Appunti del corso di Biostatistica
Laurea triennale in Scienze Biologiche

Michele Gianfelice
Dipartimento di Matematica
Università della Calabria
Campus di Arcavacata
Ponte Pietro Bucci - cubo 30B
87036 Arcavacata di Rende (CS)
gianfelice@mat.unical.it

a.a. 2007/2008

Indice

Parte I

Statistica descrittiva

0.1 Organizzazione dei dati di un campione di misure sperimentali

Al fine di estrarre da un esperimento informazioni utili alla comprensione di un fenomeno oggetto di un indagine scientifica è necessario organizzare la mole di dati prodotta dalla procedura sperimentale in modo che le informazioni in questi contenute risultino di facile lettura da parte dello sperimentatore.

0.1.1 Campioni di dati univariati

Sia X una quantità che può essere misurata sperimentalmente, ad esempio una lunghezza di un oggetto, il suo peso, ma anche il tasso di crescita di una coltura batterica od il tempo di dimezzamento di quest'ultima, e sia x il numero reale che rappresenta il valore di X misurato nel corso di un esperimento. Ripetendo più volte lo stesso esperimento, avendo cura ogni volta di riprodurre esattamente le stesse condizioni sperimentali, si ottiene una serie di misure della quantità X. La collezione di queste misure è detta *campione* di misure di X; ogni elemento del campione, ovvero ogni misura di X, è detto *dato*, mentre il numero di volte che s'è ripetuto l'esperimento volto alla misura di X, ovvero il numero di misure di X che compongono il campione, è detto *numerosità* oppure *ordine* o *ampiezza* del campione.

Pertanto nel seguito indicheremo con

$$\mathcal{C}_N(X) := \{x_1, .., x_N\} \tag{1}$$

un campione di misure di X di ordine N, che in generale è rappresentato da una tabella o da una stringa di numeri (cfr [R] fig 2.3 oppure la tabella 2).

Ogni sottoinsieme $\mathcal{C}'_M(X) := \{x_{i_1}, .., x_{i_M}\}$ di $\mathcal{C}_N(X)$ è detto sottocampione di $\mathcal{C}_N(X)$ di ampiezza M, dove evidentemente $M < N$. In generale, ogni sottoinsieme $\mathcal{A}$ di $\mathcal{C}_N(X)$ ne è un sottocampione. Nel seguito, dato un qualsiasi sottocampione $\mathcal{A}$ di $\mathcal{C}_N(X)$, ne indicheremo la numerosità col simbolo $|\mathcal{A}|$.

Esempio 1 *Se X indica i giorni di vita di una cavia esposta ad un agente patogeno e $N = 36$ è il numero della cavie, il campione $\mathcal{C}_{36}(X)$ è l'insieme dei numeri che compaiono nella seguente tabella*

$$\begin{array}{ccccccccc} 82 & 89 & 94 & 110 & 74 & 122 & 112 & 95 & 100 \\ 78 & 65 & 60 & 90 & 83 & 87 & 75 & 114 & 85 \\ 69 & 94 & 124 & 115 & 107 & 88 & 97 & 74 & 72 \\ 68 & 83 & 91 & 90 & 102 & 77 & 125 & 108 & 65 \end{array} , \tag{2}$$

pertanto, $x_1 = 82,\ x_2 = 89, ..., x_{10} = 78, ..$ *e così via.*

D'altra parte, se per esempio $\mathcal{C}'_{25}(X)$ *è il sottocampione composto dai dati di* $\mathcal{C}_{36}(X)$ *i cui valori sono minori di* 100, *la tabella ad esso associata si costruisce selezionando i dati di quella relativa a* $\mathcal{C}_{36}(X)$ *(2) cui valori risultano minori di* 100, *cioè*

$$\begin{array}{ccccccccc} \mathbf{82} & \mathbf{89} & \mathbf{94} & 110 & \mathbf{74} & 122 & 112 & \mathbf{95} & 100 \\ \mathbf{78} & \mathbf{65} & \mathbf{60} & \mathbf{90} & \mathbf{83} & \mathbf{87} & \mathbf{75} & 114 & \mathbf{85} \\ \mathbf{69} & \mathbf{94} & 124 & 115 & 107 & \mathbf{88} & \mathbf{97} & \mathbf{74} & \mathbf{72} \\ \mathbf{68} & \mathbf{83} & \mathbf{91} & \mathbf{90} & 102 & \mathbf{77} & 125 & 108 & \mathbf{65} \end{array} . \tag{3}$$

Pertanto, la tabella dei dati associata a $\mathcal{C}'_{25}(X)$ *risulta*

$$\begin{array}{ccccc} 82 & 89 & 94 & 74 & 95 \\ 78 & 65 & 60 & 90 & 83 \\ 87 & 75 & 85 & 69 & 94 \\ 88 & 97 & 74 & 72 & 68 \\ 83 & 91 & 90 & 77 & 65 \end{array} \tag{4}$$

e quindi $x_{i_1} = x_1 = 82,\ x_{i_2} = x_2 = 89,$ *ma* $x_{i_4} = x_5 = 74, ...$ *eccetera.*

Anche la collezione dei dati che compaiono nella prima riga della tabella (2) costituisce un sottocampione di $\mathcal{C}_{36}(X)$. *Denotando questa* $\mathcal{A} = \{x_1, .., x_9\}$, *ne segue che la sua numerosità è* $|\mathcal{A}| = 9$.

Definizione 2 *Un campione di misure di una sigola quantità è detto campione* univariato.

Grafici delle frequenze

Dato un campione di misure di numerosità N di una quantità X è utile conoscerne l'intervallo di variazione, ovvero il valore minimo e quello massimo dei dati che compaiono in $\mathcal{C}_N(X)$, ma soprattutto il numero di volte in cui compaiono, nella tabella associata a $\mathcal{C}_N(X)$, dati dello stesso valore numerico. Questo numero è detto *frequenza di un dato del campione.* Più precisamente,

Definizione 3 *Sia* $\mathcal{C}_N(X)$ *un campione di numerosità* N *di misure di una quantità* X *e per ogni* $i = 1, .., N$, *sia* $\mathcal{F}^i_N := \{y \in \mathcal{C}_N(X) : y = x_i\}$[1] *il sottocampione di* $\mathcal{C}_N(N)$ *composto da tutti quei dati che hanno lo stesso valore assunto dal dato* x_i. *Allora la numerosità* $|\mathcal{F}^i_N|$ *di* $\mathcal{F}^i_N$ *è detta* frequenza del valore del dato i-simo del campione.

[1] Ricordiamo che, se A è un insieme, la scrittura $y \in A$ si legge: y è elemento di A o y appartiene ad A.

Esempio 4 *Considerando il campione* $\mathcal{C}_{36}(X)$ *di cui all'Esempio 1 si ha che per* $i = 1$,

$$\mathcal{F}_{36}^{1} = \{y \in \mathcal{C}_N(X) : y = x_1 = 82\} = \{x_1\} \tag{5}$$

è il sottocampione di $\mathcal{C}_{36}(X)$ *i cui elementi assumono lo stesso valore di* x_1 *ovvero* 82. *Dunque,* $|\mathcal{F}_{36}^{1}| = 1$, *mentre, scelto per esempio* $i = 5$,

$$\mathcal{F}_{36}^{5} = \{y \in \mathcal{C}_N(X) : y = x_5 = 74\} = \{x_5, x_{26}\} \tag{6}$$

è il sottocampione di $\mathcal{C}_{36}(X)$ *i cui elementi assumono lo stesso valore di* x_5 *cioè* 74. *Allora,* $|\mathcal{F}_{36}^{5}| = 2$. *Lo stesso vale anche per* $\mathcal{F}_{36}^{13} = \{x_{13}, x_{31}\}$ *e* $\mathcal{F}_{36}^{14} = \{x_{14}, x_{29}\}$.

Osservazione 5 *Osserviamo che, se il dato* x_i *di* $\mathcal{C}_N(X)$ *appartiene al sottocampione* $\mathcal{F}_N^j$, *allora il dato* x_j *appartiene a* $\mathcal{F}_N^i$ *e viceversa, ovvero* $\mathcal{F}_N^i = \mathcal{F}_N^j$. *Nell'esempio precedente infatti, si ha che* x_{26} *appartiene a* $\mathcal{F}_{36}^{5}$, *ma poiché* $x_{26} = x_5 = 74$, *allora,* x_5 *appartiene a* $\mathcal{F}_{36}^{26} = \{x_5, x_{26}\} = \mathcal{F}_{36}^{5}$.

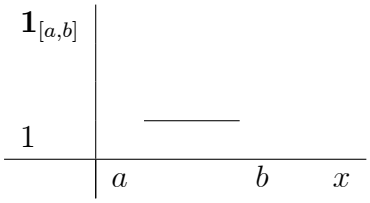
Definizione 6 *Sia* $\mathcal{C}_N(X)$ *un campione di numerosità* N *di misure di una quantità* X *e per ogni* $i = 1, .., N$, *sia* $|\mathcal{F}_N^i|$ *la ferquenza del dato* i*-simo. Il rapporto* $f_i := \frac{|\mathcal{F}_N^i|}{N}$ *è detto* frequenza relativa del valore del dato i-simo del campione.

Quindi, la frequenza relativa di uno dei valori assunti dai dati del campione non rappresenta altro che la frazione dei dati del campione che assumono tale valore. Nell'esempio precedente infatti, si ha che la frequenza relativa di 74, così come quella di 90, risulta essere $\frac{|\mathcal{F}_{36}^{5}|}{36} = \frac{2}{36} = \frac{1}{18} \cong 6\%$.

Osservazione 7 *Per ogni sottoinsieme di numeri reali* $A \subseteq \mathbb{R}$, *sia*

$$\mathbb{R} \ni x \longmapsto \mathbf{1}_A(x) \in \{0, 1\} \tag{7}$$

la funzione indicatrice di A, *ovvero la funzione la quale restituisce il valore* 1 *se valutata su un elemento* x *appartenente ad* A *e* 0 *altrimenti. Il grafico di* $\mathbf{1}_A$, *se per esempio* $A = [a, b]$, *risulta*



. (8)

Allora, se y è un qualsiasi numero reale, e $\mathcal{C}_N(X)$ è un campione di misure di una quantità X, la frequenza F_y del valore y, ovvero la numerosità del sottocampione di $\mathcal{C}_N(X)$, $\{x \in \mathcal{C}_N(X) : x = y\}$ è

$$F_y := |\{x \in \mathcal{C}_N(X) : x = y\}| = \sum_{i=1}^{N} \mathbf{1}_y(x_i) \ . \tag{9}$$

Perciò, per ogni $i = 1, .., N$ la frequenza del valore dell'i-simo dato risulta

$$F_i := F_{x_i} = \left|\mathcal{F}_N^i(X)\right| = \sum_{j=1}^{N} \mathbf{1}_{x_i}(x_j) \ . \tag{10}$$

Inoltre, la frequenza relativa di un qualsiasi valore y si può scrivere

$$f_y := \frac{F_y}{N} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_y(x_i) \tag{11}$$

e quindi, per ogni $i = 1, .., N$, quella dell'i-simo dato del campione vale

$$f_i = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_y(x_i) \ . \tag{12}$$

Caso in cui i valori assunti dai dati del campione siano pochi

Al fine di capire quali valori della quantità X in studio l'operazione di misura di X ha riprodotto più volte è utile graficare in ascissa tutti i valori assunti dai dati del campione in esame $\mathcal{C}_N(X)$, il cui numero il generale è più piccolo di N salvo quando il campione contiene dati i cui valori sono tutti diversi fra loro, ed in ordinata le frequenze di tali valori, dando origine ad un grafico a linee verticali (*bastoncini*) (cfr [R] fig. 2.1 pag.13).

Nel caso in cui si vogliano confrontare due campioni $\mathcal{C}_N(X)$ e $\mathcal{C}_M(X)$ di misure della stessa quantità X, ma di numerosità N e M diverse, è utile, per entrambi i campioni, riportare in ordinata invece che le frequenze dei valori dei dati le frequenze relative dei valori assunti dai dati di ciascun campione, in modo da poterli meglio comparare.

Un altro modo utile per presentare i dati, nel caso in cui i valori da questi assunti siano poco numerosi, è quello di riportare detti valori e le loro frequenze relative in un diagramma detto *a torta* (cfr [R] fig. 2.4 pag. 16).

Caso in cui i valori assunti dai dati del campione siano molti

Nel caso in cui l'insieme dei valori assunti dai dati di un campione di misure di una quantità X sia molto numeroso, non è spesso utile graficarne il diagramma delle frequenze, in quanto la frequenza di ciascun valore assunto dai dati del campione potrebbe essere così piccola da non far emergere tra questi alcun valore tipico, cioè un valore che si presenta più volte degli altri. Ciò che però può accadere in questi casi è che ci sia una collezione di valori, tra tutti quelli assunti dai dati del campione, i cui elementi ricorrano più volte degli altri. In tal caso, conviene suddividere opportunamente l'insieme dei valori che una misura della quantità in esame X può assumere, tipicamente un sottoinsieme dell'insieme dei numeri reali $\mathbb{R}$, in intervalli disgiunti della stessa lunghezza e conseguentemente ripartire il campione in sottocampioni, detti *classi*, in modo tale che ogni sottocampione della partizione contenga solo i dati il cui valore appartiene ad un certo intervallo della suddivisione del'insieme di variabilità di X. A questo punto, riportando in ascissa il valori estremali degli intervalli della suddivisione dell'insieme di variabilità di X ed in ordinata la frequenza di ogni classe, ovvero il numero dei dati i cui valori assunti ricadono in quella classe e, in corrispondenza di tale valore, disegnando un tratto di retta pari alla lunghezza di un intervallo in cui s'è diviso l'asse delle ascisse, si ottiene un diagramma delle frequenze delle classi detto *istogramma* (cfr [R] fig. 2.5 pag.28). Allo stesso modo, per confrontare campioni di numerosità diversa i cui dati assumono molti valori, conviene graficare l'istogramma delle frequenze relative delle classi in cui s'è diviso l'insieme dei valori assunti dai dati, ovvero riportare in ordinata le frequenze di ogni classe divise per la numerosità del campione.

Più precisamente, indicando con $\mathbb{Z}$ l'insieme dei numeri interi, cioè $\{..,-2,-1,0,1,2,..\}$, sia $\{a_k\}_{k\in\mathbb{Z}}$ una successione strettamente crescente di numeri reali, ovvero tale che, per ogni intero k, $a_k < a_{k+1}$. Allora, per ogni intero k, gli intervalli

$$(a_k, a_{k+1}] = \{x \in \mathbb{R} : a_k < x \leq a_{k+1}\} \tag{13}$$

sono a due a due disgiunti e perciò la loro collezione $\{(a_k, a_{k+1}]\}_{k\in\mathbb{Z}}$ è una partizione dell'insieme dei numeri reali $\mathbb{R}$, cioè $\mathbb{R} = \bigcup_{k\in\mathbb{Z}} (a_k, a_{k+1}]$. Quindi, poiché le misure della quantità in esame X assumono valori reali, se associamo ad ogni elemento $(a_k, a_{k+1}]$ della partizione di $\mathbb{R}$ il sottocampione di $\mathcal{C}_N(X)$

$$\mathcal{B}_N^k := \{y \in \mathcal{C}_N(X) : y \in (a_k, a_{k+1}]\}\ , \tag{14}$$

questi sottocampioni sono a due a due disgiunti, ovvero i dati che ne compongono uno non possono essere elementi di un'altro. Pertanto, la collezione

$\{\mathcal{B}_N^k\}_{k\in\mathbb{Z}}$ di questi sottocampioni realizza una partizione di $\mathcal{C}_N(X)$, cioè $\mathcal{C}_N(X) = \bigcup_{k\in\mathbb{Z}} \mathcal{B}_N^k$, i sottocampioni $\mathcal{B}_N^k$ sono le classi associate alla partizione $\{(a_k, a_{k+1}]\}_{k\in\mathbb{Z}}$, la frequenza della k-sima classe è $\left|\mathcal{B}_N^k\right| = \sum_{i=1}^N \mathbf{1}_{\mathcal{B}_N^k(X)}(x_i)$ e $\frac{\left|\mathcal{B}_N^k\right|}{N}$ ne è la frequenza relativa.

Esempio 8 *Sia $\mathcal{C}_{36}(X)$ come nell'Esempio 1. Ponendo per ogni $k \in \mathbb{Z}$, $(a_k, a_{k+1}] = (10k, 10(k+1)]$ otteniamo*

$$\mathcal{B}_{36}^k := \{y \in \mathcal{C}_{36}(X) : y \in (10k, 10(k+1)]\} . \tag{15}$$

Perciò, dalla tabella 2, segue che, per ogni $k \leq 5$, $\left|\mathcal{B}_{36}^k\right| = 0$, mentre

$$\begin{aligned} \mathcal{B}_{36}^6 &= \{y \in \mathcal{C}_{36}(X) : y \in (60, 70]\} \\ &= \{x_{12} = 60, x_{19} = 69, x_{28} = 68, x_{36} = 65\} \end{aligned} \tag{16}$$

da cui segue $|\mathcal{B}_{36}^6| = 4$.

$$\begin{aligned} \mathcal{B}_{36}^7 &= \{y \in \mathcal{C}_{36}(X) : y \in (70, 80]\} \\ &= \{x_5 = 74, x_{10} = 78, x_{16} = 75, x_{25} = 74, x_{33} = 77\} , \end{aligned} \tag{17}$$

dunque $|\mathcal{B}_{36}^7| = 5$. $|\mathcal{B}_{36}^8| = 9$, e così via fino a $k = 12$, dopodiché, per ogni $k \geq 13$, $\left|\mathcal{B}_{36}^k\right| = 0$.

Osservazione 9 *È necessario sottolineare che la divisione in classi descritta in questa sezione dipende strettamente dal campione in esame e quindi dev'essere effettuata* ad hoc *campione per campione. Il criterio da seguire per operare una partizione dell'insieme dei valori assunti dai dati del campione utile al fine di rappresentarne la distribuzione e quindi di estrarne il maggior contenuto informativo sul fenomeno in studio, è quello di scegliere la massima ampiezza dell'intervallo della partizione dell'insieme di variabilità della quantità in esame X in modo che i dati del campione appartenenti ad ogni singola sottoclasse risultino quanto più possibile equidistribuiti, ovvero in modo tale che le frequenze relative dei valori dei dati di ogni classe risultino tra di loro quanto più possibile pressoché identiche. Infatti,*

- *scegliendo classi troppo grandi il contenuto informativo dovuto alla distribuzione dei dati all'interno di ogni classe non trasparirebbe dall'istogramma delle frequenze di tali classi;*
- *scegliendo classi troppo piccole s'incorrebbe di nuovo nell'errore che si commetterebbe graficando direttamente le frequenze dei singoli dati.*

Grafici *ramo-foglia (stem&leaf)* Dato un campione $\mathcal{C}_N(X)$ di misure della quantità in esame X, consideriamo il valore numerico assunto dai dati del campione come una stringa finita e ordinata di cifre, ovvero, per ogni $i = 1, .., N$, un dato del campione x_i, può essere così rappresentato

$$x_i = a_i^1 a_i^2 \cdot \cdot \, a_i^{K_i} \tag{18}$$

dove K_i è un numero naturale e, per ogni $j = 1, .., K_i$, a_i^j è un numero da 0 a 9. Consideriamo quindi i sottocampioni di $\mathcal{C}_N(X)$

$$\mathcal{B}_N^k(X) := \{x_i,\ i = 1, .., N : x_i = a_i^1 a_i^2 \cdot \cdot \, a_i^{K_i}\ ;\ K_i = k\}, \quad k \in \mathbb{N}, \tag{19}$$

ovvero quelli formati da dati rappresentati come nella (18), ma da una stringa composta dallo stesso numero di cifre.

Fissato un numero naturale k, se l_k, con $0 \leq l_k \leq k-1$, indica la posizione dell'ultima delle cifre della rappresentazione (18) dei dati del sottocampione $\mathcal{B}_N^k$ che sono identiche, leggendo le cifre inordine lessicografico, cioè da sinistra verso destra, si può considerare la stringa di cifre associata ad ogni dato come la giustapposizione di due sottostringhe di cifre, la prima composta dalle prime cifre comuni a tutti i dati, che chiameremo *parte comune,* e la seconda composta dal resto delle cifre, che chiameremo *resto.* Per esempio, se le prime l_k cifre delle stringhe che rappresentano i dati del sottocampione $\mathcal{B}_N^k$ sono uguali, per ogni $i = 1, .., \left|\mathcal{B}_N^k(X)\right|$, la (18) diventa

$$x_i = \underbrace{a^1 \cdot \cdot \, a^{l_k}}_{\text{parte comune}} \underbrace{a_i^{l_k+1} \cdot \cdot \, a_i^k}_{\text{resto}} \ . \tag{20}$$

Allora, un modo efficiente per generare una partizione in classi di $\mathcal{C}_N(X)$ utile a descriverne il contenuto informativo, secondo quanto affermato nell'osservazione precedente, è quello di raggruppare i dati in sottocampioni contenenti solo quelli che abbiano identiche la lunghezza della stringa di cifre che li rappresenta e la prima cifra della sottostringa resto. Ovvero, per ogni $k \in \mathbb{N}$ e $m = 0, 1, .., 9$,

$$\mathcal{B}_N^{k,m}(X) := \{x_i,\ i = 1, .., N : x_i \in \mathcal{B}_N^k(X)\ ;\ x_i = a^1 \cdot \cdot \, a^{l_k} a_i^{l_k+1} \cdot \cdot \, a_i^k\ ;\ a_i^{l_k+1} = m\}\,. \tag{21}$$

Ciò implica che, per ogni $i = 1, .., \left|\mathcal{B}_N^k(X)\right|$, l'$i$-simo dato del sottocampione $\mathcal{B}_N^k(X)$ si può rappresentare mediante la giustapposizione di due nuove sottostringhe:

- la prima, detta *ramo (stem),* composta dalle cifre componenti la sottostringa parte comune $a^1 \cdot \cdot \, a^{l_k}$ seguite dal valore di $a_i^{l_k+1}$;

- la seconda, detta *foglia (leaf),* composta dalle rimanenti cifre della rappresentazione (20).

Ovvero, per ogni $i = 1, .., \left|\mathcal{B}_N^k(X)\right|$,

$$x_i = \underbrace{a^1 \cdot \cdot\, a^{l_k} a_i^{l_k+1}}_{\text{ramo (stem)}} \underbrace{a_i^{l_k+2} \cdot \cdot\, a_i^k}_{\text{foglia (leaf)}} \quad . \tag{22}$$

Per ogni numero naturale k, consideriamo il sottocampione $\mathcal{B}_N^k(X)$. La tabella in cui nella prima colonna compaiono le cifre corrispondenti alla sottostringa ramo dei dati di $\mathcal{B}_N^k(X)$, ordinate secondo i valori dell'ultima cifra, e nelle successive colonne le cifre corrispondenti sottostringhe foglie degli stessi dati è detto grafico *ramo-foglia (stem&leaf)* del sottocampione $\mathcal{B}_N^k(X)$. Ovvero, per ogni $m = 0, .., 9$, se la m-sima classe $\mathcal{B}_N^{k,m}$ definita nella (21) corrisponde alla collezione di dati di $\mathcal{C}_N(X)$, $\{x_{i_1}, .., x_{i_M}\}$, con $M = \left|\mathcal{B}_N^{k,m}\right|$, la m-sima riga del diagramma ramo-foglia risulta essere composta proprio dagli elementi di $\mathcal{B}_N^{k,m}$. Infatti, rappresentando per ogni $j = 1, .., M$, $x_{i_j} = a^1 \cdot \cdot\, a^{l_k} m a_{i_j}^{l_k+2} \cdot \cdot\, a_{i_j}^k$, questa si rappresenta così

$$a^1 \cdot \cdot\, a^{l_k} m \mid a_{i_1}^{l_k+2} \cdot \cdot\, a_{i_1}^k \ , \ a_{i_2}^{l_k+2} \cdot \cdot\, a_{i_2}^k \ , \ \ldots \ , \ a_{i_M}^{l_k+2} \cdot \cdot\, a_{i_M}^k \ . \tag{23}$$

Il diagramma ramo-foglia relativo a tutto il campione si ottiene incolonnando in ordine crescente rispetto all'indice k, dall'alto verso il basso, i diagrammi ramo-foglia dei sottocampioni $\mathcal{B}_N^k(X)$.

Esempio 10 *Consideriamo il campione di dati* $\mathcal{C}_{25}(X)$ *descritto dalla tabella 4.* $k = 2$, $l_2 = 0$, *allora,* $a_i^{l_2+1} = a_i^1 = m$ *assume solo i valori* $6, 7, 8, 9$. *Perciò, per ogni* $m = 0, .., 5$, $\left|\mathcal{B}_{25}^{2,m}\right| = 0$,

$$\begin{aligned}
\mathcal{B}_{25}^{2,6} &= \{x_7 = 65, x_8 = 60, x_{14} = 69, x_{20} = 68, x_{25} = 65\} \ , \\
\mathcal{B}_{25}^{2,7} &= \{x_4 = 74, x_6 = 78, x_{12} = 75, x_{18} = 74, x_{19} = 72, x_{23} = 77\} \ , \\
\mathcal{B}_{25}^{2,8} &= \{x_1 = 82, x_2 = 89, x_{10} = 83, x_{11} = 87, x_{13} = 85, x_{16} = 88, x_{20} = 83\} \ , \\
\mathcal{B}_{25}^{2,8} &= \{x_3 = 94, x_5 = 95, x_9 = 90, x_{15} = 94, x_{17} = 97, x_{21} = 91, x_{22} = 90\}
\end{aligned} \tag{24}$$

e il diagramma ramo-foglia di $\mathcal{C}_{25}(X)$ *risulta essere il seguente*

$$\begin{array}{c|ccccccc}
5 & & & & & & & \\
6 & 0, & 5, & 5, & 8, & 9 & & \\
7 & 2, & 4, & 4, & 5, & 7, & 8 & \\
8 & 2, & 3, & 3, & 5, & 7, & 8, & 9 \\
9 & 0, & 0, & 1, & 4, & 4, & 5, & 7
\end{array} \tag{25}$$

Se invece consideriamo tutto il campione $\mathcal{C}_{36}(X)$, *dalla tabella 2 segue:* $\mathcal{B}_{36}^{2}(X) = \mathcal{C}_{25}(X)$ *e*

$$\begin{aligned} \mathcal{B}_{36}^{3}(X) &= \{x_4, x_6, x_7, x_9, x_{17}, x_{21}, x_{22}, x_{23}, x_{32}, x_{34}, x_{35}\} \\ &= \{110, 122, 112, 100, 114, 124, 115, 107, 102, 125, 108\}\ . \end{aligned} \tag{26}$$

Allora, poiché in questo caso $k = 3$, *si ha* $l_3 = 1$ *e* $a_i^{l_3+1} = a_i^2 = m$ *assume i valori* $0, 1, 2$. *Dunque,*

$$\mathcal{B}_{36}^{3,0}(X) = \{x_9 = 100, x_{22} = 107, x_{23} = 102, x_{35} = 108\}, \tag{27}$$

$$\mathcal{B}_{36}^{3,1}(X) = \{x_4 = 110, x_7 = 112, x_{17} = 114, x_{22} = 115\}, \tag{28}$$

$$\mathcal{B}_{36}^{3,2}(X) = \{x_6 = 122, x_{21} = 124, x_{34} = 125\} \tag{29}$$

e il diagramma ramo-foglia di $\mathcal{B}_{36}^{3}(X)$ *risulta*

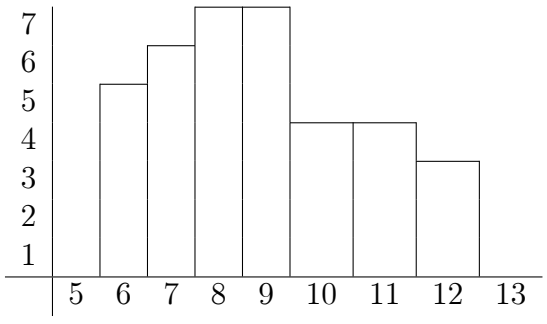
$$\begin{array}{c|llll} 10 & 0, & 2, & 7, & 8 \\ 11 & 0, & 2, & 4, & 5 \\ 12 & 2, & 4, & 5 & \end{array} \tag{30}$$

Quindi il diagramma ramo-foglia di $\mathcal{C}_{36}(X)$ *risulta*

$$\begin{array}{c|lllllll} 6 & 0, & 5, & 5, & 8, & 9 & & \\ 7 & 2, & 4, & 4, & 5, & 7, & 8 & \\ 8 & 2, & 3, & 3, & 5, & 7, & 8, & 9 \\ 9 & 0, & 0, & 1, & 4, & 4, & 5, & 7 \\ 10 & 0, & 2, & 7, & 8 & & & \\ 11 & 0, & 2, & 4, & 5 & & & \\ 12 & 2, & 4, & 5 & & & & \end{array}\ . \tag{31}$$

(cfr anche [R] Esempio 2.3.9 pag. 29).

Osservazione 11 *Notiamo che, dato un campione* $\mathcal{C}_N(X)$ *di misure di una quantità* X, *di cui è possibile graficare il diagramma ramo-foglia, l'istogramma delle frequenze delle classi* $\{\mathcal{B}_N^{k,m}\}_{k\in\mathbb{N},\, m=0,..,9}$ *può ottenersi da quest'ultimo semplicemente immaginando di ruotarlo di* $90°$ *in senso antiorario. Infatti, nel caso dell'esempio prededente si ottiene*



. (32)

Statistiche

Le informazioni rilevanti per la comprensione del fenomeno in studio possono essere sintetizzate da quantità dipendenti dai dati del campione. Tali quantità sono dette *statistiche* ed in questa sezione definiremo le pricipali. Analizzeremo inoltre la relazione che intercorre tra le statistiche di due campioni di dati i cui valori siano legati da una relazione funzionale lineare con coefficienti indipendenti dai dati stessi.

Ci sono molte ragioni che sottolineano l'importanza di studiare il comportamento delle statistiche sotto trasformazioni lineari dei dati di un campione, una di queste è che diverse relazioni funzionali tra due quantità X, Y possono essere ridotte a relazioni lineari tra altre quantità Z, W ad esse legate. Per esempio:

- se $Y = Ae^{BX}$, con A e B costanti note, cioè indipendenti sia da X che da Y, si ha che la quantità $Z = \log Y$ risulta legata alla quantità X dalla relazione lineare $Z = BX + C$, dove $C = \log A$;
- se $Y = AX^B$, con A e B costanti note, cioè indipendenti sia da X che da Y, si ha che la quantità $Z = \log Y$ risulta legata alla quantità $W = \log X$ dalla relazione lineare $Z = BW + C$, dove $C = \log A$.

Altre motivazioni per lo studio delle statistiche di dati linearmente dipendenti gli uni dagli altri saranno introdotte nel seguito.

Mediana campionaria Sia $\mathcal{C}_N(X)$ un campione univariato e siano: $\min \mathcal{C}_N(X)$ il più piccolo valore assunto dai dati del campione e

$$\iota := \min\{1 \leq i \leq N : x_i = \min \mathcal{C}_N(X)\} \ . \tag{33}$$

Indichiamo con $\hat{x}_1 = x_\iota$ il primo elemento del sottocampione

$$\mathcal{C}'_{M_1}(X) := \{\acute{x} \in \mathcal{C}_N(X) : \acute{x} = \min \mathcal{C}_N(X)\} \tag{34}$$

preso in ordine lessicografico, ovvero il primo dato del campione di misure di X, tra quelli che assumono il valore $\min \mathcal{C}_N(X)$, che compare nella tabella rappresentativa di $\mathcal{C}'_{M_1}(X)$ leggendo i dati da destra verso sinistra.

Esempio 12 *Considerando il campione di dati di cui all'Esempio 1,* $\min \mathcal{C}_{36}(X) = 65$, $M_1 = 2$ *e* $\mathcal{C}'_2(X) = \{x_{11}, x_{36}\}$. *Perciò, in questo caso,* $\iota = 11$ *e* $\hat{x}_1 = x_{11}$.

Sia allora $\widehat{\mathcal{C}}_N(X) := \{\hat{x}_1, .., \hat{x}_N\}$, dove per ogni $i = 2, .., N$, $\hat{x}_i$ è il primo elemento del sottocampione

$$\mathcal{C}'_{M_i}(X) := \{\acute{x} \in \mathcal{C}_N(X) \setminus \{\hat{x}_1, .., \hat{x}_{i-1}\} : \acute{x} = \min \mathcal{C}_N(X) \setminus \{\hat{x}_1, .., \hat{x}_{i-1}\}\} \tag{35}$$

preso in ordine lessicografico, ovvero il primo dato del campione di misure di X, tra quelli che assumono il valore $\min(\mathcal{C}_N(X) \setminus \{\hat{x}_1, .., \hat{x}_{i-1}\})$, che compare nella tabella rappresentativa di $\mathcal{C}'_{M_i}(X)$ leggendo i dati da destra verso sinistra. Ovvero, $\widehat{\mathcal{C}}_N(X)$ è il campione di misure ottenuto da $\mathcal{C}_N(X)$ riordinando i dati in ordine di valore crescente, cioè dal più piccolo al più grande, tenendo conto dell'ordinamento con cui compaiono nella tabella rappresentativa di $\mathcal{C}_N(X)$.

Definizione 13 *Si definisce* mediana campionaria *il valore*

$$\hat{x} := \begin{cases} \frac{\left(\hat{x}_{\frac{N}{2}} + \hat{x}_{\frac{N}{2}+1}\right)}{2} & \text{se } N \text{ è pari} \\ \hat{x}_{\frac{N+1}{2}} & \text{se } N \text{ è dispari} \end{cases} . \tag{36}$$

Proposizione 14 *Sia $\mathcal{C}_N(X)$ un campione di numerosità N di misure di una data quantità X e sia Y una quantità legata ad X dalla relazione lineare $Y = AX + B$, con A e B costanti indipendenti dai dati del campione $\mathcal{C}_N(X)$. Se $\mathcal{C}_N(Y)$ è il campione di misure di Y generato dai dati di $\mathcal{C}_N(X)$, la mediana campionaria di $\mathcal{C}_N(Y)$ risulta*

$$\hat{y} = A\hat{x} + B . \tag{37}$$

Dimostrazione. Sia $\widehat{\mathcal{C}}_N(X)$ il campione di misure di X in cui i dati compaiono in ordine crescente. Distinguiamo due casi, il primo in cui $A \geq 0$ e il secondo in cui $A < 0$.

1. $A \geq 0$.

   Poiché $A \geq 0$, per ogni $i = 1, .., N$,

   $$\hat{x}_i \leq \hat{x}_{i+1} \Longrightarrow {}^2 A\hat{x}_i \leq A\hat{x}_{i+1} \Longrightarrow A\hat{x}_i + B \leq A\hat{x}_{i+1} + B . \tag{38}$$

   Dunque, ponendo per ogni $i = 1, .., N$, $\hat{y}_i := A\hat{x}_i + B$, la (38) implica che, se $\hat{x}_i \leq \hat{x}_{i+1}$, allora $\hat{y}_i \leq \hat{y}_{i+1}$. Perciò,

[2] L'espressione $A \Longrightarrow B$ indica che l'affermazione A implica l'affermazione B.

$\widehat{\mathcal{C}}_N(X) := \{\hat{y}_1, .., \hat{y}_N\}$ risulta essere il campione di misure di Y, ottenute da quelle di X, in cui i dati compaiono in ordine crescente. Quindi,

$$\hat{y} = \begin{cases} \frac{\left(\hat{y}_{\frac{N}{2}} + \hat{y}_{\frac{N}{2}+1}\right)}{2} = \frac{\left(A\hat{x}_{\frac{N}{2}} + B + A\hat{x}_{\frac{N}{2}+1} + B\right)}{2} = A\frac{\left(\hat{x}_{\frac{N}{2}} + \hat{x}_{\frac{N}{2}+1}\right)}{2} + B & se\, N\, è\, pari \\ \hat{y}_{\frac{N+1}{2}} = A\hat{x}_{\frac{N+1}{2}} + B & se\, N\, è\, dispari \end{cases} , \tag{39}$$

ovvero la tesi.

2. $A < 0$.

   Poiché $A < 0$, per ogni $i = 1, .., N$,

$$\hat{x}_i \leq \hat{x}_{i+1} \Longrightarrow A\hat{x}_i \geq A\hat{x}_{i+1} \Longrightarrow A\hat{x}_i + B \geq A\hat{x}_{i+1} + B\ . \tag{40}$$

   Dunque, ponendo per ogni $i = 1, .., N$, $\hat{y}_i := A\hat{x}_{N-i+1} + B$, la (38) implica che se $\hat{x}_{N-i} \leq \hat{x}_{N-i+1}$ allora $\hat{y}_i \leq \hat{y}_{i+1}$. Perciò, $\widehat{\mathcal{C}}_N(X) := \{\hat{y}_1, .., \hat{y}_N\}$ risulta essere il campione di misure di Y, ottenute da quelle di X, in cui i dati compaiono in ordine crescente. Quindi,

$$\begin{aligned} \hat{y} &= \begin{cases} \frac{\left(\hat{y}_{\frac{N}{2}} + \hat{y}_{\frac{N}{2}+1}\right)}{2} = \frac{\left(A\hat{x}_{N-\frac{N}{2}+1} + B + A\hat{x}_{N-\left(\frac{N}{2}+1\right)+1} + B\right)}{2} \\ \hat{y}_{\frac{N+1}{2}} = A\hat{x}_{N-\frac{N+1}{2}+1} + B = A\hat{x}_{\frac{N+1}{2}} + B \end{cases} \\ &= \begin{cases} \frac{\left(A\hat{x}_{\frac{N}{2}+1} + B + A\hat{x}_{\frac{N}{2}} + B\right)}{2} = A\frac{\left(\hat{x}_{\frac{N}{2}} + \hat{x}_{\frac{N}{2}+1}\right)}{2} + B & se\, N\, è\, pari \\ A\hat{x}_{\frac{N+1}{2}} + B & se\, N\, è\, dispari \end{cases} , \end{aligned} \tag{41}$$

   ovvero di nuovo la tesi.

■

Valori modali campionari Consideriamo l'istogramma delle frequenze dei valori dei dati di un campione $\mathcal{C}_N(X)$ di misure di una certa quantità X.

Definizione 15 *I valori dei dati di un campione che hanno frequenza, e quindi anche frequenza relativa, massima sono detti* valori modali campionari.

Nel caso ne esista uno solo di tali valori, questo è detto moda campionaria.

Nel seguito, se il campione $\mathcal{C}_N(X)$ ha K valori modali, con $1 \leq K \leq N$, denoteremo tali valori $\{\tilde{x}_1, .., \tilde{x}_K\}$ e la moda campionaria $\tilde{x}$.

Proposizione 16 *Sia $\mathcal{C}_N(X)$ un campione di numerosità N di misure di una data quantità X e sia $\{\tilde{x}_i\}_{i=1,..,K}$, $1 \leq K \leq N$, la collezione dei valori modali campionari di $\mathcal{C}_N(X)$. Se Y è una quantità legata ad X dalla relazione lineare $Y = AX + B$, con A e B costanti indipendenti dai dati del campione $\mathcal{C}_N(X)$, i valori modali campionari $\{\tilde{y}_i\}_{i=1,..,K}$, del campione $\mathcal{C}_N(Y)$ di misure di Y generato dai dati di $\mathcal{C}_N(X)$ risultano essere*

$$\tilde{y}_i = A\tilde{x}_i + B \qquad i = 1, .., K \ . \tag{42}$$

Dimostrazione. Notiamo che poichè i dati del campione $\mathcal{C}_N(Y)$ si ottengono da quelli del campione $\mathcal{C}_N(X)$ tramite la relazione

$$y_i = Ax_i + B \qquad i = 1, .., N \ , \tag{43}$$

per ogni $i = 1, .., N$, la frequenza del dato y_i è la stessa di quella del dato x_i. Infatti, poiché per ogni $i, j = 1, .., N$,

$$y_j = y_i \iff {}^{3} Ax_i + B = Ax_j + B \iff x_i = x_j \tag{44}$$

dall'Osservazione 7 , segue

$$\frac{F_{y_i}}{N} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{y_i}(y_j) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{x_i}(x_j) = f_i \ . \tag{45}$$

Dunque, i valori modali campionari di $\mathcal{C}_N(Y)$ sono quelli che corrispondono ai valori modali campionari di $\mathcal{C}_N(X)$ secondo la relazione lineare $y = Ax + B$, ovvero $\{A\tilde{x}_i + B\}_{i=1,..,K}$. ■

Media campionaria

Definizione 17 *Sia $\mathcal{C}_N(X)$ un campione di numerosità N di misure di una data quantità X. La statistica*

$$\bar{x} := \frac{1}{N} \sum_{i=1}^{N} x_i \tag{46}$$

è detta media campionaria *o* empirica.

[3] L'epressione $A \iff B$ indica che $A \implies B$ e che $B \implies A$.

Osservazione 18 *Notiamo che, se* $\{\mathcal{B}_N^k\}_{k\in\mathbb{Z}}$ *è una partizione in classi del campione del tipo di quella definita nella (14), si ha*

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{k\in\mathbb{Z}} \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathcal{B}_N^k(X)}(x_i)\, x_i = \sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{B}_N^k(X)\right|}{N} \times \tag{47}$$

$$\times \frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1}^{N} \mathbf{1}_{\mathcal{B}_N^k(X)}(x_i)\, x_i = \sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{B}_N^k(X)\right|}{N} \frac{1}{\left|\mathcal{B}_N^k(X)\right|} \sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)} x_i$$

$$= \sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{B}_N^k(X)\right|}{N} \bar{x}_k \ ,$$

ovvero $\bar{x}$ *si può scrivere come la somma dei prodotti delle frequenze relative delle classi* $\frac{\left|\mathcal{B}_N^k(X)\right|}{N}$ *per la media empirica dei valori dei dati di ciascuna classe*

$$\bar{x}_k := \frac{1}{\left|\mathcal{B}_N^k(X)\right|} \sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)} x_i \ . \tag{48}$$

Nel caso in cui i dati del campione assumano un numero finito o numerabile di valori $\{v_k\}_{k=1,..,K}$, *con* $K \in \mathbb{N}^+$ *(cioè* K *intero positivo), rimpiazzando nella (47) le classi* $\mathcal{B}_N^k(X)$ *con i sottocampioni* $\mathcal{F}_N^k(X)$ *associati ai valori* v_k *definiti nella Definizione 3 , ovvero*

$$\mathcal{F}_N^k(X) = \{y \in \mathcal{C}_N(X) : y = v_k\} \ , \qquad k = 1,..,K \ , \tag{49}$$

si ottiene

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{k\in\mathbb{Z}} \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathcal{F}_N^k(X)}(x_i)\, x_i \\ &= \sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{F}_N^k(X)\right|}{N} \frac{1}{\left|\mathcal{F}_N^k(X)\right|} \sum_{i=1,..,N\,:\,x_i\in\mathcal{F}_N^k(X)} x_i = \sum_{k\in\mathbb{Z}} f_k v_k \ , \end{aligned} \tag{50}$$

dove $f_k := \frac{\left|\mathcal{F}_N^k(X)\right|}{N}$ *è la frequenza relativa del* k*-simo valore (cioè* v_k*) assunto dai dati del campione.*

Proposizione 19 *Sia* $\mathcal{C}_N(X)$ *un campione di numerosità* N *di misure di una data quantità* X *e sia* Y *una quantità legata ad* X *dalla relazione lineare* $Y = AX + B$, *con* A *e* B *costanti indipendenti dai dati del campione* $\mathcal{C}_N(X)$. *Se* $\mathcal{C}_N(Y)$ *è il campione di misure di* Y *generato dai dati di* $\mathcal{C}_N(X)$, *la media campionaria di* $\mathcal{C}_N(Y)$ *risulta*

$$\bar{y} = A\bar{x} + B \ . \tag{51}$$

Dimostrazione. Poiché $\frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$ e $\frac{1}{N}\sum_{i=1}^{N} 1 = 1$ si ha

$$\begin{aligned} \bar{y} &= \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}(Ax_i + B) = \frac{1}{N}\sum_{i=1}^{N} Ax_i + \frac{1}{N}\sum_{i=1}^{N} B \qquad (52) \\ &= A\frac{1}{N}\sum_{i=1}^{N} x_i + B\frac{1}{N}\sum_{i=1}^{N} 1 = A\bar{x} + B \ . \end{aligned}$$

■

Varianza e deviazione standard campionarie Dato un campione $\mathcal{C}_N(X)$ di misure di una quantità X, notiamo che ogni dato x_i del campione può essere espresso come la somma della media campionaria più un resto. Ovvero,

$$x_i = \bar{x} + (x_i - \bar{x}) \ , \quad i = 1, .., N \, . \qquad (53)$$

La differenza tra il valore di ciascun dato e la media campionaria è detta *scarto.* Notiamo che dalla definizione di media campionaria (46) segue che la somma degli scarti è nulla come pure quindi la loro media empirica, cioè

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) = \bar{x} - \frac{1}{N}\sum_{i=1}^{N}\bar{x} = 0 \, . \qquad (54)$$

Dunque, per capire come si distribuiscono i valori dei dati intorno alla loro media campionaria sarebbe utile calcolare la media empirica dei valori assoluti degli scarti, ovvero

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}| \ , \qquad (55)$$

cioè la distanza media dei valori dei dati dalla loro media campionaria. Per ragioni che sono parte dell'oggetto di un corso di Statistica più avanzato, risulta più utile calcolare la media empirica dei quadrati degli scarti, ovvero

$$\bar{s}^2_X := \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \ , \qquad (56)$$

detta *scarto quadratico medio,* anziché la (55). D'altra parte, per la concavità della funzione radice quadrata e per disuguaglianza di Jensen[4] si ha

$$\frac{1}{N}\sum_{i=1}^{N}|x_i-\bar{x}| = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i-\bar{x})^2} \leq \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2} = \sqrt{\bar{s}_X^2}\,. \qquad (57)$$

Perciò la radice quadrata dello scarto quadratico medio, rappresenta una sovrastima della media empirica dei valori assoluti degli scarti.

Definizione 20 *Sia $\mathcal{C}_N(X)$ un campione di numerosità N di misure di una data quantità X. La statistica*

$$s_X^2 := \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2 \qquad (58)$$

è detta varianza campionaria *o* empirica.

Osservazione 21 *Talvolta, in letteratura, la varianza campionaria e lo scarto quadratico medio vengono confuse. Tuttavia, poiché*

$$\frac{N-1}{N} = 1-\frac{1}{N} < 1 \quad e \quad \frac{N}{N-1} = 1+\frac{1}{N-1} > 1\,, \qquad (59)$$

$$\begin{aligned}\bar{s}_X^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 = \frac{N-1}{N}\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2 \qquad (60)\\ &= \left(1-\frac{1}{N}\right)s_X^2 < s_X^2\,,\end{aligned}$$

cioè la varianza campionaria è una sovrastima dello scarto quadratico medio. Inoltre, la loro differenza percentuale è inversamente proporzionale alla numerosità del campione e tende a zero nel limite di $N \longrightarrow \infty$, in quanto

$$\left|\frac{s_X^2-\bar{s}_X^2}{s_X^2}\right| = \frac{1}{N} \quad , \quad \left|\frac{s_X^2-\bar{s}_X^2}{\bar{s}_X^2}\right| = \frac{1}{N-1}\,. \qquad (61)$$

Tuttavia, si può dimostrare (cfr [R] par. 6.4) che la varianza campionaria gode di proprietà ulteriori oltre a quelle di cui gode lo scarto quadratico medio.

[4]La disuguaglianza di Jensen in questo caso afferma che, se f è una funzione concava, allora

$$\frac{1}{N}\sum_{i=1}^{N}f(x_i) \leq f\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)\,.$$

Osservazione 22 *Sviluppando il quadrato che compare nella definizione della varianza empirica e ricordando che* $\sum_{i=1}^{N} x_i = N\bar{x}$ *otteniamo*

$$\begin{aligned}(N-1)\,s_X^2 &= N\bar{s}_X^2 = \sum_{i=1}^{N}(x_i-\bar{x})^2 = \sum_{i=1}^{N}\left(x_i^2 - 2x_i\bar{x} + x_i^2\right) \qquad (62)\\ &= \sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} x_i\bar{x} + \sum_{i=1}^{N}\bar{x}^2\\ &= \sum_{i=1}^{N} x_i^2 - 2\bar{x}\sum_{i=1}^{N} x_i + N\bar{x}^2\\ &= \sum_{i=1}^{N} x_i^2 - 2N\bar{x}^2 + N\bar{x}^2\end{aligned}$$

e dunque

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i^2 - \bar{x}^2\right), \qquad (63)$$

$$\bar{s}_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\,. \qquad (64)$$

Osservazione 23 *Analogamente all'Osservazione 18, notiamo che, se* $\{\mathcal{B}_N^k\}_{k\in\mathbb{Z}}$ *è una partizione in classi del campione del tipo di quella definita nella (14), si ha*

$$\begin{aligned}(N-1)\,s_X^2 &= N\bar{s}_X^2 = \sum_{i=1}^{N}(x_i-\bar{x})^2 = \sum_{k\in\mathbb{Z}}\sum_{i=1}^{N}\mathbf{1}_{\mathcal{B}_N^k(X)}(x_i)\,(x_i-\bar{x})^2 \qquad (65)\\ &= \sum_{k\in\mathbb{Z}}\left|\mathcal{B}_N^k(X)\right|\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1}^{N}\mathbf{1}_{\mathcal{B}_N^k(X)}(x_i)\,(x_i-\bar{x})^2\\ &= N\sum_{k\in\mathbb{Z}}\frac{\left|\mathcal{B}_N^k(X)\right|}{N}\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1}^{N}\mathbf{1}_{\mathcal{B}_N^k(X)}(x_i)\,(x_i-\bar{x})^2\\ &= N\sum_{k\in\mathbb{Z}}\frac{\left|\mathcal{B}_N^k(X)\right|}{N}\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}(x_i-\bar{x})^2\,.\end{aligned}$$

Notiamo che, per la (48), si ha

$$\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}(x_i-\bar{x})^2=\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}\left[(x_i-\bar{x}_k)+(\bar{x}_k-\bar{x})\right]^2$$

$$=\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}\left[(x_i-\bar{x}_k)^2+2\,(x_i-\bar{x}_k)\,(\bar{x}_k-\bar{x})+(\bar{x}_k-\bar{x})^2\right]\ ,$$

ma poiché $\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}=1$,

$$\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}(\bar{x}_k-\bar{x})^2=\frac{(\bar{x}_k-\bar{x})^2}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}=(\bar{x}_k-\bar{x})^2\ . \tag{66}$$

Inoltre, dalla (48), segue

$$\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}(x_i-\bar{x}_k)\,(\bar{x}_k-\bar{x})=\frac{(\bar{x}_k-\bar{x})}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}(x_i-\bar{x}_k) \tag{67}$$

$$=(\bar{x}_k-\bar{x})\left[\left(\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}x_i\right)-\bar{x}_k\right]=0\ .$$

Quindi,

$$\begin{aligned}(N-1)\,s_X^2 &= N\bar{s}_X^2=\sum_{i=1}^{N}(x_i-\bar{x})^2 \qquad (68)\\ &= N\sum_{k\in\mathbb{Z}}\frac{\left|\mathcal{B}_N^k(X)\right|}{N}\left[(\bar{x}_k-\bar{x})^2+\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}(x_i-\bar{x}_k)^2\right]\ .\end{aligned}$$

Ovvero,

$$\begin{aligned}s_X^2 &= \frac{N}{N-1}\sum_{k\in\mathbb{Z}}\frac{\left|\mathcal{B}_N^k(X)\right|}{N}\left[(\bar{x}_k-\bar{x})^2+\frac{1}{\left|\mathcal{B}_N^k(X)\right|}\sum_{i=1,..,N\,:\,x_i\in\mathcal{B}_N^k(X)}(x_i-\bar{x}_k)^2\right]\\ &= \left(1+\frac{1}{N-1}\right)\sum_{k\in\mathbb{Z}}\frac{\left|\mathcal{B}_N^k(X)\right|}{N}\left[\bar{s}_k^2+(\bar{x}_k-\bar{x})^2\right] \qquad (69)\end{aligned}$$

si può scrivere come $1+\frac{1}{N-1}$ *volte la somma dei prodotti delle frequenze relative delle classi* $\frac{\left|\mathcal{B}_N^k(X)\right|}{N}$ *per la somma dello scarto quadratico medio dei*

valori dei dati di ciascuna classe

$$\bar{s}_k^2 := \frac{1}{\left|\mathcal{B}_N^k(X)\right|} \sum_{i=1,..,N\,:\,x_i \in \mathcal{B}_N^k(X)} (x_i - \bar{x}_k)^2 \tag{70}$$

e del quadrato dello scarto dei valori delle medie empiriche di ciascuna classe dalla media compionaria $(\bar{x}_k - \bar{x})^2$. *Analogamente,*

$$\bar{s}_X^2 = \sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{B}_N^k(X)\right|}{N} \left[\bar{s}_k^2 + (\bar{x}_k - \bar{x})^2\right] \ . \tag{71}$$

Nel caso in cui i dati del campione assumano un numero finito o numerabile di valori $\{v_k\}_{k=1,..,K}$, *con* $K \in \mathbb{N}^+$ *(cioè* K *intero positivo), rimpiazzando nella (65) le classi* $\mathcal{B}_N^k(X)$ *con i sottocampioni* $\mathcal{F}_N^k(X)$ *associati ai valori* v_k *definiti nella Definizione 3 e nella (49), si ha*

$$\begin{aligned} s_X^2 &= \frac{1}{N-1}\sum_{i=1}^N (x_i - \bar{x})^2 = \sum_{k\in\mathbb{Z}} \frac{1}{N-1}\sum_{i=1}^N \mathbf{1}_{\mathcal{F}_N^k(X)}(x_i)\,(x_i - \bar{x})^2 \qquad (72)\\ &= \frac{N}{N-1}\sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{F}_N^k(X)\right|}{N} \frac{1}{\left|\mathcal{F}_N^k(X)\right|} \sum_{i=1,..,N\,:\,x_i\in\mathcal{F}_N^k(X)} (x_i - \bar{x})^2 \\ &= \left(1 + \frac{1}{N-1}\right)\sum_{k\in\mathbb{Z}} f_k\,(v_k - \bar{x})^2 \ . \end{aligned}$$

Proposizione 24 *Sia* $\mathcal{C}_N(X)$ *un campione di numerosità* N *di misure di una data quantità* X *e sia* Y *una quantità legata ad* X *dalla relazione lineare* $Y = AX + B$, *con* A *e* B *costanti indipendenti dai dati del campione* $\mathcal{C}_N(X)$. *Se* $\mathcal{C}_N(Y)$ *è il campione di misure di* Y *generato dai dati di* $\mathcal{C}_N(X)$, *la varianza campionaria di* $\mathcal{C}_N(Y)$ *risulta*

$$s_Y^2 = A^2 s_X^2 \tag{73}$$

come pure il suo lo scarto quadratico medio.

Dimostrazione. Dalla (51) segue

$$\sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N \left[Ax_i + B - (A\bar{x} + B)\right]^2 = A^2 \sum_{i=1}^N (x_i - \bar{x})^2 \ . \tag{74}$$

Perciò, dalla (58) e dalla (56), si ha

$$\begin{aligned} (N-1)\,s_Y^2 &= A^2\,(N-1)\,s_X^2\,, \qquad (75)\\ N\bar{s}_Y^2 &= A^2 N \bar{s}_X^2\,. \qquad (76) \end{aligned}$$

■

Osservazione 25 *Notiamo che, nel caso in cui Y rappresenti la stessa quantità rappresentata da X, ma misurata in un differente sistema di unità di misura, la varianza campionaria di Y sarà proporzionale a quella di X con costante di proporzionalità pari al quadrato del fattore di conversione tra le misure nei due sistemi di unità di misura.*

Per esempio, nel caso in cui X e Y rappresentino la velocità di un oggetto misurata rispettivamente in metri al secondo e in chilometri all'ora, la varianza campionaria di Y sarà pari a 6^{-4} volte quella di X.

Definizione 26 *Dato un campione $\mathcal{C}_N(X)$ di misure di numerosità N di una quantità X, la statistica*

$$s_X := \sqrt{s_X^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{77}$$

è detta deviazione standard campionaria *o* empirica.

Osservazione 27 *Dalle (57), (60) segue che*

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}| \leq \sqrt{\bar{s}_X^2} < \sqrt{s_X^2} = s_X\,, \tag{78}$$

ovvero la deviazione standard campionaria rappresenta una sovrastima della media empirica dei valori assoluti degli scarti.

Inoltre, dalla Proposizione 24 , si ha che se Y è una quantità legata ad X dalla relazione lineare $Y = AX + B$, con A e B costanti indipendenti dai dati del campione $\mathcal{C}_N(X)$, e $\mathcal{C}_N(Y)$ è il campione di misure di Y generato dai dati di $\mathcal{C}_N(X)$, la deviazione standard campionaria di $\mathcal{C}_N(Y)$ risulta

$$s_Y = \sqrt{s_Y^2} = \sqrt{A^2 s_X^2} = |A|\, s_X\,. \tag{79}$$

Diamo ora una stima per difetto della frequenza relativa dei dati di un campione $C_N(X)$ di misure di una quantità X che cadono in un intorno della media empirica di raggio proporzionale alla deviazione standard.

Proposizione 28 (Disuguglianza di Chebychev) *Sia $\mathcal{C}_N(X)$ un campione di misure di numerosità N di una quantità X e*

$$\mathcal{S}_N^{\alpha}(X) := \{x \in \mathcal{C}_N(X) : |x - \bar{x}| < \alpha s_X\}\,, \quad \alpha \geq 1\,, \tag{80}$$

sia, fissato α, il sottocampione di $\mathcal{C}_N(X)$ contenente tutti i dati il cui valore è compreso tra $\bar{x} - \alpha s_X$ e $\bar{x} + \alpha s_X$. Allora,

$$\frac{|\mathcal{S}_N^{\alpha}(X)|}{N} > 1 - \frac{1}{\alpha^2}\,. \tag{81}$$

Dimostrazione. Per ogni $\alpha \geq 1$, se $\mathcal{C}_N(X) \setminus \mathcal{S}_N^\alpha(X)$ è il sottocampione di $\mathcal{C}_N(X)$ composto dai dati che non appartengono a $\mathcal{S}_N^\alpha(X)$, ovvero

$$\mathcal{C}_N(X) \setminus \mathcal{S}_N^\alpha(X) := \{x \in \mathcal{C}_N(X) : |x - \bar{x}| \geq \alpha s_X\}, \tag{82}$$

si ha

$$\begin{aligned}(N-1)\, s_X^2 &= \sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1,..,N\,:\, x_i \in \mathcal{S}_N^\alpha(X)} (x_i - \bar{x})^2 + \\ &+ \sum_{i=1,..,N\,:\, x_i \in \mathcal{C}_N(X) \setminus \mathcal{S}_N^\alpha(X)} (x_i - \bar{x})^2 .\end{aligned} \tag{83}$$

Poiché $\sum_{i=1,..,N\,:\, x_i \in \mathcal{S}_N^\alpha(X)} (x_i - \bar{x})^2$, come somma di quadrati di numeri, è una quantità non negativa,

$$(N-1)\, s_X^2 \geq \sum_{i=1,..,N\,:\, x_i \in \mathcal{C}_N(X) \setminus \mathcal{S}_N^\alpha(X)} (x_i - \bar{x})^2 . \tag{84}$$

Inoltre, poiché se $x_i \in \mathcal{C}_N(X) \setminus \mathcal{S}_N^\alpha(X)$ dalla (82) segue che $(x_i - \bar{x})^2 \geq \alpha^2 s_X^2$,

$$\begin{aligned}(N-1)\, s_X^2 &\geq \sum_{i=1,..,N\,:\, x_i \in \mathcal{C}_N(X) \setminus \mathcal{S}_N^\alpha(X)} \alpha^2 s_X^2 \\ &= |\mathcal{C}_N(X) \setminus \mathcal{S}_N^\alpha(X)|\, \alpha^2 s_X^2 \\ &= (N - |\mathcal{S}_N^\alpha(X)|)\, \alpha^2 s_X^2 .\end{aligned} \tag{85}$$

Perciò,

$$\begin{aligned}(N-1) &\geq (N - |\mathcal{S}_N^\alpha(X)|)\, \alpha^2 , \\ \frac{|\mathcal{S}_N^\alpha(X)|}{N} &\geq 1 - \frac{1}{\alpha^2}\frac{N-1}{N} > 1 - \frac{1}{\alpha^2} .\end{aligned} \tag{86}$$

■

Analisi qualitativa della forma degli istogrammi delle frequenze dei dati di un campione

Definizione 29 *Un campione sufficientemente numeroso di misure di una quantità X, $\mathcal{C}_N(X)$, è detto* normale *se l'istogramma delle frequenze relative delle classi in cui può essere suddiviso ha le seguenti proprietà:*

- *esiste un'unica classe di frequenza relativa massima, la quale corrisponde all'intervallo di valori dei dati in cui ricade la mediana campionaria;*

- *l'istogramma delle frequenze relative delle classi risulta piuttosto simmetrico rispetto alla mediana campionaria e le frequenze relative delle classi decrescono fino al valore* 0, *man mano che il valore assoluto della differenza tra i valori dei dati e la mediana campionaria cresce.*

Definizione 30 *Un campione sufficientemente numeroso di misure di una quantità* X, $\mathcal{C}_N(X)$, *è detto* distorto a destra, *se esiste un'unica classe di frequenza relativa massima, ma la mediana campionaria è sensibilmente più grande dei valori dei dati che appartengono a questa classe.*

Definizione 31 *Un campione sufficientemente numeroso di misure di una quantità* X, $\mathcal{C}_N(X)$, *è detto* distorto a sinistra, *se esiste un'unica classe di frequenza relativa massima, ma la mediana campionaria è sensibilmente più piccola dei valori dei dati che appartengono a questa classe.*

Nel caso in cui $\mathcal{C}_N(X)$ sia un campione normale di numerosità N di misure di una quantità X, è noto empiricamente che:

- circa il 68% dei dati di $\mathcal{C}_N(X)$ cade in un intorno di raggio s_X da $\bar{x}$, ovvero nell'intervallo $[\bar{x} - s_X, \bar{x} + s_X]$;
- circa il 95% dei dati di $\mathcal{C}_N(X)$ cade in un intorno di raggio $2s_X$ da $\bar{x}$, ovvero nell'intervallo $[\bar{x} - 2s_X, \bar{x} + 2s_X]$;
- circa il $99,7\%$ dei dati di $\mathcal{C}_N(X)$ cade in un intorno di raggio $3s_X$ da $\bar{x}$, ovvero nell'intervallo $[\bar{x} - 3s_X, \bar{x} + 3s_X]$.

Queste stime sono in perfetto accordo con quella che s'ottiene dalla disuguaglianza di Chebichev e sono anche più precise, sebbene valgano esclusivamente nel caso di campioni normali.

Osserviamo inoltre che, se Y è una quantità legata alla X da una relazione lineare del tipo $Y = AX + B$, con A eB costanti indipendenti dai dati di $\mathcal{C}_N(X)$, anche il campione $\mathcal{C}_N(Y)$, generato di dati di $\mathcal{C}_N(X)$ tramite la relazione lineare che lega la Y alla X, sarà normale e le sue statistiche principali saranno legate a quelle del campione tramite le (37), (42), (51), (73) e (79).

Definizione 32 *Un campione sufficientemente numeroso di misure di una quantità* X, $\mathcal{C}_N(X)$, *è detto* multimodale, *se non esiste un'unica classe di frequenza relativa massima.*

In generale un campione multimodale è indice di una popolazione campionaria non omogenea, ovvero quando i dati sono relativi a misure in cui una o più caratteristiche sperimentali sono mutate, per esempio l'apparato di misura, e quindi rappresenta la collezione di più campioni omogenei *unimodali* del tipo di quelli appena descritti, ovvero normali e/o distorti a destra o sinistra (cfr [R] fig. 2.12). Per esempio, nel caso si tratti di misure prodotte da due distinti apparati di misura che restituiscono valori sensibilmente distinti, si otterrebbe un istogramma bimodale risultato della composizione dei due campioni di dati ciascuno composto da misure ottenute con lo stesso apparato.

0.1.2 Campioni di dati bivariati

Alcuni esperimenti sono spesso volti ad analizzare il comportamento di una quantità Y in relazione alla variazione di una quantità misurata X. Ad esempio, si vuole studiare il comportamento del volume di un gas al variare della temperatura, o la quantità d'ossigeno consumato da una persona che cammina in corrispondenza della sua andatura. I dati generati da esperimenti di questo tipo sono dunque coppie di numeri reali (x, y) dove x rappresenta il valore della misura della quantità X e y quello della quantità Y. Un campione di numerosità N di tali dati si indicherà

$$\mathcal{C}_N(X, Y) := \{(x_1, y_1), .., (x_N, y_N)\} \tag{87}$$

e sarà detto *campione bivariato.*

I campioni $\mathcal{C}_N(X)$ e $\mathcal{C}_N(Y)$ composti rispettivamente dalle misure delle quantità X e Y, sono detti *campioni univariati associati* o *marginali.* Dunque, ogni campione bivariato si può rappresentare tramite una tabella composta da due colonne, nella prima delle quali compaiono le misure della quantità X, ovvero i dati di $\mathcal{C}_N(X)$, e nella seconda le misure della quantità Y, ovvero i dati di $\mathcal{C}_N(Y)$ (cfr [R] tab. 2.8 pag. 36).

Diagrammi di dispersione

Per analizzare la relazione che intercorre tra le misure delle quantità X e Y è utile graficare i dati della tabella associata a $\mathcal{C}_N(X, Y)$ come punti del piano cartesiano, ponendo in ascissa i valori delle misure della quantità X, cioè i valori dei dati di $\mathcal{C}_N(X)$, ed in ordinata i corrispondenti valori delle misure della quantità Y, cioè quelli che compaiono nella tabella relativa a $\mathcal{C}_N(X, Y)$ affianco di ciascun dato di $\mathcal{C}_N(X)$. Questo grafico è noto come *diagramma di dispersione* (cfr [R] figg. 2.13, 2.14, 2.15).

Correlazione tra le misure delle due quantità rappresentate dai dati di un campione bivariato Volendo conoscere se al variare della quantità X vari anche la quantità Y, è utile osservare se al valore del segno dello scarto di una misura di X dalla media campionaria di $\mathcal{C}_N(X)$, cioè $(x_i - \bar{x})$, corrisponda un valore del segno dello scarto della corrispondente misura di Y dalla media campionaria di $\mathcal{C}_N(Y)$, ovvero $(y_i - \bar{y})$, che risulti spesso identico per tutti i valori di $(x_i - \bar{x})$ dello stesso segno. In altre parole, è possibile che vi sia una dipendenza funzionale di Y da X se a valori positivi degli scarti di $\mathcal{C}_N(X)$ corrispondano valori degli scarti di $\mathcal{C}_N(Y)$ aventi spesso tutti stesso segno e, viceversa, a valori negativi degli scarti di $\mathcal{C}_N(X)$ corrispondano valori degli scarti di $\mathcal{C}_N(Y)$ aventi spesso valori di segno opposto a quello assunto nel caso di scarti di $\mathcal{C}_N(X)$ positivi. In tal caso, il prodotto degli scarti $(x_i - \bar{x})(y_i - \bar{y})$ avrebbe spesso segno definito e la somma di questi

$$\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \tag{88}$$

risulterebbe sempre diversa da zero.

Definizione 33 *Sia $\mathcal{C}_N(X,Y)$ un campione di numerosità N di misure delle quantità X, Y e $\mathcal{C}_N(X), \mathcal{C}_N(Y)$ i campioni univariati associati. La statistica*

$$s_{X,Y} := \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \tag{89}$$

è detta covarianza campionaria *e, nel caso in cui $s_{X,Y}$ risulti positiva, le quantità X e Y si dicono essere* positivamente correlate. *Al contrario, nel caso in cui $s_{X,Y}$ risulti negativa, le quantità X e Y si dicono essere* negativamente correlate, *mentre si dicono essere* scorrelate *o* incorrelate *nel caso in cui $s_{X,Y}$ risulti nulla.*

Osservazione 34 *Lo svantaggio della covarianza campionaria è che fisicamente risulta una quantità la cui dimensione è il prodotto delle dimensioni delle quantità X e Y. Per esempio se X fosse una pressione e Y una temperatura, le dimensioni di $s_{X,Y}$ sarebbero quelle di una forza per una temperatura divise per una superfice, quindi scalerebbe in maniera complicata cambiando sistema di unità di misura. Ad ulteriore riprova di ciò, calcoliamo $s_{Z,W}$ se $Z = AX + B$ e $W = CY + D$, con A, B, C, D costanti indipendenti dai dati.*

Allora, per la (51)

$$s_{Z,W} = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{z})(w_i - \bar{w}) = \frac{1}{N-1}\sum_{i=1}^{N}(Ax_i + B - (A\bar{x} + B)) \times \tag{90}$$

$$\times (Cy_i + D - (C\bar{y} + D)) = \frac{AC}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = ACs_{X,Y}\,.$$

Per ovviare al problema descritto nella precedente osservazione, s'introduce una statistica che ha le stesse caratteristiche della covarianza campionaria, ma che risulta indipendente dal sistema di unità di misura in uso.

Definizione 35 *Sia $\mathcal{C}_N(X,Y)$ un campione di numerosità N di misure delle quantità X, Y e $\mathcal{C}_N(X), \mathcal{C}_N(Y)$ i campioni univariati associati. La statistica*

$$r_{X,Y} := \frac{s_{X,Y}}{s_X s_Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{91}$$

è detto coefficiente di correlazione campionaria.

Osservazione 36 *Notiamo che essendo s_X e s_Y quantità positive, nel caso in cui $r_{X,Y}$ sia positivo, le quantità X e Y risultano essere positivamente correlate. Al contrario, nel caso in cui $r_{X,Y}$ sia negativo, le quantità X e Y risultano essere negativamente correlate, mentre risultano essere scorrelate o incorrelate nel caso in cui $r_{X,Y}$ sia nullo.*

Proposizione 37 *Sia $\mathcal{C}_N(X,Y)$ un campione di numerosità N di misure delle quantità X, Y il coefficiente di correlazione campionaria è un numero puro sempre compreso tra -1 e 1, ovvero $r_{X,Y} \in [-1,1]$. I valori estremali sono raggiunti solo se tra X ed Y sussiste una relazione funzionale lineare.*

Dimostrazione. Per la disuguaglianza di Schwarz[5]

$$\begin{aligned}\left|\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})\right| &\leq \sum_{i=1}^{N}|(x_i-\bar{x})(y_i-\bar{y})| \\ &\leq \sum_{i=1}^{N}|(x_i-\bar{x})|\,|(y_i-\bar{y})| \\ &\leq \sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}\,.\end{aligned} \tag{92}$$

Perciò $|r_{X,Y}| \leq 1$.

Se $Y = AX + B$, con A e B costanti indipendenti dai dati di $\mathcal{C}_N(X)$, per la (58),

$$s_{X,Y} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(Ax_i+B-(A\bar{x}+B)) = As_X^2\,, \tag{93}$$

che, unitamente alla (79) e alla (91), implica

$$r_{X,Y} = \frac{A}{|A|} = \begin{cases} 1 & se\, A>0 \\ 0 & se\, A=0 \\ -1 & se\, A<0 \end{cases}\,. \tag{94}$$

Se invece $X = AY + B$, con A e B costanti indipendenti dai dati di $\mathcal{C}_N(Y)$, scambiando X con Y nelle precedenti espressioni di $s_{X,Y}$ e $r_{X,Y}$ si ottiene lo stesso risultato, da cui segue che il coefficiente di correlazione campionaria raggiunge i suoi valori estremali 1 e -1 quando le quantità X e Y sono una funzione lineare dell'altra.

Se inoltre poniamo $Z = AX + B$ e $W = CY + D$, con A, B, C, D costanti indipendenti dai dati di $\mathcal{C}_N(X, Y)$. Allora, dalle (90), (79) si ha che

$$r_{Z,W} = \frac{s_{Z,W}}{s_Z s_W} = \frac{ACs_{X,Y}}{|A|\,s_X\,|C|\,s_Y} = \frac{A}{|A|}\frac{C}{|C|}r_{X,Y}\,, \tag{95}$$

cioè il coefficiente di correlazione campionaria è un numero puro, ovvero è indipendente dal sistema di unità di misura in uso. ■

[5]La disuguaglianza di Schwarz afferma che date due collezioni di numeri reali $\{a_i\}_{i=1,..,N}$, $\{b_i\}_{i=1,..,N}$,

$$\sum_{i=1}^{N}a_ib_i \leq \sqrt{\sum_{i=1}^{N}a_i^2}\sqrt{\sum_{i=1}^{N}b_i^2}\,.$$

Pertanto, se il coefficiente di correlazione tra le misure delle quantità X e Y che formano il campione $\mathcal{C}_N(X,Y)$ assume valori prossimi a 1, i punti rappresentati nel diagramma di dipersione associato a $\mathcal{C}_N(X,Y)$ dovrebbero essere più o meno allineati lungo una retta crescente e viceversa. Se invece $r_{X,Y}$ assume valori prossimi a -1, i punti rappresentati nel diagramma di dipersione associato a $\mathcal{C}_N(X,Y)$ dovrebbero essere più o meno allineati lungo una retta decrescente e viceversa. Quindi, se i punti del diagramma di disperione risultano molto sparsi nel piano cartesiano, il coefficiente di correlazione tra X e Y dovrebbe assumere in modulo valori prossimi a zero.

Osservazione 38 *È importante rilevare che l'esistenza di una correlazione tra due quantità misurate non è in generale indice dell'esistenza di una relazione funzionale tra queste. Al contrario, se due quantità sono legate da una relazione funzionale allora queste saranno necessariamente correlate.*

Una spiegazione di ciò riposa sul concetto di dipendenza stocastica che esula dai limiti di questo corso. Per un ulteriore approfondimento su questo fatto si rimanda a [R] Esempio 2.6.2, pag.40 e successivo commento.

Parte II

Elementi di Statistica inferenziale

0.2 Distribuzione limite delle frequenze relative campionarie e suo utilizzo in statistica

Consideriamo un campione molto numeroso $\mathcal{C}_N(X)$ di misure di una quantità X, le quali possono assumere valori appartenenti ad un intervallo di numeri reali $I_X \subseteq \mathbb{R}$ eventualmente illimitato, com'è il caso in cui X sia la concentrazione di una sostanza, $I_X = [0,1]$, o la resistenza ohmica di un materiale $I_X = [0,+\infty)$. Come osservato nella sottosezione relativa alla descrizione della distribuzione dei dati di campioni di quantità le cui misure assumono una gran quantità di valori, per graficarne l'istogramma delle frequenze relative è opportuno suddividere il campione in classi legate ad una opportuna partizione dell'insieme di variabilità delle misure della quantità in esame X, scelta in modo tale che i dati del campione risultino quanto più possibile equidistribuiti all'interno di ogni classe. In generale, maggiore è la numerosità del campione $\mathcal{C}_N(X)$ di misure di X, più fine deve essere la partizione dell'insieme di variabilità delle misure di X a cui associare le classi in cui ripartire $\mathcal{C}_N(X)$. Pertanto, consideriamo, al variare della numerosità del campione N, una collezione di partizioni $\{\left(a_k^{(N)}, a_{k+1}^{(N)}\right]\}_{k\in\mathbb{Z}}$ di $\mathbb{R}$ tali che:

- per ogni $k \in \mathbb{Z}$, $a_{k+1}^{(N)} - a_k^{(N)} = \Delta_N$ è indipendente da k;
- $\Delta_N > \Delta_{N+1}$.

Ad esempio si può scegliere $\{\left(\frac{k}{N}, \frac{k+1}{N}\right]\}_{k\in\mathbb{Z}}$ per cui $\Delta_N = \frac{1}{N}$, oppure $\{\left(\frac{k}{2^N}, \frac{k+1}{2^N}\right]\}_{k\in\mathbb{Z}}$ per cui $\Delta_N = \frac{1}{2^N}$. Allora, indipendentemente dal valore di N, ricordando la definizione della classe $\mathcal{B}_N^k(X)$ (14),

$$1 = \sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{B}_N^k(X)\right|}{N} = \sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{B}_N^k(X)\right|}{N} \frac{\Delta_N}{\Delta_N} = \sum_{k\in\mathbb{Z}} \varphi_k^{(N)}(X)\,\Delta_N \ , \tag{96}$$

dove

$$\varphi_k^{(N)}(X) := \frac{\left|\mathcal{B}_N^k(X)\right|}{N\Delta_N} \tag{97}$$

è la *frequenza relativa della k-esima classe per unità di lunghezza della partizione.* Ponendo,

$$\mathbb{R} \ni x \longmapsto \phi_X^{(N)}(x) := \sum_{k\in\mathbb{Z}} \varphi_k^{(N)}(X)\,\mathbf{1}_{\left(a_k^{(N)}, a_{k+1}^{(N)}\right]}(x) \in \mathbb{R} \ , \tag{98}$$

poiché

$$\int_{\mathbb{R}} dx \mathbf{1}_{\left(a_k^{(N)}, a_{k+1}^{(N)}\right]}(x) = \Delta_N \ , \tag{99}$$

la serie $\sum_{k\in\mathbb{Z}} \varphi_k^{(N)}(X)\,\Delta_N$ non è altro che l'integrale della funzione $\phi_X^{(N)}$. Sia allora,

$$\mathbb{R} \ni x \longmapsto F_X^{(N)}(x) := \int_{-\infty}^{x} dy \phi_X^{(N)}(y) \in [0,1] \tag{100}$$

la frequenza relativa dei dati che cadono nell'insieme di valori $(-\infty, x]$. Si può dimostrare, ma esula dai limiti del corso, che, sotto opportune ipotesi, fissato un qualsiasi numero reale x, esiste il limite della successione numerica $\{F_X^{(N)}(x)\}_{N\in\mathbb{N}^+}$ che denotiamo $F_X(x)$. Perciò, ponendo per ogni intervallo $(a,b] \subset \mathbb{R}$,

$$\mathbb{P}_X(a,b] := F_X(b) - F_X(a) \ , \tag{101}$$

che è detta *probabilità dell'intervallo* $(a,b]$ *per i valori delle misure* X, si può ragionevolmente assumere che la frequenza relativa di un qualsiasi intervallo $[a,b]$ di valori assunti dai dati di un campione molto numeroso $\mathcal{C}_N(X)$ di misure della quantità X risulti

$$\frac{|\{x \in \mathcal{C}_N(X) : x \in [a,b]\}|}{N} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[a,b]}(x_i) = \mathbb{P}_X(a,b] + \varepsilon_N \ , \tag{102}$$

dove ε_N è l'errore che si commette approssimando tale frequenza relativa con $\mathbb{P}_X(a,b]$, il quale è tanto più piccolo quanto più N è grande, ovvero $\lim_{N\uparrow\infty} \varepsilon_N = 0$.

In diversi casi si può dimostrare anche che esiste il limite della successione di funzioni $\{\phi_X^{(N)}\}_{N\in\mathbb{N}^+}$, che denotiamo $\phi_X(x)$ e che è detta *densità di probabilità delle misure di* X, per cui, per ogni coppia di numeri reali a, b tali che $a < b$, gli intervalli (a,b), $(a,b]$, $[a,b)$ e $[a,b]$ hanno tutti la stessa probabilità

$$\mathbb{P}_X(a,b) = \int_a^b dx \phi_X(x) \ . \tag{103}$$

Perciò, per ogni numero reale x, il limite della frequenza relativa campionaria dell'intervallo infinitesimo $(x, x+dx]$ di valori delle misure di X risulta essere $\mathbb{P}_X(x, x+dx] = \phi_X(x)\,dx$.

In pratica, quando la forma funzionale di ϕ_X è nota esplicitamente, la dipendenza da X di ϕ_X, si traduce nella dipendenza esplicita di questa da un certo insieme di parametri Θ. Ovvero, per ogni numero reale x,

$$\phi_X(x) = g(x|\Theta) \ . \tag{104}$$

Esempio 39 *L'unico esempio di questo caso trattato in questo corso, è la densità di probabilità gaussiana, ovvero*

$$g\left(x|\mu,\sigma\right) := \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \ , \tag{105}$$

dove μ è un numero reale e σ è un numero reale strettamente positivo, cioè $\sigma > 0$. Inoltre valgono le seguenti identità

$$\mu = \int_{\mathbb{R}} dx x g\left(x|\mu,\sigma\right) = \int_{\mathbb{R}} dx x \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \tag{106}$$

$$\begin{aligned} \sigma^2 &= \int_{\mathbb{R}} dx \left(x-\mu\right)^2 g\left(x|\mu,\sigma\right) = \int_{\mathbb{R}} dx \left(x-\mu\right)^2 \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \\ &= \int_{\mathbb{R}} dx x^2 \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} - \mu^2 \end{aligned} \tag{107}$$

(cfr [R] fig. 5.7 pag. 171 e [T] figg. 5.11 e 5.12). Pertanto, se le misure di X hanno densità di probabilità gaussiana, esistono due particolari valori delle costanti μ e σ, legati alla natura della quantità X, tali che $\phi_X = g\left(\cdot|\mu,\sigma\right)$. Dunque, nel caso gaussiano, $\Theta = \{\mu,\sigma\}$.

Se la densità di probabilità delle misure di X non esiste, quanto sopra affermato circa l'insieme di parametri Θ che specifica la dipendenza da X vale per il limite delle frequenze relative campionarie di un intervallo qualsiasi di valori delle misure di X, $\mathbb{P}_X$. Nel seguito comunque ci restringeremo al caso in cui ϕ_X esiste.

Osservazione 40 *Notiamo che, nel caso esista ϕ_X, dalla (47) e dalla (97) si ha che, per un campione $\mathcal{C}_N\left(X\right)$ molto numeroso di misure di X,*

$$\bar{x} = \sum_{k\in\mathbb{Z}} \frac{\left|\mathcal{B}_N^k\left(X\right)\right|}{N} \bar{x}_k = \sum_{k\in\mathbb{Z}} \Delta_N \varphi_k^{(N)} \bar{x}_k \ . \tag{108}$$

Ma da questa espressione segue che $\bar{x}$ dipende, com'è ovvio, dalla numerosità del campione N. Allora, esplicitando questa dipendenza e ponendo dunque $\bar{x} = \bar{x}_N$, passando al limite per N che tende all'infinito si può dimostrare che la successione numerica $\left\{\bar{x}_N\right\}_{N\in\mathbb{N}^+}$ converge a $\int_{\mathbb{R}} dx x \phi_X\left(x\right)$, che, nel caso gaussiano, coincide esattamente col membro destro della (106). Allo stesso modo, usando la (65), la (63) e la (97), si può dimostrare che, nello stesso limite, la successione delle varianze campionarie converge a $\int_{\mathbb{R}} dx x^2 \phi_X\left(x\right) - \left(\int_{\mathbb{R}} dx x \phi_X\left(x\right)\right)^2$, coincidendo questa espressione, nel caso gaussiano, esattamente col membro destro della (107).

0.2.1 Metodo della massima verosimiglianza

Le caratteristiche qualitative dell'istogramma delle frequenze relative di un campione $\mathcal{C}_N(X)$ di misure di numerosità finita di una quantità X, possono suggerire quale forma può avere il grafico della densità di probabilità delle misure di X. Per esempio, se il campione $\mathcal{C}_N(X)$ è normale, ci si può aspettare che ϕ_X sia gaussiana e dunque, per la (102), la (103) e la (105), approssimare, per un qualsiasi intervallo $[a,b]$ di valori delle misure di X, $\frac{|\{x\in\mathcal{C}_N(X):x\in[a,b]\}|}{N}$ con $\int_a^b dx g(x|\mu,\sigma)$ a meno di un errore tanto più piccolo quanto più grande è il valore di N.

Una volta appurato che la distribuzione di probabilità ϕ_X delle misure di X esiste, possiamo dire che, se ripetessimo infinite misure di X, la frequenza relativa campionaria dei valori compresi nell'intervallo $(x, x+dx]$, cioè la probabilità che questi ricadano all'interno di tale intervallo, sarebbe proprio $\phi_X(x)\,dx$. Quindi è lecito chiedersi quale sarebbe, nel limite di infinite misure, la probabilità di osservare un campione $\mathcal{C}_N(X)=\{x_1,..,x_N\}$ di misure di X di numerosità N finita.

Se i dati del campione $\mathcal{C}_N(X)$ sono frutto di misure sperimentali della quantità X in modo tale che, all'atto di effettuare ciascuna di queste misure, le condizioni sperimentali risultano identiche per ciascuna di esse, si può applicare lo stesso ragionamento intuitivo che porta a considerare, per esempio, che la probabilità di ottenere una collezione di N numeri $\{n_1,..,n_N\}$ estratti lanciando un dado a 6 facce, e quindi il cui valore n_i varia da 1 a 6 per ogni $i=1,..,N$, sia $p_{n_1}\cdot\cdot\,p_{n_N}$, dove p_k è la probabilità di ottenere il numero k. Allora, la probabilità di osservare la collezione di misure $\{x_1,..,x_N\}$ che compongono il campione $\mathcal{C}_N(X)$, e che, per brevità di notazione, indichiamo col simbolo $\mathbb{P}_X(\mathcal{C}_N(X))$, per la (104) risulta

$$\begin{aligned}\mathbb{P}_X(\mathcal{C}_N(X)) &= \mathbb{P}_X(x_1,x_1+dx_1]\times\cdot\cdot\times\mathbb{P}_X(x_N,x_N+dx_N] \qquad (109)\\ &= \phi_X(x_1)\,dx_1\times\cdot\cdot\times\phi_X(x_N)\,dx_N=\prod_{i=1}^{N}g(x_i|\Theta)\,dx_i\ .\end{aligned}$$

Sfortunatamente, i parametri che compongono l'insieme Θ e che identificano la densità di probabilità delle misure di X tra tutte quelle con le stesse caratteristiche restano incogniti, pertanto vanno stimati a partire dai dati che compongono il campione di misure $\mathcal{C}_N(X)$ in nostro possesso.

Una strategia plausibile per stimare tali parametri è quella di rimpiazzarli con opportune funzioni dei dati del campione che abbiamo a disposizione in modo che, inseriti nella precedente espressione di $\mathbb{P}_X(\mathcal{C}_N(X))$, questa risulti la più grande possibile. Ciò è equivalente a calcolare il massimo di $\prod_{i=1}^{N}g(x_i|\Theta)$

come funzione degli elementi di Θ, che è nota col nome di *verosimiglianza* del campione $\mathcal{C}_N(X)$. Perciò, se fissato un qualsiasi x la funzione $g(x|\Theta)$ è differenziabile come funzione degli elementi di Θ, nel caso in cui per esempio $\Theta = \{\theta_1, \theta_2\}$, e quindi

$$\prod_{i=1}^{N} g(x_i|\Theta) = \prod_{i=1}^{N} g(x_i|\theta_1, \theta_2), \tag{110}$$

è necessario:

1. calcolare le derivate parziali in θ_1 e θ_2 della funzione verosimiglianza e porle uguali a zero;

2. ricavare dalla coppia di equazioni

$$\begin{cases} \frac{\partial}{\partial \theta_1} \prod_{i=1}^{N} g(x_i|\theta_1, \theta_2) = 0 \\ \frac{\partial}{\partial \theta_2} \prod_{i=1}^{N} g(x_i|\theta_1, \theta_2) = 0 \end{cases} \tag{111}$$

   delle espressioni $\check{\theta}_1(x_1, .., x_N), \check{\theta}_2(x_1, .., x_N)$ di θ_1 e θ_2 in funzione dei dati del campione;

3. selezionare, tra le coppie di valori $\left(\check{\theta}_1, \check{\theta}_2\right)$ ottenuti al punto precedente, quella (spesso nella pratica se ne trova solo una) che rappresenta un punto di massimo funzione verosimiglianza.

Osservazione 41 *Poiché la funzione* $\log x$ *è monotona crescente, piuttosto che procedere alla ricerca del massimo della funzione verosimiglianza è più comodo provare a calcolare quello del suo logaritmo, dato che la monotonia della funzione* $\log x$ *garantisce la coincidenza dei punti di massimo di entrambe le funzioni.*

Il metodo appena descritto per stimare i valori incogniti degli elementi di Θ che specificano la distribuzione di probabilità delle misure di una quantità X, se esiste, in funzione dei dati di un campione $\mathcal{C}_N(X)$ di misure di X è detto *principio di massima verosimiglianza.*

Caso della distribuzione gaussiana: metodo dei minimi quadrati

Trattiamo ora esplicitamente il caso in cui la densità di probabilità delle misure di una quantità X sia gaussiana di parametri μ e σ. Dalla (105) la funzione verosimiglianza di un campione $\mathcal{C}_N(X)$ di misure di X risulta

$$\prod_{i=1}^{N} g(x_i|\mu,\sigma) = \prod_{i=1}^{N} \frac{e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} = \frac{e^{-\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}}}{(2\pi)^{\frac{N}{2}}\,\sigma^N} \ . \tag{112}$$

Come sottolineato alla fine della sezione precedente, conviene cercare il massimo del logaritmo della funzione verosimiglianza,

$$\log \prod_{i=1}^{N} g(x_i|\mu,\sigma) = -\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} - N\log\sigma - \frac{N}{2}\log(2\pi) \tag{113}$$

dato che questo è raggiunto nello stesso punto in cui è raggiunto il massimo della verosimiglianza. Ma il membro a destra dell'uguale nella precedente uguaglianza è massimo quando la funzione

$$h_N(\mu,\sigma) = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} + N\log\sigma \tag{114}$$

assume il suo valore minimo. Pertanto, ponendo uguali a zero le espressioni delle derivate parziali in μ e σ si ottiene il sistema di equazioni

$$\begin{cases} \frac{\partial}{\partial\mu} h_N(\mu,\sigma) = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = 0 \\ \frac{\partial}{\partial\sigma} h_N(\mu,\sigma) = -\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^3} + \frac{N}{\sigma} = 0 \end{cases} \tag{115}$$

da cui segue che le stime di masima verosimiglianza $\check{\mu}(x_1,..,x_N)$, $\check{\sigma}(x_1,..,x_N)$ dei parametri μ e σ risultano

$$\begin{cases} \check{\mu}(x_1,..,x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x} \\ \check{\sigma}^2(x_1,..,x_N) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\check{\mu})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 = \bar{s}_X^2 \end{cases} \ , \tag{116}$$

ovvero la media campionaria e lo scarto quadratico medio.

Poiché per calcolare gli stimatori di massima verosimiglianza di μ e σ si è cercato il minimo della funzione $h_N(\mu,\sigma)$, nel caso gaussiano il principio di massima verosimiglianza prende il nome di *principio dei minimi quadrati.*

Regressione lineare univariata

Dato un campione $\mathcal{C}_N(X,Y)$ bivariato di misure delle quantità X e Y, supponiamo che il diagramma di dispersione associato risulti approssimativamente lineare o, ciò che risulta quantitativamente più significativo, che il cofficiente di correlazione campionaria $r_{X,Y}$ delle misure delle due quantità risulti in modulo approssimativamente pari al valore 1, allora è ragionevole aspettarsi che le due quantità risultino legate da una relazione funzionale lineare, ad esempio $Y = AX + B$.

Poiché le costanti A e B che compaiono nell'espressione di Y come funzione di X sono incognite, è necessario stimarle in funzione dei dati del campione $\mathcal{C}_N(X,Y)$. A questo scopo, consideriamo l'errore quadratico medio che si è commesso avendo misurato per Y durante l'i-simo esperimento, per $i = 1, .., N$, il valore y_i invece che il valore $Ax_i + B$, ovvero

$$e_N^2(A,B) := \frac{1}{N}\sum_{i=1}^{N}\left[y_i - (Ax_i + B)\right]^2 \ . \tag{117}$$

Una stima ragionevole di A e B è quella che si ottiene minimizzando la quantità (117) come funzione di questi due parametri. A tal fine è necessario e sufficiente porre uguali a zero le derivate parziali di e_N^2 ottenendo così un sistema di equazioni in A e B che sono dette *equazioni normali*:

$$\begin{cases} \frac{\partial}{\partial A}e_N^2(A,B) = \frac{2}{N}\sum_{i=1}^{N}(y_i - Ax_i - B)\,x_i = 0 \\ \frac{\partial}{\partial B}e_N^2(A,B) = \frac{2}{N}\sum_{i=1}^{N}(y_i - Ax_i - B) = 0 \end{cases} . \tag{118}$$

Ricordando le definizioni di media campionaria (46), varianza campionaria (58) e di covarianza campionaria (89), si ottiene

$$\begin{cases} \frac{\partial}{\partial A}e_N^2(A,B) = \frac{1}{N}\sum_{i=1}^{N}(y_i x_i - Ax_i^2) - B\bar{x} = 0 \\ \frac{\partial}{\partial B}e_N^2(A,B) = \bar{y} - A\bar{x} - B = 0 \end{cases} \Longrightarrow \tag{119}$$

$$\begin{cases} \frac{1}{N}\sum_{i=1}^{N}(y_i x_i - Ax_i^2) - \bar{y}\bar{x} + A\bar{x}^2 = 0 \\ B = \bar{y} - A\bar{x} \end{cases} \Longrightarrow$$

$$\begin{cases} A\sum_{i=1}^{N}x_i^2 - N\bar{x}^2 = \sum_{i=1}^{N}y_i x_i - N\bar{y}\bar{x} \\ B = \bar{y} - A\bar{x} \end{cases} \Longrightarrow$$

$$\begin{cases} A\frac{\sum_{i=1}^{N}x_i^2 - N\bar{x}^2}{N-1} = \frac{\sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})}{N-1} \\ B = \bar{y} - A\bar{x} \end{cases} ,$$

da cui segue che le stime $\check{A}(x_1, .., x_N; y_1, ..y_N)$, $\check{B}(x_1, .., x_N; y_1, ..y_N)$ dei parametri A e B così calcolate risultano

$$\begin{cases} \check{A}(x_1, .., x_N; y_1, ..y_N) = \frac{s_{X,Y}}{s_X^2} \\ \check{B}(x_1, .., x_N; y_1, ..y_N) = \bar{y} - \frac{s_{X,Y}}{s_X^2}\bar{x} \end{cases} . \tag{120}$$

Osservazione 42 *Notiamo che lo stesso risultato si sarebbe raggiunto applicando il metodo della massima verosimiglianza sotto l'ipotesi che la densità di probabilità di X fosse gaussiana. Infatti, poiché si può dimostrare che se una quantità X ha densità di probabilità gaussiana allora una quantità Y funzione lineare di X ha anch'essa densità di probabilità gaussiana, in tal caso, la funzione verosimiglianza del sottocampione $\mathcal{C}_N(Y)$ di $\mathcal{C}_N(X,Y)$ risulterebbe*

$$\prod_{i=1}^{N} g(y_i|\Theta) = \prod_{i=1}^{N} g(y_i|\mu_i,\sigma) = \prod_{i=1}^{N} g(y_i|Ax_i+B,\sigma) = \frac{e^{-\sum_{i=1}^{N}\frac{[y_i-(Ax_i+B)]^2}{2\sigma^2}}}{(2\pi)^{\frac{N}{2}}\,\sigma^N}\ , \tag{121}$$

dove il parametro σ è considerato fissato.

Sottolineiamo che, fissato σ, $\Theta = \{A,B\}$ poiché, per ogni $i = 1,..,N$, i valori di μ_i, una volta noti i valori dei dati del sottocampione $\mathcal{C}_N(X)$ di $\mathcal{C}_N(X,Y)$, risultano funzioni soltanto dei parametri A e B.

Pertanto, poiché nel caso gaussiano il principio di massima verosimiglianza coincide con quello dei minimi quadrati, anche in questo caso è necessario e sufficiente porre uguali a zero le derivate parziali in A e B della funzione

$$\bar{h}_{N,\sigma}(A,B) := \sum_{i=1}^{N} \frac{[y_i-(Ax_i+B)]^2}{2\sigma^2} \tag{122}$$

e risolvere il sistema di equazioni che ne deriva, ovvero

$$\left\{\begin{array}{l} \frac{\partial}{\partial A}\bar{h}_{N,\sigma}(A,B) = \sum_{i=1}^{N}\frac{[y_i-(Ax_i+B)]x_i}{\sigma^2} = 0 \\ \frac{\partial}{\partial B}\bar{h}_{N,\sigma}(A,B) = \sum_{i=1}^{N}\frac{[y_i-(Ax_i+B)]}{\sigma^2} = 0 \end{array}\right. \Longrightarrow \tag{123}$$

$$\left\{\begin{array}{l} \sum_{i=1}^{N}[y_i-(Ax_i+B)]\,x_i = 0 \\ \sum_{i=1}^{N}[y_i-(Ax_i+B)] = 0 \end{array}\right.$$

la cui soluzione è la coppia di funzioni dei dati di $\mathcal{C}_N(X,Y)$, A,B (120) calcolate precedentemente.

Media pesata

Supponiamo che due gruppi di ricerca, che considereremo distinti dal valore dell'indice $i = 1,2$, i quali investigano sullo stesso fenomeno, producano ciascuno un campione di misure $\mathcal{C}_{N_i}(X)$ di una quantità X e le relative statistiche: media campionaria $\bar{x}_i$ e deviazione standard campionaria $s_{X,i}$. Sorge allora il problema di combinare le due stime $\bar{x}_1$ e $\bar{x}_2$ in modo da trovare la miglior stima del valore di X. Chiaramente tale stima deve avere un valore

che interpola tra i due in modo tale che, nel caso in cui si disponga solo del valore $\bar{x}_1$ la miglior stima del valore di X risulti proprio $\bar{x}_1$, viceversa, nel caso in cui si disponga solo del valore $\bar{x}_2$ la miglior stima del valore di X risulti proprio $\bar{x}_2$. Costruiamo allora la funzione

$$[0,1] \ni \lambda \longmapsto \bar{x}(\lambda) := \lambda \bar{x}_2 + (1-\lambda)\bar{x}_1 \in [\bar{x}_1, \bar{x}_2] \tag{124}$$

il cui valore, per ogni $\lambda \in [0,1]$, interpola tra i valori assunti da $\bar{x}_1$ e $\bar{x}_2$ coincidendo con $\bar{x}_1$ quando $\lambda = 0$ e con $\bar{x}_2$ quando $\lambda = 1$. Una stima ragionevole del valore di X, dati $\bar{x}_1$ e $\bar{x}_2$, è allora quella che si ottiene minimizzando la funzione

$$\bar{e}_X^2(\lambda) := \lambda^2 s_{X,2}^2 + (1-\lambda)^2 s_{X,1}^2 \tag{125}$$

che rappresenta una stima dell'errore quadratico medio che si commetterebbe nell'assumere come valore vero per X proprio $\bar{x}(\lambda)$. A tal fine è necessario e sufficiente porre uguale a zero la derivata prima di $\bar{e}_X^2$

$$\frac{d}{d\lambda}\bar{e}_X^2(\lambda) = 2\left(\lambda s_{X,2}^2 - (1-\lambda)s_{X,1}^2\right) = 0 \tag{126}$$

da cui si ricava il valore $\check{\lambda}$ che minimiza $\bar{e}_X^2$, ovvero

$$\check{\lambda} = \frac{s_{X,1}^2}{s_{X,1}^2 + s_{X,2}^2} = \frac{\frac{1}{s_{X,2}^2}}{\frac{1}{s_{X,2}^2} + \frac{1}{s_{X,1}^2}} \tag{127}$$

perciò la miglior stima di X risulta

$$\breve{x} = \bar{x}\left(\check{\lambda}\right) = \frac{\frac{1}{s_{X,2}^2}\bar{x}_2 + \frac{1}{s_{X,1}^2}\bar{x}_1}{\frac{1}{s_{X,2}^2} + \frac{1}{s_{X,1}^2}} \tag{128}$$

che è un valore medio fra i due valori disponibili $\bar{x}_1$ e $\bar{x}_2$ restituiti dall'analisi statistica sui campioni $\mathcal{C}_{N_1}(X), \mathcal{C}_{N_2}(X)$, ciascuno pesato con l'inverso della varianza capionaria del campione d'appartenenza e quindi, in definitiva, con l'inverso della sua incertezza.

Questo procedimento si può generalizzare al caso in cui si disponga di un numero arbitrario M di campioni di misure di X, $\{\mathcal{C}_{N_i}\}_{i=1,..,M}$ e relative statistiche $\{\bar{x}_i, s_{X,i}\}_{i=1,..,M}$. Allora, per ogni collezione di M numeri reali positivi $\{\lambda_i\}_{i=1,..,M}$ tali che $\sum_{i=1}^M \lambda_i = 1$, il valore interpolante le M medie campionarie $\{\bar{x}_i\}_{i=1,..,M}$ risulta

$$\bar{x}(\lambda_1, .., \lambda_M) := \sum_{i=1}^M \lambda_i \bar{x}_i \tag{129}$$

e la stima dell'errore quadratico medio da minimizzare è

$$\bar{e}_X^2\left(\lambda_1,..,\lambda_M\right) := \sum_{i=1}^{M} \lambda_i^2 s_{X,i}^2 \tag{130}$$

col vincolo che $\sum_{i=1}^{M} \lambda_i = 1$. Procedendo in questo modo si ottiene come punto di minimo per $\bar{e}_X^2$ il punto $\left(\frac{s_{X,1}^{-2}}{\sum_{i=1}^{M} s_{X,i}^{-2}},..,\frac{s_{X,M}^{-2}}{\sum_{i=1}^{M} s_{X,i}^{-2}}\right)$ e dunque come miglior stima del valore di X,

$$\breve{x} = \frac{\sum_{i=1}^{M} \frac{1}{s_{X,i}^2}\bar{x}_i}{\sum_{i=1}^{M} \frac{1}{s_{X,i}^2}} \ . \tag{131}$$

Osservazione 43 *Anche in questo caso, come in quello relativo al calcolo delle stime dei parametri di regressione lineare, si sarebbe raggiunto lo stesso risultato utilizzando il metodo della massima verosimiglianza sotto l'ipotesi di gaussianità dei campioni* $\{\mathcal{C}_{N_i}\}_{i=1,..,M}$. *Infatti, in tal caso, per ogni* $i = 1,..M$, *la probabilità di osservare una misura di* X *pari a* $\bar{x}_i$, *per la (104) e la (105), risulterebbe*

$$g\left(\bar{x}_i|\mu,\sigma_i\right) = \frac{e^{-\frac{\left(\bar{x}_i-\mu\right)^2}{2\sigma_i^2}}}{\sqrt{2\pi}\sigma_i} d\bar{x}_i \ . \tag{132}$$

Perciò, la probabilità di osservare le misure di X *rappresentate dai dati del campione composto dalle medie empiriche* $\{\bar{x}_1,..,\bar{x}_M\}$ *dei campioni* $\mathcal{C}_{N_1},..,\mathcal{C}_{N_M}$, *per la (109) risulterebbe*

$$\mathbb{P}_X\{\bar{x}_1,..,\bar{x}_M\} = \prod_{i=1}^{M} g\left(\bar{x}_i|\mu,\sigma_i\right) d\bar{x}_i = \prod_{i=1}^{M} \frac{e^{-\frac{\left(\bar{x}_i-\mu\right)^2}{2\sigma_i^2}}}{\sqrt{2\pi}\sigma_i} d\bar{x}_i \ . \tag{133}$$

Sostituendo quindi ai parametri σ_i *quelli stimati* $s_{X,i}$, *la funzione verosimiglianza del campione* $\{\bar{x}_1,..,\bar{x}_M\}$,

$$\prod_{i=1}^{M} g\left(\bar{x}_i|\mu,s_{X,i}\right) = \prod_{i=1}^{M} \frac{e^{-\frac{\left(\bar{x}_i-\mu\right)^2}{2s_{X,i}^2}}}{\sqrt{2\pi}s_{X,i}} = \frac{e^{-\sum_{i=1}^{M}\frac{\left(\bar{x}_i-\mu\right)^2}{2s_{X,i}^2}}}{\left(2\pi\right)^{\frac{M}{2}}\prod_{i=1}^{M} s_{X,i}} \ , \tag{134}$$

sarebbe dunque funzione eslusivamente del parametro μ. *Massimizzare la funzione verosimiglianza, o, equivalentemente, applicare il metodo dei minimi*

quadrati, si ridurrebbe quindi a calcolare l'unico zero $\check{\mu}$ *della derivata in* μ *della funzione* $\sum_{i=1}^{M} \frac{(\bar{x}_i-\mu)^2}{2s_{X,i}^2}$, *ovvero*

$$\sum_{i=1}^{M} \frac{(\bar{x}_i - \mu)}{s_{X,i}^2} = 0 \tag{135}$$

che restituisce per $\check{\mu}$ *proprio il valore (131).*

0.2.2 Test del χ^2 (chi quadro) semplificato: caso gaussiano

Fin'ora abbiamo supposto che le misure della quantità X oggetto della nostra indagine sperimentale avessero una certa densità di probabilità descritta da parametri incogniti, il cui valore è stato stimato a partire dai dati del campione $\mathcal{C}_N(X)$ di misure di X, prodotto dalla nostra indagine sperimentale, tramite il metodo della massima verosimiglianza.

Diamo ora, nel caso gaussiano, un criterio utile a discernere se il campione di misure $\mathcal{C}_N(X)$ in nostro posseso sia compatibile o meno con l'ipotesi di gaussianità delle misure della quantità X oggetto del nostro studio, avendo assunto come parametri della densità di probabilità $\mu = \bar{x}$ e $\sigma = s_X$, cioè proprio la media campionaria e la deviazione standard campionaria calcolate dai dati di $\mathcal{C}_N(X)$.

Suddividiamo la retta reale $\mathbb{R}$ in $2(K+1)$ intervalli disgiunti ponendo:

- $\mathbb{R} = (-\infty, \bar{x}] \cup (\bar{x}, +\infty)$;
- per ogni $k = 0, .., K-1$,

$$A_{k+1} \;:=\; (\bar{x} + \alpha_k s_X, \bar{x} + \alpha_{k+1} s_X] \ , \tag{136}$$
$$A_{-(k+1)} \;:=\; (\bar{x} - \alpha_{k+1} s_X, \bar{x} - \alpha_k s_X] \ , \tag{137}$$

  dove $\{\alpha_0, .., \alpha_K\}$ è una collezione di $K+1$ numeri reali positivi tali che $\alpha_0 = 0$ e, per ogni $k = 0, .., K-1$, $\alpha_k < \alpha_{k+1}$;

- $A_{K+1} := (\bar{x} + \alpha_K s_X, +\infty)$, $A_{-(K+1)} := (-\infty, \bar{x} - \alpha_K s_X)$.

Perciò,

$$(\bar{x}, +\infty) = \bigcup_{k=0}^{K-1} A_{k+1} \bigcup A_{K+1} \ , \ (-\infty, \bar{x}] = \bigcup_{k=0}^{K-1} A_{-(k+1)} \bigcup A_{-(K+1)} \ . \tag{138}$$

Siano allora, per ogni $l \in \{-(K+1), .., -1\} \cup \{1, .., K+1\}$,

$$O_l := |\{x \in \mathcal{C}_N(X) : x \in A_l\}| = \sum_{i=1}^{N} \mathbf{1}_{A_l}(x_i) \tag{139}$$

le frequenze dei sottocampioni di $\mathcal{C}_N(X)$ associati agli elementi della partizione $\{A_k\}_{k \in \{-(K+1), .., -1\} \cup \{1, .., K+1\}}$ di $\mathbb{R}$, cui diamo il nome di *occorrenze.*

Definiamo inoltre, per ogni $k \in \{-(K+1), .., -1\} \cup \{1, .., K+1\}$ le quantità

$$E_k := N \int_{A_k} dx g(x|\bar{x}, s_X) \ . \tag{140}$$

Osservazione 44 *Notiamo che, siccome la densità di probabilità gaussiana* $g(\cdot|\bar{x}, s_X)$ *è simmetrica rispetto alla retta parallela all'asse delle ordinate passante per il punto del piano cartesiano di coordinate* $(\bar{x}, 0)$, *ovvero, per ogni* $x \in \mathbb{R}$,

$$g(x - \bar{x}|\bar{x}, s_X) = g(-(x - \bar{x})|\bar{x}, s_X) \ , \tag{141}$$

per ogni $k = 1, .., K+1$, $E_k = E_{-k}$ *e quindi è sufficiente calcolare solo i valori di* E_k *per valori positivi dell'indice* k.

Poiché se la numerosità N dei campioni di misure di X che la nostra indagine sperimentale può restituirci è così grande tanto poter considerare valori di K anch'essi molto grandi, sotto questa ipotesi si può dimostrare che, per ogni $k \in \{-(K+1), .., -1\} \cup \{1, .., K+1\}$, la densità di probabilità della quantità O_k risulta essere concentrata sull'insieme dei numeri naturali $\mathbb{N}$ e tale che il limite della media campionaria di O_k, per valori di N molto, grandi risulti pari a E_k, come pure quello della sua varianza campionaria.

Allora, se K è scelto rispettando i seguenti criteri:

1. $K \geq 1$ e molto minore di N;

2. il valore delle occorrenze deve essere sempre positivo.

e la densità di probabilità di X è supposta gaussiana di parametri $\mu = \bar{x}$ e $\sigma = s_X$, è possibile confrontare i valori delle occorrenze O_k con i valori E_k. Definiamo quindi la quantità

$$\bar{\chi}^2 := \frac{1}{2(K+1) - 3} \sum_{k=1}^{K} \left[\frac{(O_k - E_k)^2}{E_k} + \frac{(O_{-k} - E_{-k})^2}{E_{-k}} \right] \tag{142}$$

detta *chi quadro normalizzato o ridotto.* Se la discrepanza tra il valore dell'occorrenza O_k ed il suo valore medio campionario limite E_k è molto piccola,

il valore di $\bar{\chi}^2$ sarà prosimo a 0 e l'ipotesi di gaussianità della densità di probabilità di X, con parametri $\mu = \bar{x}$ e $\sigma = s_X$, risuterà plausibile, altrimenti sarà falsa.

TEST (del chi quadro semplificato nel caso gaussiano) Dato un campione di numerosità N di misure della quantità X, $\mathcal{C}_N(X)$, formuliamo l'ipotesi che le misure costituenti i dati di $\mathcal{C}_N(X)$ abbiano densità di probabilità gaussiana di parametri μ pari alla media campionaria $\bar{x}$ e σ pari alla varianza campionaria s_X. Allora, l'ipotesi sarà da accettare se $\bar{\chi}^2 \leq 1$, altrimenti sarà rigettata.

(Per un esempio in cui $K = 1$ cfr [T] tabelle 12.2, 12.3, .12.4. Per un esempio in cui $K = 3$ cfr [T] tabella 12.7.)

Osservazione 45 *Sottolineiamo che il numero* $2(K+1) - 3$ *che compare al denominatore della definizione di* $\bar{\chi}^2$, *detto* numero di gradi di libertà di χ^2_K, *dove*

$$\chi^2_K := \sum_{k=1}^{K} \left[\frac{(O_k - E_k)^2}{E_k} + \frac{(O_{-k} - E_{-k})^2}{E_{-k}} \right] \tag{143}$$

è detto chi quadro, *rappresenta la differenza tra il numero d'insiemi in cui s'è ripartita la retta reale,* $2(K+1)$, *ed il numero di equazioni che legano le quantità stimate dai dati ed i parametri del problema. Nel caso gaussiano questi parametri sono: la numerosità del campione* N *ed i parametri che individuano la densità di probabilità delle misure di* X, *ovvero* μ *e* σ, *infatti*

$$\begin{cases} N = \sum_{k=1}^{K} (O_k + O_{-k}) \\ \mu = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \\ \sigma = s_X = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \end{cases} , \tag{144}$$

il che dimostra che il valore del numero di gradi di libertà di χ^2_K *è proprio* $2(K+1) - 3$. *Ciò implica che, dovendo questo numero essere maggiore o uguale ad uno,*

$$2(K+1) - 3 \geq 1 \Longrightarrow 2K \geq 2 \Longrightarrow K \geq 1 , \tag{145}$$

cioè K *non può essere pari a* 0 *ed il numero d'insiemi in cui ripartire* $\mathbb{R}$, $2(K+1)$, *deve essere maggiore o uguale a* 4. *Infatti, dividendo semplicemente in due la retta reale, cioè* $\mathbb{R} = (-\infty, \mu] \cup (\mu, +\infty)$, *con* $\mu = \bar{x}$, *poiché questa partizione è indipendente dal parametro* σ *che pure compare nella densità di probabilità delle misure di* X, *in definitiva non si terrebbe in considerazione questo parametro per discernere se c'è accordo o meno tra i dati del campione e l'ipotesi che questi, in quanto misure di* X, *abbiano densità di probabilità gaussiana di parametri* $\bar{x}$ *e* s_X. *Mancando questa informazione il risultato del test di cui sopra perderebbe di affidabilità.*

Bibliografia

[R] S. M. Ross *Probabilità e Statistica per l'ingegneria e le scienze* Apogeo (Milano) 2003.

[T] J. R. Taylor *INTRODUZIONE ALL'ANALISI DEGLI ERRORI Lo studio delle incertezze nelle misure fisiche. Seconda edizione* Zanichelli (Bologna) 1999.