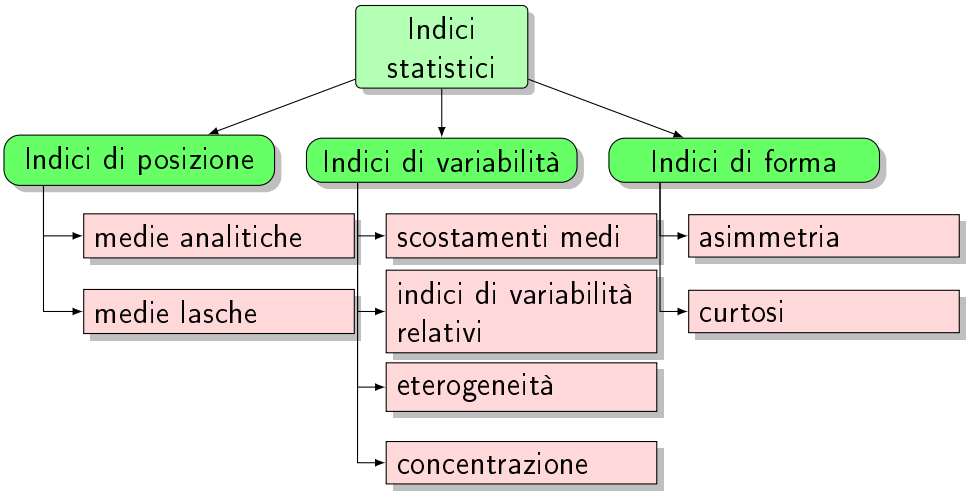
Appunti di Probabilità e Statistica

a.a. 2014/2015 C.d.L. Informatica – Bioinformatica

I. Oliva

Lezione 2

1 Indici statistici



1.1 Indici di variabilità

La variabilità si definisce come l'attitudine di un fenomeno ad assumere modalità differenti. Può essere misurata in diversi modi:

- variabilità delle singole modalità $x_1, x_2, \ldots, x_n$ rispetto ad un indice di posizione

- mutua variabilità
- variabilità delle modalità $x_1, x_2, \ldots, x_n$, ordinate in modo crescente (usando la f. di ripartizione)
- variabilità delle frequenze relative

Proprietà generali degli indici di variabilità (IV):

1. non esiste variabilità se $IV = 0$
2. un indice di variabilità deve assumere valori maggiori o uguali a 0
3. gli indici di variabilità assumono valori crescenti, all'aumentare della variabilità
4. invarianza per traslazioni ($IV(x) = IV(x + a)$)

Varianza: misura la variabilità di un carattere X rispetto alla media aritmetica. Si indica con σ^2 o con $Var(X)$ e si calcola nel modo seguente:

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \quad \text{per dati disaggregati}$$

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 n_i \quad \text{per dati in distribuzione di frequenze assolute}$$

$$\sigma^2 := \sum_{i=1}^{N}(x_i - \mu)^2 f_i \quad \text{per dati in distribuzione di frequenze relative}$$

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N}(\bar{x}_i - \mu)^2 n_i \quad \text{per dati raggruppati in classi}$$

dove n_i indica la frequenza assoluta, f_i indica la frequenza relativa, $\bar{x}_i$ indica il valore centrale della classe (c_{i-1}, c_i).

Formula alternativa per il calcolo della varianza (formula operativa):

$$\sigma^2 = \mu_{x^2} - \mu_x^2 \,,$$

dove μ_x rappresenta la media (aritmetica) rispetto alla modalità x (*momento primo*) e μ_{x^2} rappresenta la media (aritmetica) rispetto al

quadrato della modalità x (*momento secondo*). Più precisamente, si ha:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 \quad \text{per dati disaggregati} \tag{1}$$

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N} x_i^2 n_i - \mu^2 \quad \text{per dati in distr. fr. ass.} \tag{2}$$

$$\sigma^2 := \sum_{i=1}^{N} x_i^2 f_i - \mu^2 \quad \text{per dati in distr. fr. rel.} \tag{3}$$

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N} \bar{x}_i^2 - \mu^2 \quad \text{per dati raggruppati in classi} \tag{4}$$

Dimostriamo l'equazione (1).

Proof.

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i\mu_x + \mu_x^2) \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\sum_{i=1}^{N} 2x_i\mu_x + \frac{1}{N}\sum_{i=1}^{N}\mu_x^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\mu_x\frac{1}{N}\sum_{i=1}^{N} x_i + \mu_x^2 \\
&= \mu_{x^2} - 2\mu_x\mu_x + \mu_x^2 = \mu_{x^2} + \mu_x^2 \,.
\end{aligned}$$

□

Esercizio 1.1. *Verificare le relazioni* (2), (3) *e* (4).

La radice quadrata della varianza si indica con σ e si chiama *scarto quadratico medio* o *deviazione standard*, la cui utilità dipende dal fatto che la varianza si esprime nell'unità di misura *al quadrato* della variabile cui si riferisce.

Esempio 1.1. *Distribuzione delle superfici di appartementi (in mq):* $\{110, 97, 120, 125\}$

Calcoliamo la media (aritmetica):

$$\mu = \frac{110 + 97 + 120 + 125}{4} = 113 \,.$$

Calcoliamo la varianza:

$$\sigma^2 = \frac{(110-113)^2 + (97-113)^2 + (120-113)^2 + (125-113)^2}{4} = 114.5\ .$$

Quindi, $\sigma = \sqrt{\sigma^2} = \sqrt{114.5} = 10.7$

Da questi calcoli, si ricava che la superficie media degli appartamenti è pari a $114.5\,m^2$ *e, mediamente, le superfici differiscono dalla media di* $10.7\,m^2$.

Esempio 1.2. *Distribuzione dei laureati nello scorso a.a. per numero di esami sostenuti alla fine del primo anno di corso: (si conoscono solo il numero degli esami e la corrispondente frequenza assoluta) Dunque:*

Num. esami	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
0	14	0	0	0
1	41	41	1	41
2	83	166	4	332
3	116	348	9	1044
4	56	224	16	896
5	5	25	25	125
Tot	315	804	55	2438

$$\mu = \frac{0+41+83+116+56+5}{315} = 2.55$$
$$\sigma^2 = \frac{0+41+332+1044+896+125}{315} - (2.55)^2 = 1.24$$
$$\sigma = \sqrt{1.24} = 1.11$$

Questi dati ci dicono che il collettivo statistico ha sostenuto ***in media*** 2.55 *esami alla fine del primo anno di corso e che il numero di esami sostenuti al primo anno si discosta dal numero medio di* 1.11, *ossia,* 1.11 *rappresenta la dispersione dei valori intorno alla media.*

Vediamo le proprietà principali soddisfatte dalla varianza.

Proposizione 1.1. *La varianza di* X *risulta essere sempre un numero non negativo ed è pari a* 0 *se e solo se* X *una costante.*

Proposizione 1.2. *Se alla variabile* X *si aggiunge una costante, la varianza non cambia. Se si moltiplica la variabile* X *per una costante,*

la varianza coincide con il prodotto della varianza di X per il quadrato della costante. In simboli,

$$Y := aX + b \Rightarrow Var(Y) = a^2 Var(X) \ .$$

Esercizio 1.2. *Dimostrare la Prop. 1.1 e 1.2.*

Proposizione 1.3. *La situazione di massima variabilità per un collettivo con media μ si ha quando, $x_k = N\mu$ e $x_i = 0$, per ogni $i = 1, \ldots, N$, $k \neq i$. In simboli,*

$$\sigma_{max} = |\mu|\sqrt{N-1} \ .$$

Proof. Supponiamo $x_k = N\mu \neq 0$ e $x_j = 0$, per ogni $j \neq k$, con $k = 1, \ldots, N$. Allora, si ha:

$$\begin{aligned}
\sigma^2_{max} &= \frac{1}{N}\left((N-1)(0-\mu)^2 + (N\mu - \mu)^2\right) \\
&= \frac{1}{N}\left((N-1)\mu^2 + (N^2\mu^2 - 2N\mu + \mu^2)\right) \\
&= \frac{1}{N}\left((N-1)\mu^2 + \mu^2(N^2 - 2N + 1)\right) \\
&= \frac{1}{N}\mu^2\left(N - 1 + N^2 - 2N + 1\right) \\
&= \frac{1}{N}\mu^2\left(N^2 - N\right) \\
&= \frac{1}{N}N\mu^2\left(N-1\right) = \mu^2\left(N-1\right) \ .
\end{aligned}$$

Dunque, $\sigma_{max} = \sqrt{\mu^2\left(N-1\right)} = |\mu|\sqrt{N-1}$. □

Coefficiente di variazione: si indica con CV e rappresenta un indice *relativo* di variabilità, che consente il confronto fra fenomeni rilevati in momenti diversi o espressi in unità di misura diverse.

$$CV := \frac{\sigma}{\mu} \ .$$

Problemi legati a CV:

- CV definito solo per valori positivi della media
- se la media diventa molto piccola, CV assume valori molto grandi

Scostamento medio semplice: permette di determinare se le modalità sono stabili rispetto ad un indice di posizione ritenuto rappresentativo della distribuzione.

$$S_\mu = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu| \quad \text{per dati disaggregati}$$

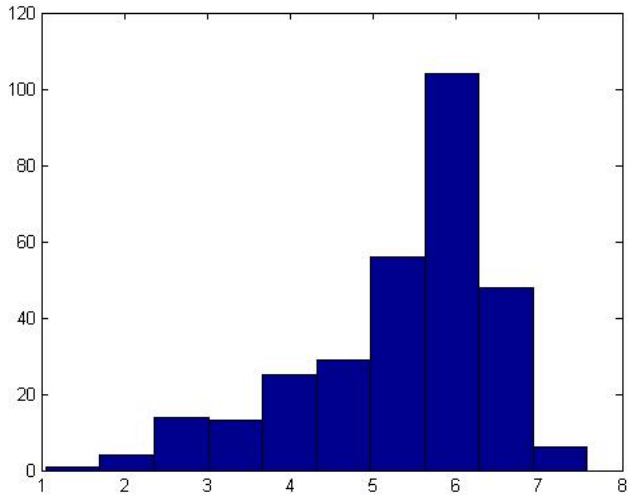
$$S_\mu = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|n_i \quad \text{per dati in distr. fr. ass.}$$

$$S_\mu = \sum_{i=1}^{N}|x_i - \mu|f_i \quad \text{per dati in distr. fr. rel.}$$

$$S_\mu = \frac{1}{N}\sum_{i=1}^{N}|\bar{x}_i - \mu|n_i \quad \text{per dati raggruppati in classi}$$

1.2 Indici di forma

Le distribuzioni aventi stessa posizione e variabilità possono differire tra loro per la forma, che dipende dal valore delle modalità più piccole (o più grandi) del valore centrale della distribuzione.



La forma di una distribuzione si misura attraverso gli indici di *asimmetria* e di *curtosi*.

Asimmetria (skewness): indice che mostra la mancanza di specularità di una distribuzione rispetto a qualsiasi asse verticale.

Per capire meglio di cosa si tratta, partiamo dalla definizione di *simmetria.* Consideriamo la seguente tabella, relativa ad una generica distribuzione con mediana Me:

Modalità	x_1	x_2	...	x_i	...	x_k	tot
Frequenza	n_1	n_2	...	n_i	...	n_k	N

Consideriamo le coppie (x_1, x_k), (x_2, x_{k-1}), (x_3, x_{k-3}) etc etc.

La distribuzione si dice *simmetrica* se, per ciascuna coppia, le modalità sono equidistanti dalla mediana ed hanno la stessa frequenza, in simboli

$$\begin{aligned}(x_i - Me) &= -(x_{k+i-1} - Me)\\ n_i &= n_{k+i-1}\ ,\end{aligned}$$

per ogni $i = 1, 2, \ldots, \frac{k-\delta}{2}$.

Proposizione 1.4. *Una distribuzione simmetrica soddisfa le seguenti proprietà:*

1. *la media aritmetica coincide con la mediana*
2. *la somma degli scarti dalla media, elevati ad una potenza dispari, è nulla*
3. *il primo ed il terzo quartile hanno la stessa distanza dalla mediana.*

Proof. Esercizio □

Una distribuzione non simmetrica si dice *asimmetrica.* In particolare, si parla di *asimmetria positiva,* se la distribuzione presenta una coda verso destra, e di *asimmetria negativa,* se la distribuzione presenta una coda verso sinistra.

Tale asimmetria si misura attraverso indici assoluti o indici relativi. Alla prima categoria appartengono gli indici ottenuti tramite confronto della media con moda e mediana:

$$\alpha_1 := \mu - Me; \quad \alpha_2 := \mu - Mo\ .$$

Ben più utile per una analisi statistica risulta essere il seguente indice relativo di asimmetria, noto come *indice di Fisher:*

$$\gamma := \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3\ .$$

Se $\gamma = 0$, allora la distribuzione è simmetrica. Se $\gamma > 0$, allora la distribuzione è asimmetrica positiva. Se $\gamma < 0$, allora la distribuzione è asimmetrica negativa.

Esempio 1.3. *Consideriamo i dati dell'Esempio 1.2: sappiamo già che* $\mu = 2.55$ *e* $\sigma = 1.11$

Num. esami	n_i	$(x_i - \mu)^3 n_i$
0	14	-232.14
1	41	-152.68
2	83	-13.81
3	116	10.57
4	56	170.72
5	5	73.53
Tot	315	-143.81

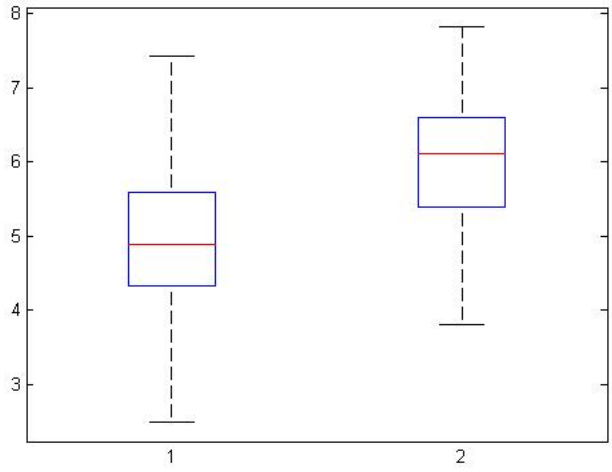
Dunque, si ha:

$$\gamma = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_i - \mu)^3 n_i}{\sigma^3} = \frac{-143.81}{315(1.11)^3} = -0.34 \, .$$

Il modo migliore per verificare se una distribuzione sia simmetrica o meno consiste nel guardare opportuni grafici. Gli istogrammi offrono un notevole aiuto in questo senso, ma la rappresentazione grafica maggiormente utile è data dal *box-plot.* Si tratta di un *grafico a scatola,* nel quale i dati presi in considerazione sono i quartili della distribuzione ed il campo di variazione (i.e., il range in cui variano i valori della distribuzione).

Si costruisce come segue:

- calcolo del valore minimo e del valore massimo della distribuzione
- calcolo del primo, secondo e terzo quartile
- costruzione di un rettangolo i cui estremi sono Q_1 e Q_3 e la cui larghezza vale $Q_3 - Q_1$
- individuazione della mediana all'interno del box (si traccia una linea in corrispondenza)
- collegamento del box ai valori minimo e massimo



È possibile che alcuni valori della distribuzione non cadano all'interno della scatola. In questo caso, si parla di *outliers* o *valori anomali.* Essi vengono determinati dal raffronto con dei valori-soglia. Più precisamente, x è un outlier se

$$x < Q_1 - 1.5 \times (Q_3 - Q_1) \text{ oppure } x > Q_3 + 1.5 \times (Q_3 - Q_1) \, .$$

Curtosi: misura la maggiore o minore sporgenza di una curva di distribuzione in prossimità del suo massimo e la maggiore o minore lunghezza delle code.

Si misura attraverso *l'indice di curtosi di Pearson:*

$$\beta := \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^4 \quad \text{per dati disaggregati}$$

$$\beta := \frac{1}{\sigma^4} \frac{\sum_{i=1}^{N} (x_i - \mu)^4 n_i}{\sum_{i=1}^{N} n_i} \quad \text{per distr. freq. ass.}$$

$$\beta := \frac{1}{\sigma^4} \frac{\sum_{i=1}^{N} (\bar{x}_i - \mu)^4 n_i}{\sum_{i=1}^{N} n_i} \quad \text{per dati raggruppati in classi}$$

A questo si aggiunge *l'indice di curtosi di Fisher* $\gamma_2 = \beta - 3$, che permette un migliore confronto dei dati. Infatti, la distribuzione non ha curtosi se $\gamma_2 = 0$, ossia, $\beta = 3$. La distribuzione di dice *leptocurtica* se $\beta > 3$ (quindi $\gamma_2 > 0$) oppure *platicurtica* se $\beta < 3$ (e $\gamma_2 < 0$).

2 Le relazioni tra i fenomeni

2.1 Distribuzioni multiple

Le distribuzioni statistiche finora analizzate prevedevano che su ciascuna delle n unità statistiche fosse rilevato un solo carattere, le cui modalità venivano descritte attraverso distribuzioni *unitarie,* ossia dati disaggregati, o attraverso distribuzioni di frequenze. In questo caso, si parla di **distribuzioni semplici.**

Quando, invece, su ciascuna delle n unità statistiche vengono rilevati due o più caratteri, si hanno **distribuzioni statistiche multiple.** Anche in questo caso, i dati vengono raccolti in tabelle a doppia entrata, chiamate *matrici di contingenza.*

Esempio 2.1. *Consideriamo la seguente distribuzione doppia, di un campione di famiglie secondo il numero di vani (Y) dell'appartamento in cui vivono ed il numero di componenti (X) il nucleo familiare:*

	2	3	4	5	6	TOT
1	2	3	0	0	0	5
2	0	4	3	0	1	8
3	0	2	2	4	0	8
4	0	1	0	3	1	5
5	0	0	0	1	1	2
TOT	2	10	5	8	3	28

Come si leggono i risultati di questa tabella?

- *Il dato in rosso rappresenta il numero di famiglie, composte da 3 elementi ciascuna, che vivono in altrettanti appartamenti da 5 vani ciascuno.*
- *Il dato in blu rappresenta il numero di famiglie costituite da un solo componente, a prescindere dal numero di vani di cui è composto il suo appartamento.*
- *Il dato in verde rappresenta il numero di famiglie che vive in appartamenti da 3 vani, senza interessarsi del numero dei componenti.*
- *Il dato in giallo rappresenta il numero totale di famiglie prese in esame.*

	y_1	y_2	$\cdots$	y_t	TOT
x_1	n_{11}	n_{12}	$\cdots$	n_{1t}	n_{10}
x_2	n_{21}	n_{22}	$\cdots$	n_{2t}	n_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_s	n_{s1}	n_{s2}	$\cdots$	n_{st}	n_{s0}
TOT	n_{01}	n_{02}	$\cdots$	n_{0t}	N

L'esempio precedente può essere generalizzato. Consideriamo due caratteri: X, con modalità $x_1, \ldots, x_s$, e Y, con modalità $y_1, \ldots, y_t$. La matrice di contingenza assume la seguente forma:

La parte centrale della tabella, ossia:

	y_1	y_2	$\cdots$	y_t	
x_1	n_{11}	n_{12}	$\cdots$	n_{1t}	
x_2	n_{21}	n_{22}	$\cdots$	n_{2t}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_s	n_{s1}	n_{s2}	$\cdots$	n_{st}	
					N

rappresenta la *distribuzione di frequenza assoluta* ***congiunta*** delle due variabili, mentre i bordi della tabella (ultima riga e ultima colonna) rappresentano la *distribuzione di frequenza assoluta* ***marginale*** delle singole variabili, dove:

- n_{ij} rappresenta il numero di unità con la modalità i-sima di X e con la modalità j-sima di Y, ossia, la modalità assoluta congiunta della coppia (x_i, y_j). Inoltre, si ha
$$\sum_{i=1}^{s}\sum_{j=1}^{t} n_{ij} = N \ .$$
- $n_{i0} := \sum_{j=1}^{t} n_{ij}, \ \forall\, i = 1, \ldots, s$ rappresenta la *distribuzione marginale* del carattere X (i.e., la distribuzione di X, indipendentemente da Y)
- $n_{0h} := \sum_{j=1}^{s} n_{jh}, \ \forall\, h = 1, \ldots, t$ rappresenta la *distribuzione marginale* del carattere Y (i.e., la distribuzione di Y, indipendentemente da X)
- $N := \sum_{i=1}^{s} n_{i0} = \sum_{h=1}^{t} n_{0h}$

- la colonna j-sima della matrice dei dati rappresenta la *distribuzione condizionata* del carattere X (i.e., la distribuzione di X, in corrispondenza di ciascuna modalità di Y) e si indica con

$$X|Y = y_j \text{ oppure } X|y_j \ .$$

- la riga i-sima della matrice dei dati rappresenta la *distribuzione condizionata* del carattere Y (i.e., la distribuzione di Y, in corrispondenza di ciascuna modalità di X) e si indica con

$$Y|X = x_i \text{ oppure } Y|x_i \ .$$

Le tabelle di contingenza possono essere costruite anche nel caso di distribuzioni di frequenza relativa:

	y_1	y_2	$\cdots$	y_t	TOT
x_1	f_{11}	f_{12}	$\cdots$	f_{1t}	f_{10}
x_2	f_{21}	f_{22}	$\cdots$	f_{2t}	f_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_s	f_{s1}	f_{s2}	$\cdots$	f_{st}	f_{s0}
TOT	f_{01}	f_{02}	$\cdots$	f_{0t}	1

dove:

- $f_{ij} = \frac{n_{ij}}{N}$ rappresenta la frequenza relativa congiunta della coppia (x_i, y_j). Inoltre, si ha

$$\sum_{i=1}^{t}\sum_{j=1}^{s} f_{ij} = 1 \ .$$

- $f_{i0} = \frac{n_{i0}}{N} := \sum_{j=1}^{t} f_{ij}, \ \forall\, i = 1, \ldots, s$ rappresenta la frequenza relativa marginale del carattere X

- $f_{0h} = \frac{n_{0h}}{N} := \sum_{i=1}^{s} n_{ih}, \ \forall\, h = 1, \ldots, t$ rappresenta la frequenza relativa marginale del carattere Y

Abbiamo detto che le frequenze marginali e condizionate di X e Y sono univariate, dunque possiamo calcolare medie e varianze marginali e condizionate:

marginali: $M(X) = \frac{1}{N}\sum_{i=1}^{t} x_i \cdot n_{i0} = \sum_{i=1}^{t} x_i \cdot f_{i0}$,
$Var(X) = \frac{1}{N}\sum_{i=1}^{t} x_i^2 \cdot n_{i0} - M^2(X)$

$M(Y) = \frac{1}{N}\sum_{j=1}^{s} y_j \cdot n_{0j} = \sum_{i=1}^{s} y_j \cdot f_{0j}$,
$Var(Y) = \frac{1}{N}\sum_{j=1}^{s} y_j^2 \cdot n_{0j} - M^2(Y)$

condizionate: $M(X|Y = y_j) = \frac{1}{n_{0j}}\sum_{i=1}^{t} x_i \cdot n_{ij}$
$Var(X|Y = y_j) = \frac{1}{n_{0j}}\sum_{i=1}^{t} x_i^2 \cdot n_{ij} - M^2(X|Y = y_j)$

$M(Y|X = x_i) = \frac{1}{n_{i0}}\sum_{j=1}^{s} y_j \cdot n_{ij}$
$Var(Y|X = x_i) = \frac{1}{n_{i0}}\sum_{j=1}^{s} y_j^2 \cdot n_{ij} - M^2(Y|X = x_i)$

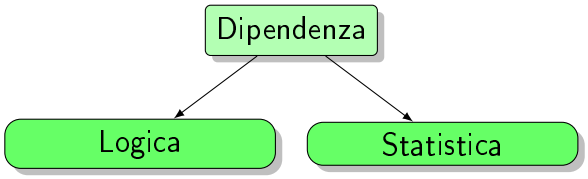
Esempio 2.2. *Consideriamo un campione di 20 studenti del Corso di Statistica ed esaminiamo i caratteri* sesso X *e* voto d'esame Y :

	26	28	30	TOT
uomo	4	2	4	10
donna	1	8	1	10
TOT	5	10	5	20

$$M(Y) = \frac{26 \cdot 5 + 28 \cdot 10 + 30 \cdot 5}{20} = 28$$
$$Var(Y) = \frac{26^2 \cdot 5 + 28^2 \cdot 10 + 30^2 \cdot 5}{20} - 28^2 = 2$$
$$M(Y|X = uomo) = \frac{26 \cdot 4 + 28 \cdot 2 + 30 \cdot 4}{10} = 28$$
$$M(Y|X = donna) = \frac{26 \cdot 1 + 28 \cdot 8 + 30 \cdot 1}{10} = 28$$
$$Var(Y|X = uomo) = \frac{26^2 \cdot 4 + 28^2 \cdot 2 + 30^2 \cdot 4}{10} - 28^2 = 3.2$$
$$Var(Y|X = donna) = \frac{26^2 \cdot 1 + 28^2 \cdot 8 + 30^2 \cdot 1}{10} - 28^2 = 0.8$$

2.2 Dipendenza tra variabili

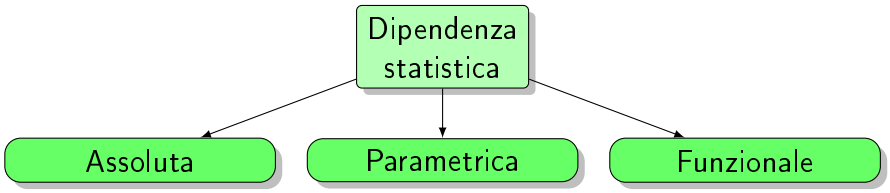
Nella pratica, si è interessati a stabilire se certi fenomeni influenzano o meno un carattere statistico ed, in caso affermativo, si vuol misurare l'intensità di tale influenza. Ciò equivale a studiare la *dipendenza* tra caratteri.



Dipendenza logica: esiste una relazione di causa/effetto tra i caratteri (e.g., la statura è influenzata dalla alimentazione)

Dipendenza statistica: tra i caratteri esistono delle regolarità nella associazione tra le modalità (e.g., reddito e spesa consumi)

Noi focalizziamo l'attenzione sulla dipendenza statistica e vogliamo *capire* se esiste dipendenza e *misurare* l'eventuale intensità.



Dip. assoluta: diciamo che un carattere X è *indipendente* da Y se assume la stessa distribuzione condizionata per ciascuna modalità di X. In simboli, si ha:

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i0}}{N}, \quad \text{per ogni } i = 1, \ldots, s,\ j = 1, \ldots, t\ .$$

Se X è indipendente da Y, allora Y è indipendente da X.

Se la condizione di indipendenza non è soddisfatta, allora si parla di caratteri *dipendenti.* Diremo che Y *dipende perfettamente* da X se, in corrispondenza di ogni modalità di X, si verifica una sola modalità di Y. In termini matematici, questo vuol dire che

$$\forall\, i,\ \exists!\, j\ :\ n_{ij} = 0\ .$$

Diremo, allora, che X è *perfettamente interindipendente* con Y se i due caratteri dipendono perfettamente l'uno dall'altro.

La connessione tra due caratteri si misura attraverso *l'indice quadratico di connessione* o *indice chi-quadrato:*

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{c_{ij}^2}{\hat{n}_{ij}} ,$$

dove:

- n_{ij}, per ogni $i = 1, \ldots, s$ e per ogni $j = 1, \ldots, t$, sono le frequenze assolute osservate
- $\hat{n}_{ij} := \frac{n_{i0} \cdot n_{0j}}{N}$, per ogni $i = 1, \ldots, t$ e per ogni $j = 1, \ldots, s$, sono le **frequenze teoriche** assolute congiunte, ossia i valori che si avrebbero se i caratteri fossero indipendenti
- $c_{ij} = n_{ij} - \hat{n}_{ij}$ sono le **contingenze,** ossia le differenze tra le frequenze effettive e quelle teoriche

Proprietà dell'indice chi-quadrato:

1. $0 \leq \chi^2 \leq N \cdot \min(s-1, r-1)$.
   In particolare, $\chi^2 = 0$ in caso di indipendenza assoluta dei caratteri e $\chi^2 = N \cdot \min(s-1, r-1)$ in caso di massima dipendenza.
2. Per poter effettivamente quantificare la dipendenza, occorre normalizzare l'indice chi-quadrato:

   $$\nu = \sqrt{\frac{\chi^2}{N \cdot \min(s-1, r-1)}} .$$

   Tale valore, che varia tra 0 e 1, prende il nome di *indice di Cramer.*
3. Chi-quadrato è un indice simmetrico, nel senso che rimane invariato se invertiamo X e Y.

Esempio 2.3. *Prendiamo in considerazione i dati relativi ai minorenni condannati secondo la tipologia di delitto (Y) ed il sesso (X) [Fonte: dati Istat, 1989]*

	Omicidio	Lesioni	Furto	Altro	TOT
uomo	37	35	913	1102	2087
donna	3	2	219	51	275
TOT	40	37	1132	1153	2362

Vogliamo misurare il grado di dipendenza tra i due caratteri. Per poter fare ciò, occorre costruire la tabella delle frequenze teoriche:

$$\hat{n}_{11} = \frac{2087 \cdot 40}{2362} = 35.34$$
$$\hat{n}_{12} = \frac{2087 \cdot 37}{2362} = 32.69$$
$$\hat{n}_{13} = \frac{2087 \cdot 913}{2362} = 1000.20$$
$$\hat{n}_{14} = \frac{2087 \cdot 1132}{2362} = 1018.76$$
$$\hat{n}_{21} = \frac{275 \cdot 40}{2362} = 4.66$$
$$\hat{n}_{22} = \frac{275 \cdot 40}{2362} = 4.31$$
$$\hat{n}_{23} = \frac{275 \cdot 40}{2362} = 131.8$$
$$\hat{n}_{24} = \frac{275 \cdot 40}{2362} = 134.24$$

	Omicidio	Lesioni	Furto	Altro	TOT
uomo	35.34	32.69	1000.20	1018.76	2087
donna	4.66	4.31	131.8	134.24	275
TOT	40	37	1132	1153	2362

Quindi, applichiamo la formula dell'indice chi-quadrato e avremo:

$$\chi^2 = \frac{(37-35.34)^2}{35.34} + \frac{(35-32.69)^2}{32.69} + \frac{(913-1000.2)^2}{1000.2} + \frac{(1102-1018.76)^2}{1018.76}$$
$$+ \frac{(3-4.66)^2}{4.66} + \frac{(2-4.31)^2}{4.31} + \frac{(219-131.8)^2}{131.8} + \frac{(51-134.24)^2}{134.24} = 125.79$$

Infine, calcoliamo l'indice di Cramer per ottenere un valore numerico

facilmente interpretabile:

$$\nu = \frac{125.79}{2362 \cdot \min(4-1, 2-1)} = \frac{125.79}{2362 \cdot \min(3,1)} = \frac{125.79}{2362} = 0.053 \ .$$

Possiamo concludere dicendo che, poiché l'indice normalizzato è molto prossimo a zero, deduciamo che i due caratteri analizzati presentano una debole dipendenza assoluta.

Dip. parametrica: permette di stabilire se e quanta influenza hanno le modalità di un carattere, rispetto ad un parametro costante di sintesi delle medesime distribuzioni. Per questo motivo, si parla di *distribuzione in media, distribuzione in varianza, distribuzione in mediana,* e così via.

Noi ci occuperemo del modello di dipendenza in media. Questa si calcola attraverso il *rapporto di correlazione di Pearson:*

$$\eta^2_{X|Y} = \frac{\sum_{j=1}^{t}[M(X|Y=y_j) - M(X)]^2 \cdot n_{0j}}{\sum_{i=1}^{s}(x_i - M(X))^2 \cdot n_{i0}}$$
$$\eta^2_{Y|X} = \frac{\sum_{i=1}^{s}[M(Y|X=x_i) - M(Y)]^2 \cdot n_{i0}}{\sum_{j=1}^{t}(y_j - M(Y))^2 \cdot n_{0j}}$$

Proprietà dell'indice di correlazione di Pearson:

1. $0 \leq \eta^2 \leq 1$
2. η^2 aumenta, al crescere della variabilità delle medie condizionate

Esempio 2.4. *In un collettivo di 420 volontari si é osservata la frequenza di attivitá di volontariato per classi di etá, ottenendo la seguente distribuzione di frequenze relative percentuali:*

$X \backslash Y$	$[14, 20]$	$(20, 35]$	$(35, 55]$	$(55, 60]$
Almeno una volta a settimana	10	15	10	5
Una o piú volte al mese	10	20	20	10

Determinare il coeff. di correlazione dell'etá dalla regolaritá del servizio di volontariato.

Si scelga X = frequenza del servizio di volontariato e Y = etá. Sia poi n la totalitá percentuale del campione preso in considerazione. Si

vede, inoltre, che la tabella presenta $S = 2$ righe e $T = 4$ colonne. Per valutare la correlazione tra due fenomeni espressi in forma di frequenze relative percentuali si fa riferimento al rapporto di correlazione (si veda Definizione 9.8 pag. 215 del libro di testo consigliato)

$$\eta^2_{Y|X} = \frac{\sum_{i=1}^{S}[(\mu(Y|X_i) - \mu(Y))^2]n_{i0}}{\sum_{j=1}^{T}[(\bar{Y}_j - \mu(Y))^2]n_{0j}},$$

dove:

- *poiché i dati sono raccolti in classi, bisogna calcolare il valore centrale di ciascuna classe, indicato con $\bar{Y}_j$, per ogni $j = 1, \ldots, T$*
- *$\mu(Y|X_i)$ é la media condizionata del carattere Y rispetto alla modalitá X_i. Si ha:*

$$\mu(Y|X_i) = \frac{1}{n_{i0}} \sum_{j=1}^{T} \bar{Y}_j n_{ij}, \quad \forall\, i = 1, 2$$

$$\mu(Y|X_1) = (17 * 10 + 27.5 * 15 + 45 * 10 + 57.5 * 5)/40 = 33$$

$$\mu(Y|X_2) = (17 * 10 + 27.5 * 20 + 45 * 20 + 57.5 * 10)/60 = 36.58$$

- *$\mu(Y)$ é la media del carattere Y. Si ha:*

$$\mu(Y) = \frac{1}{n} \sum_{j=1}^{T} \bar{Y}_j n_{0j} = (17{*}20{+}27.5{*}35{+}45{*}30{+}57.5{*}15)/100 = 35.15$$

Quindi:

$$D_S = \sum_{i=1}^{S}[(\mu(Y|X_i){-}\mu(Y))^2]n_{i0} = (33{-}35.15)^2{*}40{+}(36.58{-}35.15)^2{*}60 = 307.594$$

$$D_{tot} = \sum_{j=1}^{T}[(\bar{Y}_j - \mu(Y))^2]n_{0j} = [(17 - 35.15)^2] * 20 + [(27.5 - 35.15)^2] * 35$$
$$+ [(45 - 35.15)^2] * 30 + [(57.5 - 35.15)^2] * 15 = 19040.23$$

Dunque,

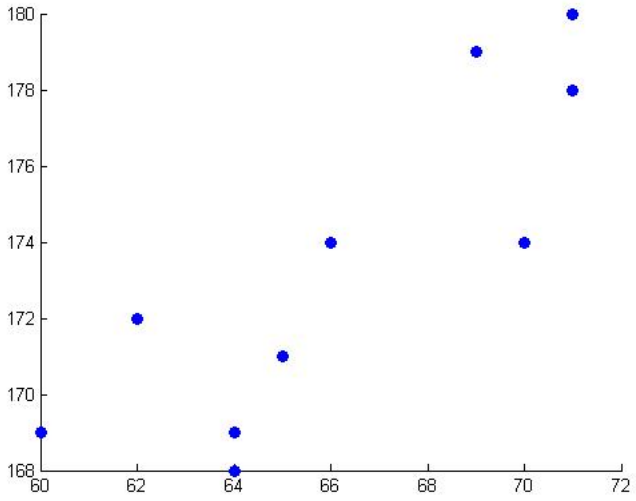
$$\eta^2_{Y|X} = \frac{307.594}{19040.23} = 0.01615.$$

Dip. funzionale: valuta la forma funzionale della dipendenza nello studio dell'associazione tra caratteri quantitativi (presenza di dipendenza, intensità della dipendenza, segno della dipendenza).

Ci concentriamo sulla dipendenza lineare, i.e., rappresentata attraverso funzioni che sono equazioni di rette.

si consideri la distribuzione di peso e altezza di 10 atleti:

atleta	peso (kg)	altezza (cm)
M	66	174
P	64	168
L	65	171
G	71	178
S	64	169
F	70	174
A	71	180
O	62	172
B	60	169
E	69	179



Le componenti della variabile doppia (X, Y) sono in generale caratterizzate da diversa posizione e variabilità, infatti $M(X) \neq M(Y)$ e $Var(X) \neq Var(Y)$.

Per poter confrontare le modalità dei due caratteri, occorre che i valori siano riferiti a delle quantità che abbiano stessa " scala di misurazione",

nel senso che si possano escludere gli effetti delle diverse medie e varianze.

L'operazione che si effettua in questi casi va sotto il nome di *standardizzazione:*

$$Z_X := \frac{X - \mu_X}{\sigma_X}, \qquad Z_Y := \frac{Y - \mu_Y}{\sigma_Y} \ .$$

Le nuove grandezze Z_X e Z_Y sono dette *standardizzate,* i cui valori sono i seguenti:

$$\begin{aligned}
\mu_X &= \frac{\sum_{i=1}^{10} x_i}{10} = 66.2 \\
\mu_Y &= \frac{\sum_{i=1}^{10} y_i}{10} = 173.4 \\
\sigma_X^2 &= \frac{\sum_{i=1}^{10} x_i^2}{10} - \mu_X^2 = 15.07 \ \Rightarrow \ \sigma_X = 3.88 \\
\sigma_y^2 &= \frac{\sum_{i=1}^{10} y_i^2}{10} - \mu_y^2 = 19.1556 \ \Rightarrow \ \sigma_X = 4.38
\end{aligned}$$

Z_X	Z_Y
-0.0515	0.1371
-0.5668	-1.2338
-0.3092	-0.5484
1.2366	1.0510
-0.5668	-1.0053
0.9790	0.1371
1.2366	1.5080
-1.0820	-0.3199
-1.5973	-1.0053
0.7214	1.2795

L'indice che misura la dipendenza lineare di Y da X si chiama *coefficiente di correlazione* ed è definito come la media aritmetica del prodotto delle standardizzate:

$$\rho_{XY} := \frac{1}{n} \sum_{i=1}^{n} z_{X,i} \cdot z_{Y,i}$$

Proposizione 2.1. *Il coefficiente di correlazione ammette la seguente formula operativa:*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \ ,$$

dove il numeratore rappresenta la covarianza *di X e Y.*

Proof. Sfruttando la definizione di standardizzata, avremo

$$\begin{aligned} \rho &= \frac{1}{n}\sum_{i=1}^{n} z_{X,i} \cdot z_{Y,i} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_X}{\sigma_X} \cdot \frac{y_i - \mu_Y}{\sigma_Y}\right) \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y} \end{aligned}$$

Guardiamo al numeratore:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i \mu_Y - \mu_X y_i + \mu_X \mu_Y) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \mu_Y - \frac{1}{n}\sum_{i=1}^{n} \mu_X y_i + \frac{1}{n}\sum_{i=1}^{n} \mu_X \mu_Y \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu_X \mu_Y = \sigma_{XY} \ . \end{aligned}$$

Sostituendo nella formula iniziale, si ottiene la tesi. □

Proprietà della correlazione

1. $-1 \leq \rho_{XY} \leq 1$ e si ha perfetta correlazione quando $\rho = \pm 1$.
2. se X e Y sono indipendenti, allora $\rho_{XY} = 0$.
3. la correlazione è invariante per trasformazioni lineari, ossia:

$$\rho(aX + B) = \rho(X)$$

4. si ha massima relazione lineare possibile in caso di correlazione di una variabile con se stessa:

$$\rho_{XX} = \frac{\sigma_{XX}}{\sigma_X^2} = \frac{\sigma_X^2}{\sigma_X^2} = 1 \ .$$

3 La regressione lineare

Per studiare la dipendenza lineare di una variabile Y rispetto ad un'altra X, si utilizza il *modello di regressione*: tale modello stabilisce, a meno di variazioni casuali, una relazione lineare tra risposta e predittore e permette di prevedere i valori della variabile di risposta (definita *variabile dipendente*) a partire da quelli di una variabile indipendente (chiamata *regressore* o *predittore*).

Da un punto di vista matematico, il modello ha la seguente forma:

$$y)f(x) + \epsilon \ ,$$

dove f è una funzione nell'incognita x e può assumere varie forme, mentre ϵ rappresenta il contributo di fattori che potrebbero influenzare la varibile di risposta e che non vengono considerati (*residui* o *scarti*).

Il caso più facile che si può presentare è quello della **regressione lineare semplice**.

lineare: la funzione che prendiamo in considerazione è l'equazione di una retta ($f(x) = mx + q$). In alternativa, si può avere regressione quadratica, esponenziale, logaritmica, etc etc.

semplice: supponiamo ci sia una sola modalità indipendente, in contrapposizione con la regressione *multipla.*

L'equazione del modello è

$$y = b_0 + b_1 x + \epsilon \ ,$$

dove:

- b_0 è l'*intercetta,* ossia l'ordinata all'origine (da un punto di vista geometrico) e rappresenta il valore medio di y quando x vale zero (da un punto di vista statistico)
- b_1 è il coefficiente angolare della retta (da un punto di vista geometrico) e ci dice come varia y in corrispondenza di una variazione unitaria di x. Il segno di b_1 ci informa circa la relazione tra la variabile risposta ed il predittore
- ϵ è la componente non osservabile del modello e supponiamo abbia media nulla

Dunque, la retta di regressione fornisce una approssimazione della dipendenza dei valori della variabile dipendente da quelli della variabile indipendente. Questo implica che la dipendenza non è esattamente descritta dalla retta di regressione. Quindi, dovremo confrontare i valori teorici, dati da

$\hat{y}_i = b_0 + b_i x_i$, con quelli osservati, che indichiamo con y_i, per ogni i. Le differenze tra i valori teorici e quelli osservati sono i residui:

$$\epsilon_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i \ .$$

Cerchiamo i valori dei parametri della retta di regressione che rendano minima la differenza tra valori teorici e valori osservati, o, equivalentemente, che rendano minimi i residui.

Il metodo utilizzato per risolvere questo problema va sotto il nome di *metodo dei minimi quadrati.* Tale metodo consiste nel determinare b_0, b_1 in modo tale che la *somma dei quadrati dei residui sia minima.*

Ma perché proprio la somma dei quadrati? Non sarebbe stato più semplice prendere semplicemente la somma dei residui?

Supponiamo di scegliere di minimizzare $\sum_{i=1}^{n} \epsilon_i$; poiché gli scarti ci dicono di quanto ci siamo spostati dal valore effettivo, tali residui potranno assumere valori positiv e/o negativi. In casi particolari, quindi, potremmo ottenere somme dei residui molto piccole, corrispondenti a rette assolutamente inadatte a descrivere i nostri punti.

Una candidata alternativa potrebbe essere la somma dei valori assoluti degli scarti, così da eliminare il problema di eventuali valori negativi dei residui.

Si sceglie proprio di minimizzare la somma dei quadrati degli scarti, in quanto, in questo modo, è soddisfatta la proprietà della media artitmetica di rendere minima la somma dei quadrati degli scarti.

Come si procede, da un punto di vista matematico? Com'è ben noto dall'analisi, un problema di ottimo si traduce nella risoluzione di un sistema di equazioni lineari, dove tali equazioni si ottengono ponendo uguali a zero le derivate della funzione da ottimizzare:

$$\begin{cases} \frac{\partial}{\partial b_0} \sum_{i=1}^{n} \epsilon_i^2 = 0 \\ \frac{\partial}{\partial b_1} \sum_{i=1}^{n} \epsilon_i^2 = 0 \end{cases}$$

Proposizione 3.1. *I parametri della retta di regressione $y = b_0 + b_1 x$ assumono le seguenti espressioni:*

$$b_1 = \frac{\sigma_{xy}}{\sigma_x^2}$$
$$b_0 = \mu_y - b_1 \mu_x$$

Proof. Ricordiamo che la somma dei quadrati degli scarti si scrive come segue:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \ .$$

Scriviamo le derivate:

$$\begin{cases} \frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0 \\ \frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0 \end{cases}$$

Per quel che riguarda la prima derivata:

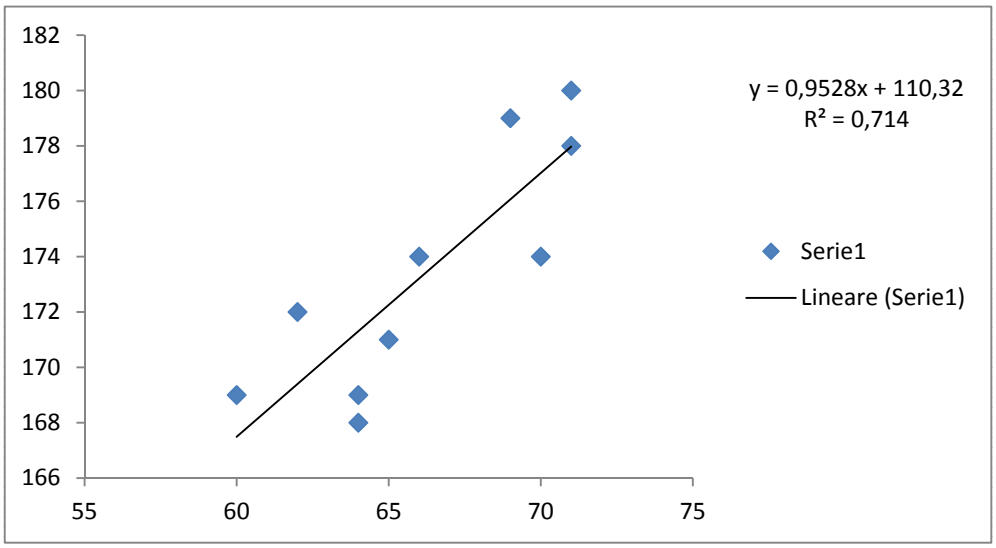
$$\begin{aligned} & \frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \\ \Rightarrow & \sum_{i=1}^{n} y_i - b_0 \sum_{i=1}^{n} - b_1 \sum_{i=1}^{n} x_i = 0 \\ \Rightarrow & \frac{1}{n} \sum_{i=1}^{n} y_i - b_0 \frac{1}{n} \sum_{i=1}^{n} - b_1 \frac{1}{n} \sum_{i=1}^{n} x_i = 0 \\ \Rightarrow & \mu_y - b_0 - b_1 \mu_x = 0 \\ \Rightarrow & b_0 = \mu_y - b_1 \mu_x \ . \end{aligned}$$

Per quel che riguarda la seconda derivata:

$$\begin{aligned} & \frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0 \\ \Rightarrow & \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0 \\ \Rightarrow & b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n} \sum_{i=1}^{n} y_i - b_1 \frac{1}{n} \sum_{i=1}^{n} x_i \right) \sum_{i=1}^{n} x_i \\ \Rightarrow & b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + b_1 \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \\ \Rightarrow & b_1 \left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n} \right] = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i \\ \Rightarrow & b_1 \left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n^2} \right] = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n^2} \\ & = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} y_i}{n} \frac{\sum_{i=1}^{n} x_i}{n} \\ \Rightarrow & b_1 \left[\mu_{x^2} - \mu_x^2 \right] = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \mu_x \mu_y \\ \Rightarrow & \sigma_x^2 b_1 = \sigma_{xy} \\ \Rightarrow & b_1 = \frac{\sigma_{xy}}{\sigma_x^2} \end{aligned}$$

□

In riferimento ai dati dell'esempio iniziale, la retta di regressione è



Una volta ottenuta le retta di regressione, occore stabilire, attraverso strumenti matematici, che essa ben si adatta ai valori osservati. In questo senso, esistono strumenti grafici e strumenti analitici. Per quel che riguarda i primi, si parla di *plot dei residui*. Nel secondo caso, si calcola il *coefficiente di determinazione lineare* o *indice di determinazione:*

$$R^2 := \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} = \rho^2 \ .$$

Per com'è definito, $0 \leq R^2 \leq 1$. In particolare, valori piccoli (vicini allo zero) indicano che la retta ottenuta non si adatta bene ai dati osservati, mentre valori grandi (vicini a uno) indicano che la retta ottenuta ben si adatta alle osservazioni.