Statistica di base

Luca Mari, versione 29.9.15

Contenuti

I principali concetti introdotti in questo capitolo

Moda

L'informazione contenuta in una distribuzione può essere sintetizzata, in particolare mediante tre indici "di posizione": la moda, la mediana, la media. Si chiama *moda* di una distribuzione la categoria (e non la frequenza) a cui corrisponde la frequenza (assoluta o relativa) massima della distribuzione

Per esempio, la moda della distribuzione di frequenze assolute:

$$\begin{bmatrix} C_j \\ f_j \end{bmatrix} = \begin{bmatrix} C_1 & C_2 & C_3 \\ 1 & 3 & 2 \end{bmatrix}$$

(tale dunque che il campione è costituito da 1 elemento nella categoria C_1, 3 elementi nella categoria C_2, 2 elemento nella categoria C_3) è la categoria C_2 (corrispondente al fatto che lo studente in questione ha preso più voti nella categoria $C_2 = \{22,...,25\}$ che nelle categorie $C_1 = \{18,...,21\}$ e $C_2 = \{26,...,30\}$.

Una distribuzione può avere più mode, e in tal caso si chiama *multimodale* (e *bimodale* nel caso particolare di due mode); altrimenti si chiama *unimodale*.

La moda viene individuata semplicemente a partire dalle frequenze della distribuzione, e quindi è sempre definita, anche nel caso in cui sull'insieme delle categorie non sia presente alcuna struttura algebrica (per esempio, data una popolazione costituita dai comuni di nascita di un insieme di persone, la moda può essere calcolata, e corrisponde al comune che ha la più elevata frequenza di nascita tra le persone in questione, benché sull'insieme dei comuni non siano definite relazioni empiricamente significative).

D'altra parte, sull'insieme delle categorie è spesso presente una struttura algebrica, e in particolare un ordine; in questi casi, proprio basandosi sull'ordinamento delle categorie è possibile adottare nella sintesi degli indici più informativi della moda.

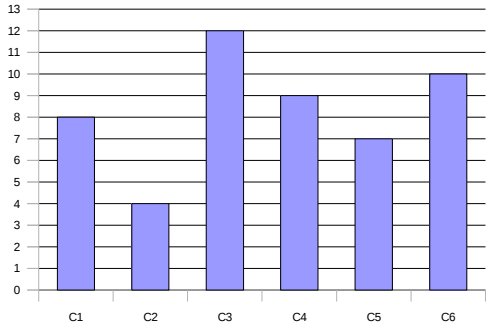
Distribuzioni cumulate

Supponiamo dunque che sull'insieme delle categorie sia presente una relazione d'ordine empiricamente significativa, $C_1<C_2<...$; è allora significativo contare il numero di elementi della successione *fino a* una data categoria inclusa. Per esempio, categorizzando i voti presi da 50 studenti mediante $C_1 = \{18,19\}$, $C_2 = \{20,21\}$, $C_3 = \{22,23\}$, $C_4 = \{24,25\}$, $C_5 = \{26,27\}$, $C_6 = \{28,30\}$, tali categorie sarebbero certamente ordinate, nel senso che un voto nella categoria C_j è migliore di uno della categoria C_i se $i<j$.

Supponiamo che la distribuzione dei voti sia:

$$\begin{bmatrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \\ 8 & 4 & 12 & 9 & 7 & 10 \end{bmatrix}$$

(e quindi la moda è la categoria C_3).

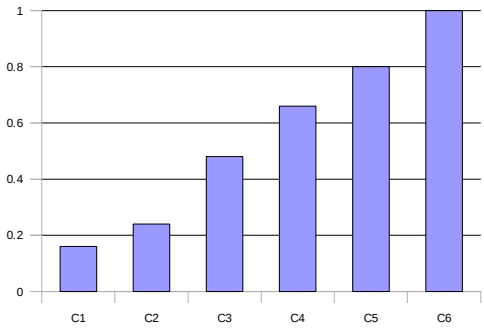


Grazie al fatto che le categorie sono ordinate, possiamo concludere che $n_1+n_2 = 8+4$ studenti hanno preso un voto fino a 21, $n_1+n_2+n_3 = 8+4+12$ un voto fino a 23 e così via. Si può allora costruire una nuova distribuzione, chiamata *cumulata*, in cui per ogni categoria si considera la frequenza dei voti presi fino a quella categoria:

$$\begin{bmatrix} C_j \\ n'_j \end{bmatrix} = \begin{bmatrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \\ 8 & 12 & 24 & 33 & 40 & 50 \end{bmatrix}$$

o in termini di frequenze relative:

$$\begin{bmatrix} C_j \\ f'_j \end{bmatrix} = \begin{bmatrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \\ 0,16 & 0,24 & 0,48 & 0,66 & 0,8 & 1 \end{bmatrix}$$



Data una distribuzione cumulata è facile ottenere la distribuzione di base, secondo la seguente logica: la frequenza (consideriamo il caso delle frequenze assolute; per le frequenze relative il discorso è lo stesso) per la prima categoria è la stessa, $n_1 = n'_1$; la frequenza per la seconda categoria è pari alla frequenza cumulata per la seconda categoria meno la frequenza cumulata per la prima categoria, $n_2 = n'_2 - n'_1$ (nell'esempio, il numero di studenti con voto 20 o 21 è pari a 12–8); in generale, dunque, $n_j = n'_j - n'_{j-1}$. Poiché, dato un campione e delle categorie, è spesso più immediato ricavarne la distribuzione cumulata, con questa semplice formula si può ottenere quindi la distribuzione di base.

Mediana, quartili, percentili

Si chiama *mediana* di una distribuzione la categoria che contiene il 50% della distribuzione cumulata. La mediana è cioè la "categoria centrale" della distribuzione una volta che i valori n_i sono stati ordinati in ordine crescente: il primo 50% della distribuzione sta entro la mediana, il secondo 50% sta oltre la mediana. Nell'esempio precedente:

$$\begin{bmatrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \\ 8 & 4 & 12 & 9 & 7 & 10 \end{bmatrix}$$

che contiene 50 voti, entro le prime 3 categorie sono contenuti 8+4+12 = 24 voti, quindi appena meno della metà: la mediana è la categoria C_4, come è chiaro anche dalla cumulata:

$$\begin{bmatrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \\ 0{,}16 & 0{,}24 & 0{,}48 & 0{,}66 & 0{,}8 & 1 \end{bmatrix}$$

Si noti perciò che la moda e la mediana (quando esiste) di una distribuzione non necessariamente coincidono. Quando la mediana è definita, si possono definire anche altri indici di posizione ordinale:

- i *quartili*: il primo / secondo / terzo quartile sono le categorie a cui corrisponde il primo 25% / 50% / 75% della distribuzione cumulata rispettivamente (ne segue che la mediana coincide con il secondo quartile); nell'esempio, i quartili sono rispettivamente C_3, C_4 e C_5; se il secondo quartile indica il "centro" della distribuzione, il primo e il terzo quartile forniscono un'informazione sulla dispersione della distribuzione stessa intorno a tale categoria centrale;
- i *percentili*: il primo / secondo / ... percentile è la categoria a cui corrisponde il primo 1% / 2% / ... della distribuzione cumulata (ne segue che la mediana coincide con il cinquantesimo percentile).

Significatività empirica degli indici ordinali

Mentre la moda di una distribuzione è sempre definita, la mediana è definita solo se sull'insieme delle categorie è presente una relazione d'ordine empiricamente significativa; ma quale criterio ci consente di stabilire se una relazione è, appunto, "empiricamente significativa"?

Consideriamo il caso di un esperimento in cui un usuale dado a 6 facce è stato lanciato 100 volte e i cui risultati sono stati sintetizzati in una distribuzione a 6 categorie, ognuna corrispondente a una faccia del dado, per esempio:

$$\begin{bmatrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \\ 18 & 11 & 22 & 15 & 22 & 12 \end{bmatrix}$$

Non ci sono problemi a identificare la moda: è la categoria corrispondente alla faccia che è uscita con maggiore frequenza (in questo caso le categorie sono due C_3 e C_5: la distribuzione è bimodale); nell'ipotesi che le categorie siano ordinate, $C_1 < C_2 < \ldots$, possiamo costruire la distribuzione cumulata:

$$\begin{bmatrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \\ 18 & 29 & 51 & 66 & 88 & 100 \end{bmatrix}$$

da cui si vede che la mediana è C_3. Ma queste categorie sono ordinate? Apparentemente sembrerebbe di sì, dato che il numero 1 che identifica una faccia è minore del numero 2 che identifica un'altra faccia, e così via. Ma supponiamo di coprire ogni faccia del dado con un'etichetta colorata, usando un colore diverso per ogni faccia, e quindi di effettuare l'esperimento, ottenendo la stessa distribuzione riportata sopra (in cui, per esempio, C_1 =etichetta rossa, C_2=etichetta verde, ...); le categorie ora non sono ordinate (rosso non è né minore né maggiore di verde) e quindi la distribuzione cumulata e la mediana non possono essere calcolate, benché la distribuzione di partenza sia la stessa.
L'applicabilità di un indice ordinale dipende dunque non dai simboli con cui si identificano le categorie della distribuzione ma dalla presenza di una relazione empirica di ordine tra le categorie.

Media

Insieme con moda e mediana, un terzo indice che sintetizza l'informazione di un campione $\langle x_i \rangle$ è la *media campionaria* $m(\langle x_i \rangle)$, definita come:

$$m(\langle x_i \rangle) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

La media è significativa solo se è empiricamente significativo sommare gli elementi della successione $\langle x_i \rangle$ (in effetti è sufficiente una condizione più debole – per esempio la media è significativa per valori di temperatura la cui somma pure non è empiricamente significativa – ma non approfondiremo questa distinzione qui).
Questa condizione è raramente soddisfatta su categorie costituite da più elementi dell'insieme supporto A – per esempio, se $C_1 = \{18,19\}$ e $C_2 = \{20,21\}$, come si calcola C_1+C_2? – e quindi la media è spesso calcolata sulla partizione più fine di A, in cui ogni categoria contiene un solo elemento (quindi in pratica su A stesso); data una distribuzione:

$$C = \begin{bmatrix} C_j \\ f_j \end{bmatrix}$$

con N categorie C_j, ognuna con frequenza assoluta n_j, la media è allora:

$$m(C) = \frac{1}{n} \sum_{j=1}^{N} C_j f_j$$

(attenzione: N è il numero delle categorie, n è il numero di elementi del campione) o anche, se si considera la distribuzione C_R delle frequenze relative r_j:

$$m(C_R) = \sum_{j=1}^{N} C_j r_j$$

Per esempio, la media del campione di voti:

$$\langle x_i \rangle = \langle 26,24,30,24,21,25,18,25,25,30 \rangle$$

si può calcolare direttamente sul campione:

$$m(\langle x_i \rangle) = \frac{1}{10}(26+24+30+...)$$

oppure a partire dalla distribuzione di frequenze assolute:

$$C = \begin{bmatrix} C_j \\ f_j \end{bmatrix} = \begin{bmatrix} 18 & 21 & 24 & 25 & 26 & 30 \\ 1 & 1 & 2 & 3 & 1 & 2 \end{bmatrix}$$

(tralasciando di indicare le categorie a frequenza nulla):

$$m(C) = \frac{1}{10}(18 \times 1 + 21 \times 1 + 24 \times 2 + ...)$$

oppure ancora a partire dalla distribuzione di frequenze relative:

$$C_R = \begin{bmatrix} C_j \\ r_j \end{bmatrix} = \begin{bmatrix} 18 & 21 & 24 & 25 & 26 & 30 \\ 0,1 & 0,1 & 0,2 & 0,3 & 0,1 & 0,2 \end{bmatrix}$$

si ha che:

$$m(C_R) = 18 \times 0,1 + 21 \times 0,1 + 24 \times 0,2 + ...$$

Ancora sulla media

La statistica:

$$m(\langle x_i \rangle) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

si chiama *media aritmetica*. Una versione più generale della media aritmetica è la *media pesata*:

$$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

scritta più semplicemente come:

$$\sum_{i=1}^{n} w_i x_i$$

nel caso in cui:

$$\sum_{i=1}^{n} w_i = 1$$

introdotta appunto per poter "pesare" in modo diverso, attraverso il vettore di pesi $\langle w_i \rangle$, il contributo alla media dei vari termini x_i. La media sulla distribuzione di frequenze assolute:

$$C = \begin{bmatrix} C_j \\ f_j \end{bmatrix}$$

è calcolata dunque come:

$$m(C) = \frac{1}{n}\sum_{j=1}^{N} C_j f_j$$

(si noti il coefficiente di normalizzazione: è uguale al numero degli elementi del campione, e non al numero di categorie della distribuzione!), mentre la media sulla distribuzione di frequenze relative:

$$C_R = \begin{bmatrix} C_j \\ r_j \end{bmatrix}$$

è pari a:

$$m(C_R) = \sum_{j=1}^{N} C_j r_j$$

Come si vede, tali medie possono essere interpretate come medie sulle categorie pesate mediante le frequenze, assolute o relative.

La media è un *operatore interno*, cioè è sempre maggiore o uguale del valore della categoria minima e sempre minore o uguale del valore della categoria massima:

$$C_1 \le m(C) \le C_N$$

Ciò fornisce un criterio di validazione (nella forma di condizione necessaria) del calcolo della media. Se la distribuzione è simmetrica, vale inoltre che mediana e media sono uguali e coincidono con la categoria centrale.

Il confronto tra mediana e media è particolarmente interessante proprio in condizioni di asimmetria della distribuzione: mentre per definizione la mediana divide il campione su cui la distribuzione è costruita in due parti della stessa numerosità, può accadere che la gran parte degli elementi del campione sia sopra, o sotto, la media. Si consideri per esempio il caso di un campione di 100 elementi, 99 dei quali a valore 2 e uno solo a valore 1. La media è dunque 1,99 e perciò il 99% degli elementi del campione è sopra la media! La conseguenza è perciò evidente: se si è interessati a una statistica che divida il campione in parti uguali, occorre impiegare la mediana, e non la media.

A una conclusione analoga si giunge prendendo in considerazione un caso complementare, in cui in un campione di 100 elementi 99 hanno valore 1 e uno ha valore, per esempio, 1000000. La media è (1·99+1000000·1)/100=10000,99, dunque ben superiore al valore del 99% degli elementi del campione, mentre la mediana (e in effetti tutti i percentili fino al novantottesimo) rimane ancorata al valore 1.

Una nota sul calcolo della media

Dato un campione $\langle x_i \rangle$, si vede facilmente che la media del campione derivato $\langle ax_i+b \rangle$ (con a e b costanti, $a \neq 0$) si può calcolare indifferentemente come:

$$m(\langle a x_i + b \rangle) = \frac{1}{n}\sum_{i=1}^{n} (a x_i + b)$$

oppure come:

$$a\,m(\langle x_i\rangle)+b=\frac{a}{n}(\sum_{i=1}^{n} x_i)+b$$

La media soddisfa dunque la proprietà per cui $m(\langle a\,x_i+b\rangle)=a\,m(\langle x_i\rangle)+b$, cioè è *un operatore lineare*. Questa proprietà fornisce un utile strumento per semplificare il calcolo della media stessa.
Un esempio, nel caso semplice in cui si considera $a=1$ e quindi $m(\langle x_i\rangle)=m(\langle x_i-b\rangle)+b$: per calcolare la media del campione $\langle x_i\rangle=\langle 26,24,30,24,22,25\rangle$ si può calcolare dapprima la media del campione derivato $\langle x_i-24\rangle=\langle 2,0,6,0,-2,1\rangle=7/6$ (dunque assumendo $b=24$) e quindi ottenere la media del campione iniziale da $\langle x_i\rangle=7/6+24$, avendo con ciò applicato la proprietà precedente come segue: dati $a=1$ e $b=24$, abbiamo calcolato la media di $\langle x_i-24\rangle$, che sappiamo essere uguale alla media di $\langle x_i-24\rangle$; ma da $m(\langle x_i-24\rangle)=m(\langle x_i\rangle)-24$ si ottiene appunto che $m(\langle x_i\rangle)=m(\langle x_i-24\rangle)+24$.
Analogamente, si può dunque facilmente mostrare che la media di un campione somma di due campioni è uguale alla somma delle medie dei due campioni:

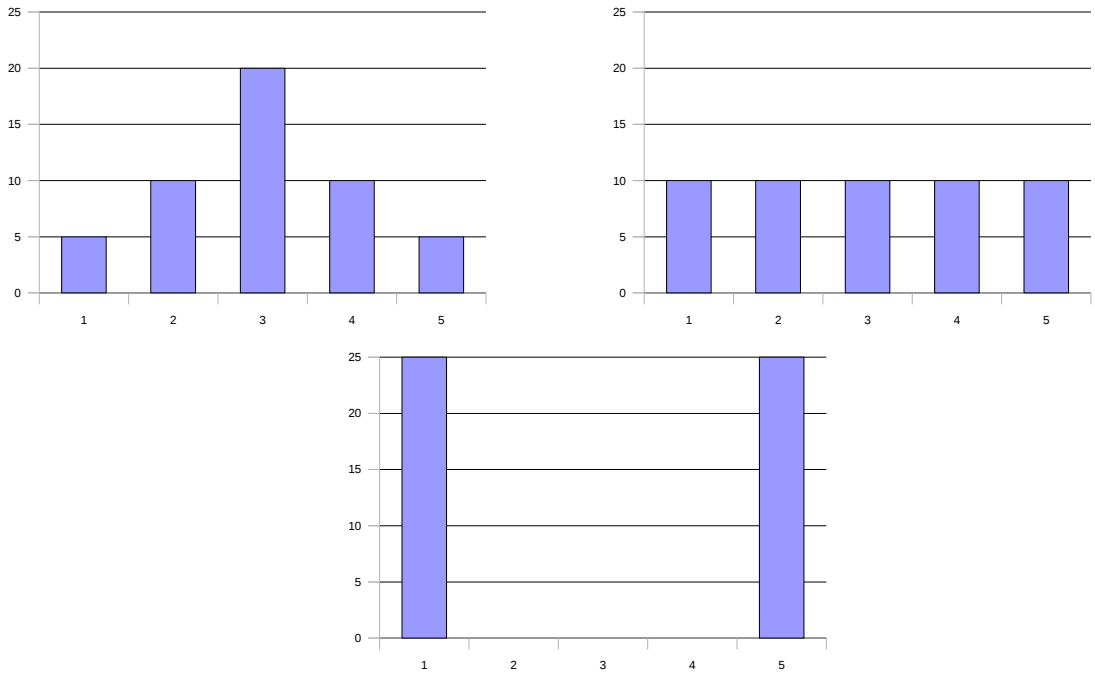
$m(\langle x_i+y_i\rangle) = m(\langle x_i\rangle)+m(\langle y_i\rangle)$

Statistiche di dispersione

Moda, mediana e media forniscono un'informazione, progressivamente sempre più specifica, sul "baricentro" della distribuzione, ma non indicano quanto i valori della distribuzione stessa sono dispersi intorno a tale baricentro:

- se tra le categorie della distribuzione non è definito nemmeno un ordine, l'unica informazione di dispersione che si può ottenere è il numero delle categorie che contengono almeno un elemento;
- se tra le categorie della distribuzione è definito un ordine ma non la somma, il primo e il terzo quartile indicano quanto la distribuzione è dispersa rispetto alla mediana, e le categorie minima e massima che contengono almeno un elemento forniscono un'indicazione della "dispersione complessiva" della distribuzione stessa.

Più interessante è il caso di distribuzioni per cui è empiricamente significativo calcolare la media
Per esempio le seguenti tre distribuzioni (ognuna di 50 elementi):



hanno la stessa media (la categoria 3) ma sono evidentemente molto diverse.

Varianza e deviazione standard

Dato che si può calcolare lo scarto $x_i - m$ di ogni elemento x_i dalla media m del campione $\langle x_i \rangle$ (per semplicità scriviamo $m_x = m(\langle x_i \rangle)$), come statistica di dispersione si potrebbe pensare di adottare la media degli scarti:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)$$

ma non è una buona idea, poiché scarti positivi e negativi si compensano. Infatti:

$$\sum_{i=1}^{n}(x_i - m_x) = \sum_{i=1}^{n} x_i - n\,m_x = \sum_{i=1}^{n} x_i - n\left(\sum_{i=1}^{n} x_i / n\right) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

Si potrebbero allora prendere in considerazione le distanze $|x_i - m_x|$, adottando quindi come statistica di dispersione la media delle distanze (e quindi la distanza media), cioè:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m_x|$$

ma non è la scelta abituale. Per varie ragioni (che non approfondiamo qui), gli scarti dalla media m_x si considerano in forma quadratica, $(x_i - m_x)^2$, e li si normalizza dividendo per $n-1$ invece che per n, ottenendo in questo modo:

$$s^2(\langle x_i \rangle) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2$$

chiamata *varianza campionaria* (nota: il fattore di normalizzazione è invece $1/n$ nel caso in cui si consideri che $\langle x_i \rangle$ è l'intera popolazione, invece di un suo campione). La varianza ha però il problema che non è dimensionalmente omogenea ai valori del campione (per esempio, se x_i, e quindi anche m_x, è misurato in metri, s^2 è misurato in metri quadrati). Per questo, invece della varianza si impiega tipicamente la sua radice quadrata:

$$s(\langle x_i \rangle) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2}$$

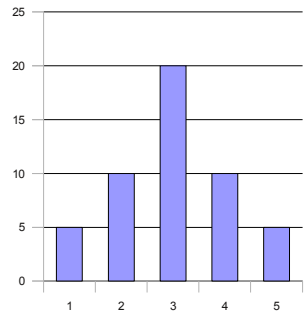
(per semplicità potremo indicare $s(\langle x_i \rangle)$ anche come s_x) chiamata *deviazione standard campionaria* (o anche *scarto tipo campionario*).

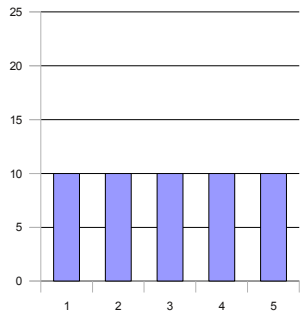
Deviazione standard

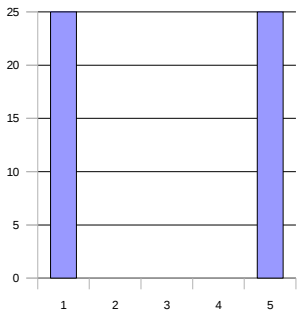
Nel caso in cui sia data una distribuzione C con N categorie C_j, ognuna con frequenza assoluta f_j, gli scarti quadratici rispetto alla media m_x sono $(C_j - m_x)^2$; la deviazione standard si calcola allora:

$$s(C) = \sqrt{\frac{1}{n-1}\sum_{j=1}^{N} f_j (C_j - m)^2}$$

Riprendendo l'esempio precedente, per le tre distribuzioni:







la deviazione standard è rispettivamente:

$$\sqrt{\frac{5\times 2^2 + 10\times 1^2 + 20\times 0^2 + 10\times 1^2 + 5\times 2^2}{49}} = 1{,}11 \qquad 1{,}43 \qquad 2{,}02$$

Si noti che, a differenza della media, la deviazione standard non è un operatore lineare, e quindi in generale:

$$s(\langle a\,x_i + b \rangle) \neq a\,s(\langle x_i \rangle) + b$$

Poiché, d'altra parte:

$$(x_i - m(\langle x_i \rangle))^2 = ((x_i + b) - m(\langle x_i + b \rangle))^2$$

cioè gli scarti dalla media sono invarianti per traslazione del campione, anche la deviazione standard è invariante per traslazione del campione:

$$s(\langle x_i+b\rangle)=s(\langle x_i\rangle)$$

Non è poi difficile dimostrare che, per $a>0$:

$$s(\langle a\,x_i\rangle)=a\,s(\langle x_i\rangle)$$

(e quindi anche che per la varianza: $s^2(\langle a\,x_i\rangle)=a^2\,s^2(\langle x_i\rangle)$). E' immediato verificare che la deviazione standard è un *operatore non negativo*, cioè:

$$s_x\geq 0$$

e in particolare che $s_x = 0$ nel caso in cui gli elementi del campione hanno tutti lo stesso valore.
Si noti infine che la deviazione standard di un campione somma di due campioni, $s(\langle x_i+y_i\rangle)$, in generale non è uguale alla somma delle deviazioni standard dei due campioni, $s(\langle x_i\rangle)+s(\langle y_i\rangle)$: come vedremo, per il calcolo di $s(\langle x_i+y_i\rangle)$ occorre tener conto anche della relazione che intercorre tra i due campioni.

Disuguaglianza di Chebyshev

L'importanza della deviazione standard è tale che la si usa spesso come unità di misura della dispersione intorno alla media, per esempio riportando i risultati di misurazioni nella forma $m_x\pm s_x$ o, più in generale, $m_x\pm k\,s_x$, per k positivo (e generalmente maggiore o uguale di 1). Al proposito è rilevante l'informazione circa quanti elementi del campione $\langle x_i\rangle$ stanno entro k, $k>0$, deviazioni standard dalla media, cioè quanti elementi sono contenuti nell'intervallo (m_x-ks_x, m_x+ks_x) .
Una soluzione a questo problema è data dalla fondamentale *disuguaglianza di Chebyshev*. Sia dato un campione $\langle x_i\rangle$ di n elementi, con media m_x e deviazione standard s_x. Sia:

$$S_k=\{x_i:|x_i-m_x|\leq ks_x\}$$

l'insieme degli elementi del campione che distano dalla media entro k deviazioni standard, e sia $\#S_k$ la cardinalità (cioè il numero degli elementi) dell'insieme S_k. La disuguaglianza di Chebyshev indica allora che la frazione $\#S_k/n$ ha un valore minimo (e quindi, inversamente, che la frazione degli elementi che distano più di k deviazioni standard da m_x ha un valore massimo), che dipende solo da k, e in particolare:

$$\frac{\#S_k}{n}\geqslant 1-\frac{1}{k^2}$$

Per esempio, entro $2s_x$ dalla media è contenuto almeno il 75% (1−1/4 = 3/4) degli elementi del campione, ed entro $3s_x$ almeno l'89% (1−1/9 = 8/9 = 89%) degli elementi.
Nota: poiché il rapporto $\#S_k/n$ è un numero non negativo, la disuguaglianza di Chebyshev è significativa solo per $k>1$.
Il fatto importante di questa disuguaglianza è che essa si applica a ogni possibile campione, a prescindere dalla forma della distribuzione (come si vedrà nel seguito trattando di particolari distribuzioni, conoscendo la forma della distribuzione tipicamente la disuguaglianza può essere resa più stringente: in particolare per campioni che seguono la distribuzione gaussiana, entro $1s_x$ dalla media è contenuto il 68% del campione, e in questo caso la disuguaglianza di Chebyshev non fornisce alcun limite, ed entro $2s_x$ e $3s_x$ sono contenuti il 95% e il 99,7% rispettivamente, ben più elevati del 75% e 89% che si ottengono dalla disuguaglianza di Chebyshev).

Disuguaglianza di Chebyshev: dimostrazione

La dimostrazione di questo teorema è semplice, e merita di essere studiata. Riprendendo la definizione di varianza, si ha che:

$$(n-1)\,s_x^2=\sum_{i=1}^{n}(x_i-m_x)^2$$

La somma a destra può essere divisa in due parti:

$$\sum_{i=1}^{n}(x_i-m_x)^2=\sum_{x_i\in S_k}(x_i-m_x)^2+\sum_{x_i\notin S_k}(x_i-m_x)^2$$

entrambe non negative, e quindi tralasciando il primo termine si ottiene:

$$(n-1)\,s_x^2\geqslant\sum_{x_i\notin S_k}(x_i-m_x)^2$$

Poiché questa somma è calcolata sugli elementi della popolazione che distano dalla media almeno k deviazioni standard, per ogni termine vale che:

$$(x_i - m_x)^2 \geqslant k^2 s_x^2$$

e quindi vale a maggior ragione che:

$$(n-1)\, s_x^2 \geqslant \sum_{x_i \notin S_k} k^2 s_x^2$$

e poiché la somma viene ripetuta $n - \# S_k$ volte:

$$(n-1)\, s_x^2 \geqslant k^2 s_x^2 (n - \# S_k)$$

Dividendo entrambi i termini per $n s_x^2 k^2$ si ottiene:

$$\frac{(n-1)}{k^2 n} \geqslant \frac{(n - \# S_k)}{n}$$

da cui:

$$\frac{1}{k^2} - \frac{1}{k^2 n} \geqslant 1 - \frac{\# S_k}{n}$$

e infine:

$$\frac{\# S_k}{n} \geqslant 1 - \frac{1}{k^2} + \frac{1}{k^2 n} \geqslant 1 - \frac{1}{k^2}$$

che è quanto si voleva dimostrare.

Disuguaglianza di Chebyshev: verifica numerica

Per un campione $\langle x_i \rangle$ dato, la verifica della validità della disuguaglianza di Chebyshev per un certo valore di k, $k>1$, richiede dunque i seguenti passi:

- calcolare la media m_x del campione;
- calcolare la deviazione standard s_x del campione;
- calcolare gli estremi dell'intervallo $m_x - k s_x$ e $m_x + k s_x$;
- calcolare il numero $\# S_k$ degli elementi del campione nell'intervallo dato, cioè tali che $m_x - k s_x < x_i < m_x + k s_x$;
- confrontare $\# S_k / n$ con $1-1/k^2$, per verificare appunto che $\# S_k / n > 1 - 1/k^2$.

Nota: la validità della disuguaglianza di Chebyshev è dimostrata (in pratica: tale disuguaglianza è la tesi di un teorema), e dunque questa verifica numerica non può fallire (o, detto altrimenti: se fallisce, cioè se risulta falso che $\# S_k / n > 1 - 1/k^2$ per qualche k, ciò significa che sono stati compiuti degli errori in uno o più dei passi indicati).

Statistiche ulteriori

In aggiunta a media e deviazione standard, altre due statistiche vengono spesso impiegate per caratterizzare la forma delle distribuzioni.

- La *asimmetria* (in inglese *skewness*) campionaria:

  $$\gamma_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - m_x}{s} \right)^3$$

  tale che:

  - $\gamma_1 = 0$ se la distribuzione è simmetrica;
  - $\gamma_1 < 0$ se la distribuzione ha la coda sinistra più lunga (e quindi "è più orientata verso" i valori minori della media);
  - $\gamma_1 > 0$ se la distribuzione ha la coda destra più lunga (e quindi "è più orientata verso" i valori maggiori della media).

- La *curtosi* (in inglese *kurtosis*) campionaria:

  $$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - m_x}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

  tale che:

  - $\gamma_2 = 0$ se la distribuzione è gaussiana;
  - $\gamma_2 < 0$ se la distribuzione è meno concentrata intorno alla media di una gaussiana (e quindi ha un picco meno stretto, come nel caso della distribuzione uniforme);

- $\gamma_2 > 0$ se la distribuzione è più concentrata intorno alla media di una gaussiana (e quindi ha un picco più stretto, come nel caso di una distribuzione in cui quasi tutti i valori sono concentrati in una sola categoria).

Un esempio: la curva di Lorenz e l'indice di Gini

Dato un campione $\langle x_i \rangle$, che supponiamo ordinato, e dunque $x_1 \le x_2 \le \ldots \le x_n$ (ipotizziamo anche $x_i \ge 0$), è interessante valutare il "grado di uniformità" degli elementi del campione, nell'ipotesi che:

- si ha uniformità massima quando tutti gli elementi del campione sono uguali;
- si ha uniformità minima quando i primi $n-1$ elementi del campione sono uguali e l'n-esimo elemento è diverso, e molto maggiore, degli $n-1$ precedenti (un campione del genere si chiama "singoletto", in inglese *singleton*).

Un modo per rappresentare graficamente questa informazione è mediante la cosiddetta *curva di Lorenz*, definita avendo come valori in ascissa:

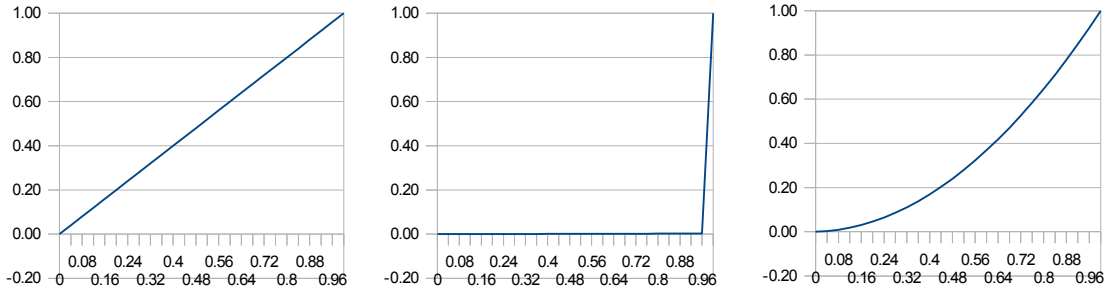
$z_0=0$, $z_i=i/n$, $i=1,\ldots,n$

e in ordinata:

$$L(z_0)=0, \quad L(z_i)=\frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{n} x_j}$$

e dunque tale che sia gli argomenti z_i sia i valori $L(z_i)$ sono nell'intervallo [0,1].

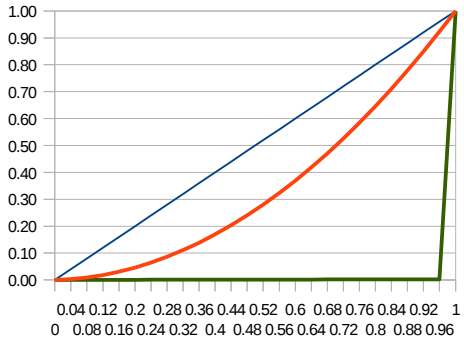
Tre esempi di questa curva sono:



curva α: uniformità massima curva β: uniformità minima curva γ: uniformità intermedia

corrispondenti rispettivamente a una condizione di uniformità massima (curva α: un campione di 25 elementi tutti uguali), a una di uniformità minima (curva β: i primi 24 elementi hanno valore 0, il venticinquesimo ha valore 100), e infine a una uniformità "intermedia" (curva γ: $x_i=i$).

L'informazione su questo concetto di grado di uniformità può essere sintetizzata in una statistica nota come *indice di Gini*. Ponendo nello stesso grafico le tre curve precedenti:



allo scopo di analizzare il campione rappresentato dalla curva γ rispetto alle due condizioni estreme, rappresentate dalle curve α e β, si può confrontare l'area delimitata dalle curve α e γ rispetto all'area delimitata dalle curve α e β; se tale rapporto è:

- pari a 0, cioè γ=α, il campione in esame è massimamente uniforme;
- pari a 1, cioè γ=β, il campione in esame è minimamente uniforme;

- maggiore di 0 e minore di 1, il campione in esame ha un'uniformità intermedia, ed è tanto più uniforme quanto minore è tale rapporto.

L'indice di Gini può essere calcolato con la seguente formula:

$$G(\langle x_i \rangle)=\frac{1}{n}\left(n+1-2\left(\frac{\sum_{i=1}^{n}(n+1-i)\,x_i}{\sum_{i=1}^{n}x_i}\right)\right)$$

Campioni multivariati

Le entità che si considerano, e da cui si ottengono i dati che sono oggetto dell'analisi statistica, sono generalmente caratterizzate da più grandezze (nel caso di oggetti fisici potrebbero essere lunghezza, massa, carica elettrica, ...), a ognuna delle quali è applicabile quanto abbiamo discusso finora: ogni grandezza ha un insieme supporto A, per ogni grandezza si può ottenere un campione, cioè una successione $\langle x_i \rangle$ di valori, ognuno appartenente ad A, e tale campione può essere quindi descritto mediante una distribuzione e sintetizzato con la moda ed eventualmente la mediana, la media e la deviazione standard. In questo modo, le grandezze si considerano indipendentemente le une dalle altre. In certi casi, si è però interessati alla *relazione tra le grandezze*: al variare di una grandezza X come varia una seconda grandezza Y? Se X cresce, cresce anche Y? oppure si riduce? oppure la variazione di Y non porta alcuna informazione sulla variazione di X?

Quando gli individui che si prendono in esame sono caratterizzati ognuno da una k-upla di valori, $k \geqslant 2$, e quindi un campione è una successione di k-uple, il campione stesso si chiama *multivariato*, e *bivariato* nel caso particolare in cui le grandezze in esame sono 2.

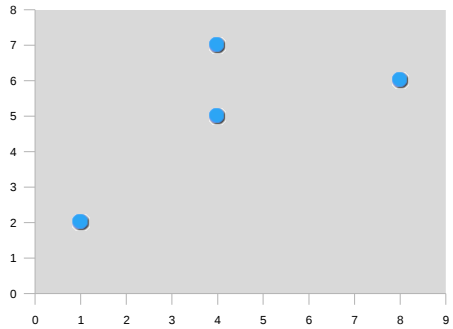
Se le grandezze fossero $k = 3$, lunghezza, massa, carica elettrica, e gli individui del campione fossero $n = 4$, il campione stesso potrebbe essere $\langle\langle l_i, m_i, q_i \rangle\rangle = \langle\langle 1,2,3 \rangle, \langle 4,5,6 \rangle, \langle 8,6,4 \rangle, \langle 4,7,2 \rangle\rangle$, a indicare che, per esempio, la massa del secondo individuo vale 5 e la carica elettrica del terzo individuo vale 4.

Introduciamo qui alcuni principi di base dell'*analisi bivariata*, con cui si studiano le relazioni campionarie tra coppie di grandezze, dunque nel caso $k = 2$.

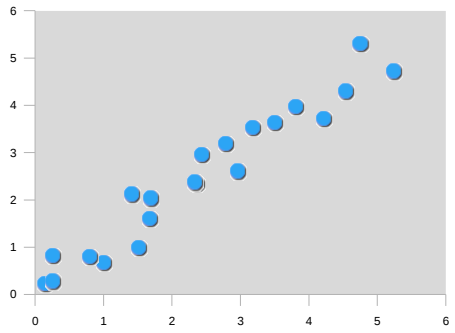
Campioni bivariati

Consideriamo due grandezze, X e Y, che possono essere contemporaneamente valutate su un insieme di entità, e sia $\langle\langle x_i, y_i \rangle\rangle$ il campione bivariato ottenuto, essendo $\langle x_i \rangle$ il sottocampione dei valori di X e $\langle y_i \rangle$ il sottocampione dei valori di Y. Ipotizzando che gli insiemi supporto per X e Y siano almeno ordinati, ogni coppia $\langle x_i, y_i \rangle$ può essere appropriatamente rappresentata come un punto su un piano $X \times Y$.

Considerando solo le prime due grandezze nel campione dell'esempio precedente, si otterrebbe un grafico come quello seguente, che si chiama *diagramma di dispersione* (in inglese *scatter diagram*; nei fogli di calcolo anche "diagramma X-Y"):



Un caso molto semplice di relazione tra X e Y si presenta quando ogni valore x_i è legato al corrispondente valore y_i attraverso un'espressione della forma $y_i = a\,x_i + b$ con $a \neq 0$, corrispondente nel diagramma di dispersione a una retta di pendenza crescente se $a>0$ o decrescente se $a<0$: è questa una situazione di relazione deterministica tra le due grandezze, nel senso che il valore x_i determina il valore y_i (e viceversa, data l'invertibilità della funzione). Più interessante dal punto di vista statistico è una situazione come quella illustrata in questo diagramma:



in cui si mostra ancora la presenza di una relazione tra le due grandezze – al crescere di una, cresce, generalmente, anche l'altra – ma senza che ciò sia strettamente deterministico: i punti si dispongono intorno a una retta, ma non su di essa. Potrebbe essere, per esempio, il caso della relazione tra altezza e peso di un insieme di persone: benché non in modo deterministico, all'aumentare dell'altezza aumenta generalmente anche il peso. In casi di questo genere si può dunque cercare di valutare non solo se le due grandezze hanno una relazione, ma anche, e più specificamente, come e quanto, in senso statistico, sono in relazione l'una con l'altra, in breve quanto sono *correlate*. La presenza di correlazione fornisce un'informazione di carattere statistico: la conoscenza del valore di X fornisce un'informazione almeno parziale sul valore di Y.

Distribuzioni congiunte, condizionali e marginali: dipendenza statistica

Quanto considerato a proposito della relazione tra campioni e distribuzioni nel caso univariato si applica anche a campioni bivariati: per ognuna delle due grandezze X e Y si può definire un insieme di categorie, $\{X_j\}$ e $\{Y_k\}$, così che ogni elemento $\langle x_i , y_i \rangle$ appartiene a una e una sola coppia di categorie $\langle X_j , Y_k \rangle$. La distribuzione che si ottiene corrisponde a una tabella del tipo (nel caso di frequenze assolute):

		Y			
		Y_1	...	Y_k	...
X	X_1	$n_{1,1}$		$n_{1,k}$	
	...				
	X_j	$n_{j,1}$		$n_{j,k}$	
	...				...

in cui $n_{j,k}$ è il numero di elementi del campione bivariato tali che x_i appartiene alla categoria X_j e y_i appartiene alla categoria Y_k. Questa tabella descrive la *distribuzione congiunta* delle grandezze X e Y rispetto alle categorie date.
La generica riga j-esima della tabella contiene la distribuzione dei valori di Y calcolata sulle sole coppie in cui il valore di X appartiene alla categoria X_j, e chiamata perciò *distribuzione condizionale* di Y data la categoria X_j, scritto in breve distribuzione di $Y \mid X_j$, dove dunque il simbolo '|' è da leggere "dato", cioè appunto "condizionato a":

		Y			
		Y_1	...	Y_k	...
X	X_1	$n_{1,1}$		$n_{1,k}$	
	...				
	X_j	$n_{j,1}$	...	$n_{j,k}$	...
	...				

Analogamente per ogni categoria Y_k si può considerare la distribuzione condizionale di $X \mid Y_k$:

		Y			
		Y_1	...	Y_k	...

X	X_1	$n_{1,1}$		$n_{1,k}$	
	...			...	
	X_j	$n_{j,1}$		$n_{j,k}$	
	...			...	

Ancora a partire da questa tabella si possono poi introdurre i totali parziali per ogni riga e per ogni colonna:

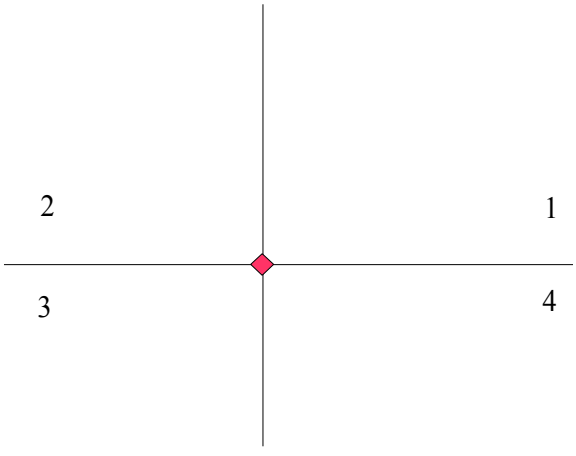
		Y				
		Y_1	...	Y_k	...	Σ
X	X_1	$n_{1,1}$		$n_{1,k}$		$\Sigma_k n_{1,k}$
	...					...
	X_j	$n_{j,1}$		$n_{j,k}$		$\Sigma_k n_{j,k}$
	...					...
	Σ	$\Sigma_j n_{j,1}$	...	$\Sigma_j n_{j,k}$	...	

Si ottengono le *distribuzioni marginali* per X (ultima colonna) e per Y (ultima riga).
Data la distribuzione congiunta di XY, le distribuzioni condizionali e le distribuzioni marginali consentono di cominciare a studiare la possibile dipendenza tra la grandezza X e la grandezza Y. L'idea è semplice. Supponiamo di voler investigare se per esempio i valori di Y dipendono dai valori di X, cioè, concretamente, se la conoscenza del fatto che un certo individuo ha un certo valore X_j per X porta informazione sul valore che quello stesso individuo ha per Y: se la risposta fosse positiva dovremmo osservare che le distribuzioni di $Y \mid X_j$ e Y (opportunamente normalizzate) sono diverse tra loro. In tal caso si dice che Y ha una *dipendenza statistica* da X_j. Viceversa, se le distribuzioni di $Y \mid X_j$ e Y sono uguali allora sapere che X vale X_j non è informativo relativamente a Y, e perciò Y è indipendente statisticamente da X_j.

Covarianza e coefficiente di correlazione campionaria

Al fine di giungere a caratterizzare quantitativamente il *grado di correlazione* tra due grandezze X e Y, che continuiamo a supporre essere più che solo ordinali, calcoliamone i valori medi, m_x e m_y (il rombo rosso nel diagramma sotto), e consideriamo gli scarti $(x_i - m_x)(y_i - m_y)$:

- se $x_i > m_x$ e $y_i > m_y$ oppure $x_i < m_x$ e $y_i < m_y$ (cioè se il punto $\langle x_i, y_i \rangle$ sta nei quadranti 1 o 3), il prodotto è positivo;
- se invece i segni sono discordi, $x_i > m_x$ e $y_i < m_y$ oppure $x_i < m_x$ e $y_i > m_y$ (cioè se il punto $\langle x_i, y_i \rangle$ sta nei quadranti 2 o 4), il prodotto è negativo.



Prendiamo ora in esame la somma:

$\sum_{i=1}^{n}(x_i-m_x)(y_i-m_y)$:

- se, come in questo caso, i punti si addensano nei quadranti 1 e 3, tale somma è positiva e si dice che c'è *correlazione positiva* tra le grandezze X e Y del campione: al crescere di una, cresce anche l'altra;
- se i punti si addensano nei quadranti 2 e 4, la somma è negativa e si dice che c'è *correlazione negativa* tra le grandezze X e Y del campione: al crescere di una, l'altra decresce;
- se infine i punti sono più o meno omogeneamente sparsi nei quattro quadranti, la somma contiene termini positivi e termini negativi, che si compensano e quindi producono un valore vicino all'origine del grafico, e in tal caso si dice che c'è *correlazione bassa*, e al limite nulla, tra le grandezze X e Y del campione.

La statistica che si ottiene se si normalizza la somma precedente dividendola per $n-1$:

$$\mathrm{cov}(\langle\langle x_i, y_i\rangle\rangle)=\frac{\sum_{i=1}^{n}(x_i-m_x)(y_i-m_y)}{n-1}$$

si chiama *covarianza campionaria*, in analogia con la varianza campionaria, che infatti può essere scritta:

$$s_x^2=\frac{\sum_{i=1}^{n}(x_i-m_x)(x_i-m_x)}{n-1}$$

Si noti che la covarianza campionaria ha unità di misura uguale al prodotto delle unità di misura di X e Y. Dividendo ancora per il prodotto delle deviazioni standard campionarie s_x e s_y, si ottiene perciò un coefficiente adimensionale:

$$R(\langle\langle x_i, y_i\rangle\rangle)=R_{x,y}=\frac{\sum_{i=1}^{n}(x_i-m_x)(y_i-m_y)}{(n-1)s_x s_y}=\frac{\mathrm{cov}(\langle\langle x_i, y_i\rangle\rangle)}{s_x s_y}$$

chiamato *coefficiente di correlazione campionaria*, che assume valori tra −1 (completa correlazione lineare negativa: i punti del campione bivariato sono disposti lungo una retta a pendenza negativa) e 1 (completa correlazione lineare positiva: i punti del campione bivariato sono disposti lungo una retta a pendenza positiva), con il valore centrale, 0, che corrisponde all'assenza di correlazione.

Infine, come avevamo mostrato che la media di un campione somma di due campioni è uguale alla somma delle medie dei due campioni:

$m(\langle x_i+y_i\rangle) = m(\langle x_i\rangle)+m(\langle y_i\rangle)$

possiamo chiederci a questo punto come si possa calcolare la deviazione standard di un campione somma di due campioni, $s(\langle x_i+y_i\rangle)$. Al proposito vale l'interessante risultato:

$s^2(\langle x_i+y_i\rangle) = s^2(\langle x_i\rangle)+s^2(\langle y_i\rangle)+2\mathrm{cov}(\langle\langle x_i,y_i\rangle\rangle)$

(naturalmente, per poter sommare gli elementi di due campioni essi devono avere la stessa unità di misura, e quindi la formula risulta corretta dal punto di vista dimensionale). Dunque, quando i due campioni hanno covarianza nulla la varianza di un campione somma di due campioni è uguale alla somma delle varianze dei due campioni; nel caso generale, occorre invece tener conto anche delle relazioni tra i due campioni, formalizzate appunto mediante la covarianza.

Correlazione e causalità

Nell'analisi della correlazione tra grandezze è necessario evitare di confondere correlazione e causalità; date due grandezze X e Y:

se X implica Y (cioè il valore di X determina il – è causa del – valore di Y) allora c'è correlazione tra X e Y

D'altra parte:

se c'è correlazione tra X e Y, non necessariamente X implica Y o Y implica X

(la situazione di correlazione senza causalità potrebbe essere dovuta a ciò che a volte si chiama "semplice coincidenza" – ma cosa sono le coincidenze? –, oppure per esempio al fatto che sia X sia Y sono gli effetti di una stessa causa, che dunque determina la loro correlazione senza renderli causalmente dipendenti).
La differenza tra questi due concetti si nota anche dal fatto che:

- la correlazione è simmetrica (se X è correlato con Y allora Y è ugualmente correlato con X, poiché $R_{x,y}=R_{y,x}$);

- la causalità è antisimmetrica (se X è causa di Y allora Y non può essere causa di X).

Si noti che questa considerazione si applica anche al caso deterministico in cui $R_{x,y}=1$, situazione che non implica che tra X e Y ci sia una relazione causale.
Qualche esempio (da wikipedia). Qual è l'errore?

- Si osserva che c'è una forte correlazione tra numero di pompieri dedicati a spegnere un incendio e dimensioni dell'incendio, e se ne conclude che il numero di pompieri dedicati determina le dimensioni degli incendi.
- Si osserva che c'è una forte correlazione tra dormire con le scarpe ai piedi e svegliarsi con il mal di testa, e se ne conclude che dormire con le scarpe ai piedi causa il mal di testa.
- Si osserva che c'è una forte correlazione tra numero di gelati venduti e numero di morti per annegamento, e se ne conclude che i gelati sono un'importante causa di annegamento.
- Si osserva che c'è una forte correlazione inversa tra numero di assalti di pirati a navi ed effetti del riscaldamento globale, e se ne conclude che la mancanza di pirati causa riscaldamento globale.
- ... e come interpretare in termini causali le relazione (di correlazione deterministica) tra pressione e temperatura in un gas perfetto?

Campioni di statistiche campionarie

Supponiamo che almeno in linea di principio l'intera popolazione di riferimento $\langle x_i \rangle$, costituita da n elementi, sia disponibile e che dunque, sempre almeno in linea di principio, sia possibile calcolare su di essa le statistiche di base, e in particolare la media m_x. Supponiamo inoltre che per qualche ragione (per esempio per ridurre i costi di acquisizione dei dati) si consideri opportuno operare non sull'intera popolazione, ma solo su uno o più suoi campioni $\langle y_i \rangle_j$, ognuno di n_1, $1 \le n_1 \le n$, elementi. Le statistiche calcolate su tali campioni forniscono un'informazione sulle corrispondenti statistiche, ignote, della popolazione.
In particolare, di ogni campione è possibile calcolare la media, m_j; il problema che ci poniamo è allora:

che relazione c'è tra le medie campionarie m_j e la media della popolazione m_x?

La risposta nei due casi estremi è semplice:

- se $n_1 = n$ (cioè se il campione coincide con la popolazione), $m_j = m_x$;
- se $n_1 = 1$ (cioè se il campione coincide con un singolo elemento dell'insieme supporto), m_j non porta alcuna informazione diversa dal campione stesso.

Nei casi intermedi, $1<n_1<n$, m_j porta un'informazione su m_x pur senza coincidere, in generale, con m_x stessa: se il campione $\langle y_i \rangle_j$ è costituito di elementi scelti senza alcun criterio definito (e quindi si potrebbe dire: "in modo casuale", benché cosa ciò significhi non sia così chiaro) dalla popolazione $\langle x_i \rangle$, allora al variare del campione anche la media campionaria m_j varierà in modo casuale "intorno a m_x", e quanto maggiore è la numerosità n_1 del campione tanto più sarà stabile, cioè tanto meno varierà, m_j.
In generale, si dice a questo proposito che *m_j è uno stimatore di m_x* (che tradizionalmente verrebbe chiamata la "media vera"), e che è in generale uno stimatore tanto migliore (nel senso di: tanto più stabilmente vicino a m_x) quanto maggiore è il numero n_1 di elementi del campione. E' fondamentale notare qui che, dato il fatto che ogni campione $\langle y_i \rangle_j$ è ottenuto "in modo casuale", le medie dei vari campioni, m_1, m_2, ... non coincideranno, pur essendo tutte stimatori della stessa media della popolazione m_x. Si genera in questo modo un campione $\langle m_j \rangle$ di medie campionarie.
Questa situazione può essere formalizzata in modo elegante come segue.
Supponiamo che dalla popolazione $\langle x_i \rangle$ siano ottenuti n_2 campioni $\langle y_i \rangle_j$ ognuno costituito da n_1 elementi (dunque i=1, ..., n_1 e j=1, ..., n_2), una situazione sintetizzabile nella tabella:

	campione 1	...	campione j	...	campione n_2
elemento 1	$y_{1,1}$		$y_{1,j}$		y_{1,n_2}
...					
elemento i	$y_{i,1}$		$y_{i,j}$		y_{i,n_2}
...					
elemento n_1	$y_{n_1,1}$		$y_{n_1,j}$		y_{n_1,n_2}
media campionaria	m_1		m_j		m_{n_2}

Consideriamo, innanzitutto, un generico campione $\langle y_i \rangle$, dunque una colonna della tabella: si può supporre che ognuno dei suoi elementi, y_1, ..., y_{n_1} sia il valore assunto da una *variabile casuale* (o anche: "variabile aleatoria"), Y_1, ..., Y_{n_1}; in simboli:

$$Y_i = y_i$$

(si noti la notazione, abituale in statistica: la lettera maiuscola indica una variabile – potrebbe essere per esempio una grandezza fisica –, mentre la lettera minuscola indica un valore della variabile). Allora l'ipotesi che tutti gli elementi del campione $\langle y_i \rangle$ siano ottenuti, senza polarizzazione, dalla stessa popolazione corrisponde all'ipotesi che tutte le variabili casuali Y_i abbiano la stessa distribuzione. Ne segue dunque che se si ripete l'operazione di campionamento, e quindi si ottengono n_2 campioni $\langle y_i \rangle_1$, ..., $\langle y_i \rangle_{n_2}$, ogni n_2-upla $\langle y_{i,1}, ..., y_{i,n_2} \rangle$ contiene n_1 valori della stessa variabile casuale Y_i.
Ogni media campionaria m_j è perciò uno stimatore della variabile $(Y_1+...+Y_{n_1})/n_1$.
Che dire allora di $m(\langle m_j \rangle) = m_m$, cioè la media del campione delle medie campionarie? Per definizione:

$$m_m = m\left(\frac{Y_1 + ... + Y_{n_1}}{n_1}\right)$$

Per la linearità della media:

$$m\left(\frac{Y_1 + ... + Y_{n_1}}{n_1}\right) = \frac{m_1 + ... + m_{n_1}}{n_1}$$

e per l'ipotesi che le variabili casuali Y_i abbiano la stessa distribuzione, e quindi $m_1 = m_2 = ... \; m_x$:

$$\frac{m_1 + ... + m_{n_1}}{n_1} = \frac{n_1 m_x}{n_1} = m_x$$

La media delle medie campionarie approssima ("stima" appunto) la media della popolazione: al crescere della cardinalità n_1 dei campioni $\langle y_i \rangle_j$, la media delle medie m_m è sempre più simile a m_x.
Riprendendo in considerazione il campione delle medie campionarie $\langle m_j \rangle$, si può studiare ora la sua deviazione standard, $s(\langle m_j \rangle) = s_m$, per chiedersi in particolare se essa coincida, a meno di approssimazione, con la deviazione standard campionaria s_j dei campioni $\langle y_i \rangle_j$. Secondo la formalizzazione appena introdotta, si tratta dunque di studiare:

$$s_m = s\left(\frac{Y_1 + ... + Y_{n_1}}{n_1}\right)$$

Si può dimostrare che, nel caso in cui le n_1 variabili casuali Y_i abbiano correlazione nulla, vale che:

$$s_m^2 = s^2\left(\frac{Y_1 + ... + Y_{n_1}}{n_1}\right) = \frac{s_1^2 + ... + s_{n_1}^2}{n_1^2}$$

e poiché si assume che le variabili casuali Y_i abbiano la stessa distribuzione:

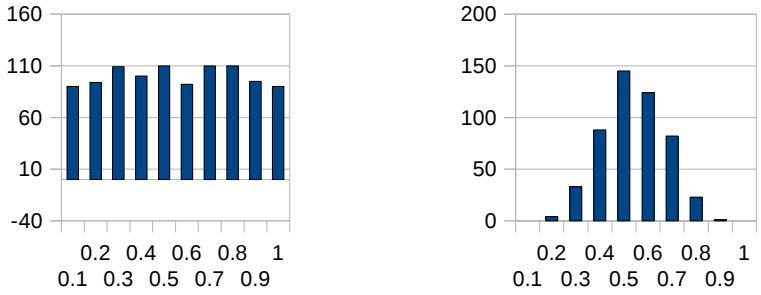
$$s_m^2 = \frac{n_1 s_x^2}{n_1^2} = \frac{s_x^2}{n_1}$$

Ricordando poi che la deviazione standard s_x della popolazione è stimata dalle deviazioni standard campionarie, cioè $s_x \approx s_j$, vale allora che:

$$s_m \approx \frac{s_j}{\sqrt{n_1}}$$

Il campione delle medie campionarie ha una deviazione standard approssimativamente pari alla deviazione standard di un generico campione divisa per la radice quadrata del numero degli elementi del campione stesso.
Mettiamo alla prova questo risultato fondamentale con un esempio. Supponiamo che la distribuzione della popolazione $\langle x_i \rangle$ sia (a meno di approssimazioni) uniforme, cioè che le frequenze delle categorie siano almeno approssimativamente uguali, e che i campioni $\langle y_i \rangle_j$ siano scelti in modo da seguire anch'essi una distribuzione uniforme (è quello che ci si aspetta, in questo caso, da un campionamento "casuale", che non produca distorsioni): a causa della maggiore stabilità del campione $\langle m_j \rangle$ rispetto a ognuno dei campioni $\langle y_i \rangle_j$, si può constatare che il campione $\langle m_j \rangle$ non è più uniforme, ma ha una forma "a campana" (chiamata *gaussiana* o anche "normale"), simmetrica e centrata in m_x.

Per esempio, da una popolazione di $n = 1000$ valori scelti uniformemente nell'intervallo [0,1] (istogramma a sinistra) è ottenuto un certo numero di campioni $\langle y_i \rangle_j$, ognuno di $n_1 = 5$ valori, e se ne ricava la distribuzione delle medie campionarie $\langle m_j \rangle$ (a destra):



Come si vede, il campione $\langle m_j \rangle$ è molto più concentrato intorno a m_x di quanto non lo siano la popolazione $\langle x_i \rangle$ e quindi i campioni $\langle y_i \rangle_j$: ci si può dunque aspettare che la deviazione standard del campione delle medie campionarie, s_m, sia minore delle deviazioni standard dei campioni, s_j.

Il teorema del limite centrale

Il fatto che in condizioni come quelle specificate per l'esempio precedente le medie campionarie (e quindi anche le somme di elementi di campioni) siano approssimativamente distribuite secondo una gaussiana è un caso particolare di un risultato più generale, formalizzato nel *teorema del limite centrale*: la somma di k variabili casuali indipendenti ottenute da una stessa distribuzione, con media m_x e deviazione standard s_x, approssima sempre meglio, al crescere di k, una distribuzione gaussiana, e ciò indipendentemente dalla forma della distribuzione da cui le variabili casuali sono ottenute; tale distribuzione gaussiana "attrattore" ha media km_x e deviazione standard $(k\,s_x)/\sqrt{k} = \sqrt{k}\,s_x$. In conseguenza, se invece della somma delle k variabili casuali si considera la loro media, la distribuzione gaussiana "attrattore" ha media m_x e deviazione standard $s_x/\sqrt{k}$, esattamente come ottenuto sopra. Questo teorema rende conto del fatto che in molte situazioni sperimentali le distribuzioni sono approssimabili mediante gaussiane, e quindi giustifica la scelta – già citata – di chiamare anche "normali" tali distribuzioni.