DEFINIZIONI E CLASSIFICAZIONE DEI CARATTERI

Con il termine statistica descrittiva si intende un insieme di tecniche e strumenti finalizzati ad assolvere uno dei principali compiti assegnati alla Statistica:

> *descrivere, rappresentare e sintetizzare in maniera opportuna un insieme di dati relativamente ad una o più* caratteristiche *di una* popolazione *di interesse.*

Per *popolazione* (o collettivo statistico) si intende la totalità dei casi, ovvero dei *membri* (o unità statistiche) sui quali e possibile *rilevare* uno o più *caratteri* che rivestono particolare importanza per il fenomeno che si sta studiando.

Esempio 1
Durante il semestre viene proposto agli studenti dei corsi di Laurea offerti in Ateneo il questionario sulla valutazione della struttura didattica nella sua complessità.

Qui la *popolazione* è costituita dagli studenti con o senza obbligo di frequenza ciascuno dei quali è un *membro*. Per lo studio in questione ci si serve di un *questionario* avente un certo numero di domande raggruppate per sezioni. In questo contesto ciascuna domanda del questionario corrisponde ad un *carattere*. Per le maggior parte delle domande lo studente può rispondere scegliendo una tra 4 possibili risposte tra loro alternative:

Decisamente no, *Più no che sì*, *Più sì che no*, *Decisamente sì.*

Per completezza si riportano alcune di queste domande.

Le aule dove si svolgono le lezioni sono adeguate? Le modalità con le quali si è svolto l'insegnamento (lezioni, diapositive, audiovisivi, ecc.) sono soddisfacenti? Il carico di studio dell'insegnamento è proporzionale ai crediti assegnati? Il docente è reperibile per chiarimenti e spiegazioni?

◊

Esempio 2
In un allevamento di bufale da latte si vuole mettere in relazione la produzione giornaliera di latte con la grandezza delle mammelle e con lo stato di salute delle stesse.

Qui la *popolazione* è costituita da tutte le bufale dell'allevamento ciascuna delle quali è un *membro*. Per lo studio in questione i *caratteri* da *rilevare* su ciascun membro sono: la produzione di latte, una misura lineare delle mammelle, lo stato di salute delle mammelle. Per la rilevazione sono necessari due *strumenti di misura* e la diagnosi di un veterinario determinata su una prestabilita *scala* di valori.

◊

Le diverse espressioni con le quali si manifesta un carattere si chiamano *modalità*.

Nell'Esempio 1 le modalità, per ciascun carattere, sono rappresentate dalle 4 risposte alternative tra loro. Ciascuno dei caratteri, ovvero una domanda proposta nel questionario, è di tipo *qualitativo ordinale*. Infatti, si osservi che esse sono delle etichette e che in relazione al gradimento espresso dallo studente risulta:

Decisamente no < Più no che sì < Più sì che no < Decisamente sì.

Invece nell'Esempio 2, i caratteri "produzione giornaliera di latte" e "grandezza delle mammelle" sono *quantitativi continui* in quanto per la loro determinazione sono necessari uno strumento di misurazione di una capacità (volume) e uno strumento per la misurazione di una lunghezza e pertanto i dati ottenuti sono numeri decimali appartenenti ad un conveniente intervallo. Nello stesso esempio, il carattere "stato di salute delle mammelle" è invece di tipo qualitativo ordinale e le modalità sono i diversi valori della scala prescelta (ad esempio, mammelle sane, infiammazione lieve, infiammazione moderata, infiammazione grave). Il fatto stesso che si parla di scala comporta la presenza di un ordinamento tra le etichette:

sana < infiammazione lieve < infiammazione moderata < infiammazione grave.

Per un esempio di carattere *qualitativo nominale* si pensi al *gruppo sanguigno* che, prescindendo dal fattore Rh, si manifesta con le modalità: A, B, AB e 0.

Per un esempio di carattere *quantitativo discreto* si pensi al numero delle persone presenti nello stato di famiglia dei residenti nel comune di Napoli alla data del più recente censimento ISTAT.

DISTRIBUZIONI DI FREQUENZA

Si può senz'altro pensare ad una popolazione come ad un insieme. La cardinalità della popolazione rilevata è detta *taglia*; essa, di solito, si designa con la lettera N. Un carattere viene designato con una lettera latina maiuscola mentre i valori rilevati vengono rappresentati con la stessa lettera ma in minuscolo. Se, allora, Y rappresenta il carattere sotto studio non continuo, la sequenza

$$\underline{y} = (y_1, y_2, \cdots, y_N)$$

rappresenta l'intera rilevazione dei dati. Denotiamo ora con $x_1,\ x_2,\ \cdots, x_k$ (con $k \leq N$) le modalità del carattere Y. Per $j = (1,\ 2,\ \cdots,\ k)$, il numero n_j che rappresenta quante volte è presente la modalità x_j in $\underline{y}$ è detto *frequenza assoluta* (o semplicemente *frequenza*) della modalità x_j. In aggiunta,

$$j = (1,2,\cdots,k), \qquad f_j = \frac{n_j}{N}$$

è detto *frequenza relativa* della modalità x_j.

La frequenza relativa è più informativa della frequenza assoluta in quanto tiene conto anche della taglia. È del tutto ovvio che la somma delle k frequenze assolute vale N mentre la somma delle k frequenze relative vale 1.

Infine,

$$j = (1,2,\cdots,k), \qquad F_j = f_1 + f_2 + \cdots + f_j = F_{j-1} + f_j$$

è detta *frequenza assoluta cumulata* di x_j. È del tutto ovvio che $F_k = 1$.

La rappresentazione tabellare di quanto appena esposto è detta *distribuzione di frequenza* della rilevazione dati $\underline{y} = (y_1,\ y_2,\ \cdots, y_N)$:

modalità del carattere	frequenza assoluta	frequenza relativa	frequenza relativa cumulata
x_1	n_1	$f_1 = n_1 / N$	$F_1 = f_1$
x_2	n_2	$f_1 = n_1 / N$	$F_2 = F_1 + f_2$
x_{k-1}	n_{k-1}	$f_{k-1} = n_{k-1} / N$	$F_{k-1} = F_{k-2} + f_{k-1}$
x_k	n_k	$f_k = n_k / N$	1
	———	———	
	N	1	

Per un carattere quantitativo continuo è necessario dapprima procedere alla suddivisione in *classi di modalità* dell'intervallo nel quale si manifesta il carattere stesso. Qui bisogna fare attenzione a rendere le classi contigue ma senza ingenerare dubbi di collocazione di un dato nelle classi stesse. Allo scopo, se i dati $(y_1,\ y_2,\ \cdots, y_N)$ sono riportati con s cifre decimali, è sufficiente tenere conto dell'operazione di arrotondamento e rappresentare gli estremi delle classi con $s + 1$ cifre decimali. Ad esempio, supponiamo che una rilevazione di un peso è effettuata con bilancia digitale precisa all'ettogrammo. Sia 3,2 kg il peso minore rilevato: 3,2 rappresenta tutte le misurazioni comprese nell'intervallo $[3{,}15;\ 3{,}25[$. Allo stesso modo sia 4,4 kg il peso maggiore rilevato: 4,4 rappresenta tutte le misurazioni comprese nell'intervallo $[4{,}35;\ 4{,}45[$. Quindi se è vero che l'intervallo nel quale si osservano i dati è $[3{,}2;\ 4{,}4]$ è a maggior ragione vero che senza l'operazione di arrotondamento esso sarebbe stato $[3{,}15;\ 4{,}45[$. Allora, è quest'ultimo intervallo che deve essere suddiviso nel numero desiderato di classi e queste devono avere come estremi dei numeri aventi due cifre decimali. Dopo di ciò, per un carattere quantitativo, la prima colonna contiene le classi di modalità così individuate.

MODA E QUARTILI

Definizione
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.

◊

Definizione
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) oppure *secondo quartile* (Q_2) della rilevazione dati. La mediana suddivide la rilevazione dati ordinata

$$\left(y_{(1)}, y_{(2)}, \cdots, y_{(N)}\right) \quad \text{con} \quad y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$$

in due parti: i dati minori della mediana sono nello stesso numero dei dati maggiori della mediana.

◊

Definizione
Si consideri un carattere (non qualitativo nominale). La modalità corrispondente alla più piccola frequenza relativa cumulata maggiore o uguale a $0{,}25$ viene detta *primo quartile* (Q_1) della rilevazione dati. Il primo quartile corrisponde anche alla mediana dei dati minori della mediana.

◊

Definizione
Si consideri un carattere (non qualitativo nominale). La modalità corrispondente alla più piccola frequenza relativa cumulata maggiore o uguale a $0{,}75$ viene detta *terzo quartile* (Q_3) della rilevazione dati. Il terzo quartile corrisponde anche alla mediana dei dati maggiori della mediana.

◊

Definizione
Si consideri un carattere (non qualitativo nominale). Il dato più piccolo $y_{(1)}$ è il *quartile di ordine zero* (Q_0) della rilevazione dati.

◊

Definizione
Si consideri un carattere (non qualitativo nominale). Il dato più grande $y_{(N)}$ è il *quarto quartile* (Q_4) della rilevazione dati.

◊

RAPPRESENTAZIONI GRAFICHE

Le distribuzioni di frequenza possono essere rappresentate in forma grafica con scelta eseguita opportunamente rispetto al tipo di carattere. Per i caratteri qualitativi e quantitativi discreti, oltre al *diagramma circolare*, la rappresentazione grafica più usata è quella del *diagramma a barre verticali*: ogni barra verticale è centrata attorno ad una modalità ed ha altezza pari alla frequenza assoluta (oppure relativa).

Per i soli caratteri quantitativi continui è opportuno utilizzare la rappresentazione grafica detta *istogramma*.

La differenza qualitativa tra un istogramma ed un diagramma a barre verticali è che nell'istogramma le barre verticali devono essere contigue (in effetti sono dei rettangoli e, quindi, dotati di base e altezza). Ma la differenza sostanziale è che nell'istogramma la frequenza della classe di modalità rappresenta l'area del rettangolo (e non l'altezza come nel diagramma a barre).

Pertanto, per qualsiasi $j \in \{1,\ 2,\ \cdots,\ k\}$, se b_j rappresenta l'ampiezza della j-ma classe di modalità, l'altezza h_j del relativo rettangolo è ottenuto mediante la formula inversa per l'area e quindi:

$$j = (1,2,\cdots,k), \qquad h_j = \frac{n_j}{b_j}.$$

Alcuni esempi di rappresentazioni grafiche con ulteriori dettagli sono stati forniti nella dispensa Statistica_Descrittiva_1.

INDICI DI POSIZIONE

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 quantitativi è possibile fare ricorso ai due indici media e mediana. Nel caso di 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. 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 indeterminata.

Per i caratteri quantitativi ci sono almeno altri due approcci teorici in grado di far ottenere indici di tendenza centrale:

- le *medie analitiche*;
- i *centri*.

MEDIE ANALITICHE

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 $\underline{y}$ deve coincidere con la valutazione di C su $\underline{y}^*$. In simboli:

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

Esempio: *media aritmetica*
Si scelga come funzione di circostanza C la somma dei dati:

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

Dopo di ciò,

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

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

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

In definitiva, la soluzione dell'equazione di circostanza è la *media aritmetica* dei dati:

$$\bar{y} = \frac{y_1 +\ y_2 +\ \cdots + y_N}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

◊

Esempio: *media geometrica*
Si scelga come funzione di circostanza C il prodotto dei dati (che devono essere rilevati da un carattere positivo):

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

Dopo di ciò,

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

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

In definitiva, la soluzione dell'equazione di circostanza è la *media geometrica* dei dati:

$$M_g = \sqrt[N]{y_1 \cdot y_2 \cdots y_N} = \sqrt[N]{\prod_{i=1}^{N} y_i}.$$

◊

Esempio: *media armonica*

Si scelga come funzione di circostanza C la somma dei reciproci dei dati (che devono essere da un carattere non nullo):

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

Dopo di ciò,

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

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

$$\Leftrightarrow \left(\frac{1}{y_1} + \frac{1}{y_2} + \ \cdots + \frac{1}{y_N}\right) = \frac{N}{y}.$$

In definitiva, la soluzione dell'equazione di circostanza è la *media armonica* dei dati:

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

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

Si può allora dire che *la media armonica M_a di dati non nulli è uguale al reciproco della media aritmetica dei loro reciproci*.

◊

CENTRI

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

$$\forall x \in \mathbb{R}, \quad d\left(x, \underline{y}\right) \geq 0$$

una funzione che si ritiene adatta a rappresentare la *distanza* di un generico valore reale x da tutti gli elementi della rilevazione dati $\underline{y}$. Si definisce *centro* di una rilevazione dati $\underline{y}$, e lo si indica con $\xi\left(\underline{y}\right)$, il punto di minimo assoluto della funzione $d\left(x, \underline{y}\right)$; in simboli:

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

In particolare, sono molto spesso considerate le seguenti funzioni di distanza di *tipo potenze*:

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

e

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

La funzione $d_r\left(x, \underline{y}\right)$ è detta anche *distanza di ordine* r e il suo punto di minimo assoluto è detto *centro di ordine* r. Quindi il centro di ordine r di una rilevazione dati $\underline{y}$ è il numero reale $\xi_r\left(\underline{y}\right)$ che rende minima la distanza di ordine r. In simboli:

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

Teorema 1

Il centro di ordine 0 *della rilevazione dati* $\underline{y}$*, ovvero* $\xi_0\left(\underline{y}\right)$*, coincide con la moda della rilevazione dati.*

Dimostrazione

Per definizione, la funzione

$$d_0\left(x, \underline{y}\right) = \frac{1}{N}\sum_{i=1}^{N} |x - y_i|^0$$

rappresenta la distanza di ordine 0 di x dall'intera rilevazione dati $\underline{y}$, mentre $|x-y_i|^0$ rappresenta la distanza tra x e il generico dato y_i.

Se ne ricava che quando x coincide con y_i la distanza è nulla ovvero l'addendo i-mo non porta contributo alla distanza complessiva. Pertanto, si ha:

$$d_0\left(x,\underline{y}\right)=\begin{cases}1, & \text{se } x\notin\{y_1,y_2,\cdots,y_n\},\\ 1-\dfrac{n_x}{N}<1, & \text{se } x\in\{y_1,y_2,\cdots,y_n\}.\end{cases}$$

Nella precedente formula n_x rappresenta il numero delle volte che si presenta il dato x. Allora, il minimo si trova tra gli elementi di $\underline{y}$ e precisamente è quel dato al quale compete la frequenza maggiore che per definizione è la moda della rilevazione dati:

$$\xi_0\left(\underline{y}\right)=\underset{x\in\mathbb{R}}{\operatorname{argmin}}\frac{1}{N}\sum_{i=1}^{N}|x-y_i|^0=M_0.$$

Teorema 2 (senza dimostrazione)

Il centro di ordine 1 *della rilevazione dati* $\underline{y}$*, ovvero* $\xi_1\left(\underline{y}\right)$*, coincide con la mediana:*

$$\xi_1\left(\underline{y}\right)=\underset{x\in\mathbb{R}}{\operatorname{argmin}}\,d_1\left(x,\underline{y}\right)=\underset{x\in\mathbb{R}}{\operatorname{argmin}}\frac{1}{N}\sum_{i=1}^{N}|x-y_i|=M_1\equiv Q_2.$$

In più, il valore minimo assunto dalla distanza di ordine 1 vale:

$$d_1\left(Q_2,\underline{y}\right)=\frac{1}{N}\sum_{i=1}^{N}|Q_2-y_i|.$$

◊

Teorema 3

Il centro di ordine 2 *della rilevazione dati* $\underline{y}$*,* $\xi_2\left(\underline{y}\right)$*, coincide con la media aritmetica* $\bar{y}$*. In più, il valore minimo assunto dalla distanza di ordine 2:*

$$d_2\left(\bar{y},\underline{y}\right)=\frac{1}{N}\sum_{i=1}^{N}(\bar{y}-y_i)^2.$$

Dimostrazione

Per definizione,

$$d_2\left(x,\underline{y}\right) = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{N}|x-y_i|^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x-y_i)^2}.$$

D'altra parte, dal momento che la funzione radice quadrata è strettamente crescente nel suo dominio $[0,+\infty[$, il minimo della distanza di ordine due viene raggiunto in corrispondenza del minimo del radicando, ovvero della funzione:

$$x \in \mathbb{R}, \quad f(x) = \frac{1}{N}\sum_{i=1}^{N}(x-y_i)^2.$$

Pertanto,

$$\xi_2\left(\underline{y}\right) = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\, d_2\left(x,\underline{y}\right) = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\, f(x).$$

Ricerca dei punti stazionari di f(x):

$$x \in \mathbb{R}, \quad f'(x) = \frac{2}{N}\sum_{i=1}^{N}(x-y_i) \Longrightarrow \quad f'(x) = 0 \Leftrightarrow \sum_{i=1}^{N}(x-y_i) = 0$$

$$\Leftrightarrow x = \frac{1}{N}\sum_{i=1}^{N} y_i \equiv \bar{y}.$$

Ricerca dei punti di minimo e massimo relativo di $f(x)$:

$$f''(x) = \frac{2}{N}\sum_{i=1}^{N} 1 = 2 \Longrightarrow \quad f''(\bar{y}) = 2 > 0,$$

la qual cosa implica che $\bar{y}$ è un punto di minimo relativo per $f(x)$.

Ricerca del minimo assoluto di $f(x)$:

$$\bar{y} = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\, f(x)$$

in quanto $\lim_{x\to+\infty} f(x) = \lim_{x\to-\infty} f(x) = +\infty$, la derivata prima di $f(x)$ è definita in $\mathbb{R}$, e $f(x)$ ammette un unico punto di minimo relativo.

In definitiva,

$$\xi_2\left(\underline{y}\right) = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\, d_2\left(x,\underline{y}\right) = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\, f(x) = \bar{y}.$$

Definizione

Il *centro di ordine infinito* di una rilevazione dati $\underline{y}$ si indica con il simbolo $\xi_\infty\left(\underline{y}\right)$ ed è definito dalla posizione:

$$\xi_\infty\left(\underline{y}\right) \coloneqq \lim_{r\to+\infty} \xi_r\left(\underline{y}\right).$$

◊

Definizione

Il *valore centrale* di una rilevazione dati $\underline{y}$ è la semisomma tra il dato minimo e il dato massimo:

$$\frac{y_{(1)} + y_{(N)}}{2} = \frac{Q_{(0)} + Q_{(4)}}{2}$$

◊

Teorema (senza dimostrazione)

Il centro di ordine infinito della rilevazione dati $\underline{y}$ coincide con il valore centrale:

$$\xi_\infty\left(\underline{y}\right) = \frac{y_{(1)} + y_{(N)}}{2}.$$

◊

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 da una 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.

Sia $\underline{y} = (y_1,\ y_2,\ \cdots, y_N)$ una rilevazione dati su un carattere quantitativo. Sono indici di dispersione i seguenti.

a) La differenza tra il dato più grande da quello più piccolo (*campo o intervallo di variazione*):

$$\Gamma = y_{(N)} - y_{(1)} = Q_4 - Q_0.$$

Il campo di variazione è associato al valore centrale $\frac{y_{(1)}+y_{(N)}}{2}$.

b) La differenza tra il terzo e il primo quartile (*differenza interquartilica*):

$$\gamma = Q_3 - Q_1.$$

La differenza interquartilica è associata alla mediana Q_2.

c) La media del valore assoluto delle differenze dei dati dalla loro mediana Q_2 (*scarto mediano assoluto*):

$$S_{Q_2} = \frac{1}{N}\sum_{i=1}^{N}|y_i - Q_2|.$$

Lo scarto mediano assoluto è associato alla mediana Q_2: si veda il Teorema 2 in questa stessa dispensa.

d) La media del valore assoluto delle differenze dei dati dalla loro media aritmetica $\bar{y}$ (*scarto medio assoluto*):

$$S_{\bar{y}} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}|.$$

Lo scarto medio assoluto è associato alla media aritmetica $\bar{y}$.

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

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2.$$

La varianza è associata alla media aritmetica $\bar{y}$: si veda il Teorema 3 in questa stessa dispensa

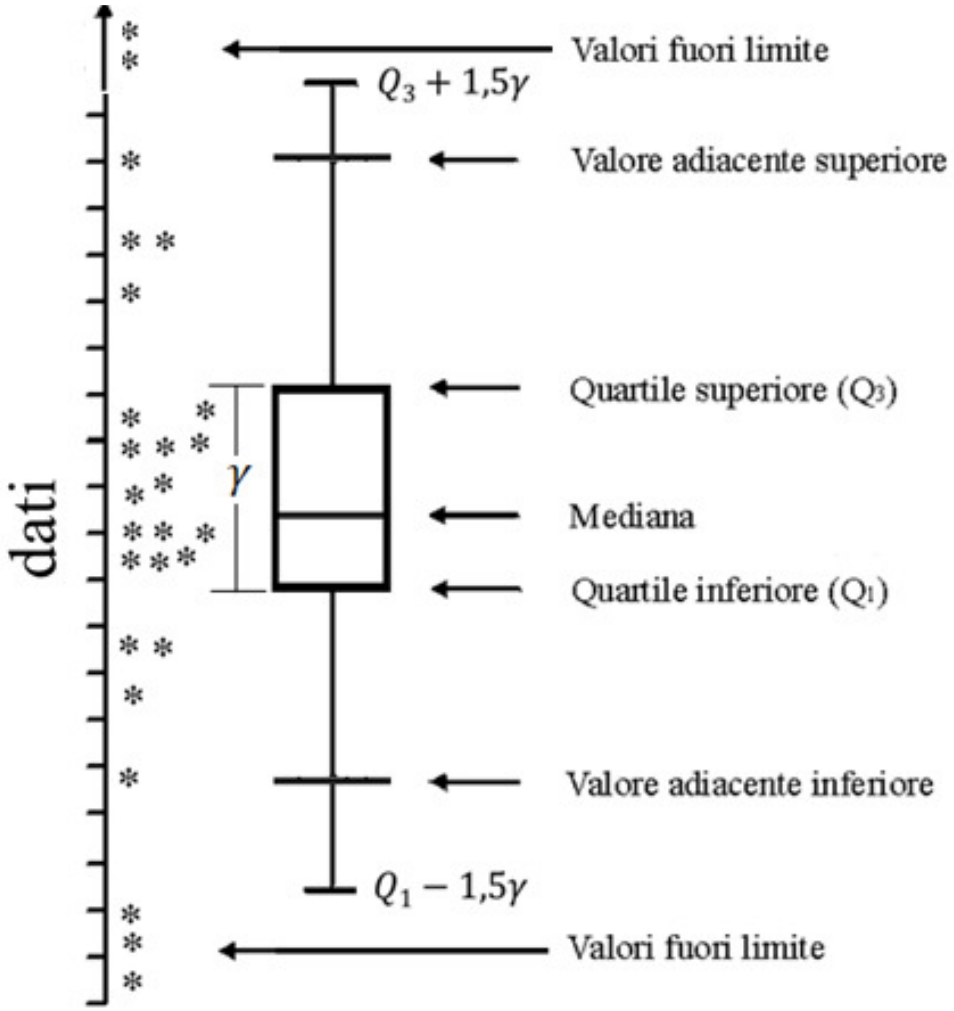
f) La radice quadrata della varianza (*scarto tipo* o *deviazione standard*):

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

La deviazione standard è associata alla media aritmetica $\bar{y}$.

DIAGRAMMA SCATOLA CON BAFFI

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



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. Il 50% dei dati si trovano comprese tra questi due valori. 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*. 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 fuori limiti oppure valori anomali*), vengono segnalati individualmente nel box-plot 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.