UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

APPUNTI DI MATEMATICA E STATISTICA
(per il Corso di Laurea in Scienze Nutraceutiche)

STATISTICA DESCRITTIVA
(continuazione)

Prof. Aniello Buonocore

Dipartimento di Matematica e Applicazioni "Renato Caccioppoli"

Scuola Politecnica e delle Scienze di Base

STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

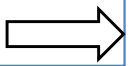
Si vuole ora considerare il problema di *sintetizzare* la rilevazione dati $\underline{y} = (y_1, y_2, \cdots, y_N)$ con un unico valore di *tendenza centrale*, ossia un valore che fornisca un'indicazione di massima sulla localizzazione di $\underline{y}$.

Ciò è utile non solo per una più immediata comprensione dei risultati dell'indagine ma anche per istituire un confronto del fenomeno studiato con altri fenomeni dello stesso tipo.

Per i caratteri qualitativi è possibile fare ricorso ai due indici media e mediana di cui si ricorda le definizioni.

DEFINIZIONE 1

Si consideri un carattere (di qualsiasi tipo). La modalità corrispondente alla frequenza (assoluta o relativa) più grande viene detta *moda* (M_0) della rilevazione dati.

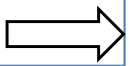


STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

DEFINIZIONE 2

Si consideri un carattere (non qualitativo nominale). La modalità corrispondente alla più piccola frequenza relativa cumulata maggiore o uguale a 0,5 viene detta *mediana* (M_1) della rilevazione dati.

Nel caso di una taglia N dispari, un modo pratico per ottenere la mediana senza costruire la distribuzione di frequenza è quello di ordinare i dati dal più piccolo al più grande e poi, ricorsivamente, depennare il minimo e il massimo fino a quando resta un unico elemento che è la mediana.

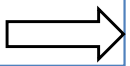


STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

Se invece N è pari, alla fine di tutti i depennamenti restano due elementi. In tal caso, bisogna separare il caso (i) i due elementi restanti sono uguali (il loro valore comune coincide con la mediana) dal caso (ii) i due elementi restanti sono diversi.

Nel caso (ii) con un carattere quantitativo la mediana è la semisomma dei due elementi restanti. Nel caso (ii) con un carattere qualitativo ordinale la mediana è, invece, indeterminata.

Per i caratteri quantitativi, invece, ci sono altri due metodi teorici in grado di far ottenere indici di tendenza centrale, ovvero, le *medie analitiche* di Chisini e i *centri*.



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

LE MEDIE ANALITICHE DI CHISINI

Sia $\underline{y} = (y_1, y_2, \cdots, y_N)$ una rilevazione dati su un carattere quantitativo. Per ottenere una *media analitica* bisogna dapprima specificare un criterio C (o *funzione di circostanza*) rispetto al quale si vuole ottenere la valutazione di tendenza centrale. Dopo di ciò bisogna determinare un numero reale y per il quale, indicata con $\underline{y}^* = (y, y, \cdots, y)$ una rilevazione dati (fittizia) di taglia N aventi tutti gli elementi uguali a y, la valutazione della funzione di circostanza C su $\underline{y}$ coincide con la valutazione di C su $\underline{y}^*$. In simboli,

$$C(y_1, y_2, \cdots, y_N) = C(y, y, \cdots, y),$$

che è chiamata *equazione di circostanza*.

STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

La media aritmetica

Si consideri la funzione di circostanza *somma dei dati*:

$$C_1(y_1, y_2, \cdots, y_N) = y_1 + y_2 + \cdots + y_N.$$

Dall'equazione di circostanza si ottiene

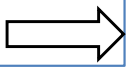
$$C_1(y_1, y_2, \cdots, y_N) = C_1(y, y, \cdots, y)$$

$$\Leftrightarrow y_1 + y_2 + \cdots + y_N = \underbrace{y + y + \cdots + y}_{N-volte}$$

$$\Leftrightarrow y_1 + y_2 + \cdots + y_N = N \cdot y$$

$$\Leftrightarrow y = \frac{y_1 + y_2 + \cdots + y_N}{N} =: M_2,$$

e la media corrispondente è detta *aritmetica*.



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

Mediante il simbolo di *sommatoria* la media aritmetica

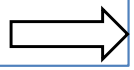
$$M_2 = \frac{y_1 + y_2 + \cdots + y_N}{N},$$

si può scrivere nella forma compatta:

$$M_2 = \frac{1}{N}\sum_{i=1}^{N} y_i .$$

Se ciascuno dei dati viene trasformato mediante una funzione composta $T(x)$ allora si può calcolare la media aritmetica dei valori ottenuti:

$$\frac{1}{N}\sum_{i=1}^{N} T(y_i) .$$



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

La media armonica

Si consideri la funzione di circostanza *somma dei reciproci dei dati* (che quindi devono essere non nulli):

$$C_2(y_1, y_2, \cdots, y_N) = \frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}.$$

Dall'equazione di circostanza si ottiene

$$C_2(y_1, y_2, \cdots, y_N) = C_2(y, y, \cdots, y)$$

$$\Leftrightarrow \frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N} = \underbrace{\frac{1}{y} + \frac{1}{y} + \cdots + \frac{1}{y}}_{N-volte}$$

$$\Leftrightarrow \left(\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}\right) = \frac{N}{y} \Leftrightarrow y = \frac{N}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}} =: M_{-1},$$

e la media corrispondente è detta *armonica*. ⇨

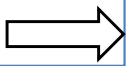
STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

In più, risultando

$$M_{-1} = \frac{N}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}} = \left(\frac{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}}{N} \right)^{-1}$$

$$= \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{y_i} \right)^{-1},$$

si può allora dire che *la media armonica di dati rilevati da un carattere non nullo è uguale al reciproco della media aritmetica dei reciproci dei dati* (si usa la trasformazione $T(x) = 1/x$).



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

La media geometrica

Si consideri la funzione di circostanza *prodotto dei dati* (che devono essere assunti positivi):

$$C_3(y_1, y_2, \cdots, y_N) = y_1 \cdot\ y_2 \cdots y_N.$$

Dall'equazione di circostanza si ottiene

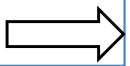
$$C_3(y_1, y_2, \cdots, y_N) = C_3(y, y, \cdots, y)$$

$$\Leftrightarrow y_1 \cdot y_2 \cdot \cdots \cdot y_N = \underbrace{y \cdot y \cdots y}_{N-volte}$$

$$\Leftrightarrow y_1 \cdot y_2 \cdots y_N = y^N \Leftrightarrow y = \sqrt[N]{y_1 \cdot y_2 \cdots y_N} =: M_g,$$

$$\Leftrightarrow y_1 \cdot y_2 \cdots y_N = y^N \Leftrightarrow y = \sqrt[N]{y_1 \cdot y_2 \cdots y_N} =: M_g,$$

e la media corrispondente è detta *geometrica*.



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

Con il simbolo di *produttoria*, la media geometrica

$$M_g = \sqrt[N]{y_1 \cdot y_2 \cdot \cdots \cdot y_N}$$

si può scrivere nella forma compatta:

$$M_g = \sqrt[N]{\prod_{i=1}^{N} y_i}\,.$$

Si può allora dire che *la media geometrica di dati rilevati da un carattere positivo è uguale alla radice di indice N della produttoria dei dati*.

Si fa esplicitamente notare che la moda e la mediana non sono medie analitiche. ⇨

STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

I CENTRI

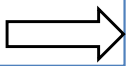
Sia $\underline{y} = (y_1, y_2, \cdots, y_N)$ una rilevazione dati su un carattere quantitativo.

Per ogni assegnato reale positivo r si ponga

$$\forall x \in \mathbb{R}, \qquad d_r\left(x, \underline{y}\right) \equiv f_r(x) = \left(\frac{1}{N}\sum_{i=1}^{N}|x - y_i|^r\right)^{\frac{1}{r}},$$

mentre per $r = 0$ si ponga

$$\forall x \in \mathbb{R}, \qquad d_0\left(x, \underline{y}\right) \equiv f_0(x) = \frac{1}{N}\sum_{i=1}^{N}|x - y_i|^0 .$$



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

In altri termini, per qualunque $r \geq 0$ la funzione

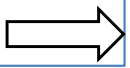
$$d_r\left(x, \underline{y}\right) \equiv f_r(x)$$

rappresenta la *distanza* (di ordine r) tra l'argomento x e la rilevazione dati $\underline{y}$.

Si definisce *centro di ordine* r il punto di minimo assoluto della funzione f_r.

Quindi il centro di ordine r di una rilevazione dati $\underline{y}$ è il numero reale $c_r\left(\underline{y}\right)$ che rende minima la distanza di ordine r. In simboli:

$$c_r\left(\underline{y}\right) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} f_r(x)\,.$$



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

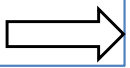
TEOREMA 1 (senza dimostrazione)

Il centro di ordine 1 *della rilevazione dati* $\underline{y}$, *ovvero* $c_1\left(\underline{y}\right)$, *coincide con la mediana:*

$$c_1\left(\underline{y}\right) = \operatorname*{argmin}_{x\in\mathbb{R}} f_1(x)$$

$$= \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{i=1}^{N} |x - y_i|$$

$$\equiv M_1.$$



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

TEOREMA 2

Il centro di ordine 0 *della rilevazione dati* $\underline{y}$*, ovvero* $c_0\left(\underline{y}\right)$*, coincide con la moda.*

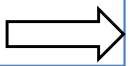
DIMOSTRAZIONE

Siano $x_1 < x_2 < \cdots < x_k$ le modalità del carattere Y. Per definizione,

$$f_0(x) = \frac{1}{N}\sum_{i=1}^{N}|x - y_i|^0$$

rappresenta la distanza di ordine 0 tra x e l'intera rilevazione dati $\underline{y}$.

D'altra parte $|x - y_i|^0$ rappresenta la distanza tra x e il generico dato y_i.



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

Se ne ricava che quando x coincide con y_i (la loro distanza è nulla) l'addendo i-mo non porta contributo a $f_0(x)$.

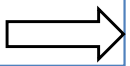
Pertanto

$$f_0(x) = \begin{cases} 1, & \text{se } x \notin \{x_1, x_2, \cdots, x_k\} \\ 1 - n_j/N, & \text{se } x = x_j \text{ e } j \in \{1,2, \cdots, k\}. \end{cases}$$

Di conseguenza $f_0(x)$ assume il suo minimo assoluto in corrispondenza della modalità alla quale compete la frequenza (relativa) maggiore che per definizione è la moda della distribuzione di frequenza:

$$c_0(\underline{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} f_0(x) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{N} |x - y_i|^0$$

$$\equiv M_0.$$



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

TEOREMA 3

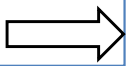
Il centro di ordine 2 della rilevazione dati $\underline{y}$, $c_2\left(\underline{y}\right)$, *coincide con la media aritmetica.*

DIMOSTRAZIONE

Per definizione,

$$f_2(x) = \left(\frac{1}{N}\sum_{i=1}^{N}|x - y_i|^2\right)^{\frac{1}{2}} = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{N}|x - y_i|^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x - y_i)^2}\,.$$



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

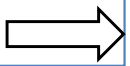
D'altra parte essendo la funzione radice quadrata crescente nel suo dominio $[0, +\infty[$ il minimo della funzione viene raggiunto in corrispondenza del minimo della funzione:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}(x - y_i)^2$$
$$= (x - y_1)^2/N + \cdots + (x - y_N)^2/N.$$

Risultando

$$f'(x) = [2(x - y_1) + 2(x - y_2) + \cdots + 2(x - y_N)]/N$$
$$= 2[(x - y_1) + (x - y_2) + \cdots + (x - y_N)]/N,$$

dalla condizione necessaria per la ricerca dei punti di minimo relativo si ricava:



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

$$f'(x) = 0 \Leftrightarrow (x - y_1) + (x - y_2) + \cdots + (x - y_N) = 0$$

$$\Leftrightarrow \underbrace{x + x + \cdots + x}_{N-volte} = y_1 + y_2 + \cdots + y_N$$

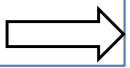
$$\Leftrightarrow x = \frac{y_1 + y_2 + \cdots + y_N}{N}.$$

D'altra parte risultando

$$f''(x) = \underbrace{2/N + 2/N + \cdots + 2/N}_{N-volte} = 2 > 0,$$

il punto stazionario determinato è un punto di minimo relativo.

Inoltre, la funzione f è definita e derivabile in $\mathbb{R}$ e infinitamente grande intorno a $+\infty$ e a $-\infty$.



STATISTICA DESCRITTIVA: INDICI DI TENDENZA CENTRALE

Se ne conclude che il punto stazionario

$$x = \frac{y_1 + y_2 + \cdots + y_N}{N} \equiv M_2,$$

della rilevazione dati, è il punto di minimo assoluto sia per la funzione f che per la funzione f_2: allora $c_2\left(\underline{y}\right)$ coincide con la media aritmetica.

Il $\lim_{r\to+\infty} c_r\left(\underline{y}\right) \equiv c_\infty\left(\underline{y}\right)$ è detto centro di ordine infinito; per esso sussiste il seguente risultato.

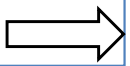
TEOREMA 4 (senza dimostrazione)
Il centro di ordine infinito della rilevazione dati $\underline{y}$ *coincide con il valore centrale:* $c_\infty\left(\underline{y}\right) = \frac{x_1 + x_k}{2}$.

STATISTICA DESCRITTIVA: INDICI DI DISPERSIONE

Un *indice di dispersione* (o indicatore di dispersione o indice di variabilità o indice di variazione) serve per descrivere sinteticamente la misura con la quale una rilevazione dati di un carattere quantitativo è distante dalla sua tendenza centrale.

La dispersione esprime la bontà o la inadeguatezza di un indice di tendenza centrale quale descrittore di una distribuzione di frequenza.

Per i caratteri qualitativi si usano gli *indici di diversità* dei quali quello maggiormente usato è l'*indice di ricchezza* che opera un semplice conteggio del numero delle modalità presenti nella rilevazione dati.



STATISTICA DESCRITTIVA: INDICI DI DISPERSIONE

Sia $\underline{y} = (y_1, y_2, \cdots, y_N)$ una rilevazione dati su un carattere quantitativo e siano $x_1 < x_2 < \cdots < x_k$ le modalità del carattere.

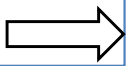
Sono indici di dispersione:

a) la differenza tra la modalità più grande da quella più piccola (*campo o intervallo di variazione*):

$$\Gamma = x_k - x_1;$$

a) la differenza tra il terzo e il primo quartile (*differenza interquartilica*):

$$\gamma = Q_3 - Q_1;$$



STATISTICA DESCRITTIVA: INDICI DI DISPERSIONE

c) la media aritmetica del valore assoluto delle differenze dei dati dalla loro media aritmetica M_2 (*scarto medio assoluto*):

$$S_{M_2} = \frac{1}{N}\sum_{i=1}^{N}|y_i - M_2|$$

(nella quale si usa la trasformazione $T(x) = |x - M_2|$);

d) la media aritmetica del valore assoluto delle differenze dei dati dalla loro mediana M_1 (*scarto mediano assoluto*):

$$S_{M_1} = \frac{1}{N}\sum_{i=1}^{N}|y_i - M_1|\,;$$

(nella quale si usa la trasformazione $T(x) = |x - M_1|$); ⟹

STATISTICA DESCRITTIVA: INDICI DI DISPERSIONE

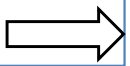
e) la media del quadrato delle differenze dei dati dalla loro media aritmetica M_2 (*varianza*):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - M_2)^2 \geq 0;$$

(si usa la trasformazione $T(x) = (x - M_2)^2$;

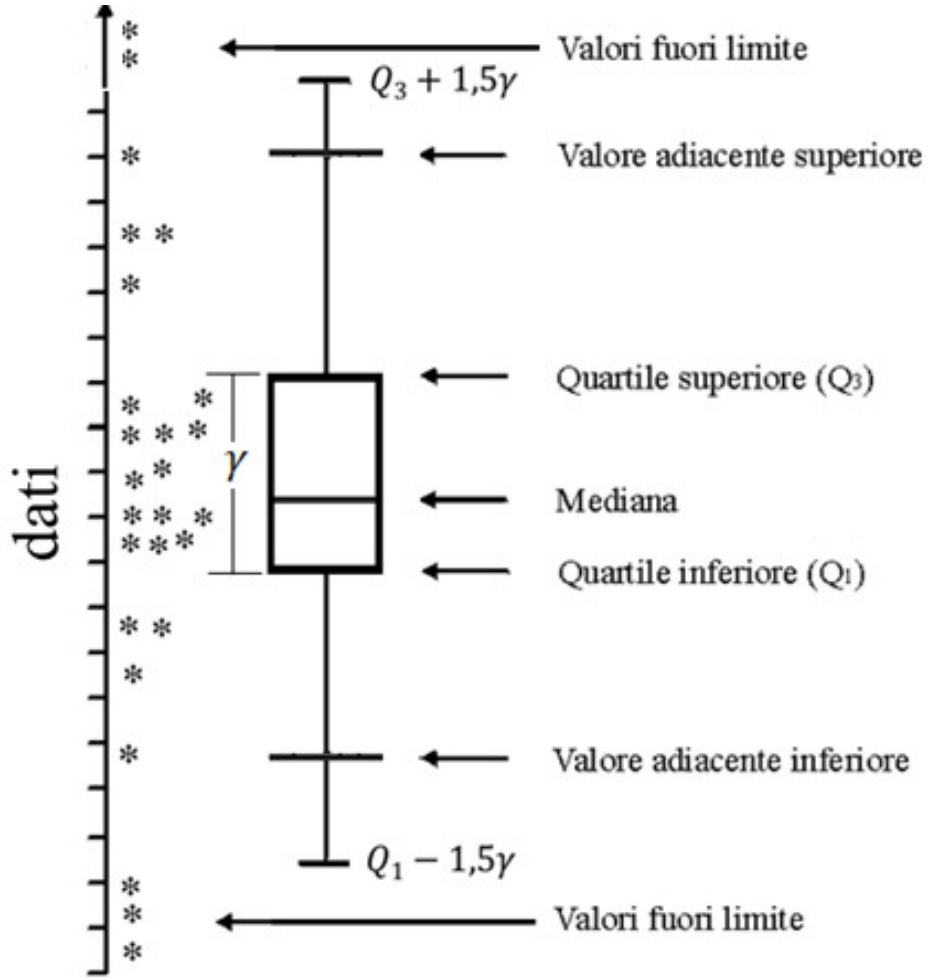
f) la radice quadrata della varianza (*scarto tipo* o *deviazione standard*):

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - M_2)^2}\,.$$



STATISTICA DESCRITTIVA: INDICI DI DISPERSIONE

Un metodo grafico per rappresentare una distribuzione di frequenze che mette in risalto anche la dispersione intorno alla mediana è la *scatola con baffi* (*boxplot*) di Tukey:



STATISTICA DESCRITTIVA: INDICI DI DISPERSIONE

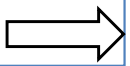
La linea interna alla scatola rappresenta la *Mediana* della distribuzione.

Le linee estreme della scatola rappresentano il primo ed il terzo quartile.

La distanza interquartilica γ, è una misura della **dispersione** della distribuzione. Tra questi due valori si trova una metà dei dati. Se l'intervallo interquartilico è piccolo, tale metà delle osservazioni si trova fortemente concentrata intorno alla mediana; all'aumentare della distanza interquartilica aumenta la dispersione del 50% dei dati centrali intorno alla mediana.

Le distanze tra ciascun quartile e la mediana forniscono informazioni relativamente alla **forma** della distribuzione. Se una distanza è diversa dall'altra allora la distribuzione è asimmetrica.

Le linee che si allungano dai bordi della scatola (*baffi*) individuano gli intervalli in cui sono posizionati i valori rispettivamente minori di Q_1 e maggiori di Q_3; i punti estremi dei "baffi" evidenziano i *valori adiacenti*.



STATISTICA DESCRITTIVA: INDICI DI DISPERSIONE

Il *valore adiacente inferiore* (VAI) è il valore più piccolo tra i dati che risulta maggiore o uguale a $Q_1 - 1{,}5\gamma$.

Il *valore adiacente superiore* (VAS), invece, è il valore più grande tra i dati che risulta minore o uguale a $Q_3 + 1{,}5\gamma$.

I valori esterni ai valori adiacenti (chiamati in genere *valori anomali*), vengono segnalati individualmente nel grafico scatola con baffi per meglio evidenziarne la presenza e la posizione. Questi valori infatti costituiscono una "anomalia" rispetto alla maggior parte dei valori osservati e pertanto è necessario identificarli per poterne analizzare le caratteristiche e le eventuali cause che li hanno determinati. Essi forniscono informazioni ulteriori sulla dispersione e sulla forma della distribuzione.

Quando il valore adiacente superiore coincide con il dato più grande e il valore adiacente inferiore coincide con il dato più piccolo, allora non comparirà alcun valore anomalo.