01 – Forma di una distribuzione

Una volta studiata la distribuzione di un carattere attraverso il calcolo delle misure di centralità e variabilità abbiamo delle informazioni sintetiche per poter comprendere il comportamento di tale caratteristica rispetto al collettivo oggetto di studio

La centralità e la variabilità di una distribuzione in alcuni casi non sono però esaustive per poter interpretare come il carattere si manifesta

Abbiamo bisogno di un altro elemento per meglio definire le caratteristiche della distribuzione:

- due variabili possono avere infatti, ad esempio, la stessa media/mediana e la stessa variabilità ma differire per il peso dei valori più grandi o più piccoli rispetto al valore centrale, a causa del comportamento differenziato delle “code” della distribuzione, cioè delle parti più esterne dell’insieme ordinato dei dati

Tale studio può essere effettuato considerando la cosiddetta **forma della distribuzione**

02 – Gli intervalli di variabilità

Data la distribuzione unitaria di un carattere X ordinata in senso crescente

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$$

è possibile individuare 5 modalità rappresentative:

$x_{(1)} = x_{min}$ è il valore più piccolo della distribuzione

Q_1 = primo quartile (25° percentile)

Me = mediana (50° percentile)

Q_3 = terzo quartile (75° percentile)

$x_{(N)} = x_{max}$ è il valore più grande della distribuzione

Attraverso tali valori possiamo costruire i cosiddetti **intervalli di variabilità** della distribuzione

03 – Centralità e variabilità — Unità n° 07

Dai valori x_{min}, Q_1, Q_3 e x_{max} è possibile ottenere due misure di posizione e due di variabilità

posizione

$$\text{MidRange} = \frac{x_{min} + x_{max}}{2}$$

$$\text{Media Interquartile} = \frac{Q_1 + Q_3}{2}$$

variabilità

$$\text{Campo di var.ne} = x_{max} - x_{min}$$

$$\text{Differenza Interquartile} = Q_3 - Q_1$$

Le misure forniscono delle indicazioni di massima sulla distribuzione dei dati e analizzare la forma della distribuzione di X (anche se le prime sono influenzate dai valori anomali e le seconde considerano invece solamente la metà delle osservazioni a disposizione)

04 – La sintesi a cinque

Utilizzando i cinque valori rappresentativi è possibile studiare la distribuzione di un carattere in un collettivo, osservando:

- la distanza tra il primo quartile e la mediana e tra la mediana e il terzo quartile
- la distanza tra x_{min} e il primo quartile e tra il terzo quartile e x_{max}
- la relazione tra la mediana, la media interquartile e il midrange

La distribuzione si dice **simmetrica** se:
- la distanza tra primo quartile e mediana e tra mediana e terzo quartile è uguale
- la distanza tra x_{min} e primo quartile e tra terzo quartile e x_{max} è uguale
- la mediana, la media interquartile e il midrange coincidono

In questo caso anche la moda e la media aritmetica coincidono con la mediana

La distribuzione si dice **asimmetrica** se:
- la distanza tra primo quartile e mediana e tra mediana e terzo quartile è diversa
- la distanza tra x_{min} e primo quartile e tra terzo quartile e x_{max} è diversa
- la mediana, la media interquartile e il midrange non coincidono

05 – Asimmetria positiva e negativa

In generale si distingue tra una asimmetria **positiva** e una asimmetria **negativa**

La distribuzione si dice **asimmetrica negativa** (o "obliqua a sinistra") se:
- la distanza tra x_{min} e primo quartile è maggiore di quella tra terzo quartile e x_{max}
- la mediana è maggiore della media interquartile, la media interquartile è maggiore del midrange

➡ In questo caso si osservano più frequentemente modalità con *alta* intensità e più raramente modalità con *bassa* intensità, quindi in generale (ma non sempre): moda>mediana>media

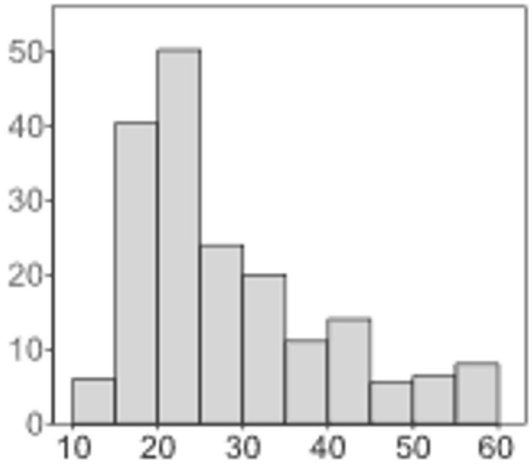
La distribuzione si dice **asimmetrica positiva** (o "obliqua a destra") se:
- la distanza tra x_{min} e primo quartile è minore di quella tra terzo quartile e x_{max}
- la mediana è minore della media interquartile, la media interquartile è minore del midrange

➡ In questo caso si osservano più frequentemente modalità con *bassa* intensità e più raramente modalità con *bassa* intensità, quindi in generale (ma non sempre): moda<mediana<media

06 – Rappresentazione grafica

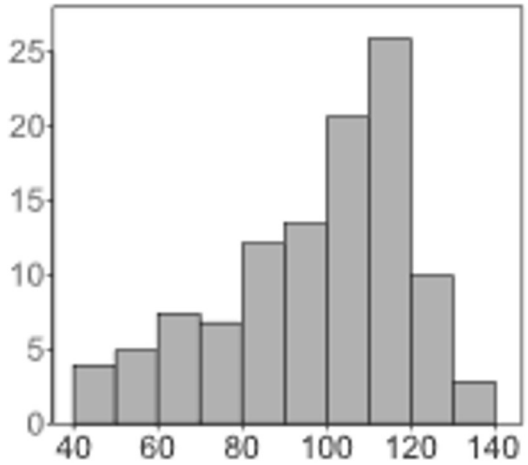
Unità n° 07

Possiamo studiare la forma di una distribuzione di frequenza o in classi osservando il corrispondente diagramma a barre o istogramma



Distribuzione asimmetrica positiva

i valori più piccoli sono più frequenti e la moda è minore del centro della distribuzione: abbiamo valori alti che "disturbano" la distribuzione



Distribuzione asimmetrica negativa

i valori più grandi sono più frequenti e la moda è maggiore del centro della distribuzione:abbiamo valori bassi che "disturbano" la distribuzione

07 – Un diverso modo di studiare la forma della distribuzione — Unità n° 07

Possiamo ricorrere ai soli intervalli di variabilità per descrivere graficamente la distribuzione

La rappresentazione ottenuta è detta **box plot** (diagramma a scatola e baffi)

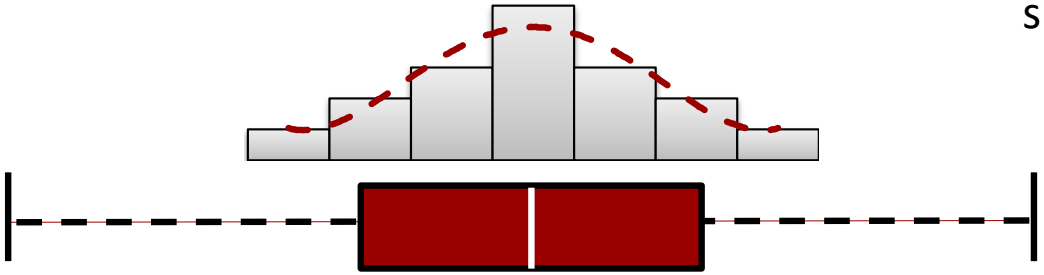
Il box-plot è un grafico caratterizzato da tre elementi:

1) un rettangolo (box) la cui dimensione indica la variabilità dei valori "prossimi" al centro della distribuzione
2) una linea o punto, che indica la posizione del centro della distribuzione
3) due segmenti che partono dal rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione

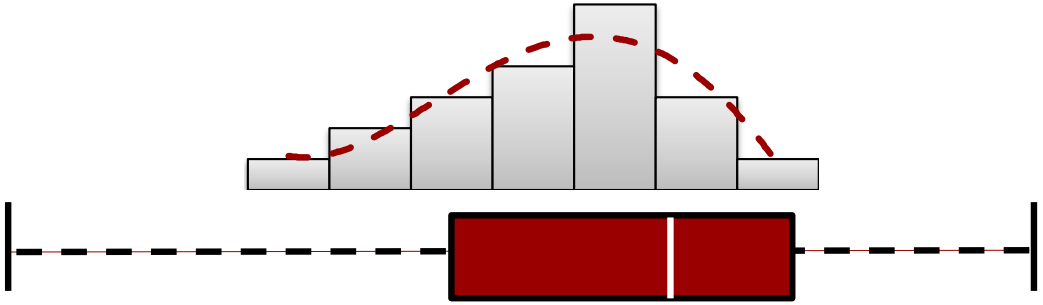
→ Generalmente come valore centrale si considera la mediana, come altezza/larghezza della scatola la distanza interquartile e come estremi dei segmenti il valore minimo e massimo della distribuzione

08 – Box plot e forma della distribuzione

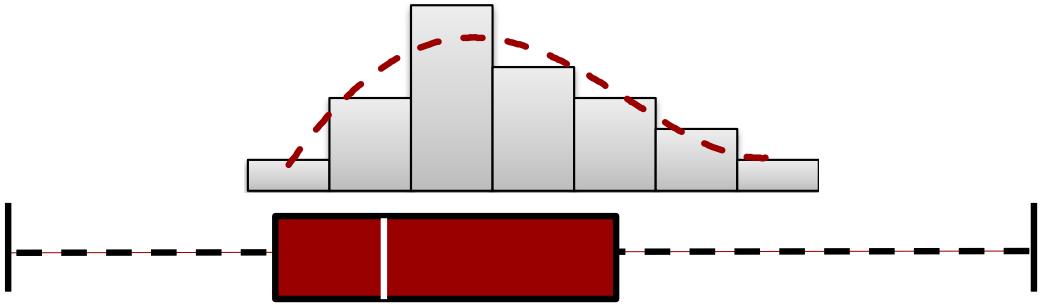
Dal Box plot possiamo dedurre informazioni su variabilità e forma della distribuzione di X



Distribuzione simmetrica



Distribuzione Asimmetrica negativa

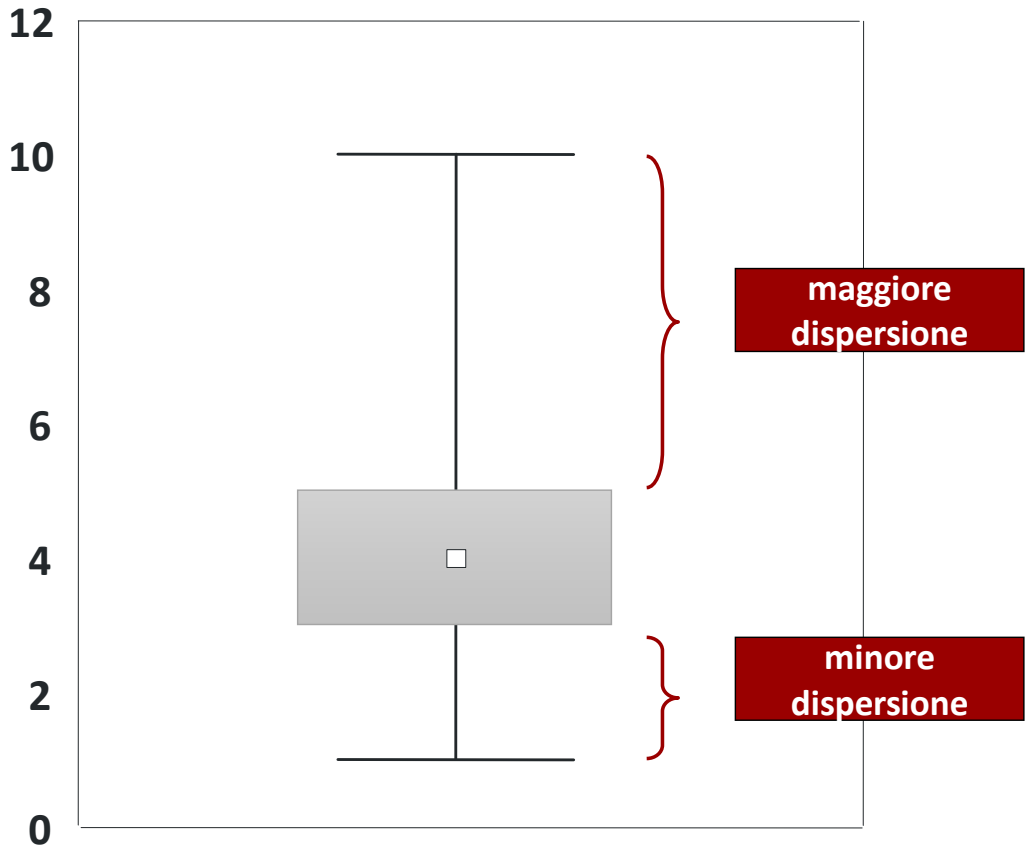


Distribuzione Asimmetrica positiva

09 – Esempio

N°atti aggressivi	1	2	3	4	5	6	7	8	9	10
Bambini	3	8	30	45	22	12	10	5	2	1

Studio sull'aggressività infantile (138 bambini)



Max = 10
Min = 1

Q_3=5
Q_1=3

Valore mediano:
Me=4

Dall'analisi del box plot si evince come ci sia una maggior frequenza di valori medio-bassi, il che spiega lo spostamento verso il basso della scatola (o verso sinistra se consideriamo un box plot orizzontale)

10 – Box plot e valori anomali

Unità n° 07

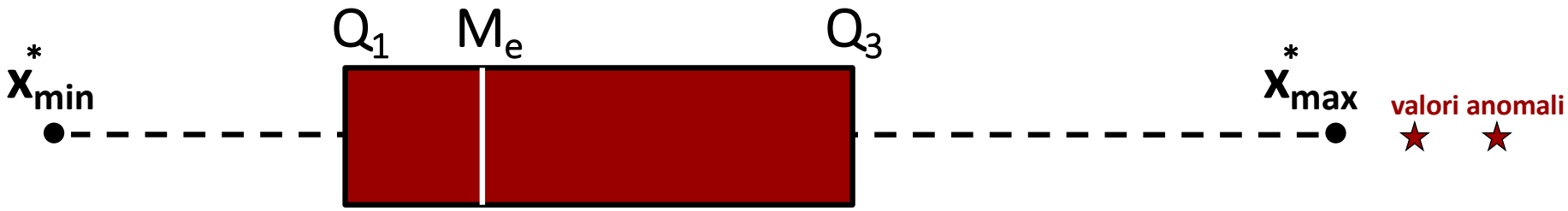
Attraverso il box plot è possibile evidenziare la presenza di eventuali valori anomali. Abbiamo già detto che un valore anomalo è un valore molto più piccolo o molto più grande rispetto ai valori della distribuzione: per poter evidenziare tali modalità particolari è necessario calcolare i cosiddetti valori minimo e massimo "teorici" e confrontarli con quelli effettivamente osservati

E' possibile considerare come minimo e massimo della distribuzione i valori così ottenuti:

$$x^*_{min} \Rightarrow \text{valore più grande tra } x_{min} \text{ e } [Q_1 - 1.5(Q_3 - Q_1)]$$

$$x^*_{max} \Rightarrow \text{valore più piccolo tra } x_{max} \text{ e } [Q_3 + 1.5(Q_3 - Q_1)]$$

Gli eventuali valori esterni a tali valori sono considerati anomali



11 – Esempio

Consideriamo la distribuzione dell'età degli studenti iscritti a un Master post-laurea

Età	Studenti

Età studenti del Corso	Frequenze assolute (n_i)
25	3
26	6
27	8
28	5
29	5
30	3
61	1
	31

$$M(X) = \frac{(25\times 3)+(26\times 6)+(27\times 8)+(28\times 5)+(29\times 5)+(30\times 3)+(61\times 1)}{31}$$

$$= \frac{75+156+216+140+145+90+61}{31} = \frac{883}{31} = 28,48$$

Mediana e Quartili

3 6 8 5 5 3 1

25 25 25 | 26 26 26 26 26 26 | 27 27 27 27 27 27 27 27 | 28 28 28 28 28 | 29 29 29 29 29 | 30 30 30 | 61

Q_1 Q_2=Me Q_3

$Q_1 = 26$

$Me = 27$

$Q_3 = 29$

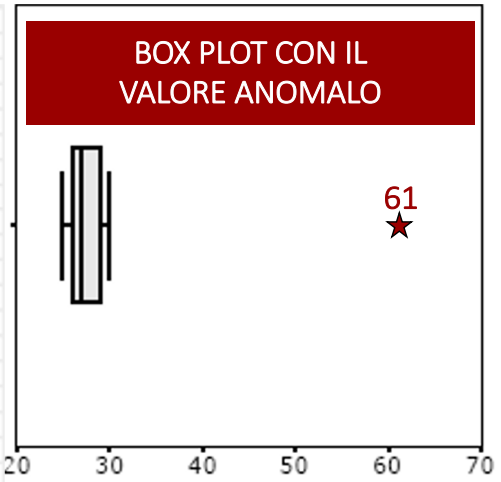
$Q_3 - Q_1 = 29-26 = 3$

Dati anomali:

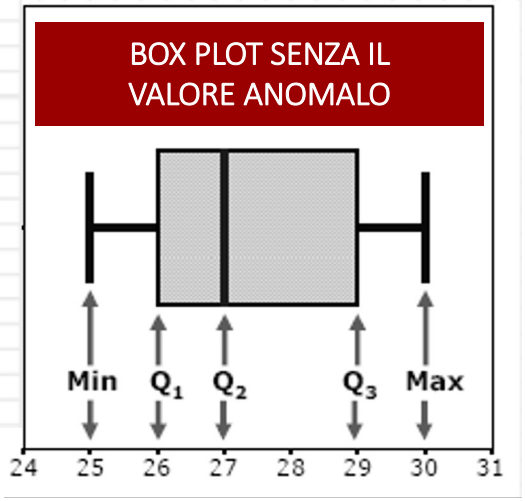
a) $Q_3 + 1,5\,(Q_3-Q_1) = 29 + 1,5\times 3 = 33,5$

b) $Q_1 - 1,5\,(Q_3-Q_1) = 26 - 1,5\times 3 = 21,5$

BOX PLOT CON IL VALORE ANOMALO

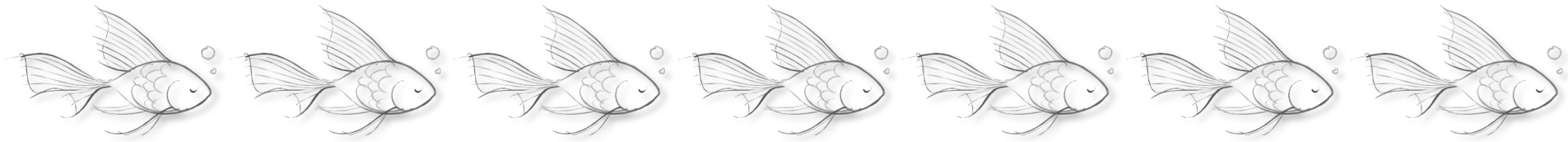


BOX PLOT SENZA IL VALORE ANOMALO



Dal box plot si rileva che 61 è un valore anomalo!

12 – Esercizio



Si consideri la distribuzione del peso di 10 maschi e 10 femmine (in Kg) di una particolare specie di pesce

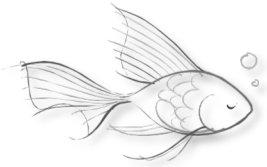
M	1.2	3.0	5.2	4.0	3.5	4.3	3.3	4.8	3.8	3.2
F	1.3	2.2	1.5	2.3	1.8	1.7	2.1	2.0	1.9	2.1

1) Calcolare per ciascuna sottopopolazione il peso medio e la deviazione standard

2) Confrontare la variabilità del peso di maschi e femmine con il coefficiente di variazione

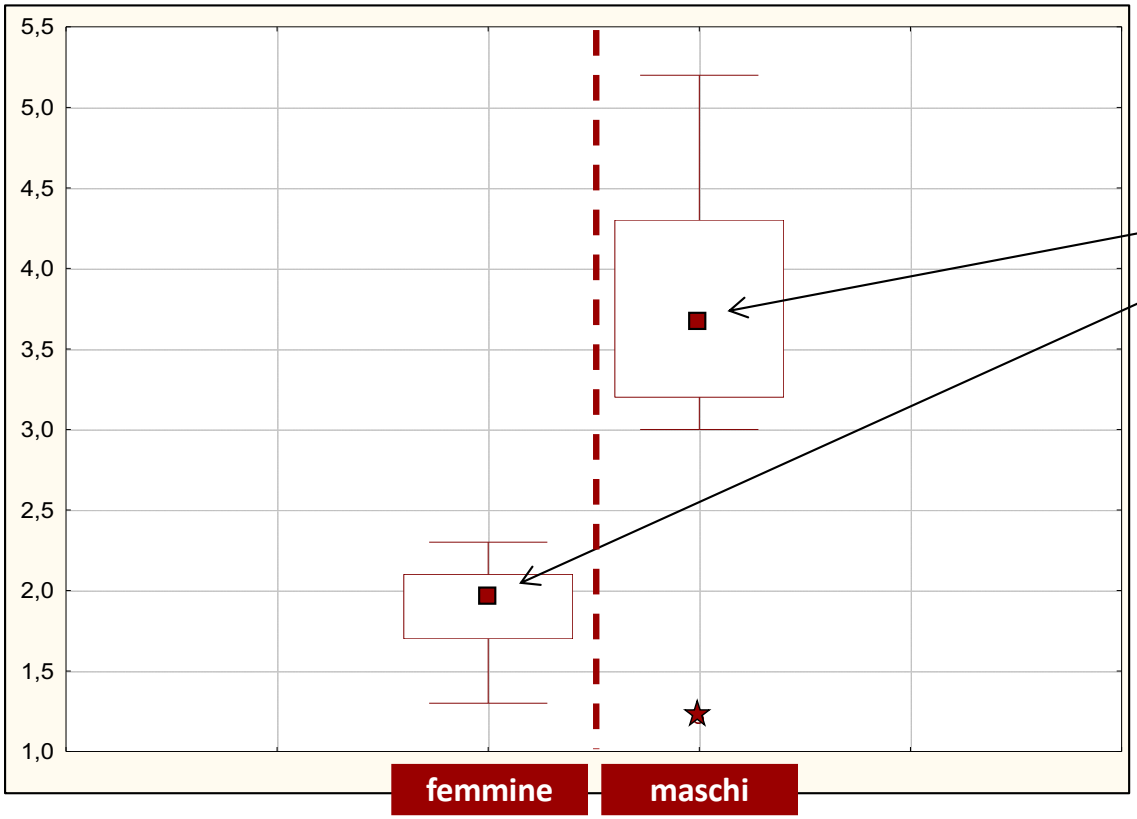
3) Costruire e commentare le rappresentazioni box plot

13 – Soluzione (2) e (3)



La variabilità del peso è maggiore nei maschi rispetto alle femmine

$CV_f = 17\%$
$CV_m = 31\%$



In media gli esemplari maschi pesano più degli esemplari femmine

Osserviamo come ci sia una maggiore dispersione nel peso degli esemplari maschi rispetto agli esemplari femmine. Rispetto alla forma delle diverse distribuzioni si vede come nel caso delle femmine ci sia una lieve asimmetria negativa, mentre nel caso dei maschi l'asimmetria è positiva

14 – Indici di forma basati sul confronto di indici di centralità — Unità n° 07

Come già visto per gli altri aspetti caratteristici della distribuzione, esistono degli indici che consentono di valutare sinteticamente la asimmetria di una distribuzione

Due misure assolute di asimmetria basate sul confronto delle misure di centralità sono date da

$$a_{Me} = \bar{x} - Me$$

$$a_{Mo} = \bar{x} - Mo$$

Tali quantità assumono valori nulli, positivi o negativi a seconda che la distribuzione sia simmetrica, asimmetrica positiva o asimmetrica negativa

È possibile ricavare anche delle misure relative, ottenute rapportando le quantità precedenti allo scarto quadratico medio:

$$A_{Me} = (\bar{x} - Me) / \sigma$$

$$A_{Mo} = (\bar{x} - Mo) / \sigma$$

Indice di asimmetria di Pearson

Si dimostra che il denominatore è il massimo valore che può assumere il numeratore, quindi questi indici variano sempre tra -1 e 1

15 – Esempio

x_i	n_i	f_i	F_i
0	5781	0,372	0,372
1	6781	0,436	0,808
2	1644	0,106	0,914
3	1123	0,072	0,986
4	220	0,014	1

Consideriamo la distribuzione del numero di figli per famiglia di un Comune calabrese nell'anno 2021

Il numero medio di figli per famiglia è:

$$\bar{x} = 0,92$$
$$Mo = 1$$
$$Me = 1$$

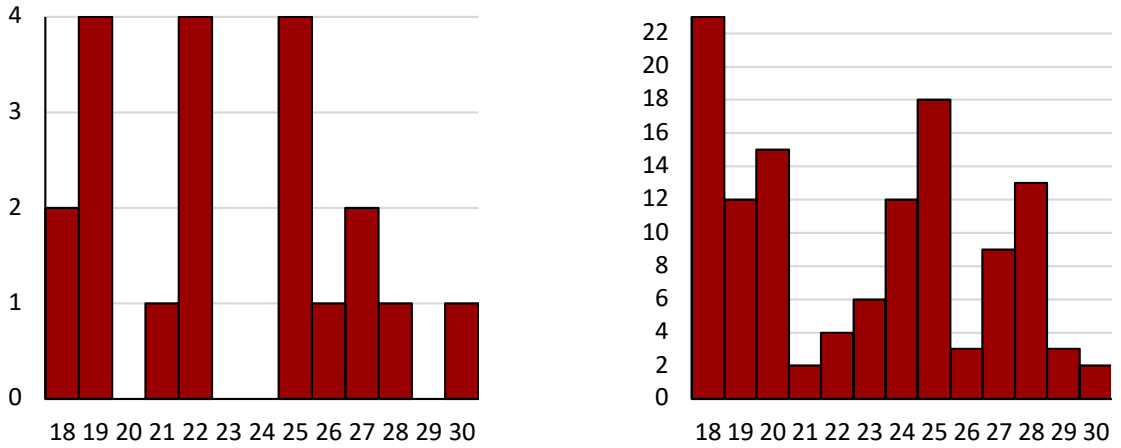
Se studiamo la variabilità del numero di figli per famiglia, otteniamo che:

$$\sigma^2 = 0,886$$
$$\sigma = \pm 0,941$$

Per quanto riguarda la forma della distribuzione otteniamo che $A_{Me} = A_{Mo} = -0,085$, concludendo che la distribuzione dei figli per famiglia è asimmetrica negativa

16 – Rendere comparabili due o più distribuzioni

Spesso è utile svincolare la distribuzione di un carattere quantitativo dalla sua tendenza centrale (*effetto media*) e dall'ordine di grandezza di ciò che è rilevato (*effetto scala di misura*)



L'esigenza si determina ad esempio quando dobbiamo confrontare due o più distribuzioni: tale operazione è detta **standardizzazione** ed è una particolare trasformazione lineare dei dati in cui si ha una **traslazione del sistema di riferimento** (nella media) ed un **mutamento** della scala di misura

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

La distribuzione della risultante variabile **Z** ha sempre media nulla e varianza unitaria

17 – Indici di forma con dati standardizzati

L'**indice di asimmetria di Fisher** è la media cubica dei valori standardizzati di una distribuzione:

$$\gamma_1 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

Tale indice varia tra $-\infty$ e $+\infty$:

- quando è minore di 0 allora abbiamo asimmetria negativa
- quando è maggiore di 0 allora abbiamo asimmetria positiva
- quando è uguale a 0 allora abbiamo simmetria

ES.: se utilizziamo l'indice di Fisher per la distribuzione del numero di figli per famiglia, otteniamo γ_1= 1,081

la distribuzione è asimmetrica positiva