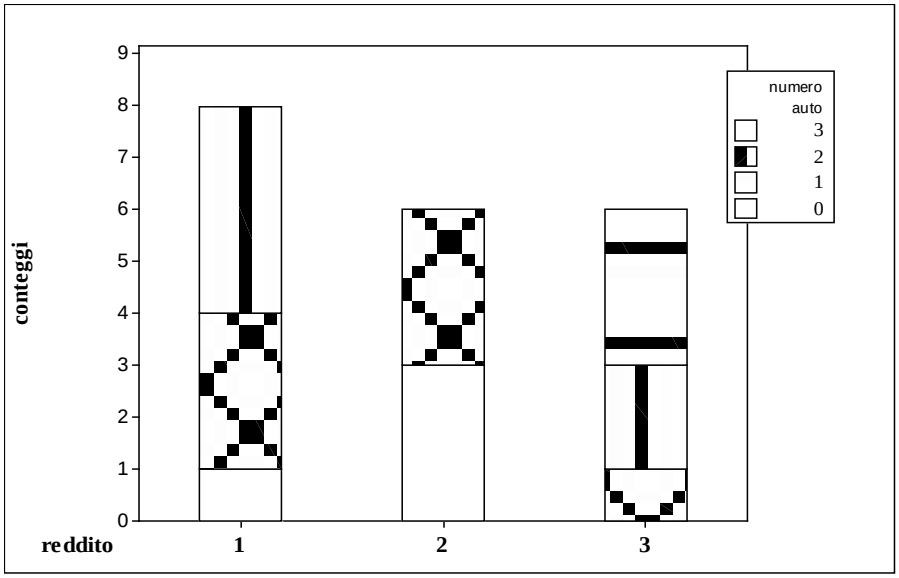
Tabella dei conteggi

SESSO/FUMO	*SI*	*NO*	*Totale*
F	5	6	11
M	2	7	9
Totale	7	13	20

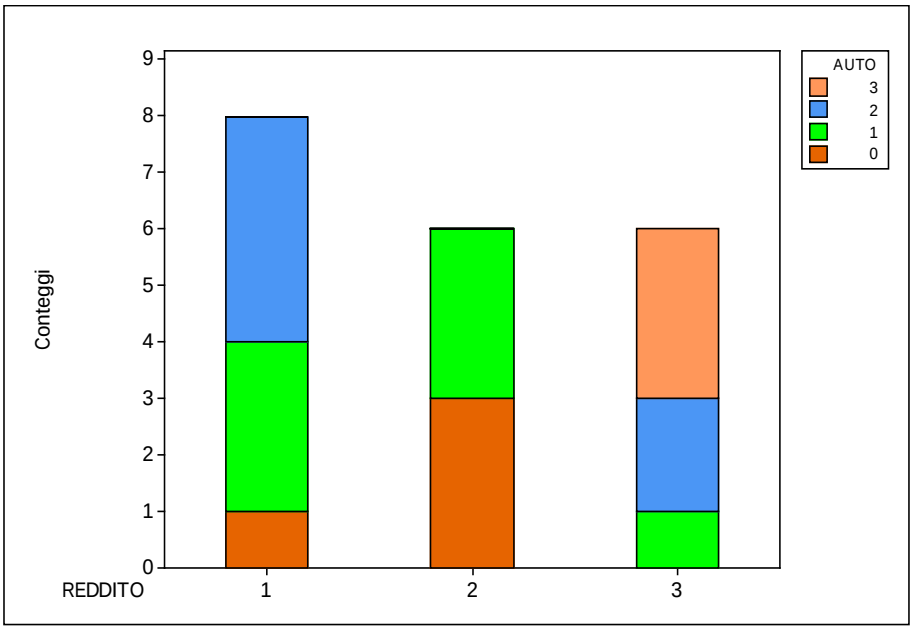
Tabella percentuale

SESSO/FUMO	*SI*	*NO*	*Totale*
F	25	30	55
M	10	35	45
Totale	35	65	100

Punto 6



	Numero auto				
Reddito	**0**	**1**	**2**	**3**	**Totale**
1	1	3	4	0	8
2	3	3	0	0	6
3	0	1	2	3	6
Totale	4	7	6	3	20



3. Statistica descrittiva per variabili quantitative

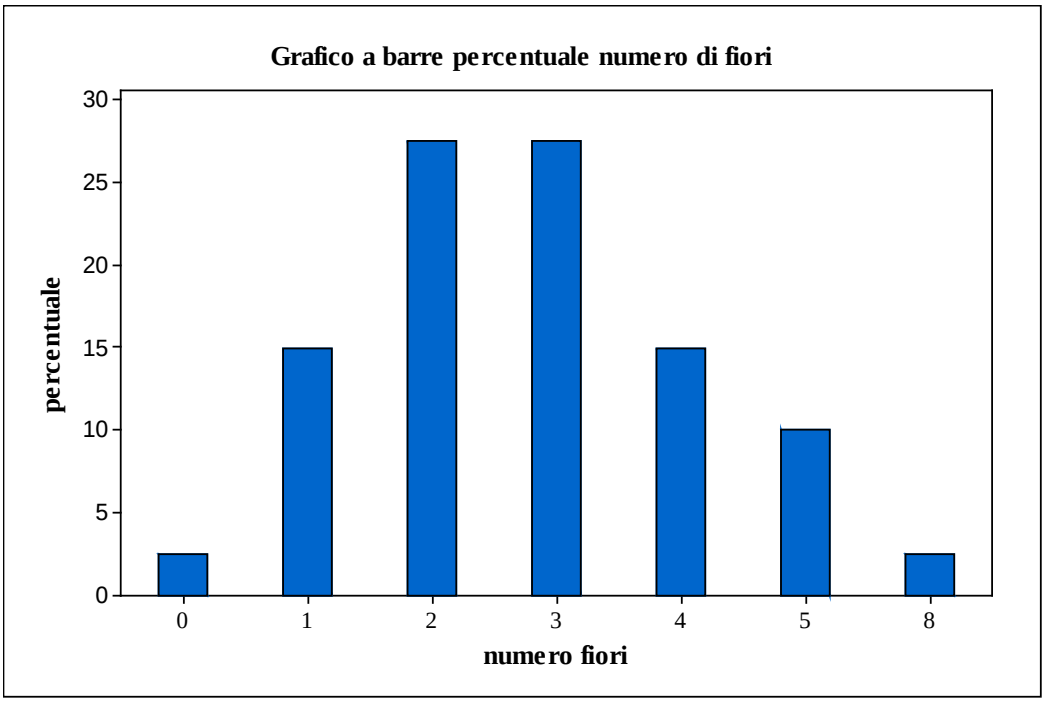
Si effettuano n = 40 misurazioni di una variabile quantitativa (numero di fiori di una pianta rilevati ogni 24 ore) e si ottengono i seguenti risultati (l'ordine di lettura e' da sinistra a destra e dall'alto in basso) :

0	2	1	4	3	1	2	3	8	2
5	2	1	3	3	1	3	2	2	5
4	4	4	2	3	5	5	1	1	2
2	4	4	2	3	3	3	3	3	2

Questa variabile puó assumere solo valori interi . La tabella seguente rappresenta la distribuzione di frequenza:

num. fiori	*Freq. Assoluta*	*Freq. Relativa*	*Freq. percentuale*	*Freq. % cumulata*
0	1	0.025	2.5	2.5
1	6	0.15	15	17.50
2	11	0.272	27.5	45
3	11	0.275	27.5	72.50
4	6	0.15	15	87.50
5	4	0.1	10	97.50
8	1	0.025	2.5	100
totale	40	1	100	*

Per variabili che hanno valori non solo interi (ad esempio altezza e peso) la scelta dei valori della prima colonna della tabella deve essere effettuata da chi fa l'analisi.

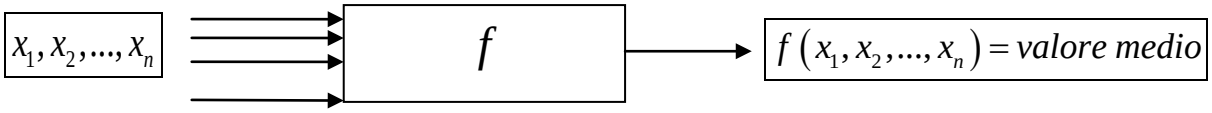


4. Indici di posizione

Si indica con $x_1, x_2, ..., x_n$ il valore di una variabile quantitativa rilevata su n individui (osservazioni).

Nelle distribuzioni di variabili quantitative si possono utilizzare indici sintetici medi che servono a stabilire quale valore meglio rappresenta la variabile quantitativa.

Possiamo pensare ai valori medi come a output di funzioni che hanno in ingresso i valori della variabile quantitativa.



La scelta della funzione *f* dovrà portare ad un valore tale che, sostituendolo a ciascuno dei valori $x_1, x_2, \ldots, x_n$, rimanga invariato l'aspetto che interessa del problema.

Se si è interessati a mantenere invariata la somma dei valori, si deve utilizzare la media aritmetica $\bar{x}$ in quanto si deve avere $n \cdot \bar{x} = x_1 + x_2 + ... + x_n$ e quindi la media aritmetica vale $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Definizione 4.1: dato un insieme finito di dati quantitativi $x_1, x_2, ..., x_n$ si definisce ***media aritmetica*** la quantità $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Se si è interessati a mantenere invariato il prodotto dei valori, si deve utilizzare la media geometrica $\bar{x}_G$ in quanto si deve avere $\left(\bar{x}_G\right)^n = x_1 \cdot x_2 \cdot ... \cdot x_n$ e quindi la media geometrica vale $\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n}$.

Definizione 4.2: dato un insieme finito di dati quantitativi positivi $x_1, x_2, ..., x_n$ si definisce ***media geometrica*** la quantità $\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Esempio 4.1

Dato un rettangolo di dimensioni x_1 e x_2 :

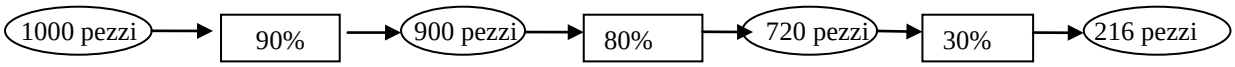
- se si vuole determinare un valore L_1 tale che il perimetro di un quadrato con questo lato sia uguale a quello del rettangolo si usa la ***media aritmetica***: $4L_1 = 2x_1 + 2x_2$ da cui $L_1 = \frac{x_1 + x_2}{2}$;
- se si vuole determinare un valore L_2 tale che la superficie di un quadrato con questo lato sia uguale a quella del rettangolo si usa la ***media geometrica***: $L_2^2 = x_1 \cdot x_2$ da cui $L_2 = \sqrt{x_1 \cdot x_2}$.

Esempio 4.2

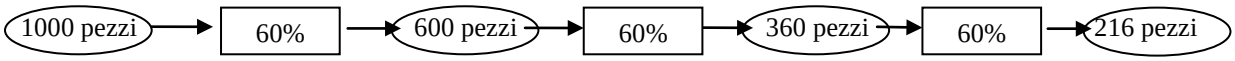
In un'azienda lavorano *n* persone ciascuna con uno stipendio $x_i, i = 1,...,n$. La somma di tali stipendi vale $S = \sum_{i=1}^{n} x_i$. Se l'azienda, avendo a disposizione la somma *S*, volesse che l'importo degli stipendi fosse uguale per tutti, dovrebbe assegnare a ciascuno uno stipendio pari alla media aritmetica degli stipendi, cioè $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Esempio 4.3

In un ciclo di lavorazione tre apparecchiature lavorano in serie: la prima ha un rendimento del 90%, la seconda dell'80% e la terza del 30%. Quale è il rendimento percentuale medio complessivo?



Quindi da 1000 pezzi teorici si ottengono in realtà, con l'intervento in serie delle tre macchine, 216 pezzi. Il problema è allora quello di trovare tre macchine, tutte con lo stesso rendimento *x*, che, messe in serie, diano partendo da 1000 pezzi, alla fine dei tre passaggi 216 pezzi.
Il valore di *x* si ottiene moltiplicando i rendimenti delle tre macchine : $0,9 \cdot 0,8 \cdot 0,3 = 0,216$ e risolvendo l'equazione $x^3 = 0,216$, e quindi $x = \sqrt[3]{0,216} = 0,60$.
In altre parole se sostituissimo alle tre macchine iniziali tre macchine uguali, tutte con un rendimento del 60%, otterremmo la stessa percentuale di pezzi in uscita.



Un altro indice che viene utilizzato è la **media armonica**. Un esempio di utilizzo della media armonica è il seguente : se un'auto si muove da A a B, che distano 80 km, alla velocità di 80 km/h e successivamente da B ad A alla velocità di 20 km/h, quale sarà la velocità media sull'intero tragitto andata-ritorno? Se si facesse la media delle due velocità si otterrebbe 50 km/h, risultato non corretto. Infatti si deve notare che l'intero tragitto è di 160 km, che viene percorso in 5 ore (1 all'andata e 4 al ritorno): la velocità media corretta è dunque 160/5, cioè 32 km/h, ben diversa dai 50 calcolati con la media aritmetica delle velocità.

In realtà la distanza di 80 km non conta nulla, come mostra il seguente ragionamento. Se si indicano con *d* la distanza AB, con t_1 e t_2 i tempi di andata e ritorno, con v_1 e v_2 le velocità all'andata e al ritorno, allora la media $\bar{x}_A$ cercata è:

$$\bar{x}_A = \frac{2d}{t_1 + t_2} = \frac{2d}{\frac{d}{v_1} + \frac{d}{v_2}} = \frac{2v_1 v_2}{v_1 + v_2}$$, che è proprio la media armonica delle due velocità.

Questa conclusione è legata al fatto che il reciproco della velocità rappresenta il tempo necessario a percorrere l'unità di spazio (una velocità di 80 km/h significa che mi serve 1/80 di ora per fare un km!). Se si tiene conto che la velocità media richiesta si può anche scrivere come $\overline{x}_A = \dfrac{d}{\frac{t_1+t_2}{2}}$, il problema potrebbe essere formulato nel seguente modo: a *quale velocità costante deve muoversi il corpo se si vuole che il tempo di andata e ritorno sia uguale ed uguale alla media degli effettivi tempi t_1 e t_2?*

Nota. In questo caso in ogni tragitto lo spazio percorso è lo stesso mentre il tempo impiegato cambia e quindi l'indice da utilizzare è la media armonica delle due velocità. Invece se, in ogni tragitto, lo spazio percorso cambiasse e il tempo rimanesse uguale allora l'indice da utilizzare sarebbe la media aritmetica delle due velocità.

Ad esempio, se un'auto si muove per un'ora ad una velocità di 80 km/h e per un'altra ora ad una velocità di 20 km/h, la sua velocità media si ottiene calcolando la media delle due velocità, cioè

$$\left(\frac{80+20}{2}\right)\text{km/h}=50\text{ km/h}.$$

In altre parole, dati due numeri *a* e *b*, la loro ***media armonica*** $\overline{x}_A$ è definita in modo che la sua inversa $\dfrac{1}{\overline{x}_A}$ sia la media aritmetica degli inversi di *a* e *b*; si pone cioè: $\dfrac{1}{\overline{x}_A} = \dfrac{1}{2}\left(\dfrac{1}{a}+\dfrac{1}{b}\right)$ o, il che è lo stesso, $\dfrac{2}{\overline{x}_A} = \dfrac{1}{a}+\dfrac{1}{b}$. É immediato che la media armonica può essere calcolata anche con la formula: $\overline{x}_A = \dfrac{2ab}{a+b}$.

Un secondo esempio è legato al concetto di potere di acquisto della moneta. Se un certo oggetto costa oggi 20€ e in un tempo successivo 80€, *a quale prezzo dovrebbe essere venduto se si vuole che il potere di acquisto della moneta sia uguale ed uguale alla media dei due poteri di acquisto?*

Per rispondere alla domanda ricordiamo che il potere di acquisto della moneta può essere definito come il reciproco del prezzo: 1/20 (cioè 0.05) rappresenta quanta parte dell'oggetto posso oggi comprare con 1€, 1/80 (cioè 0.0125) rappresenta la frazione di oggetto che posso comprare nel momento successivo sempre con 1€. Per avere il prezzo medio devo dunque calcolare intanto la media dei poteri di acquisto: (0.05+0.0125)/2=0.03125 e poi farne il reciproco, ottenendo 32, risultato che coincide esattamente con la media armonica dei prezzi.

<u>Definizione 4.3</u>: dato un insieme finito di dati quantitativi positivi $x_1, x_2, ..., x_n$ si definisce ***media armonica*** la quantità $\overline{x}_A = \dfrac{n}{\frac{1}{x_1}+\frac{1}{x_2}+...+\frac{1}{x_n}} = \dfrac{1}{\frac{\frac{1}{x_1}+\frac{1}{x_2}+...+\frac{1}{x_n}}{n}}$

Altri indici di posizione sono la moda e i percentili (o quantili).

Definizione 4.4: dato un insieme finito di dati quantitativi $x_1, x_2, ..., x_n$ si definisce ***moda*** il dato che si verifica con frequenza maggiore.

Esempio 4.4 : Nella tabella seguente è riportata la distribuzione delle famiglie per numero di componenti in un Comune italiano. Il valore della moda è 4.

Numero componenti	*Famiglie*
1	153
2	225
3	335
4	564
5	346
6	133
7	75
8	49

Definizione 4.5: dato un insieme finito di dati quantitativi ordinati $x_1, x_2, ..., x_n$ si definisce ***quantile*** di ordine p, con p compreso fra 0 e 1 e si indica con Q_p , il minimo valore per cui vale almeno p la frequenza dei dati minori o uguali ad esso.

Se si vuole calcolare il quantile di ordine ***p*** si può, ad esempio, fra i molti algoritmi di calcolo possibili, procedere nel seguente modo:

a) ordinare i valori delle osservazioni in senso crescente.
b) calcolare il valore np.
c) se np é un intero k il quantile di ordine p é la media fra x_k e x_{k+1}.
d) se np non é intero e k é l'intero successivo a np allora il quantile di ordine p é x_k.

Il p-esimo quantile viene detto anche 100p-esimo percentile.
Il 25°, 50° e 75° percentile, detti primo, secondo e terzo quartile, vengono indicati con Q1,Q2 e Q3 rispettivamente.

Esempio 4.5 n=10 osservazioni ordinate

1	1.8	2.3	2.9	2.9	3	5.2	5.3	5.4	7.3

percentile	*25*	*50*	*75*
valore	2.3	2.95	5.3

Il quantile Q2 è detto ***Mediana*** ed è il valore che divide in parti uguali i dati ordinati: cioe' meta' di essi sono minori o uguali a Q2 e meta' sono maggiori o uguali.

In particolare, per calcolare il valore della mediana :

- si ordinano in modo crescente i dati;
- se n é dispari la mediana é il valore centrale;
- se n é pari la mediana é la media dei due valori in posizione centrale.

Esempio 4.6 : n=10 osservazioni

5.2	2.3	1	5.3	3	1.8	7.3	2.9	2.9	5.4

Dati ordinati:

1	1.8	2.3	2.9	2.9	3	5.2	5.3	5.4	7.3

***Mediana* :** $\frac{2.9+3}{2} = 2.95$

Esempio 4.7 : n=9 osservazioni

5.2	2.3	1	5.3	1.8	7.3	2.9	2.9	5.4

Dati ordinati:

1	1.8	2.3	2.9	2.9	5.2	5.3	5.4	7.3

***Mediana* : 2.9**

5. Indici di dispersione

- ***Varianza campionaria***

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \quad \text{oppure} \quad \sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\overline{x}\right)^2$$

- ***Scarto quadratico medio σ***

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\overline{x}\right)^2}$$

Si noti che si utilizza il valore medio $\overline{x}$ in quanto soddisfa la seguente proprietà :

$\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \le \sum_{i=1}^{n}\left(x_i - a\right)^2$ per ogni valore di *a* reale.

- ***Differenza fra quantili***

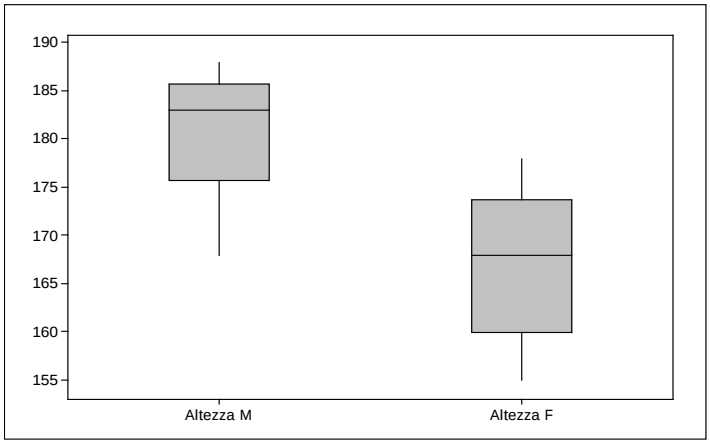
Si definisce ***distanza interquartile*** e la si indica con IQR l'ampiezza dell'intervallo delimitato dal 1° e 3° quartile [IQR=Q3- Q1]. Fra Q1 (25° percentile) e Q3 (75° percentile) è contenuto il 50% dei dati.
Nell'esempio seguente si vuole mettere in evidenza il significato di IQR quale indice di variabilità.

Esempio *(calcoli fatti "a mano")*

A fianco sono riportati valori delle altezze in cm rilevati su 20 uomini e 20 donne.

Di seguito sono riportati i valori dei tre quartili, di IQR e una semplice rappresentazione grafica.

Variabile	Q1	Mediana	Q3	IQR
Altezza M	176,5	183,0	185,5	9,0
Altezza F	160,0	168,0	173,5	13,5



Il fatto che per i maschi il valore di IQR sia di 9 cm, per le femmine di 13,5 cm significa che il 50% dei valori dell'altezza si disperde su un intervallo di ampiezza maggiore per le femmine rispetto ai maschi.

Per esercizio provate a rifare i calcoli eliminando l'osservazione numero 20.

Obs	Altezza Maschi	Altezza Femmine
1	168	155
2	168	157
3	170	157
4	175	159
5	175	160
6	178	160
7	180	165
8	180	166
9	183	168
10	183	168
11	183	168
12	183	168
13	183	170
14	185	173
15	185	173
16	186	174
17	187	175
18	188	175
19	188	175
20	188	178

- ***Scarto assoluto medio***

✓ rispetto alla media : $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

✓ rispetto alla mediana: $\frac{1}{n}\sum_{i=1}^{n}|x_i - Q2|$

Si noti che si utilizza il valore di *Q2* nella formula che permette di calcolare lo scarto assoluto medio rispetto alla mediana in quanto tale valore soddisfa la seguente proprietà:

$\sum_{i=1}^{n}|x_i - Q2| \leq \sum_{i=1}^{n}|x_i - a|$ per ogni *a* reale.

<u>Dimostrazione.</u>

Si indichino con $x_1, x_{2,} x_3, ...x_n$ i dati originari e con $x_{(1)}, x_{(2),} x_{(3)}, ...x_{(n)}$ i dati ordinati in senso non decrescente.

Sia *F(a)* una funzione definita, per ogni $a \in \mathbb{R}$, nel modo seguente : $F(a) = \sum_{i=1}^{n} |x_i - a|$

Si vuole provare che *F(a)* è minima per $a = Q2$.

Suddividiamo la dimostrazione in due casi :

- ✓ ***caso 1: n pari , n = 2k con k intero***

Sia $F_i(a) = |x_{(i)} - a| + |x_{(n-i+1)} - a|$ per ogni $i = 1,2,.....k$

È facile vedere che risulta: $F(a) = \sum_{i=1}^{k} F_i(a)$.

Ad esempio, se n = 4, k = 2, si ha:

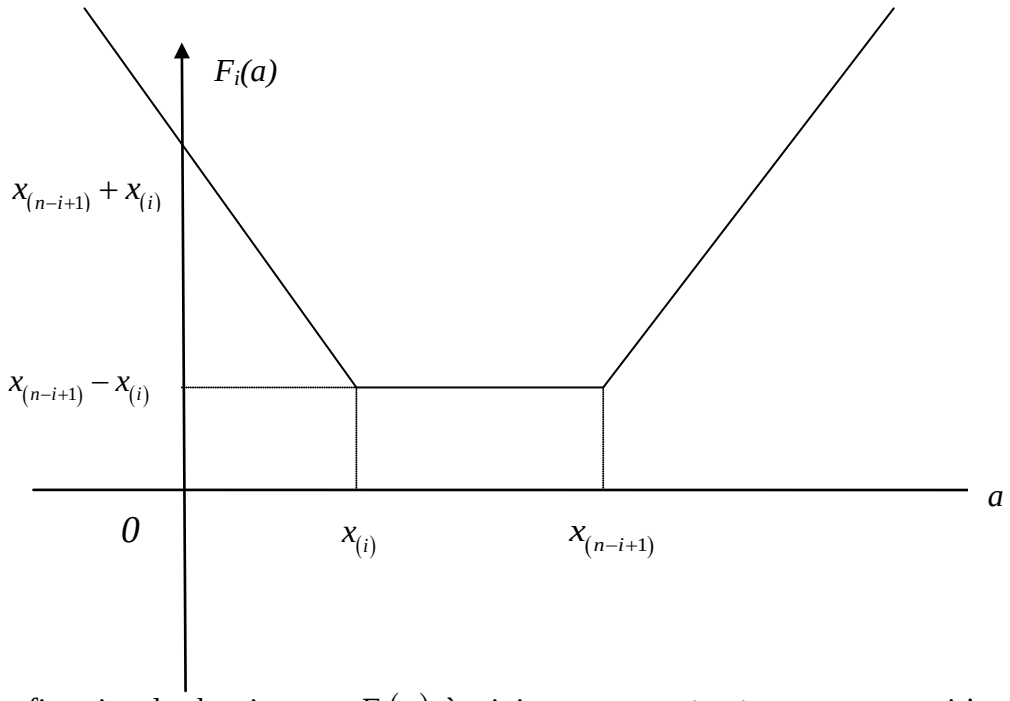
$$F_1(a) = |x_{(1)} - a| + |x_{(4)} - a|$$

$$F_2(a) = |x_{(2)} - a| + |x_{(3)} - a|$$

$$F(a) = F_1(a) + F_2(a)$$

Di seguito sono riportati l'espressione analitica e il relativo grafico di ciascuna funzione $F_i(a)$:

$$F_i(a) = |x_{(i)} - a| + |x_{(n-i+1)} - a| = \begin{cases} x_{(i)} + x_{(n-i+1)} - 2a & se \quad a < x_{(i)} \\ x_{(n-i+1)} - x_{(i)} & se \quad x_{(i)} \le a \le x_{(n-i+1)} \\ 2a - \left(x_{(i)} + x_{(n-i+1)}\right) & se \quad a > x_{(n-i+1)} \end{cases}$$



Dal grafico si vede che ciascuna $F_i(a)$ è minima per $x_{(i)} \le a \le x_{(n-i+1)}$, per ogni *i = 1,2,….k.*

Allora il punto $a_{\min}$ di minimo di $F(a)$ deve appartenere a $\bigcap_{i=1}^{k}\left[x_{(n-i+1)} - x_{(i)}\right]$,

cioè: $x_{(n/2)} \le a_{\min} \le x_{(n/2+1)}$

Quindi $a_{\min}$ è proprio, per definizione, la mediana dei dati $x_1, x_{2,} x_3, ...x_n$.

Per evitare l'indeterminazione, si sceglie poi come valore per la mediana il valore

$$Q2 = \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right).$$

Infatti precedentemente abbiamo detto che:

se *n* é pari la mediana é la media dei due valori in posizione centrale e tali valori sono proprio

$$x_{(n/2)} \text{ e } x_{(n/2+1)}$$

- ✓ ***caso 2 : n dispari, n = 2k-1 con k intero***

La dimostrazione è identica al caso precedente, con l'unica differenza che: $\bigcap_{i=1}^{k}\left[x_{(i)}, x_{(n-i+1)}\right]$ coincide con il solo valore $x_{(n+1/2)}$, che è, per definizione, la mediana dei dati $x_1, x_{2,} x_3, ...x_n$.

Infatti precedentemente abbiamo detto che: se n é dispari la mediana é il valore centrale e tale valore è proprio $x_{(n+1/2)}$.

CVD

Esercizio 5.1

La tabella riportata sotto si riferisce alle seguenti variabili riferite a n=20 individui:

- ✓ Pulse1 = pulsazioni cardiache rilevate a riposo
- ✓ Pulse2 = pulsazioni cardiache rilevate dopo aver percorso 500m a passo veloce
- ✓ Fumo : 1=fumatore 2=non fumatore
- ✓ Altezza = altezza in cm
- ✓ Peso = peso in kg
- ✓ Sport (attività sportiva) : 1=bassa 2=media 3=alta

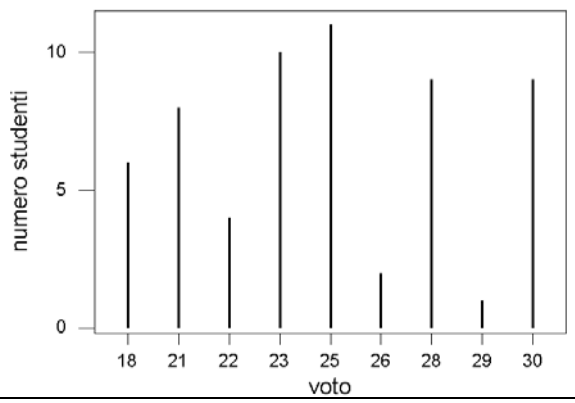
Pulse1	Pulse2	Fumo	Altezza	Peso	Sport
64	88	2	170	64	2
58	70	2	183	66	2
62	76	1	187	73	3
66	78	1	185	86	1
64	80	2	175	70	2
74	84	2	185	75	1
84	84	2	183	68	3
68	72	2	188	86	2
62	75	2	183	89	2
76	118	2	180	63	2
90	94	1	188	73	1
80	96	2	183	70	2
92	84	1	178	69	3
68	76	2	170	66	2
60	76	2	180	77	3
62	58	2	183	79	3
66	82	1	175	79	2
70	72	1	185	77	3
68	76	1	188	82	2
72	80	2	170	61	3

1. Calcolare media e varianza per le variabili quantitative.
2. Calcolare Q1, Q2 e Q3 per le variabili quantitative.
3. Costruire una tabella percentuale per ognuna delle variabili qualitative.
4. Calcolare media e varianza per Puls1 e Puls2 per ognuno dei valori di Fumo.

6. Calcolo di media e varianza per dati raggruppati in classi

Nella tabella seguente sono riportati i dati dei risultati di un esame universitario e il relativo istogramma:

voto	*Numero studenti*
18	6
21	8
22	4
23	10
25	11
26	2
28	9
29	1
30	9



Come posso calcolare la media dei voti avendo a disposizione unicamente questi dati e non le singole osservazioni?

Si devono “pesare” i voti rispetto al numero di studenti che hanno ottenuto quel voto.

La formula per il calcolo della media sarà quindi :

$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n}$	dove : ✓ x_i rappresenta l’i-esimo valore della variabile ✓ k rappresenta il numero dei valori assunti dalla variabile ✓ n_i rappresenta il numero di elementi che assumono il valore x_i ✓ n é il numero totale di osservazioni

Nel caso dei dati riportati in tabella il valore di k é 9, il numero di studenti vale 60 e si avrà :

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{18*6+21*8+22*4+23*10+25*11+26*2+28*9+29*1+30*9}{60} \cong 24.5$$

Il caso esaminato sopra é il piú semplice in quanto la variabile VOTO assume solo valori interi.

Come si calcola il valore della media se la variabile é quello della tabella dell’esercizio precedente e la variabile é Altezza ?

In questo caso, avendo a disposizione i valori di tutte e 20 le osservazioni si ha che:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{20} x_i}{20} = 180.95$$

Se invece i dati sono forniti attraverso la seguente tabella come posso calcolare la media ?

Altezza	x_i
170-174	172
174-178	176
178-182	180
182-188	184

Non é possibile calcolare il valore esatto della media ma si puó calcolarne una approssimazione costruendo una nuova colonna che contiene i valori centrali di ognuna delle classi di Altezza e poi procedere come nel caso precedente.

In questo caso la nuova tabella diventa :

Altezza	x_i	n_i	$x_{i*}\ n_i$
170-174	172	3	516
174-178	176	3	528
178-182	180	2	360
182-188	184	12	2208
totale		20	3612

E il valore della media é

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{3612}{20} = 180.6$$

Si osservi che i valori della media calcolata sui dati singoli o sui dati in tabella sono differenti.

Anche il calcolo della varianza nel caso di variabili divise in classi fornisce un risultato approssimato e lo si effettua utilizzando la formula seguente :

$\sigma^2 = \frac{1}{n}\left(\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i\right) = \frac{1}{n}\left(\sum_{i=1}^{k} x_i^2 n_i\right) - (\overline{x})^2$	dove : ✓ x_i rappresenta l'i-esimo valore della variabile ✓ k rappresenta il numero delle classi in cui é divisa la variabile ✓ n_i rappresenta il numero di elementi che assumono il valori nella classe i-esima ✓ n é il numero totale di osservazioni

Il modo piú semplice di effettuare i conti é costruire una nuova tabella :

Altezza	$x_i * n_i$	x_i^2	n_i	$x_i^2 * n_i$
172	516	29584	3	**88752**
176	528	30976	3	**92928**
180	360	32400	2	**64800**
184	2208	33856	12	**406272**
totale	**3612**	**126816**	**20**	**652752**

Nel caso dei dati non divisi in classi si ha che la varianza vale circa 34.8.

Se si utilizza la tabella con i dati divisi in classi si ottiene :

$$\sigma^2 = \frac{1}{n}\left(\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i\right) = \frac{1}{n}\left(\sum_{i=1}^{k} x_i^2 n_i\right) - (\overline{x})^2 = \frac{1}{20}(652752) - (180.6)^2 = 21.24$$

Esercizio 6.1

La tabella seguente riporta i dati di peso e altezza (kg e cm) di n=30 bambini .

Peso	Altezza
27	125
20	131
21	108
14	102
28	116
18	118
18	108
15	104
15	106
19	108
20	114
19	111
23	125
19	118
25	118
18	108
21	118
20	114
32	116
14	108
21	123
16	111
19	117
15	106
18	116
18	105
20	114
19	111
23	110
17	103

1. Suddividere le variabili in 4 classi di uguale ampiezza e costruire la corrispondente tabella.
2. Calcolare media e varianza delle due variabili prima con i dati della tabella di partenza e poi con quella della tabella con quattro classi.
3. Commentare i risultati ottenuti.

Peso	n_i
Totale	30

Altezza	n_i
Totale	30

	Dati originali	Dati divisi in classi
Media		
Varianza		

Esercizio 6.2

Sono riportati di seguito le durate, in anni, degli studi compiuti da 20 persone:
13, 18, 18, 13, 8, 8, 13, 8, 8, 8, 13, 19, 14, 8, 8, 14, 8, 13, 20, 8.

- ✓ Costruire l'istogramma degli anni di studio

- ✓ Dire quante persone hanno studiato almeno 13 anni.
- ✓ Completare la tabella seguente:

Anni di studio x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
8				
13				
14				
18				
19				
20				
Totali				

- ✓ Utilizzando i valori riportati nella tabella precedente, calcolare:

il valore medio del numero di anni di studio (con una cifra decimale); $m=$	lo scarto quadratico medio del numero di anni di studio (con due cifre decimale) $\sigma =$

Esercizio 6.3

Per un periodo di 20 giorni, il numero di scosse del terreno rilevate da un sismografo; i dati sono i seguenti:

8, 8, 6, 7, 9, 9, 8, 10, 8, 9, 1, 9, 9, 6, 6, 9, 7, 9, 6, 9

- ✓ Tracciare, in un opportuno sistema d'assi, l'istogramma dei dati rilevati.

- ✓ Per quanti giorni si sono rilevate più di 5 e meno di 10 scosse?

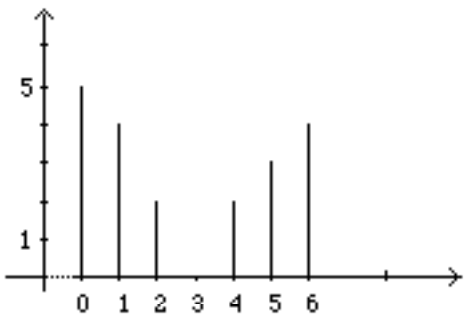
- ✓ Costruire la tabella sotto riportata:

x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
1				
6				
7				
8				
9				
10				
Totali				

- ✓ Utilizzando i valori determinati nella tabella precedente, calcolare la media e lo scarto quadratico medio del numero di scosse al giorno del terreno.

Esercizio 6.4

La figura riporta l'istogramma del numero di puntate di un cartone animato viste da 20 bambini.



- ✓ Dire quanti bambini:

- hanno visto 1 puntata;
- hanno visto 3 puntate;
- hanno visto più di 3 puntate.

- ✓ Costruire la tabella seguente.

x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
0				
1				
2				
3				
4				
5				
6				
Totali				

- ✓ Utilizzando i valori determinati nella tabella precedente, calcolare la media e lo scarto quadratico medio del del numero di puntate viste dai bambini;

7. Relazioni fra variabili quantitative

Le relazioni fra variabili qualitative vengono esaminate costruendo tabelle congiunte o grafici a barre. In che modo puó essere studiata la relazione fra due variabili quantitative X e Y misurate su n unità statistiche?

✓ ***Covarianza***

$$\operatorname{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)=\left(\frac{1}{n}\sum_{i=1}^{n}x_iy_i\right)-\overline{xy}$$

✓ ***Coefficiente di correlazione lineare di Pearson***

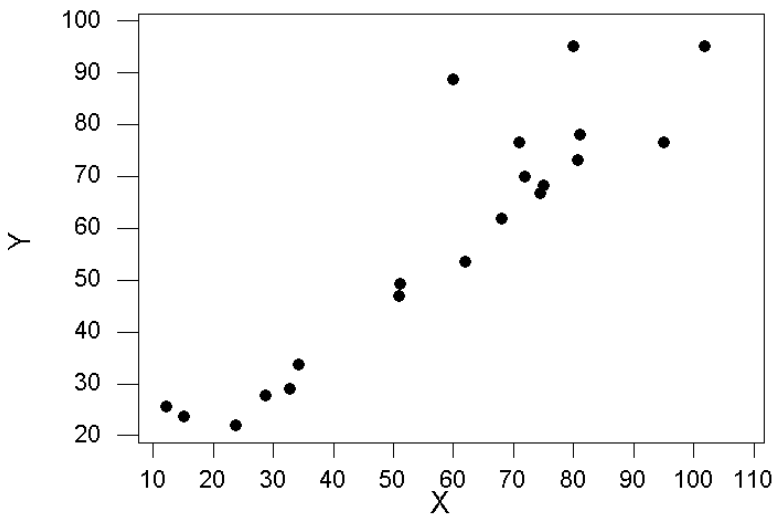
$$\rho(X,Y)=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}}=\frac{Cov(X,Y)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}}$$

La covarianza e la correlazione hanno lo stesso segno. La differenza consiste nel fatto che la correlazione é un indice, cioé un valore indipendente dalle unità di misura delle variabili X e Y.

La covarianza ha invece quale unità di misura il prodotto delle unità di misura di X e Y.

Si puo' inoltre verificare che $-1\le\rho(X,Y)\le 1$.

<u>Esempio 7.1</u>

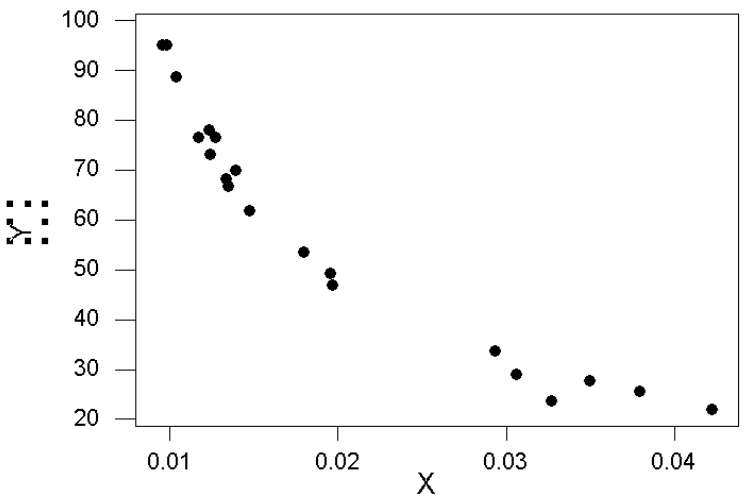


Covarianza positiva

Correlazione= 0.89

Media di X = 58.1 m

Media di Y = 58.6 m

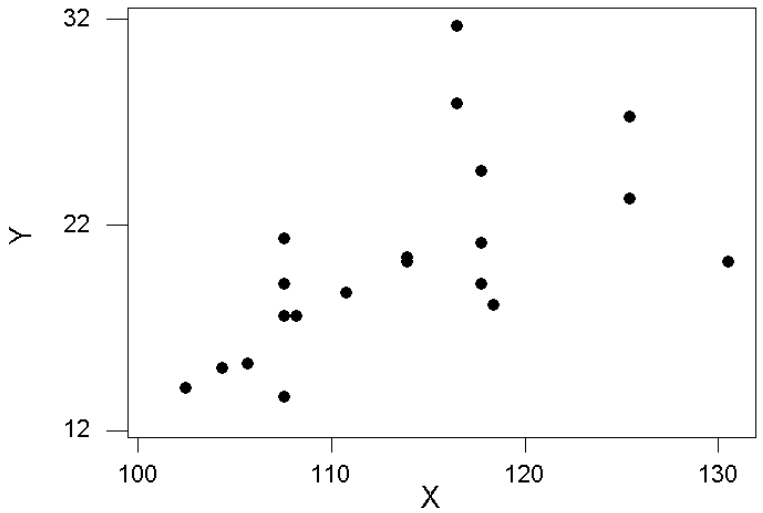


Covarianza negativa

Correlazione= -0.92

Media di X = 0.018 kg

Media di Y = 59 m



Covarianza positiva

Correlazione= 0.59

Media di X = 20.3cm

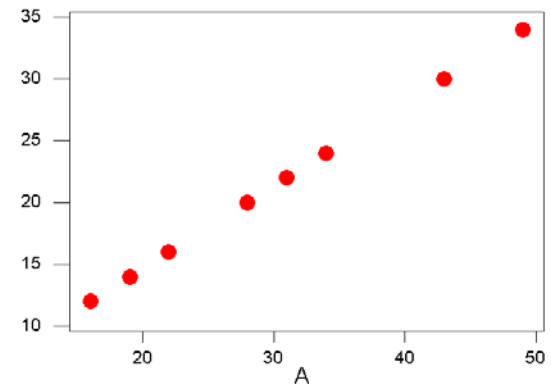
Media di Y = 113.8 dl

Esempio 7.2

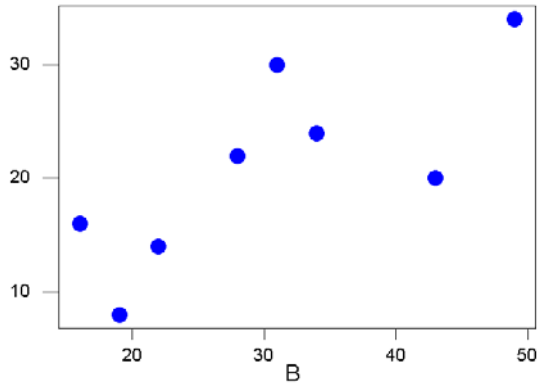
Nella tabella seguente sono riportati i valori di n=9 coppie (x_1,x_2) in cui sono stati cambiati gli appaiamenti fra x_1 e x_2.

A		B		C		D	
x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
10	8	10	12	10	12	10	16
16	12	16	16	16	24	16	14
19	14	19	8	19	8	19	24
22	16	22	14	22	20	22	20
28	20	28	22	28	14	28	22
31	22	31	30	31	16	31	34
34	24	34	24	34	34	34	12
43	30	43	20	43	22	43	30
49	34	49	34	49	34	49	8

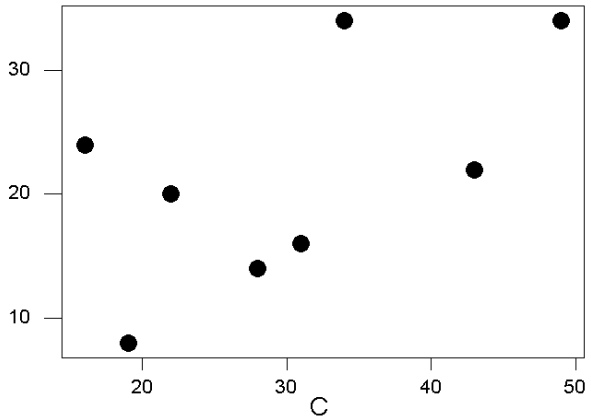
I grafici seguenti rappresentano gli scatterplot di x1 e x2 e il coefficiente di correlazione.



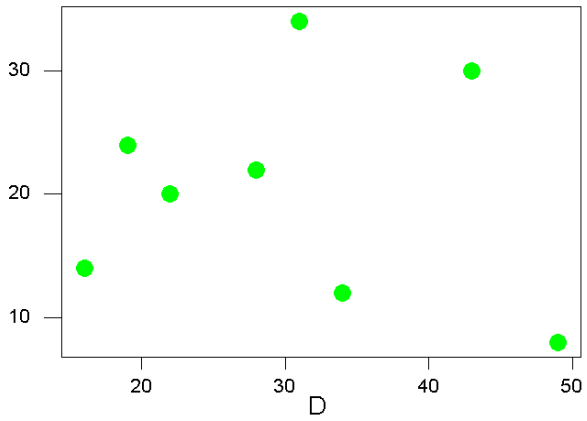
covarianza=86.4
correlazione = 1



covarianza=68.4
correlazione = 0.79



covarianza=51.6
correlazione = 0.60



covarianza=0.0
correlazione = 0.0

<u>Esercizio 7.1</u>

I dati seguenti si riferiscono ai costi di produzione di un prodotto rispetto alla quantita' giornaliera prodotta :

Costi (Y)	12	18	6	14	4	6
n.pezzi (X)	20	24	10	22	14	12

- ✓ Calcolare il valore medio e la varianza di X e Y.
- ✓ Disegnare il grafico delle variabili.
- ✓ Calcolare il valore della covarianza e della correlazione fra X e Y.

8. Relazioni fra variabili qualitative ordinali

- ***Il rango***

Sia *X* una variabile statistica i cui valori sono stati rilevati rispetto ad *n* unità, ottenendo le osservazioni $x_1, x_{2,} x_3, \ldots . x_n$. E' possibile associare a ciascuna osservazione il rango in modo da ottenere una graduatoria.

Il rango di un'osservazione è la posizione che essa occupa nella sequenza ordinata dei dati.

Esempio 8.1

La tabella sottostante riporta il punteggio, indicato mediante la variabile statistica *X*, ottenuto da quattro studenti in un test. Nella terza colonna sono riportati i ranghi delle osservazioni, ossia la posizione che ciascuna di esse occupa nella sequenza ordinata.

Unità	X	Rango
Paolo	7	2
Lucia	12	3
Francesca	24	4
Fabio	3	1

Mediante i ranghi si ottiene la graduatoria fra le unità: Fabio al primo posto, Paolo al secondo, Lucia al terzo e Francesca al quarto.

Se a due o più unità è associato lo stesso valore della variabile *X*, è prassi definire il rango di queste unità come la media aritmetica delle posizioni occupate dai valori uguali.

Esempio 8.2

La tabella sottostante riporta il punteggio, indicato mediante la variabile statistica X, ottenuto da cinque studenti in un test. Lucia e Luca hanno conseguito lo stesso punteggio e quindi occupano, rispetto alla sequenza ordinata dei valori, la terza e la quarta posizione.
Ad entrambe queste unità è associato rango 3.5.

Unità	X	Rango
Paolo	7	2
Lucia	12	3.5
Francesca	24	5
Fabio	3	1
Luca	12	3.5

- ***Coefficiente di correlazione di Spearman.***

Il coefficiente di correlazione dei ranghi di Spearman consente di confrontare due graduatorie al fine di verificare se vi è associazione e, in caso positivo, se vi è concordanza oppure discordanza. Si supponga che in base a due variabili *X* e *Y* siano state definite due graduatorie sulle stesse *n* unità statistiche.
I dati sono riportati come avviene nella tabella sottostante , dove $r(x_i)$ e $r(y_i)$ indicano il rango della *i*-esima unità rispettivamente nella prima e nella seconda graduatoria.

Unità	X	Y
1	$r(x_1)$	$r(y_1)$
2	$r(x_2)$	$r(y_2)$
…	…	…
n	$r(x_n)$	$r(y_n)$

Per ottenere il coefficiente di correlazione dei ranghi è necessario calcolare, per ciascuna unità, le differenze fra i ranghi nelle due graduatorie.
Ad ogni unità è associata la differenza

$$d_i = r(x_i) - r(y_i)$$

Il ***coefficiente di correlazione dei ranghi di Spearman*** (dal nome dello psicologo Charles Spearman che lo ideò nel 1904 avendo quale scopo anche una semplificazione dei calcoli) è dato da:

$$\boxed{\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}}$$

Il coefficiente di correlazione dei ranghi assume valori nell'intervallo $[-1,1]$.
Il coefficiente di correlazione per ranghi può anche essere calcolato utilizzando la formula del coefficiente di correlazione di Pearson dove i valori vengono convertiti in ranghi prima di calcolare il coefficiente.

$$\boxed{\rho_P = \frac{\sum_{i=1}^{n}\left(r(x_i)-\overline{r}_X\right)\left(r(y_i)-\overline{r}_Y\right)}{\sqrt{\sum_{i=1}^{n}\left(r(x_i)-\overline{r}_X\right)^2}\sqrt{\sum_{i=1}^{n}\left(r(y_i)-\overline{r}_Y\right)^2}}, \text{ con } \overline{r}_X = \frac{\sum_{i=1}^{n} r(x_i)}{n} \text{ e } \overline{r}_Y = \frac{\sum_{i=1}^{n} r(y_i)}{n}}$$

Si può dimostrare che $\rho_S = \rho_P$ quando non ci sono dati ripetuti oppure $\rho_S \cong \rho_P$ se il numero di dati ripetuti è piccolo rispetto al numero totale *n* dei dati [vedi allegato n.1]

- ✓ <u>Perfetta concordanza fra le graduatorie</u> $\Rightarrow \rho_S = +1$ **:**

quando il coefficiente di correlazione dei ranghi vale 1 vi è perfetta concordanza fra le graduatorie. Ciò significa che l'unità classificata come prima nella graduatoria rispetto ad X lo è anche nella graduatoria rispetto a *Y*, l'unità classificata come seconda nella graduatoria rispetto ad *X* lo è anche nella graduatoria rispetto a *Y* e così via : $r(x_i) = r(y_i)$ per $i = 1, 2, ..., n$.

- ✓ <u>Perfetta discordanza fra le graduatorie (in assenza di dati ripetuti)</u> $\Rightarrow \rho_S = -1$:

  quando il coefficiente di correlazione dei ranghi vale −1 vi è perfetta discordanza fra le graduatorie. Ciò significa che l'unità classificata come prima nella graduatoria rispetto ad *X* è ultima della graduatoria rispetto a *Y*, l'unità classificata come seconda nella graduatoria rispetto ad *X* è penultima nella graduatoria rispetto a *Y* e così via : $r(x_i) = n - r(y_i) + 1$ per $i = 1, 2, ..., n$.

  Nel caso di dati ripetuti sia ρ_S che ρ_P possono non essere uguali a -1.

- ✓ <u>Assenza di associazione fra le graduatorie</u> $\Rightarrow \rho_S = 0$:

se il coefficiente di correlazione dei ranghi vale zero, ciò indica che non vi è associazione fra le graduatorie.

<u>Esempio 8.3</u> ($\rho_S = \rho_P$ quando non ci sono dati ripetuti)

X	$r(x_i)$	Y	$r(y_i)$
5	1	20	4
10	2	15	3
15	3	10	2
20	4	5	1

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6\left((1-4)^2 + (2-3)^2 + (3-2)^2 + (4-1)^2\right)}{4(16-1)} = -1$$

$$\rho_P = \frac{\sum_{i=1}^{n}\left(r(x_i) - \overline{r}_X\right)\left(r(y_i) - \overline{r}_Y\right)}{\sqrt{\sum_{i=1}^{n}\left(r(x_i) - \overline{r}_X\right)^2}\sqrt{\sum_{i=1}^{n}\left(r(y_i) - \overline{r}_Y\right)^2}} = \qquad [essendo\ \ \overline{r}_X = \overline{r}_Y = 2,5]$$

$$= \frac{(1-2,5)(4-2,5) + (2-2,5)(3-2,5) + (3-2,5)(2-2,5) + (4-2,5)(1-2,5)}{\sqrt{(1-2,5)^2 + (2-2,5)^2 + (3-2,5)^2 + (4-2,5)^2}\sqrt{(4-2,5)^2 + (3-2,5)^2 + (2-2,5)^2 + (1-2,5)^2}} =$$

$$= -1$$

<u>Esempio 8.4</u> ($\rho_S \neq \rho_P$ quando ci sono dati ripetuti)

x_i	$r(x_i)$	y_i	$r(y_i)$
5	1	20	5
10	2,5	12	4
10	2,5	10	2,5
12	4	10	2,5
20	5	5	1

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6\left((1-5)^2 + (2,5-4)^2 + (2,5-2,5)^2 + (4-2,5)^2 + (5-1)^2\right)}{5(25-1)} = -0.825$$

$$\rho_P = \frac{\sum_{i=1}^{n}\left(r(x_i) - \overline{r}_X\right)\left(r(y_i) - \overline{r}_Y\right)}{\sqrt{\sum_{i=1}^{n}\left(r(x_i) - \overline{r}_X\right)^2}\sqrt{\sum_{i=1}^{n}\left(r(y_i) - \overline{r}_Y\right)^2}} = \qquad [essendo\ \ \overline{r}_X = \overline{r}_Y = 3]$$

$$= \frac{(1-3)(5-3) + (2,5-3)(4-3) + (2,5-3)(2,5-3) + (4-3)(2,5-3) + (5-3)(1-3)}{\sqrt{(1-3)^2 + (2,5-3)^2 + (2,5-3)^2 + (4-3)^2 + (5-3)^2}\sqrt{(5-3)^2 + (4-3)^2 + (2,5-3)^2 + (2,5-3)^2 + (1-3)^2}} =$$

$$= -0,921$$

Esempio 8.5 ($\rho_S \cong \rho_P$ quando i valori ripetuti non sono frequenti rispetto alla numerosità campionaria)

I dati seguenti si riferiscono a valori dei punteggi di due gare (A e B) riferiti a 8 atleti.

Atleti	*Gara A*	*Gara B*	*Rango Gara A*	*Rango Gara B*
1	66	102	2	4,0
2	70	94	4	1,0
3	96	140	7	8,0
4	62	100	1	2,5
5	78	104	5	5,0
6	82	100	6	2,5
7	100	115	8	7,0
8	68	112	3	6,0

- ✓ $\rho_S = 0,542$, $\rho_P = 0,539$

Come si può osservare i valori di ρ_S *e* ρ_P differiscono di poco essendo la differenza pari a 0,003.

Esempio 8.6 ($\rho_S \neq \rho_P$ quando i valori ripetuti sono frequenti rispetto alla numerosità campionaria)

Studenti	Voti materia A	Voti materia B	Rango voti materia A	Rango voti materia B
1	4	5	2,5	4,5
2	4	5	2,5	4,5
3	4	5	2,5	4,5
4	4	5	2,5	4,5
5	5	5	6,5	4,5
6	5	5	6,5	4,5
7	5	5	6,5	4,5
8	5	5	6,5	4,5
9	6	7	11,5	17,5
10	6	7	11,5	17,5
11	6	7	11,5	17,5
12	6	7	11,5	17,5
13	6	7	11,5	17,5
14	6	7	11,5	17,5
15	7	6	17,5	11,5
16	7	6	17,5	11,5
17	7	6	17,5	11,5
18	7	6	17,5	11,5
19	7	6	17,5	11,5
20	7	6	17,5	11,5

- ✓ $\rho_S = 0,652$
- ✓ $\rho_P = 0,616$

Come si può osservare i valori di ρ_S *e* ρ_P differiscono abbastanza essendo la differenza pari a 0,036.

Esercizio 8.1

Nella seguente tabella (estratta dal sito www.morningstar.it) sono riportati il rendimento YTD% (Year To Date) e il rischio Dev Std (3a) (calcolato come la deviazione standard dei rendimenti negli ultimi tre anni) al 03 maggio 2010 di alcuni fondi azionari gestiti da Generali.

- ✓ Costruire la graduatoria dei fondi rispetto al Rendimento e al Rischio.
- ✓ Costruire il grafico relativo ai ranghi delle due variabili.
- ✓ Calcolare il coefficiente di correlazione dei ranghi delle variabili del punto 1.

Nome Fondo	YTD%	Dev Std (3 a)
Generali IS AR Credit Strategies AX	9,83	0,00
Generali IS AR Credit Strategies AY	14,19	21,81
Generali IS AR Credit Strategies BX	14,19	21,69
Generali IS AR Credit Strategies CX	14,19	21,61
Generali IS AR Credit Strategies DX	14,18	21,54
Generali IS AR Credit Strategies EX	14,18	21,46
Generali IS AR Global Macro AX	3,38	0,44
Generali IS AR Global Macro BX	3,60	0,44
Generali IS AR Global Macro CX	3,57	0,28
Generali IS AR Global Macro DX	3,47	0,17
Generali IS AR Global Macro EX	3,56	0,12
Generali IS AR Multi Strategies AX	4,79	2,04

Esercizio 8.2

Nella tabella seguente sono riportati i valori del tempo in secondi sui 100m e 200m di 8 atleti in un meeeting estivo.

Atleta	100 m	200 m
A	10.1	20.11
B	9.98	19.81
C	10.2	19.64
D	10.0	20.25
E	9.89	20.02
F	9.96	19.56
G	10.3	20.12
H	10.5	20.22

- ✓ Costruire la graduatoria delle due variabili (100m e 200m).
- ✓ Costruire il grafico relativo ai ranghi delle due variabili.
- ✓ Calcolare il coefficiente di correlazione dei ranghi delle variabili del punto 1.

Allegato n.1

Dimostrazione dell'equivalenza delle formule relative ai coefficienti di correlazione per ranghi di Spearman e Pearson nel caso di dati NON ripetuti.

In questa dimostrazione si indica, per semplicità di notazione, con r_i il rango di x_i e con s_i il rango di y_i.

Se le variabili X e Y sono rilevate su N unità statistiche e non ci sono valori di x_i e y_i ripetuti, i rispettivi ranghi r_i e s_i assumono i valori interi fra 1 e N.

Essendo :

- $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$
- $\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$

resta facile calcolare:

- $\bar{r} = \bar{s} = \frac{\sum_{i=1}^{N} i}{N} = \frac{N(N+1)}{2N} = \frac{N+1}{2}$
- $\sigma_r^2 = \frac{\sum_{i=1}^{N}(r_i - \bar{r})^2}{N} = \frac{\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2}{N} = \frac{\sum_{i=1}^{N} i^2}{N} - \left(\frac{(N+1)}{2}\right)^2 = \frac{(N+1)(2N+1)}{6} - \left(\frac{(N+1)}{2}\right)^2 = \frac{N^2-1}{12}$
- $\sigma_s^2 = \frac{\sum_{i=1}^{N}(s_i - \bar{s})^2}{N} = \frac{\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2}{N} = \frac{\sum_{i=1}^{N} i^2}{N} - \left(\frac{(N+1)}{2}\right)^2 = \frac{(N+1)(2N+1)}{6} - \left(\frac{(N+1)}{2}\right)^2 = \frac{N^2-1}{12}$
- $Cov(r,s) = \frac{1}{N}\sum_{i=1}^{N}(r_i - \bar{r})(s_i - \bar{s}) = \frac{1}{N}\sum_{i=1}^{N} r_i s_i - \bar{r}\,\bar{s} = \frac{1}{N}\sum_{i=1}^{N} r_i s_i - \left(\frac{N+1}{2}\right)^2$

Quindi si ottiene

$$\rho_P(r,s) = \frac{Cov(r,s)}{\sigma_r \sigma_s} = \frac{\frac{1}{N}\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\frac{1}{N}\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{n}(s_i - \bar{s})^2}} = \frac{\frac{1}{N}\sum_{i=1}^{N} r_i s_i - \left(\frac{N+1}{2}\right)^2}{\sqrt{\frac{N^2-1}{12}}\sqrt{\frac{N^2-1}{12}}} = \frac{\frac{1}{N}\sum_{i=1}^{N} r_i s_i - \left(\frac{N+1}{2}\right)^2}{\frac{N^2-1}{12}}$$

Con la formula che calcola il coefficiente di correlazione di Spearman si ottiene :

$$\rho_S(r,s)=1-\frac{6\sum_{i=1}^{n}d_i^2}{N(N^2-1)}=1-\frac{6\sum_{i=1}^{n}\left(r_i-s_i\right)^2}{N(N^2-1)}=1-\frac{6\left(\sum_{i=1}^{n}r_i^2+\sum_{i=1}^{n}s_i^2-2\sum_{i=1}^{n}r_is_i\right)}{N(N^2-1)}=$$

$$=1-\frac{6\left(\frac{N(N+1)(2N+1)}{6}+\frac{N(N+1)(2N+1)}{6}-2\sum_{i=1}^{n}r_is_i\right)}{N(N^2-1)}=1-\frac{6\left(\frac{N(N+1)(2N+1)}{3}-2\sum_{i=1}^{n}r_is_i\right)}{N(N^2-1)}$$

Resta da dimostrare che:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}r_is_i-\left(\frac{N+1}{2}\right)^2}{\frac{N^2-1}{12}}=1-\frac{6\left(\frac{N(N+1)(2N+1)}{3}-2\sum_{i=1}^{n}r_is_i\right)}{N(N^2-1)}$$

Infatti:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}r_is_i-\left(\frac{N+1}{2}\right)^2}{\frac{N^2-1}{12}}=1-\frac{6\left(\frac{N(N+1)(2N+1)}{3}-2\sum_{i=1}^{n}r_is_i\right)}{N(N^2-1)}$$

$$\frac{\frac{4\sum_{i=1}^{N}r_is_i-N\left(N+1\right)^2}{4N}}{\frac{N^2-1}{12}}=\frac{N(N^2-1)-2\left(N(N+1)(2N+1)\right)+12\sum_{i=1}^{n}r_is_i}{N(N^2-1)}$$

$$\frac{\cancel{12\sum_{i=1}^{N}r_is_i}-3N\left(N+1\right)^2}{\cancel{N\left(N^2-1\right)}}=\frac{N\left(N^2-1\right)-2\left(N(N+1)(2N+1)\right)+\cancel{12\sum_{i=1}^{n}r_is_i}}{\cancel{N(N^2-1)}}$$

$$-3N^3-6N^2-3N=N^3-N-2(N^2+N)(2N+1)$$

$$-3N^3-6N^2-3N=N^3-N-4N^3-2N^2-4N^2-2N$$

$$-3N^3-6N^2-3N=-3N^3-6N^2-3N$$

9. Regressione lineare fra variabili quantitative

Si effettua una regressione lineare quando si vuole stabilire se esiste un rapporto di dipendenza lineare fra due variabili X e Y rilevate sulle stesse unita' statistiche.

Si vuole in altre parole stabilire se la variabilita' di Y puo' essere (anche solo in parte) imputata alla variabile X.

Si noti che in questo caso non e' di interesse stabilire se le variabili X e Y variano in modo congiunto e stabilire il loro grado di associazione.

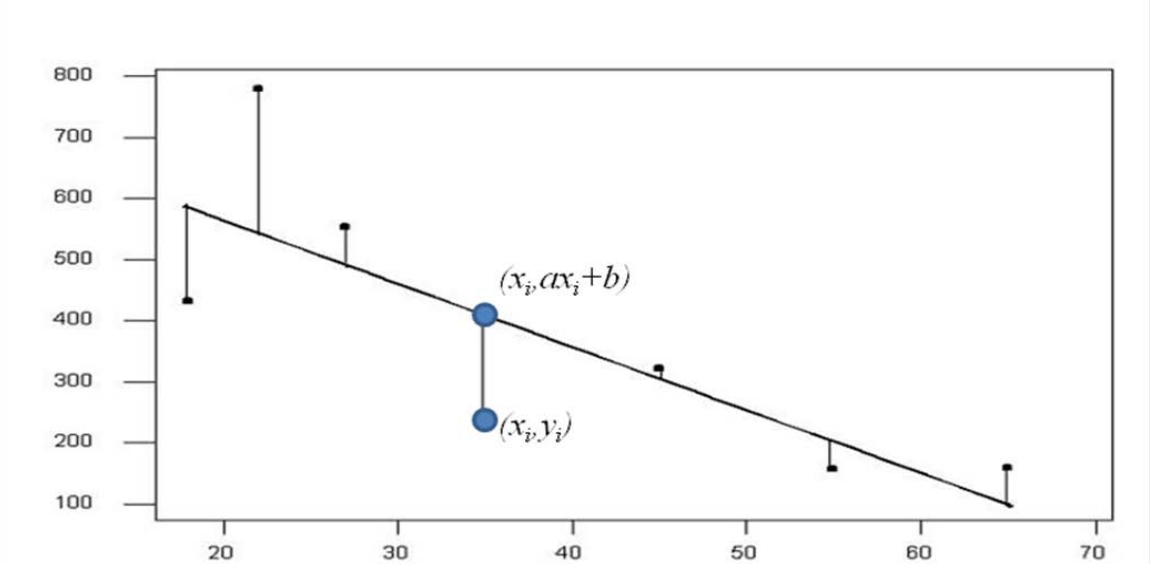
Tecnicamente il problema consiste nel cercare due valori reali a e b tali che la retta y=ax+b *passi il piu' vicino possibile a quasi tutti i punti.*

La determinazione analitica dei valori di a e b si effettua cercando di minimizzare la quantita' seguente :

$$\sum_{i=1}^{n}\left(y_i - \left(ax_i + b\right)\right)^2$$

Nota : perche' si minimizzano i quadrati delle differenze e non le differenze, e non il valore assoluto delle differenze ?

Si puo' notare che la quantita' $y_i - \left(ax_i + b\right)$ rappresenta la differenza di ordinata fra il valore della retta nel punto x_i e il punto y_i come riportato nella figura seguente.



I valori di a e b che vengono determinati in modo analitico sono :

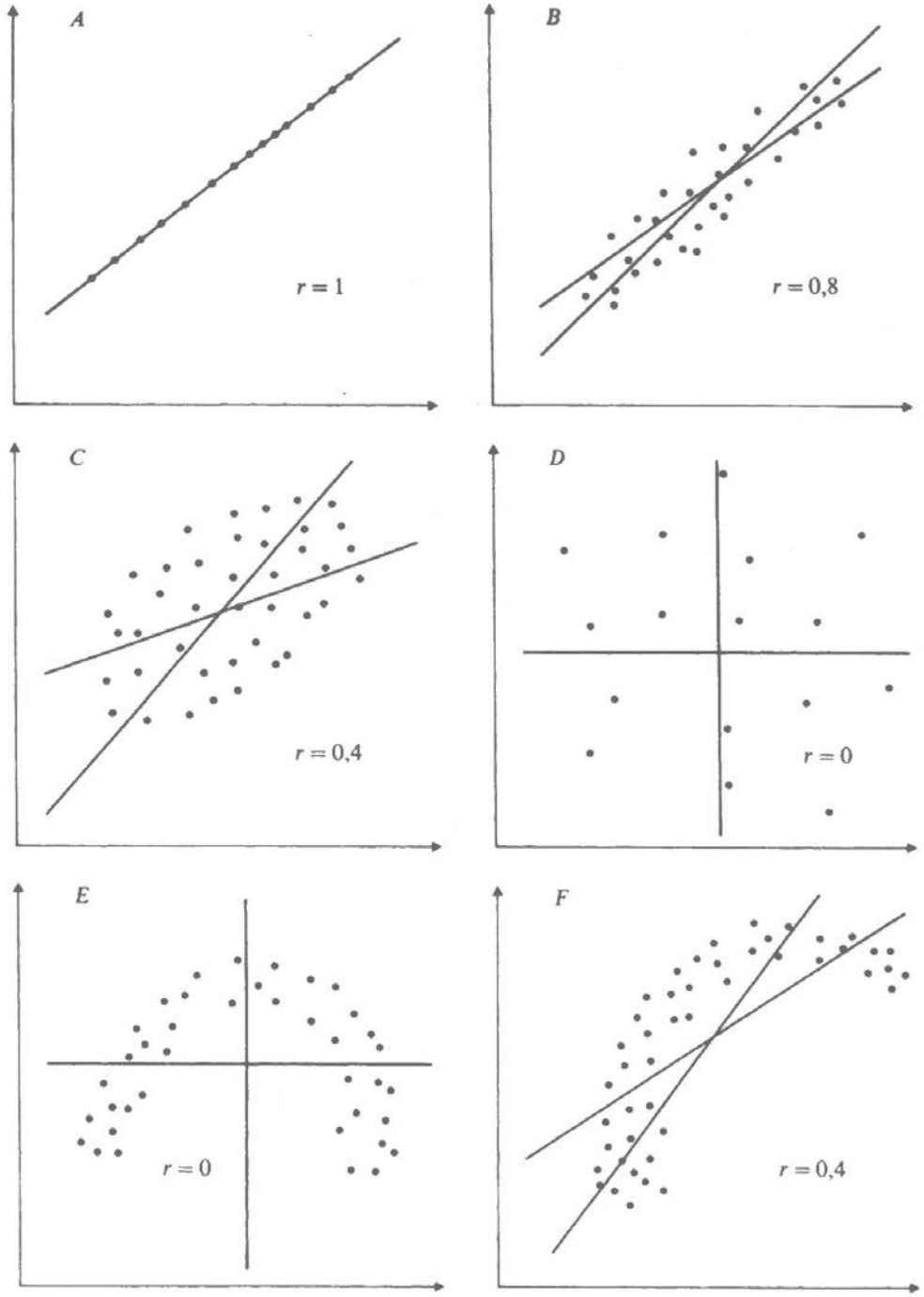
$$a = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} \qquad b = \overline{y} - \overline{x}\left(\frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}\right)$$

L'equazione della retta sara' quindi :

$$y = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}(x - \overline{x}) + \overline{y}$$

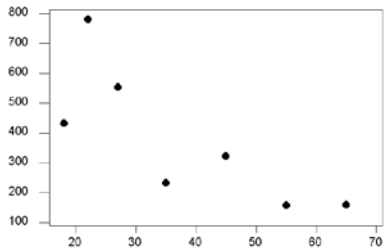
I valori che si ottengono sostituendo i valori di x_i nella formula sopra e che rappresentano i valori della variabile *y* stimati dalle retta di regressione si indicano con $\hat{y}_i$. Le quantita' $r_i = y_i - \hat{y}_i$ si chiamano residui.

I grafici seguenti rappresentano alcuni esempi di scatterplot di variabili X e Y, i coefficienti di correlazione e il grafico delle rette di regressione y in funzione di X e viceversa.



Ad esempio si vuole effettuare una regressione fra le variabili X e Y i cui valori sono riportati in tabella.

X	*Y*
18	433
22	782
27	555
35	232
45	322
55	157
65	158



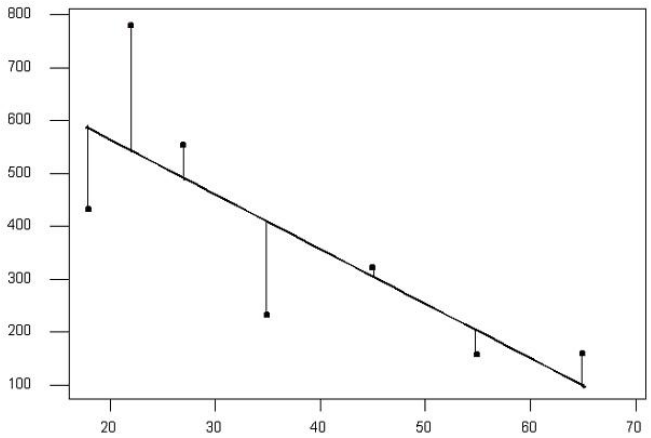
L'equazione della retta (con i valori approssimati agli interi) e' :

$$y = 771 - 10x$$

Per calcolare i valori dei residui e' bene costruire la tabella seguente :

x	y	$ax+b$	$y-(ax+b)$
18	433		
22	782		
27	555		
35	232		
45	322		
55	157		
65	158		

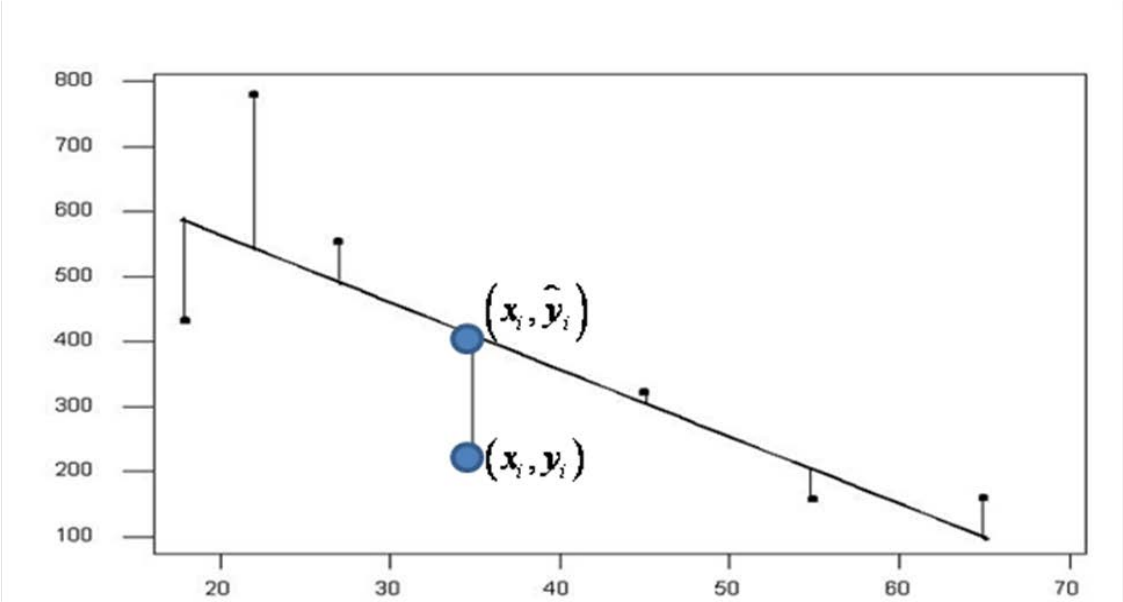
I segmenti verticali che uniscono la retta e i punti (x,y) rappresentano i valori dei residui.



9.1 Analisi della Varianza

Siano (x_1,y_1), (x_2,y_2),…,(x_n,y_n) i valori di due variabili quantitative X e Y rilevati su n unità statistiche (osservazioni) e sia $\hat{y} = \frac{\text{cov}(X,Y)}{\text{var}(X)}(x-\overline{x})+\overline{y}$ l'equazione della retta di regressione di *Y* in funzione di *X*.

I punti sperimentali si indicano con (x_i, y_i) e quelli sulla retta di regressione con $(x_i, \widehat{y_i})$.



Innanzitutto si può facilmente verificare che $\overline{\hat{y}} = \overline{y}$ se il modello di regressione presenta il termine costante, ovvero $\frac{\text{cov}(X,Y)}{\text{var}(X)}(-\overline{x})+\overline{y} \neq 0$.

Infatti $\overline{\hat{y}} = \frac{\sum_{i=1}^{n}\hat{y}_i}{n} = \frac{\sum_{i=1}^{n}\left(\frac{\text{cov}(X,Y)}{\text{var}(X)}(x_i - \overline{x})+\overline{y}\right)}{n} = \frac{\text{cov}(X,Y)}{\text{var}(X)}\left(\frac{1}{n}\sum_{i=1}^{n}\left((x_i - \overline{x})\right)\right)+\overline{y} = \overline{y}$,

essendo $\sum_{i=1}^{n}\left((x_i - \overline{x})\right) = 0$.

Si vuole dimostrare algebricamente che la varianza di Y si decompone nella somma della varianza di $\hat{Y}$ e della varianza di $r = Y - \hat{Y}$, ovvero che $\sigma_Y^2 = \sigma_{\hat{Y}}^2 + \sigma_r^2$.

Precisamente si ha:

- $Var(Y) = \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$ = varianza totale
- $Var(\hat{Y}) = \sigma_{\hat{Y}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2$ = varianza spiegata dal modello
- $Var(r) = \text{var}(Y - \widehat{Y}) = \sigma_r^2 = \frac{1}{n}\sum_{i=1}^{n}\left(r_i - \overline{r}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ = varianza residua

  essendo

  $$\overline{r} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{COV(X,Y)}{VAR(X)}\left(x_i - \overline{x}\right) - \overline{y}\right) = \overline{y} - \frac{1}{n}\frac{COV(X,Y)}{VAR(X)}\sum_{i=1}^{n}\left(x_i - \overline{x}\right) - \overline{y} = 0$$

$$Var(Y)=\sigma_Y^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i-\hat{y}_i\right)+\left(\hat{y}_i-\overline{y}\right)\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2+\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2+$$

$$+\frac{2}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\overline{y}\right)\overline{y}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2+\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2=\sigma_r^2+\sigma_{\hat{Y}}^2$$

essendo $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\overline{y}\right)=0$.

Infatti :

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\overline{y}\right)=\sum_{i=1}^{n}\left(y_i-\left(\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}(x_i-\overline{x})+\overline{y}\right)\right)\left(\left(\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}(x_i-\overline{x})+\cancel{\overline{y}}\right)-\cancel{\overline{y}}\right)=$$

$$=\sum_{i=1}^{n}\left(y_i-\overline{y}-\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}(x_i-\overline{x})\right)\left(\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}(x_i-\overline{x})\right)=$$

$$=\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)\left(x_i-\overline{x}\right)-\frac{\operatorname{cov}^2(X,Y)}{\operatorname{var}^2(X)}\sum_{i=1}^{n}(x_i-\overline{x})^2=$$

$$=\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}\left(n\operatorname{cov}(X,Y)\right)-\left(\frac{\operatorname{cov}^2(X,Y)}{\operatorname{var}^2(X)}\right)n\operatorname{var}(X)=0$$

Un indice che permette di valutare l'adeguatezza del modello è il coefficiente di determinazione $R^2=\frac{\sigma_{\hat{Y}}^2}{\sigma_Y^2}=\frac{\text{varianza spiegata}}{\text{varianza totale}}$, che rappresenta quale parte della varianza totale è spiegata dalla retta di regressione. Con semplici calcoli è facile verificare che $R^2=1-\frac{\sigma_r^2}{\sigma_Y^2}=1-\frac{\text{varianza residua}}{\text{varianza totale}}$.

Ad esempio se R^2=0,81 significa che la regressione lineare fra *X* e *Y* spiega l'81% della variabilità di *Y* attorno a $\overline{Y}$ ed esiste un residuo di variabilità pari al 19% non spiegata dalla retta di regressione ed attribuibile alla variazione casuale di r_i.

L'indice $R^2=\frac{\sigma_{\hat{Y}}^2}{\sigma_Y^2}$ può essere definito anche come $R^2=\rho^2\left(Y,\hat{Y}\right)$. Inoltre nel nostro caso, in cui vi è una sola variabile esplicativa, vale l'uguaglianza $\rho^2\left(Y,\hat{Y}\right)=\rho^2\left(X,Y\right)$.

Dimostriamo per prima cosa che $\rho^2\left(Y,\hat{Y}\right)=\rho^2\left(X,Y\right)$:

$$\rho^2\left(Y,\hat{Y}\right)=\rho^2\left(Y,aX+b\right)=\left(\frac{a}{|a|}\right)^2\rho^2\left(X,Y\right)=\rho^2\left(X,Y\right).$$

Ora proviamo che $R^2=\frac{\sigma_{\hat{Y}}^2}{\sigma_Y^2}=\rho^2(X,Y)$:

$$\sigma_r^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\overline{y}-\frac{\operatorname{cov}(X,Y)}{\sigma_X^2}(x_i-\overline{x})\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2+\frac{1}{n}\frac{\operatorname{cov}^2(X,Y)}{\sigma_X^4}\sum_{i=1}^{n}(x_i-\overline{x})^2+$$

$$-\frac{2}{n}\frac{\operatorname{cov}(X,Y)}{\sigma_X^2}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sigma_Y^2+\frac{\operatorname{cov}^2(X,Y)}{\sigma_X^2}-2\frac{\operatorname{cov}^2(X,Y)}{\sigma_X^2}=\sigma_Y^2-\frac{\operatorname{cov}^2(X,Y)}{\sigma_X^2}$$

Pertanto

$$\frac{\sigma^2_{\hat{Y}}}{\sigma^2_Y}=1-\frac{\sigma^2_r}{\sigma^2_Y}=1-\frac{\sigma^2_Y-\frac{\operatorname{cov}^2(X,Y)}{\sigma^2_X}}{\sigma^2_Y}=1-1+\frac{\frac{\operatorname{cov}^2(X,Y)}{\sigma^2_X}}{\sigma^2_Y}=\rho^2\left(X,Y\right).$$

Esempio 9.1

I dati seguenti si riferiscono a due variabili X e Y.

- ✓ Calcolare media,varianza, minimo, massimo, Q1 e Q3 di X e Y.
- ✓ Calcolare il coefficiente di correlazione fra X e Y.
- ✓ Disegnare il grafico delle variabili .
- ✓ Scrivere l'espressione della retta di regressione di Y in funzione di X e disegnarla nel grafico precedente.
- ✓ Costruire la tabella dei residui per ognuno dei valori assunti da X.

i	*X*	*Y*
1	16	18
2	28	32
3	33	36
4	31	39
5	31	34

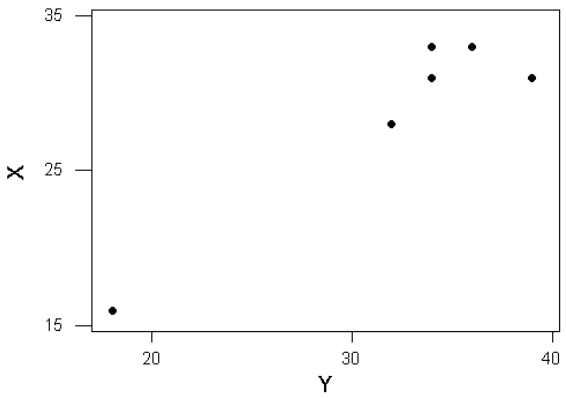
Risoluzione

Per effettuare i conti e' utile costruire la seguente tabella :

i	*X*	*Y*	*X*Y*	X^2	Y^2
1	16	18	288	256	324
2	28	32	896	784	1024
3	33	36	1188	1089	1296
4	31	39	1209	961	1521
5	31	34	1054	961	1156
6	33	34	1122	1089	1156
totale	172	193	5760	5140	6480

$\overline{x}=\frac{1}{6}\sum_{i=1}^{6}x_i=\frac{172}{6}=28.7$	$\operatorname{cov}(X,Y)=\frac{1}{6}\sum_{i=1}^{6}x_iy_i-\overline{x}\,\overline{y}=...$
$\overline{y}=\frac{1}{6}\sum_{i=1}^{6}y_i=...$	$\operatorname{var}(Y)=\frac{1}{6}\sum_{i=1}^{6}y_i^2-\left(\overline{y}\right)^2=...$
$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}}=...$	

	Media	*Varianza*	*Minimo*	*Massimo*	*Q1*	*Q3*
X	28.7	34.9	16	33	25	33
Y			18	39	28.50	

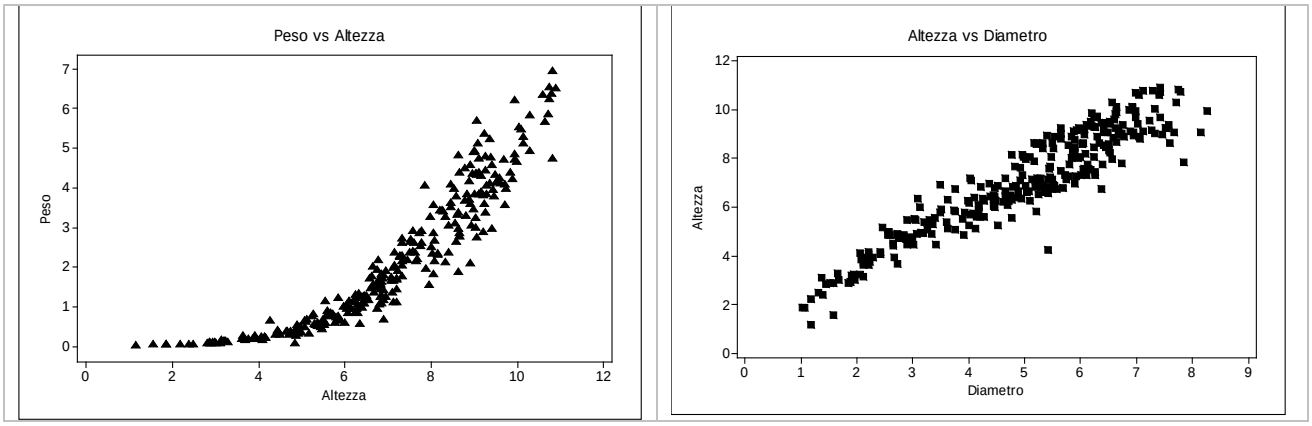


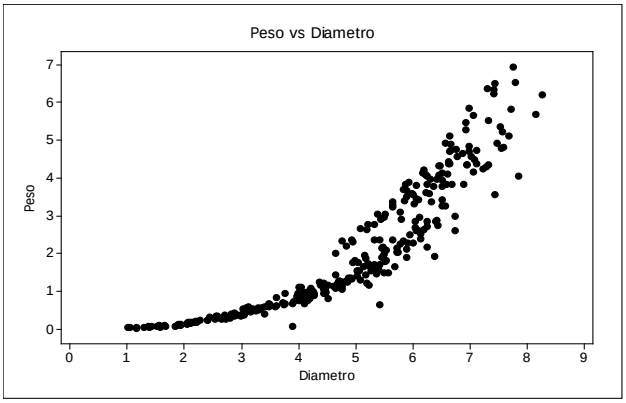
- la retta di regressione e' : $y = \frac{\text{cov}(X,Y)}{\text{var}(X)}(x - \overline{x}) + \overline{y} = ...$
- Tabella dei residui

i	*X*	*Y*	*aX+b*	*Y-(aX+b)*
1	16	18		
2	28	32		
3	33	36		
4	31	39		
5	31	34		
6	33	34		

Esempio 9.2

Allo scopo di massimizzare la produzione di una pianta (peso) dalle caratteristiche particolarmente interessanti chiamata Poplar (una specie simile al Pioppo), ricercatori della Pennsylvania State University piantarono alcuni esemplari di "Poplar Clone 252".
Le variabili esaminate furono diametro(m), altezza (m) e peso (t).
Di seguito sono riportati i grafici bidimensionali delle tre variabili.





Osservando i grafici si nota che le variabili che si possono ipotizzare legate da una relazione lineare sono Altezza e Diametro. Si effettua allora una regressione lineare dell'altezza in funzione del diametro.

Di seguito è riportata una parte dell'output Minitab

Altezza = 1,273 + 1,160 Diametro
R^2 = 87,8%

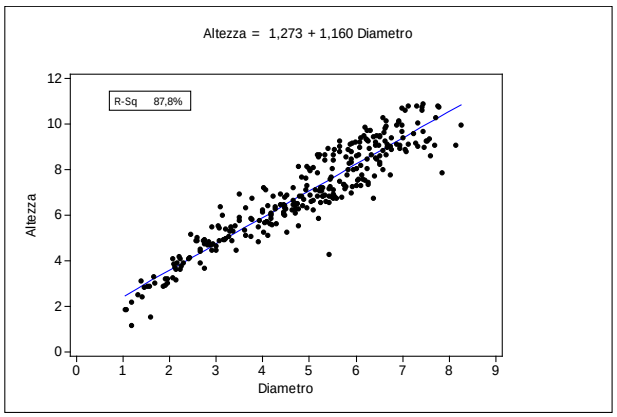
Source	SS
Regression	1146,68
Error	159,12
Total	1305,79

Quindi:

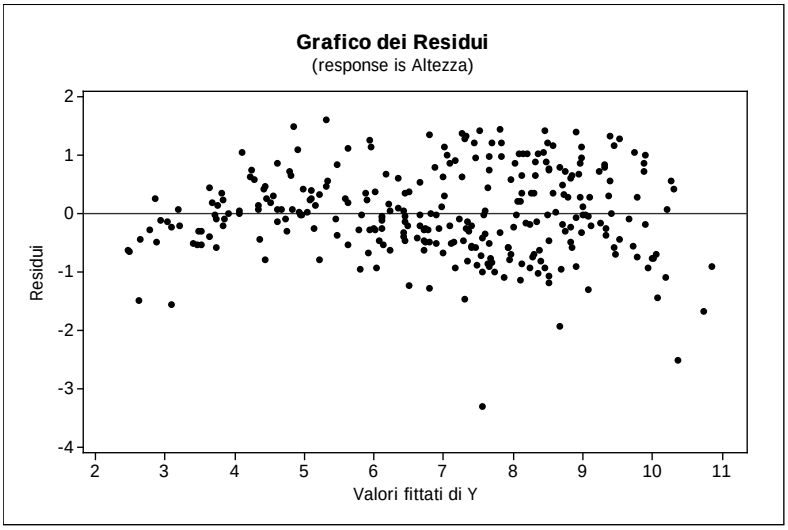
$$\sigma_Y^2 = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}{n} = \frac{SS(total)}{n} = \frac{1305,79}{295} = 4,426$$

$$\sigma_{\hat{Y}}^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2}{n} = \frac{SS(regression)}{n} = \frac{1146,68}{295} = 3,887$$

$$\sigma_r^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} = \frac{SS(error)}{n} = \frac{159,12}{295} = 0,539$$



Il grafico dei residui riportato sotto evidenzia come non siano osservabili relazioni fra le due variabili residui e valori fittati e quindi è possibile concludere che siamo in presenza di una buona approssimazione lineare fra altezza e diametro, confermata anche dal valore elevato di R^2 .

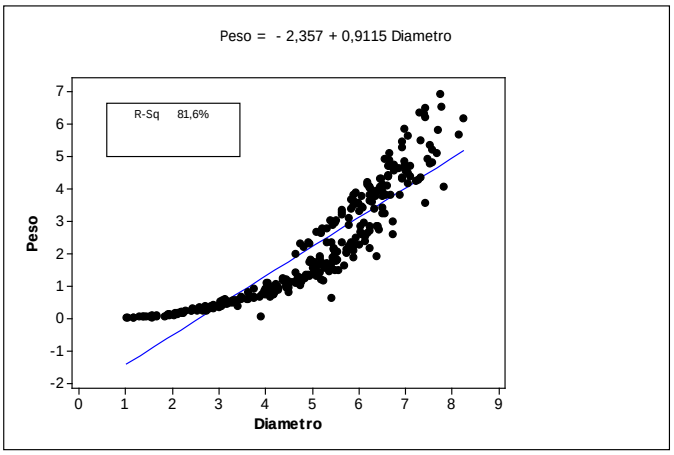


Che cosa succede se si vuole effettuare una regressione lineare se le variabili sembrano non avere una relazione di tipo lineare? Proviamo, ad esempio, ad effettuare una regressione lineare fra peso e diametro.

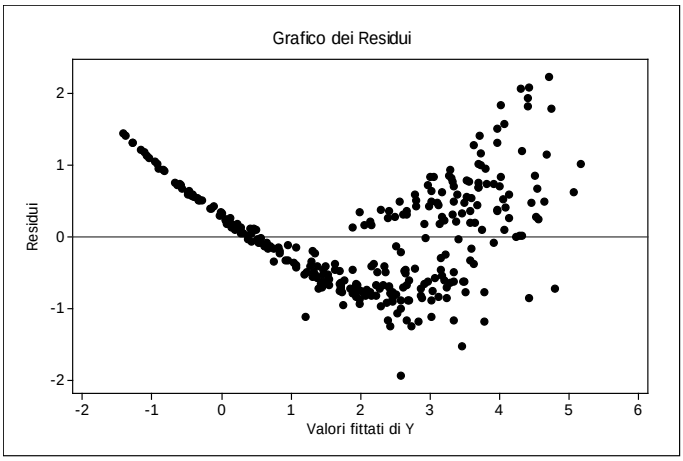
Peso = - 2,357 + 0,9115 Diametro

R^2 = 81,6%

Source	SS
Regression	707,581
Error	159,791
Total	867,372



Il grafico dei residui riportato sotto evidenzia come sia osservabile una relazione di tipo quadratico fra le due variabili residui e valori fittati e quindi è possibile concludere che non siamo in presenza di una buona approssimazione lineare fra altezza e diametro, anche se il valore di R^2 è piuttosto elevato.

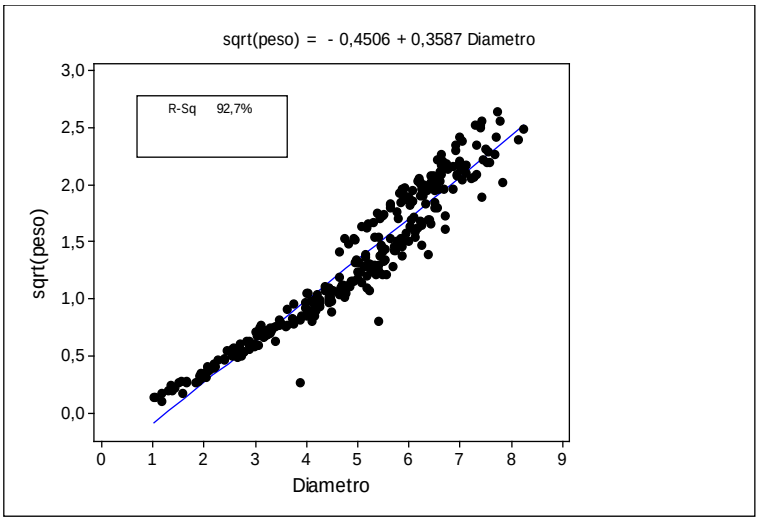


Per ovviare a ciò spesso si effettua una trasformazione della variabile risposta utilizzando l'inversa della funzione che sembra legare le due variabili. Pertanto effettuiamo una regressione lineare fra le variabili *Diametro (X) e* $\sqrt{Peso}$.
Il risultato ottenuto è:

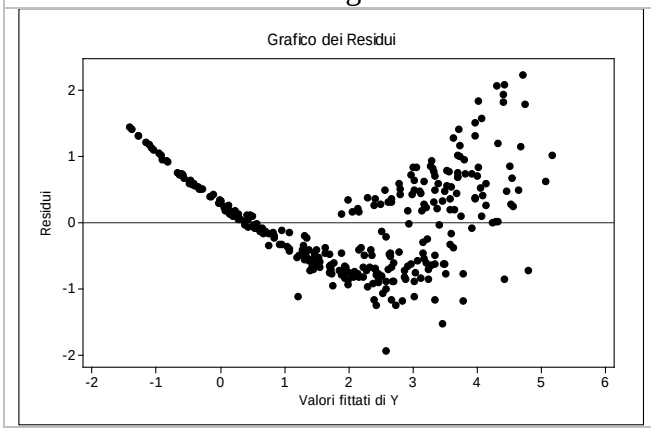
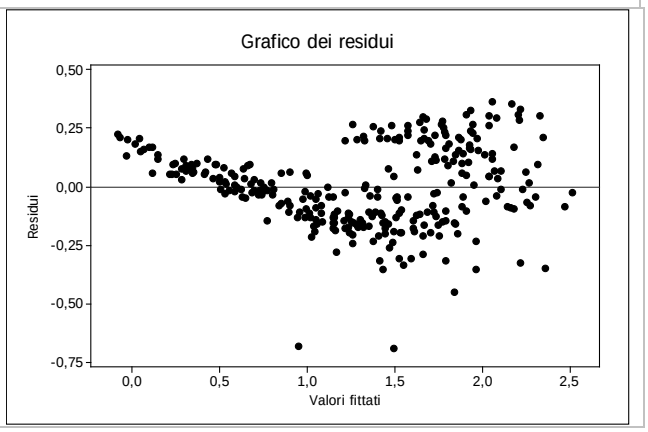
$\sqrt{Peso}$ **= - 0,4506 + 0,3587 Diametro**

R^2 = 92,7%

Source	SS
Regression	109,582
Error	8,663
Total	118,245



I grafici dei residui riportati sotto (a sinistra quello relativo ai dati originali e a destra quello relativo ai dati trasformati) evidenziano come la trasformazione abbia parzialmente eliminato la relazione quadratica fra i residui e i valori fittati dei dati originali e quindi è possibile concludere che siamo in presenza di una buona approssimazione lineare fra $\sqrt{Peso}$ e diametro, confermata anche dal valore più elevato di R^2 (da 81,6% a 92.7%).

Dati originali	Dati trasformati
	

Esercizio 9.1

I dati seguenti si riferiscono ai costi di produzione di un prodotto rispetto alla quantita' giornaliera prodotta :

Costi (Y)	2	8	6	4	4	6
n.pezzi (X)	20	24	10	22	14	12

- ✓ Calcolare il valore medio di X e Y.
- ✓ Disegnare il grafico delle variabili X e Y .
- ✓ Calcolare il valore della correlazione fra X e Y.
- ✓ Scrivere l'espressione della retta di regressione di Y in funzione di X.
- ✓ Calcolare il valore del residuo relativo al valore di x = 24

Esercizio 9.2

I dati seguenti si riferiscono a valori di un parametro chimico del sangue prima (CH1) e dopo (CH2) aver ingerito un farmaco .

n	CH1	CH2
1	66	102
2	70	94
3	96	140
4	62	100
5	78	104
6	82	100
7	100	115
8	68	112

- ✓ Calcolare il valore medio di CH1 e CH2 .
- ✓ Disegnare il grafico delle variabili CH1 e CH2 .
- ✓ Calcolare il valore della correlazione fra CH1 e CH2 .
- ✓ Scrivere l'espressione della retta di regressione di CH2 in funzione di CH1.
- ✓ Costruire la tabella dei residui per ognuno dei valori assunti da CH1.