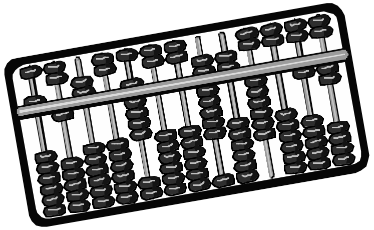
# Premessa

**La distribuzione di frequenze è il risultato della raccolta dati che in essa sono organizzati e semplificati.**



**Rimangono però ancora tante le informazioni ed è nostro obiettivo riassumerne gli aspetti salienti in pochi valori numerici (indici statistici)**

**Essi consentono il confronto tra distribuzioni relative a variabili diverse oppure relative allo stessa variabile osservate in epoche, luoghi ed occasioni diverse.**



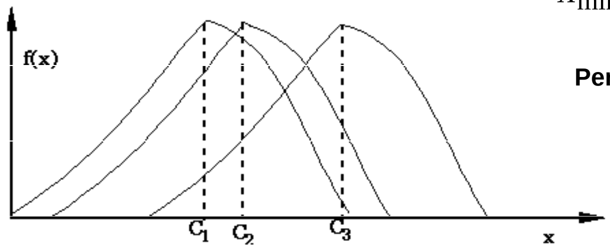
# Concetto di media

**Individua il livello di maggiore addensamento delle modalità ovvero la categoria o il valore (espresso nella stessa unità di misura) intorno alla quale sembra ruotare l'intera rilevazione.**

**Non è necessario che appartenga al dominio della variabile (può essere un valore fittizio).**

**Bisogna e basta che rispetti la condizione di internalità:**

$$X_{\min} \leq \; Media \; \leq \; X_{\max}$$




**Per scale almeno ordinali**



# Esempi

Esempio.
Carattere qualitativo

| Linguaggio | Parlanti (in milioni) |
|---|---|
| Mandarino | 740 |
| Inglese | 403 |
| Russo | 277 |
| Spagnolo | 266 |
| Indostano | 264 |
| Arabo | 160 |
| Bengalese | 155 |
| | 2265 |

Carattere quantitativo.

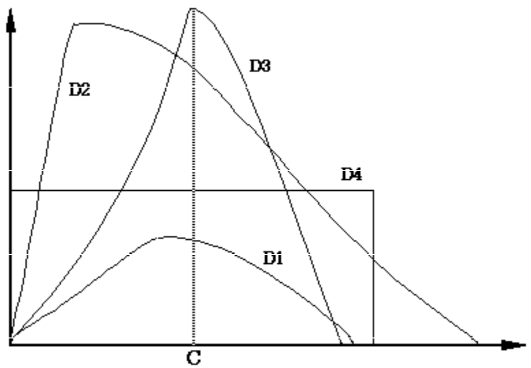
| Numeri di albini in famiglie di 5 figli | Numero di famiglie |
|---|---|
| 1 | 25 |
| 2 | 23 |
| 3 | 10 |
| 4 | 1 |
| 5 | 1 |
| | 60 |

**Nel primo caso la "media" può essere qualsiasi modalità (interna od esterna al dominio), nel secondo può essere solo un numero compreso tra 1 e 5**



# Mancanza di univocità

**Il processo di estrema sintesi che porta al collassamento della distribuzione di frequenza su di un singolo valore, costituisce il limite degli indici di posizione perché:**




**Distribuzioni molto diverse possono presentare la stessa "media". Conoscendo questa non è univocamente nota la situazione che l'ha determinata.**

# La moda

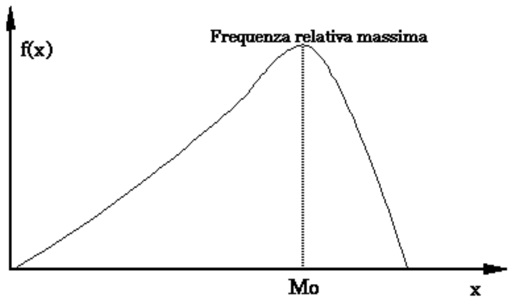
**E' l'indice di posizione più facile da calcolare, ma anche quello più grezzo. Si identifica con la modalità corrispondente alla frequenza relativa maggiore:**

$$Mo = \{ x_m \mid f_m \geq f_i \ \text{per} \ i = 1, 2, \ldots, k \}$$

**Esempio.**

**Classificazione del molare inferiore destro per numero di canali su 1000 soggetti.**

| Canali $x_i$ | Soggetti $n_i$ | Freq. Rel. $f_i$ |
|---|---|---|
| 1 | 2 | 0.0002 |
| 2 | 914 | 0.9140 |
| 3 | 76 | 0.0076 |
| 4 | 8 | 0.0008 |
| | 1000 | 1.0000 |

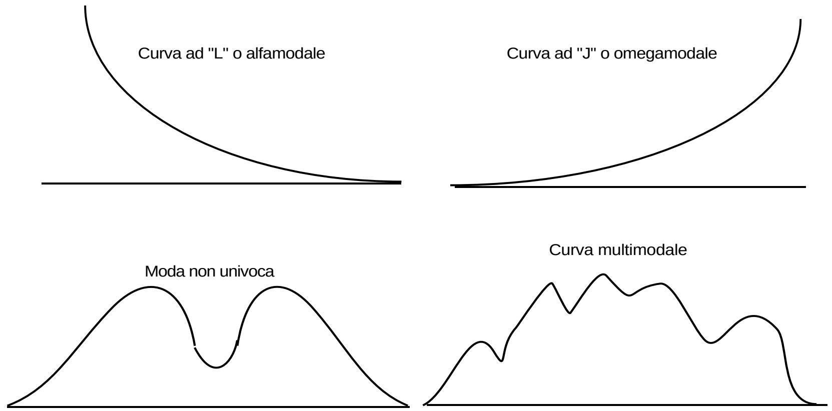



**La determinazione della Moda è influenzata solo dai rapporti ordinali tra le frequenze**



# Amodalità e multimodalità

**La moda può essere assente dalla distribuzione.**




**La frequenza massima può non corrispondere ad un valore unico.**

**Inoltre, pur esistendo, la moda può non essere significativa, cioè le frequenze relative potrebbero differire troppo poco**



# Difetto

***Poiché non usa tutti i dati la moda può dare indicazioni fuorvianti***

## ESEMPIO:

**n=20 uomini arrestati per violenza in famiglia furono sottoposti a vigilanza speciale per 2 anni. Ecco la distribuzion degli arresti alla fine del periodo:**

| Arresti | Criminali |
|---|---|
| 0 | 8 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 2 |
| 5 | 4 |
| 6 | 3 |
| | 20 |

**La moda è $M_o$=0 arresti.**

**La frequenza modale è elevata (doppia della 2ª in ordine di grandezza)**

***Tuttavia dire che zero arresti è tipico nasconderebbe un gruppo di criminali che ha reiterato il reato 5 o 6 volte***



# Modalità in classi

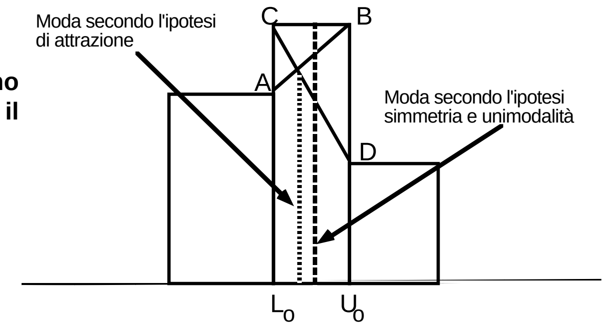
**La frequenza relativa più elevata individuerà la classe modale.**

## Metodo N. 1

**I valori della classe modale hanno uguale frequenza. La Moda è il valore centrale della classe:**

$$M_o = c_m = \frac{U_m + L_m}{2}$$

$$f_m \geq f_i \quad \forall i$$




## Metodo N. 2

**La moda è più vicina all'estremo Lm o Um che confina con la classe con più frequenza (che esercita maggiore "attrazione")**

$$M_o = L_m + \frac{(f_m - f_{m-1})}{[(f_m - f_{m-1}) + (f_m - f_{m+1})]} * (U_m - L_m)$$

# Esempio

**Ordinazioni per importi**

**Classe modale: (4 - 5.9);**

**Valore centrale classe modale:**

$$\frac{4.0+5.9}{2} \cong 5$$

| Importi | | Ordini | $f_i$ | $d_i$ | $h_i$ |
|---|---|---|---|---|---|
| 0.0 | 1.9 | 102 | 0.1726 | 1.9 | 0.0908 |
| 2.0 | 3.9 | 175 | 0.2961 | 1.9 | 0.1558 |
| 4.0 | 5.9 | 208 | 0.3519 | 1.9 | 0.1852 |
| 6.0 | 7.9 | 76 | 0.1286 | 1.9 | 0.0677 |
| 8.0 | 9.9 | 23 | 0.0389 | 1.9 | 0.0205 |
| 10.0 | 11.9 | 7 | 0.0118 | 1.9 | 0.0062 |
| | | 591 | 1.0000 | | |

**Attrazione:** $$4.0+\left[\frac{0.1852-0.1558}{(0.1852-0.1558)+(0.1852-0.0677)}\right]*1.9=4.38$$



# Ampiezze diverse

**Se le modalità sono espresse in classi si impone la considerazione della loro eventuale differenza di ampiezza.**

| $L_i$ | $U_i$ | $n_i$ | $d_i$ | $h_i$ | $c_i$ |
|---|---|---|---|---|---|
| 3 | 5 | 8 | 2 | 4 | 4 |
| 5 | 11 | 18 | 6 | 3 | 8 |

**In questi casi occorre fare riferimento non più alle frequenze relative, bensì alle densità di frequenza ovvero alle altezze già viste nella costruzione degli istogrammi:**

$$M_o = c_m = \frac{L_m + U_m}{2} \qquad \text{dove } (h_m \geq h_i \text{ per } i=1,2,\ldots,k)$$

$$M_o = L_m + \frac{(h_m - h_{m-1})}{[(h_m - h_{m-1}) + (h_m - h_{m+1})]} * (U_m - L_m)$$



# Esempio

**Una radiografia è stata scomposta in pixel e su ciascuno di questi si è misurato il tono di grigio**

**Classe modale: (31-49);**

**Valore centrale classe modale:**

$$\frac{31+49}{2}=40$$

| Riflettenza | | Pixel | $f_i$ | $d_i$ | $h_i$ |
|---|---|---|---|---|---|
| 0 | 30 | 54 | 0.0113 | 30.0 | 0.0004 |
| 31 | 49 | 613 | 0.1277 | 18.0 | 0.0071 |
| 50 | 98 | 421 | 0.0877 | 48.0 | 0.0018 |
| 99 | 127 | 716 | 0.1492 | 28.0 | 0.0053 |
| 128 | 160 | 432 | 0.0900 | 32.0 | 0.0028 |
| 161 | 191 | 798 | 0.1663 | 30.0 | 0.0055 |
| 192 | 240 | 1579 | 0.3290 | 48.0 | 0.0069 |
| 241 | 255 | 187 | 0.0390 | 14.0 | 0.0028 |
| | | 4800 | 1.0000 | | |

**Attrazione:** $$31+\left[\frac{0.0071-0.0004}{(0.0071-0.0004)+(0.0071-0.0018)}\right]*18=41.05$$



# La mediana

**La Mediana è la modalità che è preceduta e seguita dalle altre con uguale frequenza, si trova cioè in posizione centrale nella graduatoria delle modalità.**

**ESEMPIO:**

**Soldati schierati in ordine di altezza: la fila posta al centro è quella mediana**

**Ricordando che, se non altrimenti indicato, le modalità sono ordinate in senso crescente, la Mediana di "n" osservazioni è data da:**

$$M_e = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & se\ "n"\ \grave{e}\ dispari \\ \dfrac{X_{(n/2)} + X_{(n/2)+1}}{2} & se\ "n"\ \grave{e}\ pari \end{cases}$$

# Esempio

**Si considerino i seguenti dati relativi ai costi di estrazione dollaro/barile.**

| | | | |
|---|---|---|---|
| Usa/Alaska | 7.5 | Canada | 5.0 |
| Messico | 6.0 | Venezuela | 6.2 |
| Argentina | 15.1 | Medio Oriente | 2.5 |
| Indonesia | 10.7 | Africa | 7.4 |
| Nord Europa | 17.5 | URSS | 6.3 |

**I dati ordinati sono:**

$$\{2.5, 5.0, 6.0, 6.2, 6.3, 7.4, 7.5, 10.7, 15.1, 17.5\}$$

**Poichè n=10 (cioè n è pari)**

$$M_e = \frac{X_{\left(\frac{10}{2}\right)} + X_{\left(\frac{10}{2}+1\right)}}{2} = \frac{X_{(5)} + X_{(6)}}{2} = \frac{6.3 + 7.4}{2} = 6.85$$



# Proprietà della mediana

**La Mediana rende minima la somma dei moduli degli scarti delle modalità**

$$Q(A) = \sum_{i=1}^{k} |X_i - A| f_i$$

**è minima se A=M$_e$.**

$$\begin{aligned}
Q(A) &= \sum_{X_{(i)}>A} \left(X_{(i)} - A\right) f_{(i)} + \sum_{X_{(i)}\le A} \left(A - X_{(i)}\right) f_{(i)} \\
&= \sum_{X_{(i)}>A} X_{(i)} f_{(i)} - \sum_{X_{(i)}>A} A f_{(i)} + \sum_{X_{(i)}\le A} A f_{(i)} - \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} \pm \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} \\
&= \sum_{X_{(i)}>A} X_{(i)} f_{(i)} - A\left(\sum_{X_{(i)}>A} f_{(i)}\right) + A\left(\sum_{X_{(i)}\le A} f_{(i)}\right) - \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} + \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} - \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} \\
&= \sum_{X_{(i)}>A} X_{(i)} f_{(i)} + \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} - A\left[1 - F(A)\right] + AF(A) - 2\sum_{X_{(i)}\le A} X_{(i)} f_{(i)} \\
&= \mu - A\left[1 - 2F(A)\right] - 2\sum_{X_{(i)}\le A} X_{(i)} f_{(i)}
\end{aligned}$$

dove μ è non dipende da “A”.

Q(A) è decrescente per A tale che F(A)<0.5 ed è crescente per F(A)>0.5; il minimo è perciò raggiunto per F(A)=0.5 che corrisponde a A=M$_e$.



# La mediana per dati raggruppati

**La mediana corrisponde alla modalità più piccola tra quelle che hanno frequenza relativa cumulata è maggiore o uguale a 0.5**

$$Min\{x \in S | F(x) \ge 0.5\}$$

**ESEMPIO:**

**Classificazione dei clienti di un punto vendita per numero di acquisti effettuati nel mese**

| Acquisti | Clienti | $f_i$ | $F_i$ |
|---|---|---|---|
| 0 | 40 | 0.0964 | 0.0964 |
| 1 | 69 | 0.1663 | 0.2627 |
| 2 | 95 | 0.2289 | 0.4916 |
| 3 | 111 | 0.2675 | 0.7590 |
| 4 | 74 | 0.1783 | 0.9373 |
| 5 | 26 | 0.0627 | 1.0000 |
| | 415 | 1.0000 | |

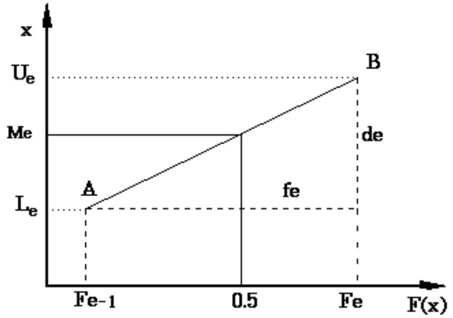
**La Mediana è "3 acquisti"**



# Modalità in classi

**E’ possibile individuare solo la classe mediana, ovvero quella cui corrisponde la frequenza relativa cumulata di 0.5.**

**Per calcolare la mediana si ipotizza che le unità siano uniformi nella classe mediana per cui i segmenti della curva di graduazione sono rette ascendenti.**




**La retta AB ha equazione**

$$X = L_e + \frac{\left(F - F_{e-1}\right)}{h_e}$$

$$M_e = L_e + \frac{\left(0.5 - F_{e-1}\right)}{h_e}; \quad per\ \text{"}e\text{"}\ tale\ che\ F_e = \underset{1\le j\le k}{Min}\left\{F | F_j \ge 0.5\right\}$$

# Esempio

**Uno studio di consulenza ha classificato le operazioni di auditing per la revisione dei conti annuali secondo la durata in giorni. Calcolo della mediana**

| Durata | | Revisioni | $f_i$ | $F_i$ |
|---|---|---|---|---|
| 5 | 7 | 5 | 0.0595 | 0.0595 |
| 8 | 10 | 9 | 0.1071 | 0.1667 |
| 10 | 14 | 14 | 0.1667 | 0.3333 |
| 15 | 19 | 18 | 0.2143 | $\boxed{0.5476}$ |
| 20 | 24 | 15 | 0.1786 | 0.7262 |
| 25 | 29 | 12 | 0.1429 | 0.8690 |
| 30 | 34 | 11 | 0.1310 | 1.0000 |
| | | 84 | 1.0000 | |

$$M_e = 15 + \frac{(0.5 - 0.3333)}{0.2143/4} = 18.11$$

**Il calcolo della mediana avviene in due passi:**

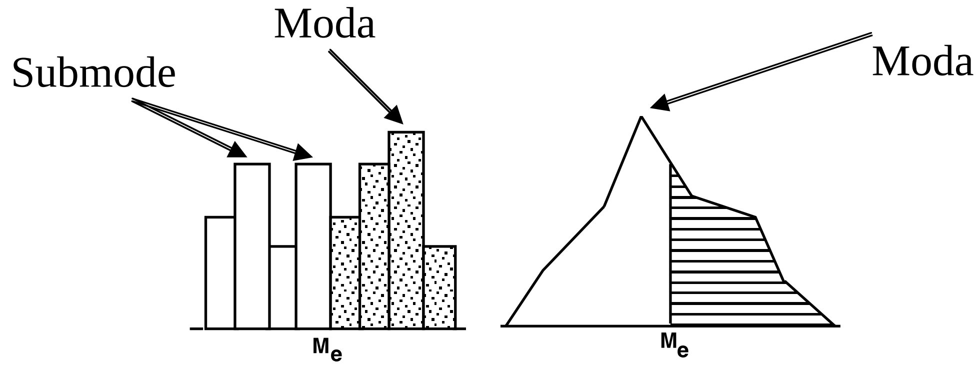
**1) Si individua la classe mediana**

**2) Si interpola per ottenere il valore puntuale.**



# Individuazione grafica

**La mediana corrisponde alla retta X=$M_e$ che separa due parti uguali dell'istogramma o dell'area sottesa al poligono di frequenza (se le classi hanno uguale ampiezza).**






# I quantili

**L'idea di valori di soglia che suddividano le modalità in particolari gruppi può essere generalizzata definendo il quantile di ordine "p"**

Modalità discrete

$$X_p = (1-\gamma)X_{(i)} + \gamma X_{(i+1)},\ \ 0<p<1;\ \ i \le np < i+1;\ \ \gamma = \begin{cases} 0.5 & se\ [np] = np \\ 1 & se\ [np] < np \end{cases}$$

Modalità continue non in classi

$$X_p = (1-\gamma)X_{(i)} + \gamma X_{(i+1)} \qquad i = [np + 0.5];\ \ \gamma = np + 0.5 - i$$

Modalità dense o continue in classi

$$X_p = (1-\gamma)L_i + \gamma U_i = L_i + \gamma(U_i - L_i);\ \ \ F_{i-1} \le p < F_i;\ \ \gamma = \frac{p - F_{i-1}}{F_i - F_{i-1}}$$

**il quantile supera il p% delle modalità ed è superato dall' (1-p)%**



# Esempi

Discrete o dense singole

**Consideriamo le n=18 rilevazioni degli arrivi di auto ad un punto di imbarco e calcoliamo il quantile del 17%.**

$$n * p = 18 * 0.17 = 3.06 \Rightarrow [np] < np \Rightarrow \gamma = 1$$

$$X_{0.17} = 0 * X_{(3)} + 1 * X_{(4)} = X_{(4)} = 744$$

| | | | | | |
|---|---|---|---|---|---|
| 612 | 623 | 666 | 744 | 883 | 898 |
| 964 | 970 | 983 | 1003 | 1016 | 1022 |
| 1029 | 1058 | 1085 | 1088 | 1122 | 1135 |

Continue singole

**Principali coltivazioni agricole delle Marche. Anno 1998. Valori in ettari. Calcolo del quantile di ordine 0.60**

$$n * p = 9 * 0.60 = 5.4 \Rightarrow i = [5.4 + 0.5] = [5.9] = 5;$$

$$\gamma = 5.9 - 5 = 0.9$$

$$X_{0.6} = 0.1 * X_{(5)} + 0.9 * X_{(6)} = 23'300.$$

| Coltivazione | Superficie |
|---|---|
| Pomodoro | 1'304 |
| Pesca | 1'486 |
| Cavolfiore | 1'967 |
| Olivo | 6'218 |
| Mais ibrido | 14'558 |
| Uva da vino | 24'272 |
| Grano tenero | 36'553 |
| Girasole | 38'281 |
| Grano duro | 123'049 |

# Esempi

Modalità in classi

**Dimensioni delle operazioni di fusione e di acquisizione in Italia per fatturato.**
**Calcolo di $X_{0.80}$.**

| Fatturato | | Operazioni | $f_i$ | $F_i$ |
|---|---|---|---|---|
| 1 | 5 | 30 | 0.2804 | 0.2804 |
| 5 | 20 | 36 | 0.3364 | 0.6168 |
| 20 | 40 | 14 | 0.1308 | 0.7477 |
| 40 | 60 | 10 | 0.0935 | 0.8411 |
| 60 | 100 | 12 | 0.1121 | 0.9533 |
| 100 | 150 | 5 | 0.0467 | 1.0000 |
| | | 107 | 1.0000 | |

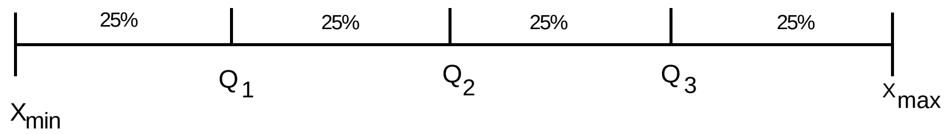
$$X_{0.80} = 40 + \frac{(0.80 - 0.7477)}{(0.0935/40)} = 51.19$$



# Uso dei quantili

**In genere, i quantili si utilizzano come soglie di separazione delle modalità in gruppi di numerosità prestabilita.**

**Fra i più usati sono da annoverare i <u>tre quartili</u> che suddividono le modalità in quattro gruppi ciascuno comprendente il 25% delle modalità:**




**Q1 supera il 25% ed è superato del 75%, Q2 coincide con la mediana ed è superato da tante unità quante ne supera esso stesso; Q3 supera il 75% delle unità ed è superato dal restante 25%.**

**Tra due soglie è sempre compreso il 25% di unità.**



# Esempio

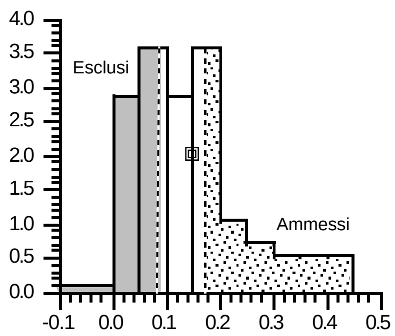
**L'esito di una selezione per l'ammissione ad un corso universitario a numero chiuso è riassunto nella tabella.**

**La commissione decide di ammettere il 40% con punteggio più alto, di escludere il 25% inferiore e di sottoporre il restante 35% a dei test suppletivi. Quali sono le soglie di divisione?**

$$X_{0.25} = 0.05 + \left(\frac{0.25 - 0.152}{0.205}\right)(0.10 - 0.05) = 0.0843;$$

$$X_{0.60} = 0.15 + \left(\frac{0.60 - 0.536}{0.143}\right)(0.20 - 0.15) = 0.1724$$

| Punteggio | | Candidati | $f_i$ | $F_i$ |
|---|---|---|---|---|
| | <0.00 | 1 | 0.009 | 0.009 |
| 0.00 | 0.05 | 16 | 0.143 | 0.152 |
| 0.05 | 0.10 | 23 | 0.205 | 0.357 |
| 0.10 | 0.15 | 20 | 0.179 | 0.536 |
| 0.15 | 0.20 | 16 | 0.143 | 0.679 |
| 0.20 | 0.25 | 20 | 0.179 | 0.857 |
| 0.25 | 0.30 | 6 | 0.054 | 0.911 |
| 0.30 | 0.35 | 4 | 0.036 | 0.946 |
| 0.35 | 0.40 | 3 | 0.027 | 0.973 |
| | >0.40 | 3 | 0.027 | 1.000 |
| | | 112 | 1.000 | |






# La media aritmetica

**E' la media per antonomasia, quella che si sottintende se non si qualifica ulteriormente il termine di media.**

$$\overline{X} = \sum_{i=1}^{k} X_i f_i = X_1 f_1 + X_2 f_2 + \ldots + X_k f_k$$

**La media aritmetica è il punto di equilibrio fisico delle modalità spaziate in base al loro valore numerico e pesate con le frequenze relative:**

| $x_i$ | $n_i$ | $f_i$ |
|---|---|---|
| -5 | 1 | 1/9 |
| -1 | 1 | 1/9 |
| 0 | 1 | 1/9 |
| 3 | 1 | 1/9 |
| 4 | 1 | 1/9 |
| 5 | 1 | 1/9 |
| 7 | 3 | 3/9 |
| | 9 | |

$$M_a = -5 \cdot \frac{1}{9} - 1 \cdot \frac{1}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{1}{9} + 5 \cdot \frac{1}{9} + 7 \frac{3}{9}$$
$$= \frac{27}{9} = 3$$

# Esempio

**Numero di figli maschi in famiglie di otto figli**

| Maschi $x_i$ | Famiglie $n_i$ | Freq. Rel. $f_i$ | Prodotti $x_i f_i$ |
|---|---|---|---|
| 0 | 161 | 0.0042 | 0.0000 |
| 1 | 152 | 0.0304 | 0.0304 |
| 2 | 3957 | 0.1043 | 0.2086 |
| 3 | 7603 | 0.2004 | 0.6012 |
| 4 | 10263 | 0.2705 | 1.0821 |
| 5 | 8498 | 0.2240 | 1.1200 |
| 6 | 4984 | 0.1314 | 0.7883 |
| 7 | 1055 | 0.0278 | 0.1947 |
| 8 | 264 | 0.0070 | 0.0557 |
| | 37937 | 1.0000 | 4.0809 |

***Da notare che si può calcolare moltiplicando le modalità per le frequenze assolute e poi dividendo l'ammontare ottenuto per il totale delle frequenze***



# Proprietà della media aritmetica

**La media aritmetica, se sostituita alle modalità, mantiene inalterato l'ammontare complessivo nella rilevazione.**

**Il totale delle modalità è infatti:** $T = \sum_{i=1}^{k} x_i n_i$

**Se al posto di $X_i$ si pone la media aritmetica si ottiene**

$$\sum_{i=1}^{n} \overline{X} n_i = \overline{X} \sum_{i=1}^{n} n_i = \overline{X} n = \frac{T}{n} n = T$$

***Quindi la media aritmetica è quella quantità che ciascuna unità avrebbe se tutte avessero la stessa parte di variabile.***



# Internalità

**Considerando le modalità in senso ascendente saranno vere le relazioni:**

$$\sum_{i=1}^{k} X_{\min} f_i \leq \sum_{i=1}^{k} X_i f_i \leq \sum_{i=1}^{k} X_{\max} f_i$$

**Ogni addendo della prima somma è inferiore o uguale ad ognuno della seconda che a loro volta sono inferiori o uguali a quelli della terza.**

**Ne consegue che:**

$$X_{\min} \sum_{i=1}^{k} f_i \leq \sum_{i=1}^{k} X_i f_i \leq X_{\max} \sum_{i=1}^{k} f_i \Rightarrow X_{\min} \leq \sum_{i=1}^{k} X_i f_i \leq X_{\max}$$



# Somma nulla degli scarti

**La media aritmetica rende nulla la somma degli scarti tra le modalità e la media.**

$$\sum_{i=1}^{k} \left(X_i - \overline{X}\right) f_i = \sum_{i=1}^{k} X_i f_i - \sum_{i=1}^{k} \overline{X} f_i = \sum_{i=1}^{k} X_i f_i - \overline{X} = \sum_{i=1}^{k} X_i f_i - \overline{X} = \overline{X} - \overline{X} = 0$$

## ESEMPIO

**Numero di dipendenti formati nei comuni e nelle province per settori.**

| Area | Dipendenti | Scarto |
|---|---|---|
| Interventi settoriali | 15913 | 5887 |
| Managerialità | 10801 | 775 |
| Organizzazione | 10711 | 685 |
| Controllo di gestione | 9362 | -664 |
| Informatizzazione | 3938 | -6088 |
| Gestione del personale | 9431 | -595 |
| Totale | 60156 | 0 |

La media aritmetica è $\overline{X} = 10026$ che, annullando la somma degli scarti, si conferma valore di equilibrio per la distribuzione.

# Minimo della somma degli scarti al quadrato

$$\sum_{i=1}^{k}(X_i - A)^2 f_i = \sum_{i=1}^{k}\left[(X_i - \overline{X}) + (\overline{X} - A)\right]^2 f_i$$

$$= \sum_{i=1}^{k}\left[(X_i - \overline{X})^2 + (\overline{X} - A)^2 + 2(\overline{X} - A)(X_i - \overline{X})\right] f_i$$

$$= \sum_{i=1}^{k}(X_i - \overline{X})^2 f_i + \sum_{i=1}^{k}(\overline{X} - A)^2 f_i + 2(\overline{X} - A)\sum_{i=1}^{k}(X_i - \overline{X}) f_i$$

$$= \sum_{i=1}^{k}(X_i - \overline{X})^2 f_i + \sum_{i=1}^{k}(\overline{X} - A)^2 f_i$$

**Il terzo termine risulta nullo per la proprietà già dimostrata della media aritmetica di annullare la somma degli scarti semplici.**

$$= \sum_{i=1}^{k}(X_i - \overline{X})^2 f_i + (\overline{X} - A)^2$$

**il primo addendo non dipende da "A" il 2° è semplicemente un quadrato che ha il minimo nello zero raggiunto per** $A = \overline{X}$



# Riproducibilità per trasformazioni lineari

**Quando la variabile x subisce una trasformazione lineare del tipo y=a + bx lo stesso succede alla media aritmetica.**

**Infatti:** $\overline{Y} = \sum_{i=1}^{k} Y_i f_i = \sum_{i=1}^{k}(a + bX_i) f_i = \sum_{i=1}^{k} a f_i + b\sum_{i=1}^{k} X_i = a + b\overline{X}$

## ESEMPIO

**Bilancio delle principali squadre di calcio di serie A (in milioni). Conversione in migliaia di dollari.**

| Squadre | M. Lire | m. Dollari |
|---|---|---|
| Juventus | 1847 | 947.18 |
| Milan | -27093 | -13893.85 |
| Inter | -21442 | -10995.90 |
| Roma | 504 | 258.46 |
| Parma | -25418 | -13034.87 |
| Lazio | 251 | 128.72 |
| Fiorentina | -10579 | -5425.13 |
| Sampdoria | -879 | -450.77 |
| Bologna | -8822 | -4524.10 |
| | -10181.22 | -5221.14 |

il rapporto di conversione tra milioni di lire e migliaia di dollari è:

$$a = 0.0, \quad b = \frac{1000}{1950} = 0.512821$$

$$\overline{X} = -10181.22, \quad \overline{Y} = 0.512821 * (-10181.22) = -5221.14$$



# Proprietà associativa

**Supponiamo di individuare "g" gruppi. Le modalità sono individuate con due indici: il primo per il gruppo ed il secondo per le unità del gruppo:**

| *Gruppi* | *Valori* | | | *Unità* |
|---|---|---|---|---|
| $G_1$ | $X_{11}$ | $X_{12}$ | $X_{1k_1}$ | $n_1$ |
| $G_2$ | $X_{21}$ | $X_{22}$ | $X_{2k_2}$ | $n_2$ |
| $\vdots$ | | | | |
| $G_g$ | $X_{g1}$ | $X_{g2}$ | $X_{gk_g}$ | $n_g$ |

$$\sum_{i=1}^{g} n_i = n$$

**Per ogni gruppo si può calcolare la propria media aritmetica**

$$\mu_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}; \quad i = 1, 2, \ldots, g$$



# Proprietà associativa/2

**La proprietà associativa consente di ricavare la media aritmetica complessiva:**

$$\mu = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i} X_{ij}}{n} = \frac{\sum_{i=1}^{g} n_i \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}}{n} = \frac{\sum_{j=1}^{g} n_i \mu_i}{n} = \sum_{j=1}^{g} \mu_i f_i$$

## ESEMPIO

**Per tre diverse aree si è considerato il consumo medio annuo di zucchero.**

| *Aree* | *Unità* | *Medie* |
|---|---|---|
| *Zona_A* | 647 | 115 |
| *Zona_B* | 173 | 80 |
| *Zona_C* | 435 | 75 |
| *totale* | 1255 | ? |

$$\overline{X} = \frac{647}{1255} * 115 + \frac{173}{1255} * 80 + \frac{435}{1255} * 75 = 96.31$$

# Modalità raggruppate in classi

E' un caso particolare della proprietà associativa.

Purtroppo le medie aritmetiche di classe non sono note ed occorre stimarle.

La tecnica più usata è quella di adoperare il valore centrale delle classi:

$$\overline{X}_i = \frac{1}{2}L_i + \frac{1}{2}U_i \qquad i = 1, 2, \ldots, k$$

**ESEMPIO**

Casi di epatite A in un comune.

Le classi estreme hanno un'ampiezza pari alla metà delle ampiezza delle classi loro continue.

| Età | | Pazienti | $c_i$ | $f_i$ | $Xif_i$ |
|---|---|---|---|---|---|
| ≤9 | | 662 | 6.75 | 0.5400 | 3.6448 |
| 10 | 19 | 420 | 14.50 | 0.3426 | 4.9674 |
| 20 | 29 | 117 | 24.50 | 0.0954 | 2.3381 |
| 30 | 39 | 18 | 34.50 | 0.0147 | 0.5065 |
| 40 | 49 | 5 | 44.50 | 0.0041 | 0.1815 |
| ≥50 | | 4 | 52.25 | 0.0033 | 0.1705 |
| | | 1226 | | 1.0000 | 11.8087 |



# Le medie di potenze

La media aritmetica rientra in una classe che fornisce una espressione dell'ordine di grandezza del fenomeno a partire da tutti i valori riscontrati:

$$M(X_1,\ldots,X_k; f_1,\ldots,f_k; \alpha) = \left\{\sum_{i=1}^{k} X_i^{\alpha} f_i\right\}^{1/\alpha}$$

Che scaturiscono dal principio di Chisini:

La misura di centralità deve lasciare invariato un particolare aspetto del fenomeno allorché al posto di ogni $X_i$ si sostituisce "M":

Ad esempio, la media aritmetica posta in vece di ciascuna modalità lascia invariato l'ammontare complessivo delle rilevazioni

$$\sum_{i=1}^{k} \overline{X} n_i = \overline{X} \sum_{i=1}^{k} n_i = \frac{\sum_{i=1}^{k} X_i n_i}{n} * n = \sum_{i=1}^{k} X_i n_i = T$$



# La media geometrica

E' una media particolarmente adatta come misura di centralità per fenomeni evolutivi che si realizzano proporzionalmente al livello già raggiunto.

**Formula classica**

$$Mg = \prod_{i=1}^{k} x_i^{f_i} = x_1^{f1} * x_2^{f2} * \ldots * x_k^{fk} \qquad \text{per } x_i > 0$$

**Formula per il calcolo**

$$M_g = e^{\left\{\sum_{i=1}^{k} f_i Ln(x_i)\right\}}$$

Esempio.

| $x_i$ | $n_i$ | $f_i$ | $Ln(x_i)$ | $f_i Ln(x_i)$ |
|---|---|---|---|---|
| 1 | 2 | 2/10 | 0.0000 | 0.0000 |
| 3 | 1 | 1/10 | 1.0986 | 0.1099 |
| 4 | 1 | 1/10 | 1.3863 | 0.1386 |
| 5 | 1 | 1/10 | 1.6094 | 0.1609 |
| 6 | 1 | 1/10 | 1.7918 | 0.1792 |
| 7 | 1 | 1/10 | 1.9459 | 0.1946 |
| 9 | 3 | 3/10 | 2.1972 | 0.6592 |
| | 10 | | | 1.4424 |

$$M_g = e^{1.4424} = 4.2308$$



# Proprietà della media geometrica

E' meno soggetta a variazioni rispetto alla loro media aritmetica:

| $X_i$ | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | | 204.6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Log(X_i)$ | 0.3010 | 0.6021 | 0.9031 | 1.2041 | 1.5051 | 1.8062 | 2.1072 | 2.4082 | 2.7093 | 3.0103 | 1.6557 | 45.2548 |

**G=45.25 supera ed è superata dallo stesso numero di valori rispetto alla media aritmetica: μ=204 che ne supera sette ed è superata solo da tre.**

# La media armonica

**Si tratta di un indice posizione da utilizzare soprattutto per misurare la centralità in situazioni in cui interessa il contributo delle modalità alla composizione di un tutto.**

$$H=\frac{1}{\sum_{i=1}^{k}\frac{f_i}{X_i}};\ per\ \ X_i\neq 0$$

| Figli | Famiglie | $f_i$ | $(1/X_i)f_i$ |
|---|---|---|---|
| 1 | 185 | 0.5362 | 0.5362 |
| 2 | 78 | 0.2261 | 0.1130 |
| 3 | 33 | 0.0957 | 0.0319 |
| 4 | 25 | 0.0725 | 0.0181 |
| 5 | 13 | 0.0377 | 0.0075 |
| 6 | 8 | 0.0232 | 0.0039 |
| 7 | 3 | 0.0087 | 0.0012 |
| | 345 | 1.0000 | 0.7119 |

Per calcolare la media armonica si calcola prima la media aritmetica di reciproci e di questa si calcola il reciproco:

$$H=\frac{1}{0.7119}=1.4047$$



# Valori anomali e medie

**La presenza di valori isolati rispetto al resto della distribuzione ha una incidenza diversa a secondo della media che si usa**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| A | 5 | 8 | 9 | 9 | 10 | 11 | 12 | 15 | 20 |
| B | 5 | 8 | 9 | 9 | 10 | 12 | 12 | 15 | $\boxed{2013}$ |

**valore isolato**

**μ passa da 11 a 210**

**La media geometrica cresce, ma meno della media aritmeticae più della armonica.**

**La mediana, non cambia (al massimo si sposta di una posizione)**

**La moda non cambia (il valore isolato per definizione non può fare moda)**



# Valori remoti (outlier)

**sono valori (uno o più) che appaiono inusuali senza che li si possa ritenere errati**

**Ciò dipende dal contesto:**

**{5, 13, 2, $\boxed{291,}$ 11, 6}**

***Non è un outlier né un errore tipografico: numero di clienti in coda ad uno sportello***

**{3.6, 2.7, 2.8, 3.9, 2.1, $\boxed{85.8,}$ 3.4, 2.8}**

***Peso alla nascita di un campione di neonati . E' anomalo se si tratta di persone***



# Valori remoti /2

**Dato un insieme di modalità quantitative, una di esse sarà maggiore delle altre ed un'altra sarà minore;**

**se gli estremi sono molto remoti nascerà il sospetto di disfunzioni:**

**Attenzione! Se un fenomeno può produrre modalità estremamente piccole e/o estremamente grandi è inevitabile che qualcuno si mostri prima o poi.**

***una ASL ha storicamente richiesto il rimborso di un numero mensile di parti cesarei con complicazioni che oscilla tra i 20 ed i 30. Un dato mese, richiede rimborsi per 120.***


***Questo non necessariamente è un dato anomalo, ma è la spia di un cambiamento nel meccanismo dei rimborsi o nel management della ASL.***