

RIASSUNTI NUMERICI DELLE CARATTERISTICHE DI UN INSIEME DI DATI

Un'attenta osservazione della distribuzione di frequenza di una variabile statistica permette di trarre informazioni su gran parte delle caratteristiche dell'insieme di dati. In particolare, possiamo:

1. cogliere informazioni sui valori tipici della distribuzione
2. avere informazioni sulla dispersione dei valori

Già tracciando un istogramma si attua un riassunto delle caratteristiche salienti di un insieme di dati. L'obiettivo ora è quello di avere riassunti ancora più sintetici delle caratteristiche principali.

INDICI DI TENDENZA CENTRALE

Occorre, come sempre, tener conto della natura delle variabili (quantitative, qualitative).

Iniziamo ad affrontare il caso di maggiore interesse relativo alle variabili QUANTITATIVE.

L'idea di fondo è quella di cercare un valore attorno al quale è concentrata l'intera distribuzione di valori: un valore “tipico” che riassume l'intero fenomeno con riferimento al suo ordine di grandezza.

COME SCEGLIERE TALE VALORE?

Le possibili soluzioni sono numerose:

Ad esempio,

- il valore intermedio tra il valore più piccolo e il valore più grande fra quelli osservati
- il valore osservato con maggiore frequenza
- un opportuno riassunto di alcuni valori compresi nel corpo centrale della distribuzione

LA MEDIANA

Una prima idea è quella di scegliere quale valore tipico, indice di tendenza centrale, quel valore che è al centro della sequenza dei dati ordinati.

ESEMPIO: Altezze rilevate su un insieme di 15 donne (in cm)

159, 156, 162, 154, 142, 166, 161, 163, 158, 164, 159,
164, 170, 168, 152

↓

mettiamo i valori in ordine crescente

1 ^o	2 ^o	3 ^o	4 ^o	5 ^o	6 ^o	7 ^o	8 ^o	9 ^o	10 ^o	11 ^o
142,	152,	154,	156,	158,	159,	159,	161,	162,	163,	164
12 ^o	13 ^o	14 ^o	15 ^o							
164,	166,	168,	170							

↓

MEDIANA

Il valore che occupa l'ottavo posto nella graduatoria è al centro della sequenza. La mediana è cioè pari a 161 cm.

N.B.: Di solito i valori ordinati della sequenza sono indicati con $x_{(i)}$. Quindi, $x_{(1)} = 142$, $x_{(2)} = 152$, $\dots$, $x_{(15)} = 170$.

In generale, la mediana è quel valore che separa la distribuzione in due parti in modo tale che metà dei valori sono inferiori alla mediana e metà sono superiori ad essa.

Se abbiamo un insieme di dati relativo a N unità e N è dispari, la mediana è il valore che occupa la posizione

$$\frac{N + 1}{2}$$

nella sequenza ordinata delle osservazioni, ossia

$$\text{Me} = x_{(\frac{N+1}{2})}$$

Cosa facciamo se N è pari?

Esempio: Altezze relative a 8 maschi (in centimetri)

188, 178, 172, 164, 171, 188, 174, 179

riordino i valori

1 ^o	2 ^o	3 ^o	4 ^o	5 ^o	6 ^o	7 ^o	8 ^o
164,	171,	172,	174,	178,	179,	186,	188
			↑↑				

Se ci chiediamo quale dei valori osservati ha la proprietà di separare in due la distribuzione osserviamo che questo non accade per nessuno di essi. Infatti, 174 cm, ad esempio, ha 4 osservazioni più grandi e solo 3 più piccole. Non è centrale. Per il valore 178 cm, vi sono 3 osservazioni più grandi e 4 osservazioni più piccole. Qualsiasi valore numerico compreso tra 174 cm e 178 cm ha invece la proprietà desiderata. Si può quindi convenire di scegliere quale valore mediano la semisomma dei due valore 174 e 178, ossia 176.

Quindi, se N è pari, la mediana è qualsiasi valore compreso tra i 2 valori che nella sequenza ordinata occupano le posizioni

$$\frac{N}{2} \quad \text{e} \quad \frac{N}{2} + 1$$

Usualmente si pone

$$\text{Me} = \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2}$$

Cosa fare quando i dati sono già accorpati in una tabella di frequenza?

Se si tratta di una variabile quantitativa discreta (con “poche” modalità) è presumibile che alcuni dei valori siano ripetuti, ossia in corrispondenza di ogni modalità vi è una frequenza osservata.

Esempio: Numero di stanze di un campione di abitazioni

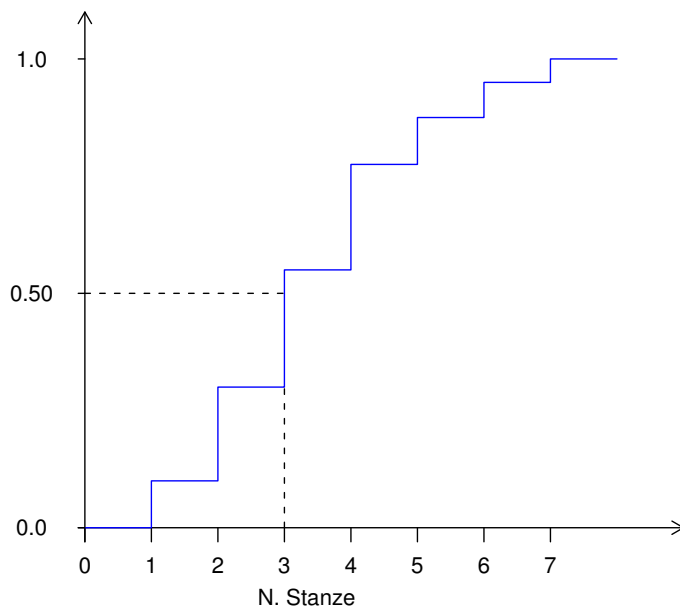
Num. stanze	n_j	f_j	N_j	F_j	
1	40	0,1	40	0,1	
2	80	0,2	120	0,3	
3	100	0,25	220	0,55	$\Longleftarrow$ Me
4	90	0,225	310	0,775	
5	40	0,1	350	0,875	
6	30	0,075	380	0,95	
7 o più	20	0,05	400	1	

$$N = 400$$

La mediana è quel valore che lascia alla sua sinistra il 50% delle unità, ossia tale che

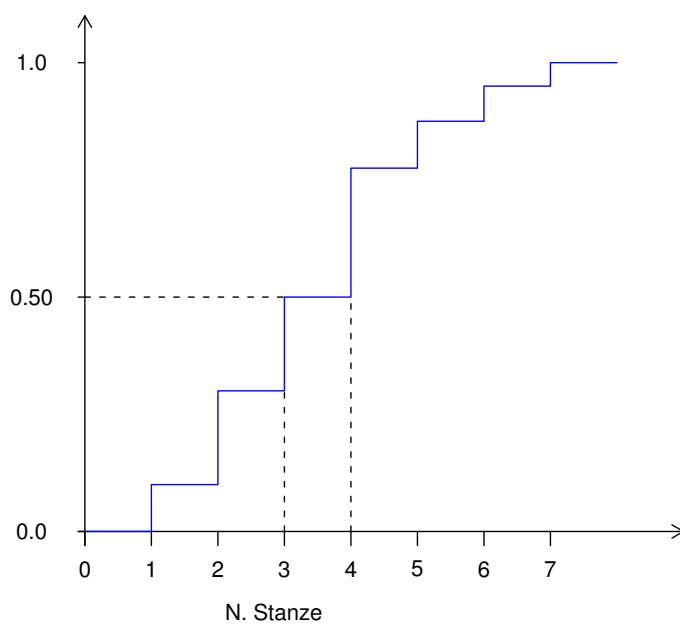
$$F(\text{Me}) = 0,5$$

Allora, costruiamo la funzione di frequenza cumulata e vediamo in corrispondenza di quale valore essa risulta pari a 0,5



La mediana è pari a 3.

Se, invece, i dati fossero tali che



allora, qualunque valore compreso tra 3 e 4 potrebbe essere la nostra mediana. Tipicamente si prende il valore intermedio $(3+4)/2=3,5$.

Possiamo da questi grafici derivare **una semplice regola** per determinare la mediana per dati discreti organizzati in una tabella di frequenza

Me è la più piccola tra le modalità a cui corrisponde $F_j > 0,5$.

Se per una modalità x_j si ha $F_j = 0,5$, allora la mediana Me è qualunque valore compreso tra x_j e x_{j+1} (la modalità immediatamente più grande di x_j). Ad esempio, possiamo prendere

$$\text{Me} = \frac{x_j + x_{j+1}}{2}$$

CALCOLO DELLA MEDIANA PER VARIABILI (CONTINUE O DISCRETE) RAGGRUPPATE IN CLASSI

Si tratta di determinare quel valore di X tale che il 50% delle osservazioni risultino inferiori ad esso.

Sotto ipotesi di distribuzione uniforme all'interno delle classi:

Classi	Freq. rel. cum.
$x_o \vdash x_1$	F_1
$\dots$	$\dots$
$x_{j-1} \vdash x_j$	$F_j < 0,5$
$x_j \vdash x_{j+1}$	$F_{j+1} > 0,5 \quad \Leftarrow \text{CLASSE MEDIANA}$
$\dots$	$\dots$
$x_{r-1} \vdash x_r$	1

E' noto che per x nella classe $x_j \vdash x_{j+1}$

$$F(x) = F_j + \frac{(x - x_j)}{(x_{j+1} - x_j)} \times (F_{j+1} - F_j)$$

e quindi

$$0,5 = F_j + \frac{(\text{Me} - x_j)}{(x_{j+1} - x_j)} \times (F_{j+1} - F_j)$$

Risolvendo la precedente equazione rispetto a Me si ottiene

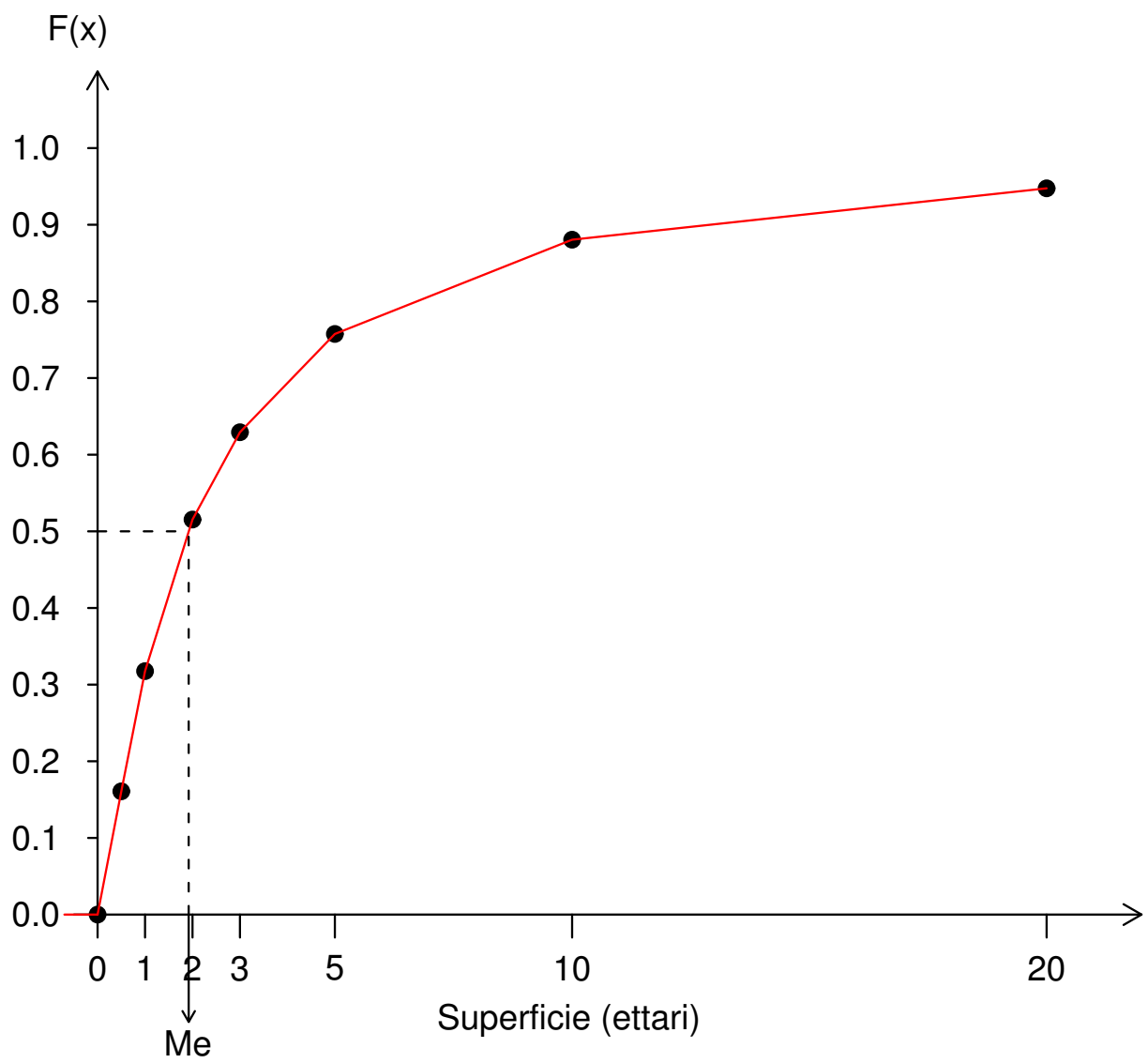
$$\text{Me} = x_j + \underbrace{\frac{(0,5 - F_j)}{(F_{j+1} - F_j)}}_{\text{frazione ampiezza classe mediana}} \times (x_{j+1} - x_j)$$

frazione ampiezza classe mediana

Esempio: Superfici delle aziende agricole (Tabella 2)

X	F_j	
$\dots$	$\dots$	
$0,5 \vdash 1$	0,3176	
$1 \vdash 2$	0,5154	$\Leftarrow$ CLASSE MEDIANA
$\dots$	$\dots$	

$$\text{Me} = 1 + \frac{(0,5 - 0,3176)}{(0,5154 - 0,3176)} \times (2 - 1) = 1,92 \text{ ettari}$$



Me=1,92 ettari

PROPRIETA' DELLA MEDIANA

Per ogni unità statistica possiamo calcolare

$$x_i - \text{Me} \quad i = 1, \dots, N$$

ossia lo scarto dalla mediana.

La mediana è quel valore tale che

$$\sum_{i=1}^N |x_i - \text{Me}| = \min$$

La somma degli scarti dalla mediana presi in valore assoluto (trascurando il segno) è la più piccola possibile. Se invece della mediana prendessimo un valore riassuntivo diverso, avremmo un valore più alto della somma degli scarti assoluti.

LA MEDIA ARITMETICA

E' senza dubbio l'indice di tendenza centrale più noto e più frequentemente utilizzato.

Nel seguito verrà indicata alternativamente con i seguenti simboli: $\bar{x}$ o M.

Media aritmetica semplice (per dati in forma grezza)

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

PROPRIETA' DELLA MEDIA ARITMETICA

Considerando per ogni osservazione lo scarto dalla media:

$$x_i - \bar{x} \quad i = 1, \dots, N$$

si ha

$$\sum_{i=1}^N (x_i - \bar{x})^2 = \min$$

ovvero la somma degli scarti al quadrato dalla media ha il valore più piccolo. Se considerassimo la stessa somma, per gli scarti da un valore diverso da $\bar{x}$ (ad esempio, Me) otterremmo un valore più elevato.

Inoltre, la somma degli scarti dalla media è nulla:

$$\begin{aligned}\sum_{i=1}^N (x_i - \bar{x}) &= \sum_{i=1}^N x_i - N\bar{x} = \sum_{i=1}^N x_i - N \frac{\sum_{i=1}^N x_i}{N} \\ &= \sum_{i=1}^N x_i - \sum_{i=1}^N x_i = 0\end{aligned}$$

La media attua un bilanciamento tra scarti positivi e scarti negativi. In tal senso, la media è una sorta di baricentro della distribuzione.

RICHIAMI DELLE PROPRIETÀ DELLE SOMMATORIE

$$\sum_{i=1}^N ax_i = a \sum_{i=1}^N x_i$$

$$\sum_{i=1}^N (x_i + a) = \sum_{i=1}^N x_i + Na$$

$$\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} x_i y_j = \sum_{i=1}^{N_1} x_i \sum_{j=1}^{N_2} y_j$$

MEDIA ARITMETICA PONDERATA

Talvolta occorre tenere conto del fatto che non tutti i valori osservati dei quali si vuole calcolare la media aritmetica hanno la stessa importanza. L'esempio di maggiore interesse è dato dal calcolo della media per dati riassunti in una tabella di frequenza.

Distribuzione di un campione di studenti del II anno per
numero di esami sostenuti

$X = \text{Num. esami}$	n_j	f_j
0	12	0,1
1	24	0,2
2	60	0,5
3	12	0,1
4	12	0,1
TOT	$N = 120$	1

I dati in forma grezza sarebbero del tipo:

0, 0, 0, 0, ..., 0, 1, 1, ..., 1, 2, 2, ..., 2, 3, 3, ..., 3, 4, 4, ..., 4

Vi sono nella sequenza 12 zeri, 24 valori pari a 1, ecc.

$$\begin{aligned}
 \bar{x} &= \frac{1}{120} \sum_{i=1}^{120} x_i = \frac{1}{120} (\underbrace{0 + \dots + 0}_{12 \text{ volte}} + \underbrace{1 + \dots + 1}_{24 \text{ volte}} + \underbrace{2 + \dots + 2}_{60 \text{ volte}} + \\
 &\quad + \underbrace{3 + \dots + 3}_{12 \text{ volte}} + \underbrace{4 + \dots + 4}_{12 \text{ volte}}) \\
 &= \frac{1}{120} (0 \times 12 + 1 \times 24 + 2 \times 60 + 3 \times 12 + 4 \times 12)
 \end{aligned}$$

ossia ogni singola modalità viene pesata in base al numero di volte che essa è stata osservata, la corrispondente frequenza assoluta.

Per calcolare la media dei valori nella tabella non sarebbe corretto considerare solo i singoli valori, ma occorre tener conto delle frequenze. Quindi,

$$\frac{(0 + 1 + 2 + 3 + 4)}{5} \quad \text{E' SBAGLIATO}$$

mentre

$$\frac{(0 \times 12 + 1 \times 24 + 2 \times 60 + 3 \times 12 + 4 \times 12)}{120} \quad \text{E' CORRETTO}$$

In formule,

$$M = \sum_{j=1}^r \frac{x_j n_j}{N} \quad r = \text{Num. di modalità}$$

Si ricordi che $f_j = n_j/N$, per cui si può scrivere anche

$$M = \sum_{j=1}^r x_j f_j$$

Più in generale, se per ogni osservazione x_i è definito un peso p_i ($i = 1, \dots, N$), la media aritmetica ponderata è

$$\frac{\sum_{i=1}^N x_i p_i}{\sum_{i=1}^N p_i}$$

Se abbiamo i dati da una tabella di frequenza relativa ad una variabile quantitativa raggrupata in classi e non si dispone più dei valori originali, per calcolare la media occorre ricorrere a qualche ipotesi semplificatrice.

Di solito si assume come valore rappresentativo di tutte le osservazioni presenti nella classe il valore centrale della classe stessa. Si calcolano i valori

$$x_j^* = \frac{x_{j-1} + x_j}{2}, \quad j = 1, \dots, r$$

dove x_{j-1} , x_j sono gli estremi della classe. Quindi si calcola la media, ponderata con le frequenze, dei nuovi valori x_j^* .

L'uso di x_j^* come valore medio della classe è coerente con l'ipotesi di distribuzione uniforme.

N.B.: Il valore della media calcolato in questo modo può differire da quello calcolato sui valori originali.

Abbiamo introdotto due candidati al ruolo di indice di tendenza centrale: mediana e media.

Quale preferire tra i due?

La mediana è meno influenzata da valori anomali, mentre la media ne risente maggiormente.

Esempio: Lo studente A ha conseguito le seguenti votazioni negli esami sostenuti. Un indice di tendenza centrale può suggerire sinteticamente il grado di riuscita.

Voti: 26, 27, 28, 28, 18

Me=27 e M=25,4

La media è molto sensibile al 18, la mediana ne è influenzata molto meno.

Se abbiamo il sospetto che fra i dati ve ne siano alcuni che sono affetti da errori, bisogna tener presente che tali errori, se sono tali da collocare il dato al di fuori del “normale” insieme di valori, influenzano in modo più pesante la media.

Il motivo di questo è la tendenza della media a bilanciare valori più alti con valori bassi (ricordate che la somma degli scarti è uguale a 0).

La mediana è più “robusta” rispetto a valori anomali e quindi a possibili errori nei dati.

D'altra parte se i dati non sono soggetti ad errori, occorre considerare che la media usa una quantità maggiore di informazioni della mediana (che dipende dalla posizione d'ordine dei valori e meno dalla loro grandezza).

La media inoltre gode di proprietà molto buone quando viene utilizzata in ambito inferenziale (come si vedrà più avanti).

LA MODA

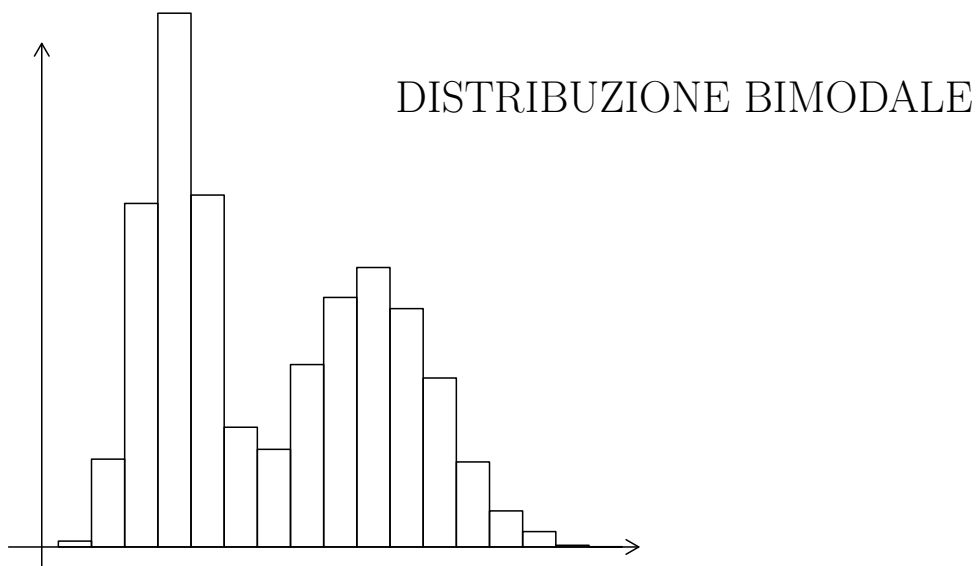
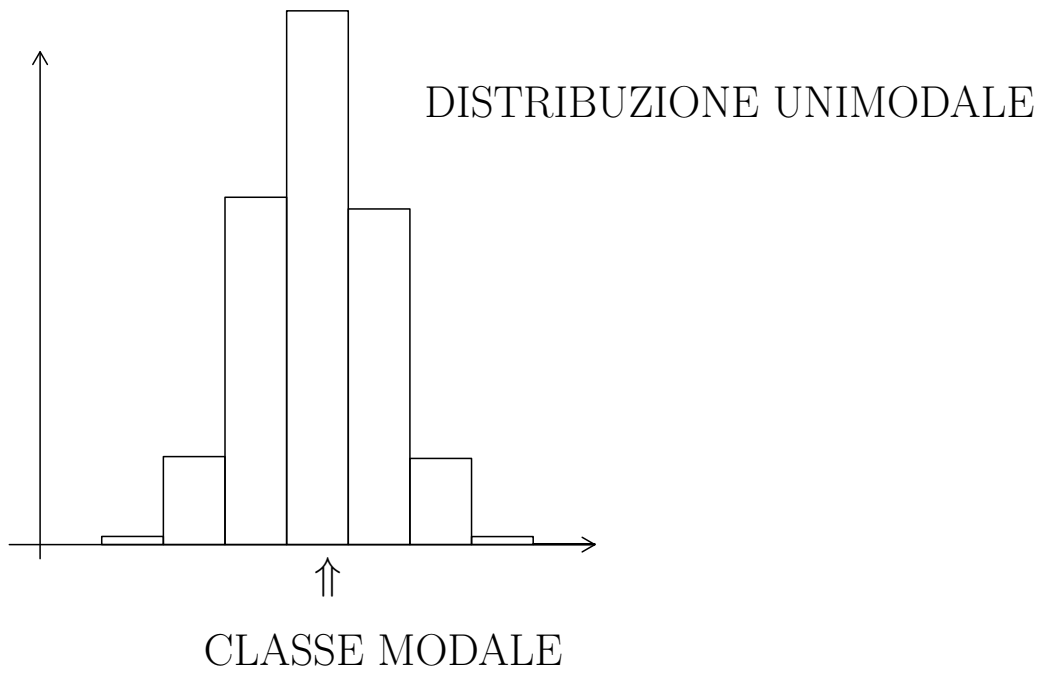
Un altro possibile indice di tendenza centrale è la **MODA**. Si tratta della modalità a cui corrisponde la frequenza più elevata.

E' l'unico indice di tendenza centrale che ha senso valutare per variabili qualitative.

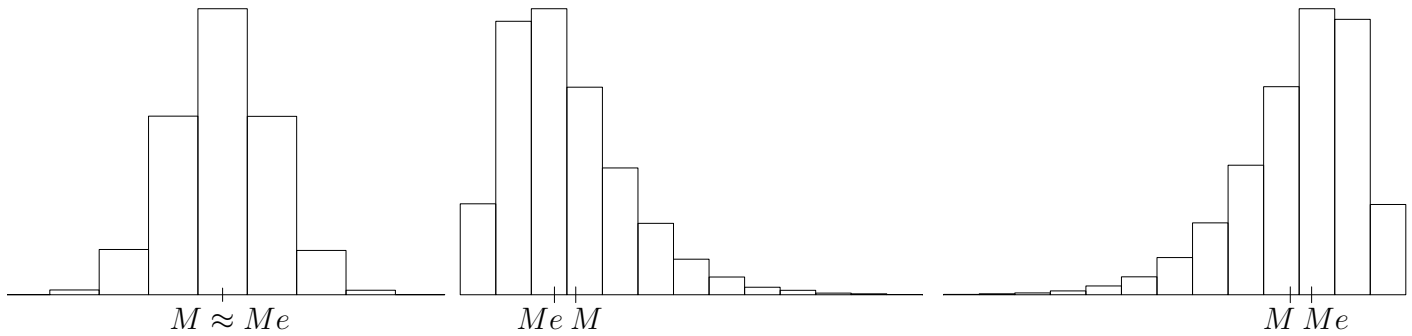
Esempio:

	Num. stanze	n_j	
	1	11	
	2	49	
	3	81	
moda $\Rightarrow$	4	86	$\Leftarrow$ freq. più elevata
	5	38	
	6 o più	20	

Per variabili quantitative raggruppate in classi si identifica la classe modale non guardando alla frequenza, ma alla densità di frequenza (frequenza/ampiezza). La classe modale è quella a cui corrisponde la densità di frequenza più alta, ossia il rettangolo più alto dell'istogramma.



LA SIMMETRIA



Distribuzione
simmetrica

Distribuzione
asimmetrica
con asimmetria
a destra

Distribuzione
asimmetrica
con asimmetria
a sinistra

Se la distribuzione è tendenzialmente simmetrica $\Rightarrow$
 $M \approx Me$

Se la distribuzione è tendenzialmente simmetrica e unimodale $\Rightarrow$
 $M \approx Me \approx \text{Moda}$

Se la distribuzione è asimmetrica con asimmetria positiva $\Rightarrow$
 $M > Me$

Se la distribuzione è asimmetrica con asimmetria negativa $\Rightarrow$
 $M < Me$

I QUANTILI

Il quantile α (x_α) è quel valore tale che la frequenza relativa di unità che risultano inferiori ad esso è pari ad α .

La mediana è quindi il quantile con α pari a 0,5.

Alcuni quantili notevoli sono quelli relativi a $\alpha = 0,25$, $\alpha = 0,5$ e $\alpha = 0,75$ che corrispondono, rispettivamente, al I, II (o mediana) e III quartile. Dividono la distribuzione in 4 parti ciascuna con frequenza relativa pari a 0,25.

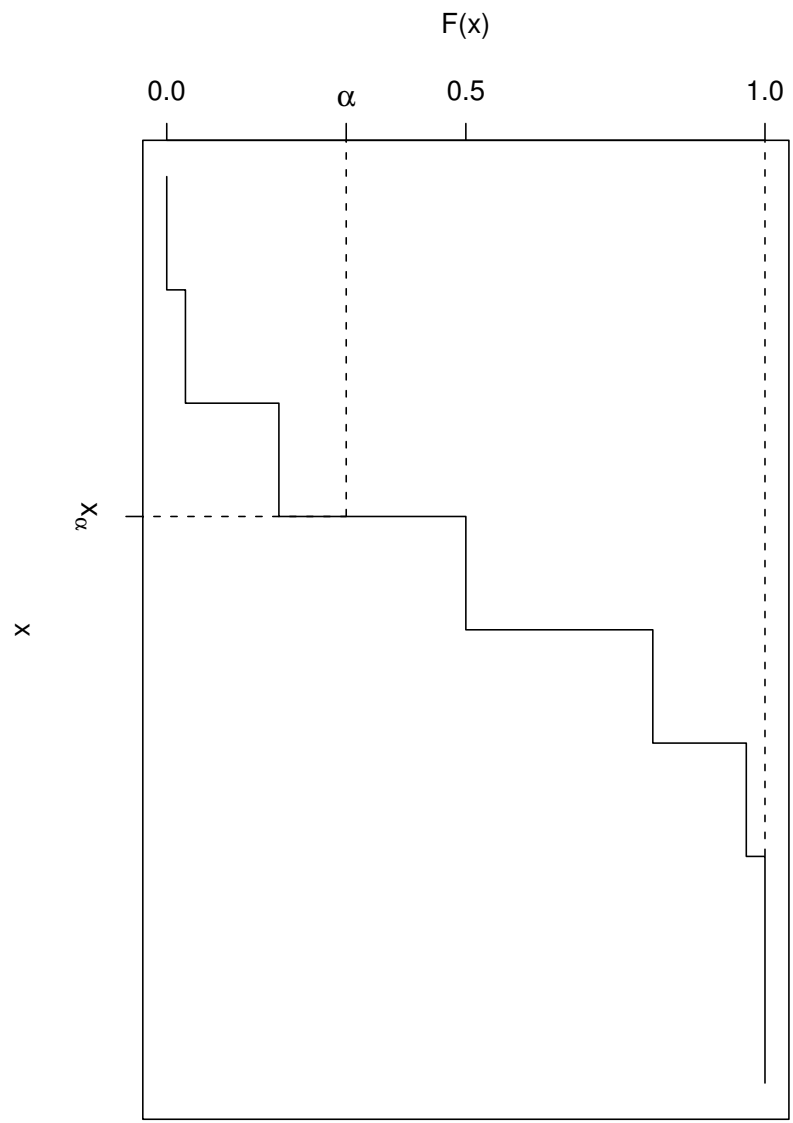
Per determinare il quantile α si segue una strada analoga a quella descritta per la mediana, cercando quel valore che lascia alla sua sinistra l' $\alpha \times 100\%$ delle unità (in sostituzione al 50% della mediana).

Per variabili quantitative discrete riassunte in una tabella di frequenza e non raggruppate in classi

$x_\alpha =$ la più piccola tra le modalità distinte a cui corrisponde $F_j > \alpha$.

Se per una qualche modalità x_j si ha $F_j = \alpha$, allora x_α è qualunque valore compreso tra x_j e x_{j+1} (la modalità immediatamente più grande di x_j). Ad esempio, possiamo prendere

$$x_\alpha = \frac{x_j + x_{j+1}}{2}$$



Per variabili raggruppate in classi si può procedere nel modo seguente:

Determinare il primo (in ordine crescente) intervallo per cui la cumulata supera α . Quindi x_α è compreso tra gli estremi della classe individuata. Sotto l'ipotesi di distribuzione uniforme all'interno della classe

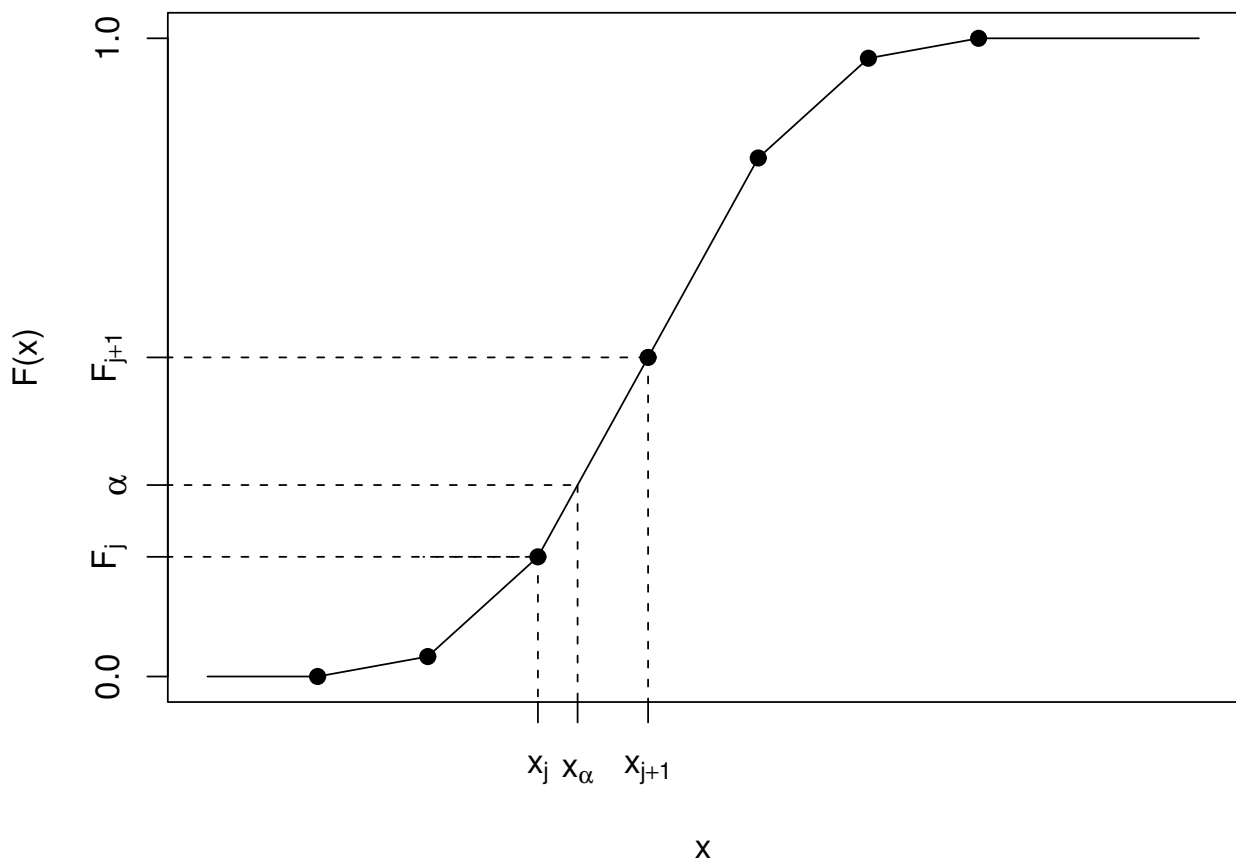
$$x_\alpha = x_j + \frac{(\alpha - F_j)}{(F_{j+1} - F_j)}(x_{j+1} - x_j)$$

x_j = estremo inferiore della classe

x_{j+1} = estremo superiore della classe

F_{j+1} = frequenza relativa cumulata di x_{j+1}

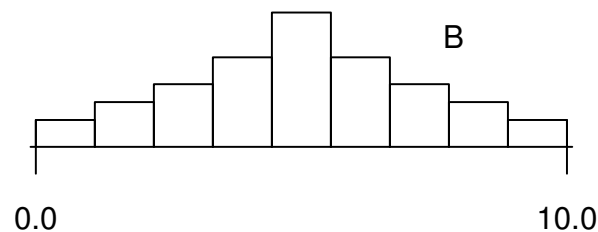
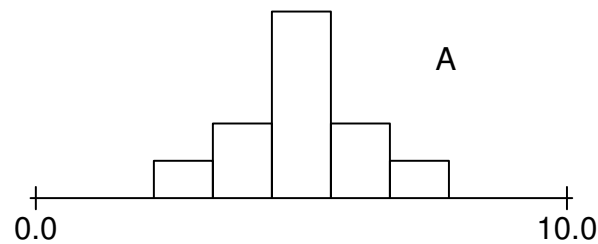
F_j = frequenza relativa cumulata di x_j .



INDICI DI DISPERSIONE

Siamo interessati a riassumere con un indice sintetico la tendenza dei dati ad essere più o meno dispersi, più o meno differenti l'uno dall'altro.

È ovvio che un indice di posizione preso isolatamente è di scarso aiuto. Infatti, potremmo avere:



Le due distribuzioni A e B hanno la stessa media (e la stessa mediana). Sintetizzare una distribuzione solo con riferimento alla sua posizione è quindi insufficiente. Le due distribuzioni differiscono soprattutto in quanto hanno una differente variabilità (B è più variabile di A).

Come per gli indici di tendenza centrale, i candidati a misurare la dispersione (o variabilità) possono essere molti e possono derivare da criteri diversi

- Potrei semplicemente considerare la differenza tra il valore massimo e il valore minimo osservati ($x_{\max} - x_{\min}$)
- o considerare le differenze tra tutti i valori osservati
- oppure considerare la dispersione delle osservazioni attorno ad un indice di tendenza centrale.

IL CAMPO DI VARIAZIONE (“RANGE”)

$$\text{campo di variazione} = x_{\max} - x_{\min} = x_{(N)} - x_{(1)}$$

Dipende solo da due valori, che fra l'altro sono quelli più estremi, di solito maggiormente soggetti ad errori di misura.

LO SCARTO INTERQUARTILE

$$Q_3 = \text{III quartile}, \quad Q_1 = \text{I quartile}$$

$$SI = Q_3 - Q_1$$

Interpretazione: ci dice quanto ampio è l'intervallo nel quale è contenuto il 50% delle osservazioni al centro della distribuzione.

E' poco influenzato dai valori estremi e tiene conto di tutti i valori (anche se solo in relazione alla loro posizione nella graduatoria ordinata).

LA VARIANZA

Abbiamo già introdotto gli scarti dalla media

$$x_i - M = r_i \quad i = 1, \dots, N$$

E' ovvio che tali scarti saranno più elevati se i dati sono più dispersi attorno alla media, ossia se c'è maggiore variabilità. Si può quindi pensare di costruire una misura di variabilità sintetizzando l'ordine di grandezza degli r_i . Tuttavia, poiché

$$\sum_{i=1}^N r_i = 0$$

un eventuale sintesi degli r_i deve prescindere dal loro segno (a valori $r_i > 0$ corrispondono valori $r_i < 0$ e questi nel complesso si bilanciano).

Possiamo agire in due modi: considerando $|r_i|$ oppure r_i^2 , liberandoci quindi del segno e concentrandoci sugli ordini di grandezza degli scarti (addirittura esaltando gli scarti come nel caso r_i^2).

Possiamo allora calcolare la media dei valori $|r_i|$ o r_i^2 , ottenendo i due indici

$$\sum_{i=1}^N \frac{|r_i|}{N} \quad \text{e} \quad \sum_{i=1}^N \frac{r_i^2}{N}$$

Il primo indice è detto **SCARTO MEDIO ASSOLUTO**.

Il secondo indice è la **VARIANZA** (indicata di solito con S^2 o V o Var).

Dei due indici è il secondo quello che ha più largo impiego.

La varianza è la media degli scarti elevati al quadrato.

Trattandosi di una media aritmetica può essere opportunamente ponderata. In particolare, è necessario usare la ponderazione quando vogliamo calcolare la varianza a partire da dati riassunti in una tabella di frequenza.

Esempio:

Num. esami	n_j	f_j	$r_j = (x_j - M)$	r_j^2
0	12	0,1	-1,9	3,61
1	24	0,2	-0,9	0,81
2	60	0,5	0,1	0,01
3	12	0,1	1,1	1,21
4	12	0,1	2,1	4,41

La media è $M=1,9$

Nell'intera sequenza di dati osserviamo 5 modalità distinte. Vi sono quindi 5 scarti e la media dei quadrati degli scarti si ottiene considerando che ciascuno scarto è osservato con la frequenza indicata. Quindi,

$$V = \frac{1}{120}(3,61 \times 12 + 0,81 \times 24 + 0,01 \times 60 + 1,21 \times 12 + 4,41 \times 12)$$

$$= 3,61 \times 0,1 + 0,81 \times 0,2 + 0,01 \times 0,5 + 1,21 \times 0,1 + 4,41 \times 0,1 = 1,09$$

ATTENZIONE: Per variabili raggruppate in classi gli scarti r_j si ottengono come differenza tra il valore centrale della classe e la media:

$$r_j = x_j^* - M = \frac{(x_{j-1} + x_j)}{2} - M.$$

UN METODO ALTERNATIVO PER CALCOLARE LA VARIANZA

$$\begin{aligned} V &= \sum_{i=1}^N \frac{(x_i - \bar{x})^2}{N} = \sum_{i=1}^N \frac{(x_i^2 + \bar{x}^2 - 2x_i\bar{x})}{N} \\ &= \frac{1}{N} \sum_{i=1}^N x_i^2 + \frac{N\bar{x}^2}{N} - \frac{2\bar{x}}{N} \sum_{i=1}^N x_i \\ &= \frac{1}{N} \sum_{i=1}^N x_i^2 + \bar{x}^2 - 2\bar{x}^2 \\ &= \frac{1}{N} \sum_{i=1}^N x_i^2 - \bar{x}^2 \end{aligned}$$

Quindi, la varianza è pari alla differenza tra la media dei valori al quadrato (anche nota con il nome di momento secondo) e la media al quadrato (o momento primo al quadrato).

In aggiunta a $\bar{x}$, occorre solo calcolare il quadrato dei singoli valori e calcolare la media di questi ultimi. Operazione più semplice della precedente.

LO SCARTO QUADRATICO MEDIO

Non è facile dare un significato fisico alla varianza. Il motivo principale è che essa considera i valori degli scarti al quadrato ed è quindi un indice espresso in una unità di misura diversa da quella originaria (se i nostri dati riguardano altezze in cm, la varianza è misurata in cm²). Un modo semplice per ricondurre l'indice all'unità di misura originaria è considerare la **radice quadrata**. Questa operazione conduce ad un altro indice: lo scarto quadratico medio

$$\text{sqm} = \sqrt{\text{varianza}}$$

spesso indicato con S.

L'indice è espresso con l'unità di misura della variabile originaria; è quindi più facile interpretarlo.

Nonostante questa operazione, è più agevole interpretare un indice come SI che non sqm. Inoltre, poiché nel calcolare la varianza abbiamo elevato al quadrato gli scarti, questa ultima (e quindi anche sqm) risulta molto influenzata dalla presenza di valori molto alti o molto bassi. Lo scarto interquartile è meno sensibile, più robusto ai valori estremi.

Tuttavia, sqm (e quindi la varianza) è l'indice più importante di dispersione, soprattutto per il ruolo fondamentale che ha nell'ambito dell'inferenza statistica.

I DIAGRAMMI A SCATOLA (BOXPLOT)

I valori dei 3 quartili (Q_1 , Me, e Q_3) contengono informazioni riassuntive sulla

- tendenza centrale (Me)
- dispersione ($Q_3 - Q_1$)
- simmetria

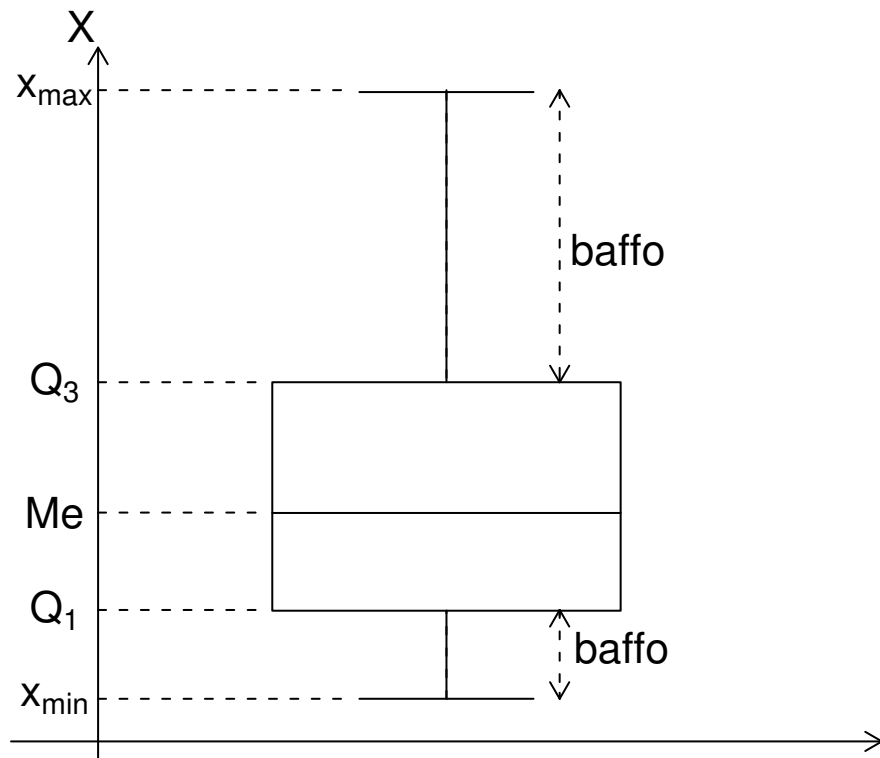
Una rappresentazione grafica che sintetizzi l'intera distribuzione può essere quindi basata sui 3 indici. Il difetto legato all'uso dei 3 indici è che non danno informazioni sulle code della distribuzione. Queste potrebbero essere riassunte attraverso i 2 valori estremi:

$$x_{\max} = x_{(N)} \quad \text{e} \quad x_{\min} = x_{(1)}$$

Il diagramma a scatola è basato sulle 5 quantità

$$(x_{\min}, \quad Q_1, \quad \text{Me}, \quad Q_3, \quad x_{\max})$$

come di seguito riportato.



1. La scatola si può pensare come il contenitore del 50% delle osservazioni che cadono nella parte centrale della distribuzione (il “grosso” della distribuzione). La lunghezza della scatola dà indicazioni sulla dispersione.
2. La posizione della scatola dà un’idea dei valori caratteristici del fenomeno. In particolare, il segmento che la divide è un indice di posizione.
3. La posizione del segmento rispetto agli estremi della scatola (assieme alla lunghezza dei baffi) è un indice della simmetria.
4. I baffi danno un’idea della lunghezza delle code.

Esempio: Precipitazioni nel mese di marzo (in pollici)

0,77	0,47	1,51	0,59	4,75	1,74
1,43	2,10	0,81	2,28	0,81	3,37
0,52	2,81	0,96	1,20	2,20	1,62
3,00	1,31	1,18	0,80	1,20	3,09
1,87	1,87	1,95	0,32	1,35	2,05

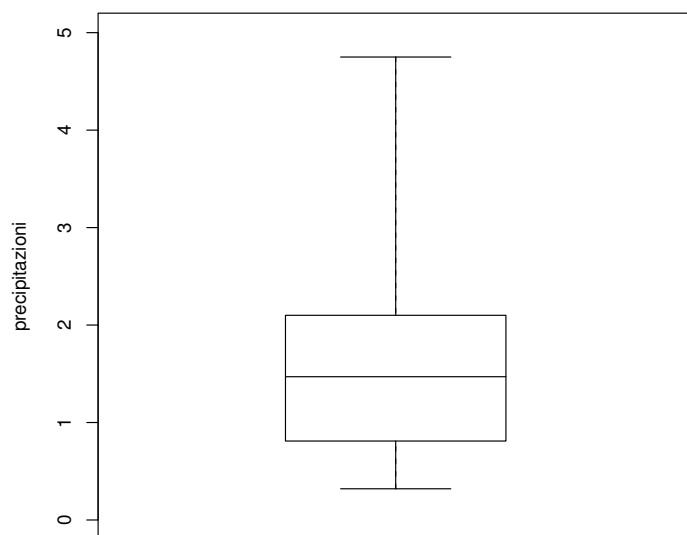
La mediana è tra il 15° e il 16° posto (nella graduatoria crescente delle osservazioni), ossia tra 1,43 e 1,51.

Quindi $Me=1,47$ pollici.

Costruendo la funzione di frequenza relativa cumulata, si vede che $Q_1=0,81$ pollici e $Q_3=2,10$ pollici.

Inoltre,

$$x_{\min} = 0,32 \quad \text{e} \quad x_{\max} = 4,75$$

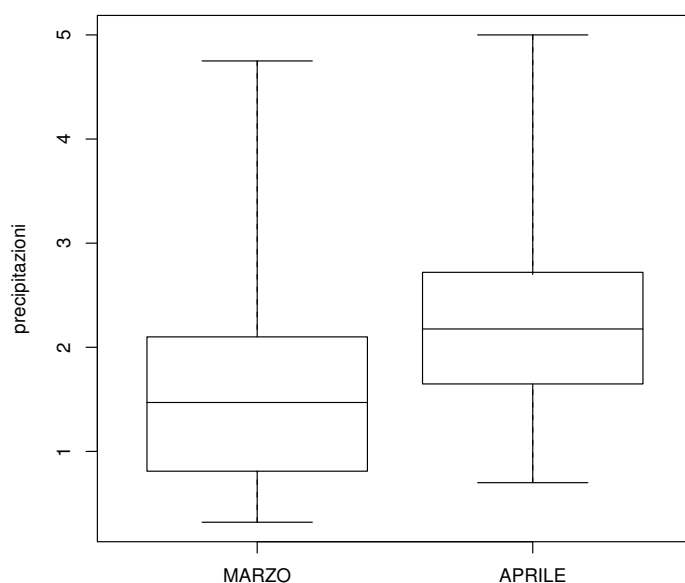


Coda destra lunga (frequenza relativamente alta di forti precipitazioni).

Asimmetria a destra.

I *boxplot* sono molto utili per il confronto tra distribuzioni di frequenza di due o più variabili omogenee: basterà affiancare i relativi *boxplot* mantenendo la stessa scala per l'asse delle y .

Si supponga, ad esempio, di disporre delle osservazioni sulle precipitazioni del mese di aprile. Volendo confrontare le distribuzioni delle precipitazioni di marzo e di aprile, possiamo ricorrere ai seguenti *boxplot* affiancati.



Dal confronto dei due *boxplot* possiamo concludere che le precipitazioni di aprile sono **statisticamente superiori** alle precipitazioni di marzo. Infatti, qualunque sia il quantile che si considera la distribuzione delle precipitazioni di aprile presenta un valore del quantile stesso superiore rispetto a quello di marzo.

Questo ragionamento può essere esteso. Se il *boxplot* della variabile X si trova “al di sopra” del *boxplot* della variabile Y (per ciascun quantile), allora X è statisticamente maggiore di Y .

TRASFORMAZIONI DEI DATI

I dati grezzi raccolti in una rilevazione possono essere opportunamente trasformati, ad esempio, per adottare un'unità di misura più idonea a rappresentare il fenomeno, o per rendere meno complicata l'analisi.

Le trasformazioni più semplici sono:

1. Cambio di origine

Si ottiene aggiungendo o togliendo una costante.

Dalla variabile X con osservazioni $x_1, x_2, \dots, x_N$ si ottiene una nuova variabile $Y = X + a$ ($a \in \mathbb{R}$) con osservazioni $y_1 = x_1 + a, y_2 = x_2 + a, \dots, y_N = x_N + a$.

Cosa succede a media e varianza in seguito ad un cambio di origine?

$$\boxed{M(Y) = M(X + a) = M(X) + a}$$

(subisce lo stesso tipo di trasformazione)

$$\boxed{V(Y) = V(X + a) = V(X)}$$

(non subisce alcun cambiamento)

Esempio: X rappresenta la variabile salario lordo mensile (in euro) osservata su un campione di 100 individui. Si sa che

$$M(X) = 1300 \text{ euro} \quad \text{e} \quad V(X) = 10000.$$

Se tutti i 100 individui ottengono un aumento del salario mensile pari a 100 euro, allora la variabile Y

che descrive il nuovo salario lordo mensile è $Y = X + 100$ con

$$M(Y) = M(X + 100) = 1300 + 100 = 1400 \text{ euro}$$

$$V(Y) = V(X + 100) = 10000.$$

2. Cambio di scala

Si ottiene moltiplicando o dividendo per una costante.

Dalla variabile X con osservazioni $x_1, x_2, \dots, x_N$ si ottiene una nuova variabile $Y = b \cdot X$ ($b \in \mathbb{R}^+$) con osservazioni $y_1 = b \cdot x_1, y_2 = b \cdot x_2, \dots, y_N = b \cdot x_N$.

Cosa succede a media e varianza in seguito ad un cambio di scala?

$$M(Y) = M(b \cdot X) = b \cdot M(X)$$

(subisce lo stesso tipo di trasformazione)

$$V(Y) = V(b \cdot X) = b^2 \cdot V(X)$$

(la costante moltiplicativa deve essere elevata al quadrato)

Esempio: Sia X la variabile definita nell'esempio precedente. Se indichiamo con Y la variabile salario lordo mensile **in lire**, allora $Y = 1936,27 \cdot X$. La media e la varianza del salario lordo mensile dei 100 individui in lire saranno, allora,

$$M(Y) = M(1936,27 \cdot X) = 1936,27 \cdot 1300 \text{ lire}$$

$$V(Y) = V(1936,27 \cdot X) = 1936,27^2 \cdot 10000.$$

3. Cambio di origine e di scala (trasformazione lineare dei dati)

Combiniamo i due tipi di trasformazione appena visti.

Dalla variabile X con osservazioni $x_1, x_2, \dots, x_N$ si ottiene una nuova variabile $Y = a + b \cdot X$ ($a \in \mathbb{R}, b \in \mathbb{R}^+$) con osservazioni $y_1 = a + b \cdot x_1, y_2 = a + b \cdot x_2, \dots, y_N = a + b \cdot x_N$.

Cosa succede a media e varianza in seguito ad un cambio di origine e di scala?

$$\boxed{M(Y) = M(a + b \cdot X) = a + b \cdot M(X)}$$

(subisce lo stesso tipo di trasformazione)

$$\boxed{V(Y) = V(a + b \cdot X) = b^2 \cdot V(X)}$$

(la costante moltiplicativa deve essere elevata al quadrato e si perde la costante additiva)

LA STANDARDIZZAZIONE

La standardizzazione è un particolare tipo di trasformazione lineare (cambio sia di origine che di scala). Consiste nel sottrarre a ciascuna osservazione di una variabile X la media $M(X)$ e dividere il tutto per lo scarto quadratico medio S :

$$Z = \frac{X - M(X)}{S}$$

Calcoliamo la media e la varianza della nuova variabile **standardizzata** Z .

Si noti che la trasformazione applicata equivale ad una trasformazione di tipo 3 con $b = 1/S$ e $a = -M(X)/S$. Allora,

$$M(Z) = M\left(-\frac{M(X)}{S} + \frac{1}{S}X\right) = -\frac{M(X)}{S} + \frac{1}{S}M(X) = 0$$

$$V(Z) = V\left(-\frac{M(X)}{S} + \frac{1}{S}X\right) = \frac{1}{S^2}V(X) = 1$$

In sintesi, una variabile standardizzata ha media 0, varianza 1 ed è adimensionale (non ha unità di misura).

Esempio: Se prendiamo le osservazioni x_i ($i = 1, \dots, 100$) sul salario lordo mensile (in euro) dei 100 individui e le trasformiamo tramite

$$z_i = \frac{x_i - 1300}{\sqrt{10000}} \quad i = 1, \dots, 100$$

otteniamo delle osservazioni adimensionali, di media nulla e varianza unitaria.

L'utilità dell'operazione di standardizzazione risulterà più chiara nel prosieguo delle lezioni.

INDICI DI DISPERSIONE RELATIVI

Abbiamo visto che lo scarto quadratico medio, come anche la varianza, è un indice di variabilità che dipende dall'unità di misura. Non possiamo pertanto utilizzare tale indice per confrontare la variabilità di fenomeni espressi in diverse unità di misura, come ad esempio centimetri e metri.

In realtà lo scarto quadratico medio, come la varianza, ha anche una altra importante limitazione: risente dell'ordine di grandezza dei dati. Non è quindi corretto confrontare tramite lo scarto quadratico medio (o la varianza) la variabilità di due fenomeni, che pur essendo espressi nella stessa unità di misura, hanno valori medi molto diversi, ad esempio il peso dei bambini e il peso degli adulti, anche se espressi entrambi in kg, oppure il reddito di paesi con diverso grado di povertà, anche se espressi nella stessa valuta.

E' opportuno allora introdurre delle misure di variabilità rapportate alla grandezza della variabile studiata: **gli indici di dispersione relativi**. Il più noto di questi indici è il **coefficiente di variazione** (CV) definito come

$$CV = \frac{S}{\bar{x}}$$

Poiché $\bar{x}$ e S hanno la stessa unità di misura, CV è un numero puro (adimensionale).

Esempio: Peso dei maschi e delle femmine (in kg)

$$\bar{x}_M = 72,6 \text{ kg} \quad \bar{x}_F = 53,3 \text{ kg}$$

$$S_M = 2,6 \text{ kg} \quad S_F = 2,1 \text{ kg}$$

Maggiore variabilità assoluta per i maschi che per le femmine.

Tuttavia,

$$CV_M = 0,036 \quad CV_F = 0,039$$

La variabilità percentuale del peso dei maschi è 3,6%; delle femmine è 3,9%. Il peso dei maschi risulta meno variabile di quello delle femmine (anche se di poco) su scala percentuale.