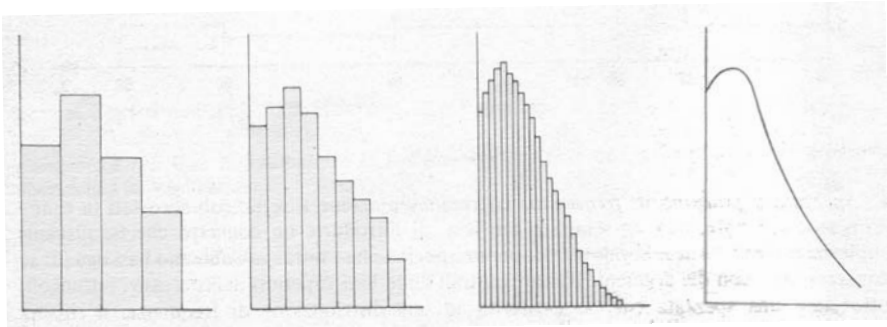


ANALISI DI UNA DISTRIBUZIONE

Indici di centralità

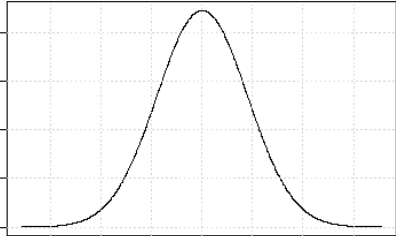
Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DI UNA DISTRIBUZIONE



The image shows four histograms in a row, illustrating the process of approximating a continuous distribution. From left to right: 1) A histogram with 4 bars of varying heights. 2) A histogram with 8 bars, showing a smoother curve. 3) A histogram with 16 bars, appearing as a dense staircase. 4) A smooth, continuous curve representing the limit of the distribution.

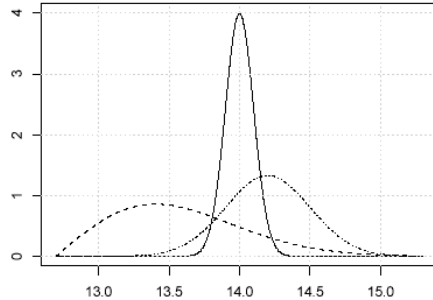
- Al crescere del numero di osservazioni, e riducendo l'ampiezza degli intervalli, l'istogramma di frequenze tende a diventare una curva, che rappresenta la forma della distribuzione della variabile
- La curva descrive la distribuzione di frequenze relative: quindi l'area totale sottesa alla curva è pari a 1.



A graph showing a smooth, bell-shaped curve (normal distribution) plotted on a grid. The curve is symmetric and centered, representing the limit of the histograms as the number of observations increases and the interval width decreases.

ANALISI DI UNA DISTRIBUZIONE

- Una distribuzione di frequenze descrive l'andamento di un "fenomeno" in una popolazione, che presenta una variabilità, o dispersione, intorno ad un valore centrale.
- Una distribuzione è caratterizzata prima di tutto dalla sua **forma**:
 - alta e stretta / larga e piatta
 - simmetrica / asimmetrica
 - uno / più "picchi"
- Possiamo individuare due principali parametri che descrivono sinteticamente una distribuzione:
 - la **tendenza centrale**
 - la **dispersione**
- Per descrivere la **tendenza centrale** di una distribuzione, ovvero la modalità intorno alla quale si concentra maggiormente il fenomeno, esistono diverse possibilità:
 - la modalità che si presenta con maggiore frequenza
 - la modalità centrale, rispetto alla quale metà delle osservazioni sono risultate minori e l'altra metà maggiori
 - il valore medio, ottenuto sommando tutti i valori osservati e dividendo il risultato per loro numero



ANALISI DI UNA DISTRIBUZIONE

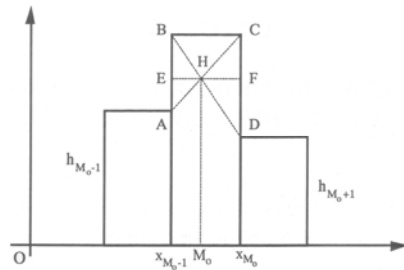
- **Moda**: la modalità in corrispondenza della quale si osserva la frequenza maggiore
 - è l'unico indice di centralità che si può determinare per le variabili qualitative *nominali*
 - possono esistere più modalità che presentano la stessa frequenza massima: se ad es. ne abbiamo due, la distribuzione si dice **bimodale**
 - nel caso di una distribuzione *uniforme*, tutte le frequenze sono uguali: la distribuzione è priva di moda
- Esempio:
 - La moda è pari a 19 per entrambe le facoltà
 - Se consideriamo la distribuzione da 20 a 30 anni:
 - per Formazione la moda risulta 20
 - per Filosofia la distribuzione risulta bimodale, con le due mode: 20 e 23

Età	Frequenze Assolute	
	Scienze Formazione	Filosofia
19	350	80
20	300	70
21	250	60
22	200	55
23	150	70
24	180	60
25	200	30
26	80	20
27	130	10
28	60	0
29	25	5
30	50	0
31	10	0
32	5	0
33	0	15
34	5	25
35	5	0
Totale	2000	500

ANALISI DI UNA DISTRIBUZIONE

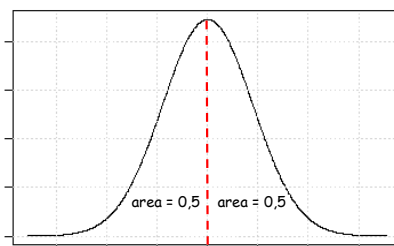
- Per le variabili quantitative *continue*, quando si ha una tabella di frequenze costruita a partire da classi di valori, si determina agevolmente la **classe modale**, cioè la classe in cui si concentra la frequenza maggiore
- Se fosse necessario determinare un valore puntuale per la moda, si porrebbe il problema di sapere come sono distribuite le frequenze all'interno della classe.
- Non conoscendo come sono distribuiti i valori all'interno della classe:
 - una soluzione alquanto approssimativa consiste nell'indicare come moda il valore centrale della classe modale
 - una soluzione più sofisticata è quella di determinare il valore per interpolazione: la moda dovrebbe essere più spostata verso la classe contigua con frequenza maggiore

Frequenze Assolute		
Età	Formazione	Filosofia
19-21	900	210
22-24	530	185
25-27	410	60
28-30	135	5
31-33	15	15
34-36	10	25
Totale	2000	500



ANALISI DI UNA DISTRIBUZIONE

- **Mediana:** è il valore centrale della serie dei dati ordinati
 - L'idea che alla base della mediana è cercare un numero che divida a metà l'insieme dei dati, cioè sia maggiore del 50% delle osservazioni e minore del restante 50% dei dati
 - E' quel valore che, in una serie di dati disposti in ordine crescente, è preceduto e seguito dallo stesso numero di osservazioni
 - E' determinabile per variabili su scala almeno ordinale
 - La mediana divide in due parti uguali l'area che sta sotto la curva che rappresenta la distribuzione



ANALISI DI UNA DISTRIBUZIONE

- Come si determina ?

Per prima cosa si ordinano i valori degli N dati osservati:

- quando N è dispari:
 - è semplicemente il termine centrale della serie ordinata: in posizione $(n+1)/2$
- quando N è pari:
 - non abbiamo un termine centrale ma due: può essere assunto come mediana qualunque valore compreso tra i due termini mediani, in posizione $(n/2)$ e $(n/2)+1$
 - convenzionalmente si assume come mediana la semisomma dei due termini mediani

Ad esempio nel grafico seguente, supponendo che le osservazioni corrispondano ai punti disegnati con una 'o', un possibile valore per la mediana è stato indicato con una 'x'. Infatti, il punto così marcato lascia sia a sinistra che a destra 6 osservazioni.



ANALISI DI UNA DISTRIBUZIONE

- Quando abbiamo dati in una tabella di frequenze, si determina la **classe mediana**, in cui cade il valore mediano :

- nella tabella delle frequenze cumulate relative si trova la classe a in cui cade il 50% delle frequenze relative
- quando i dati sono raggruppati in classi di valori, per avere un valore puntuale si può ipotizzare l'uniforme ripartizione dei valori nella classe e determinare il valore mediano per interpolazione
- ad esempio: se all'estremo inferiore della classe la frequenza relativa è 0,45 e all'estremo superiore è 0,55: la mediana cadrà esattamente al centro
- in generale, occorre fare una proporzione:

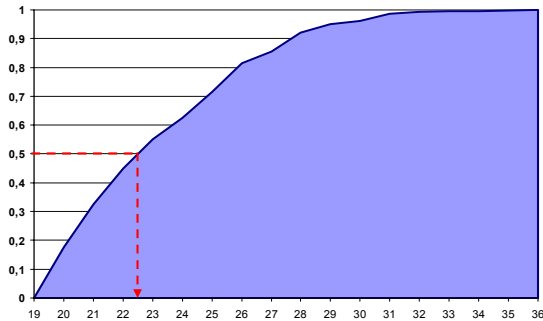
$$Me = 22 + \frac{23 - 22}{0,55 - 0,45} * (0,50 - 0,45) = 22,5$$

Scienze Formazione

Età	Frequenze Relative	Freq. Rel. Cumulate
19-20	0,1750	0,1750
20-21	0,1500	0,3250
21-22	0,1250	0,4500
22-23	0,1000	0,5500
23-24	0,0750	0,6250
24-25	0,0900	0,7150
25-26	0,1000	0,8150
26-27	0,0400	0,8550
27-28	0,0650	0,9200
28-29	0,0300	0,9500
29-30	0,0125	0,9625
30-31	0,0250	0,9875
31-32	0,0050	0,9925
32-33	0,0025	0,9950
33-34	0,0000	0,9950
34-35	0,0025	0,9975
35-36	0,0025	1,0000
Totale	1,0000	

ANALISI DI UNA DISTRIBUZIONE

- La mediana è il valore a cui corrisponde il 50% dei casi sulla Funzione di ripartizione: possiamo determinarla agevolmente proprio a partire dalla curva delle frequenze cumulate
 - sulla funzione di ripartizione, si determina il punto a cui corrisponde una frequenza cumulata pari a 0,5
 - proiettando il punto della curva sull'asse delle ascisse si ottiene il valore mediano



Scienze Formazione		
Età	Frequenze Relative	Freq. Rel. Cumulate
19-20	0,1750	0,1750
20-21	0,1500	0,3250
21-22	0,1250	0,4500
22-23	0,1000	0,5500
23-24	0,0750	0,6250
24-25	0,0900	0,7150
25-26	0,1000	0,8150
26-27	0,0400	0,8550
27-28	0,0650	0,9200
28-29	0,0300	0,9500
29-30	0,0125	0,9625
30-31	0,0250	0,9875
31-32	0,0050	0,9925
32-33	0,0025	0,9950
33-34	0,0000	0,9950
34-35	0,0025	0,9975
35-36	0,0025	1,0000
Totale	1,0000	

ANALISI DI UNA DISTRIBUZIONE

- Proprietà della mediana**
- La mediana minimizza la somma dei valori assoluti degli scarti: cioè è quel valore che rende minima la *somma* di tutti gli scarti presi in *valore assoluto*

$$\sum_{i=1}^n |x_i - Me| = \min \quad \text{ovvero:} \quad \sum_{i=1}^n |x_i - Me| < \sum_{i=1}^n |x_i - a| \quad \forall a \neq Me$$

- Se prendiamo, al posto della mediana, un qualunque altro valore costante (**a**), la somma degli scarti in valore assoluto calcolati rispetto a tale valore, $|x(i) - a|$, risulta, per qualunque insieme di numeri, maggiore della somma degli scarti in valore assoluto calcolati dalla mediana $|x(i) - Me|$
- Questa proprietà ha senso, naturalmente, solamente con riferimento a dati quantitativi (scala intervallo o rapporto)
- La mediana risente poco dei valori estremi perché dipende solo dal numero di osservazioni che cadono alla sua destra e sinistra, e non dalla loro distanza: questa caratteristica è un vantaggio in presenza di valori *anomali*
- Indicazione pratica: la mediana è *resistente* (o *robusta*) rispetto alla presenza di valori anomali (es. errori di misura)
- Vedendo le cose dall'altro punto di vista, si può dire che la mediana non è *sensibile* ai valori estremi, che possono però avere molta importanza, quando non sono anomali ma descrivono situazioni estreme reali

ANALISI DI UNA DISTRIBUZIONE

- Media (Aritmetica Semplice):** è la somma dei valori osservati divisi per il numero totale di unità

$$M(x) = \bar{x} = \mu = \frac{1}{N} \sum_{i=1}^N x_i$$

$x(i)$: valore della modalità i -esima
 N : numero totale di osservazioni

725 24 710 724 700 724 713 692 683 712 694 707 703 691 709 702 705 715
 704 705 697 725 692 719 694 717 696 707 726 703 705 712 710 697 698 694
 701 715 701 707 706 701 687 708 719 713 699 702 694 708 712 704 703 687
 709 693 715 707 710 700 718 702 718 705 723 718 701 698 692 684 716 710
 708 707 695 726 710 709 692 707 717 709 710 718 708 720 705 714 687 707
 707 723 695 676 705 684 717 719 715 710 711 696 696 715 686 702 708 713
 701 692 713 700 704 726 702 706 706 700 700 687 696 694 699 709 704 704
 715 706 688 724 713 686 697 710 704 724 721 717 690 707 713 685 706 699
 687 702 701 708 704 705 702 701 699 699 685 712 678 706 706 695 707 718
 706 716 703 721 714 704 697 693 711 697 710 713 702 715 714 716 698 714
 704 717 700 692 718 699 698 690 710 703 702 719 710 725 721 713 699 703
 698 712 714 707 691 711 712 718 702 711 709 700 719 692 716 700 707 714
 717 714 703 709 711 704 689 712 714 711 692 720 697 698 700 689 693 707
 699 704 696 708 713 714 712 708 704 720 705 703 712 719 713 716 712 703
 717 695 711 697 693 701 699 697 724 713 706 705 704 707 704 719 711 700
 694 706 705 698 697 697 700 705 722 712 703 688 694 708 703 690 706 704

- La media aritmetica è costruita come il valore che può essere sostituito ai dati osservati senza farne variare la somma
- E' calcolabile solo per le variabili quantitative (su scala *intervallo* e scala *rapporto*)
- Familiarizziamo con la notazione statistica: la "formula" è solo un modo di descrivere le operazioni di calcolo da eseguire, più veloce e sintetico rispetto alle parole:

$$\frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} (x_1 + x_2 + \dots + x_i + \dots + x_N)$$



ANALISI DI UNA DISTRIBUZIONE

- Media (Aritmetica) Ponderata:**
- Quando i dati sono organizzati in una tabella di frequenze, ciascuna modalità deve essere pesata per il numero di unità che la presentano
- Si calcola sommando i valori osservati moltiplicati per le rispettive frequenze, diviso la somma dei pesi (pari a N)

$$M(x) = \frac{\sum_{i=1}^k x_i n_i}{\sum_{i=1}^k n_i} = \frac{1}{N} \sum_{i=1}^k x_i n_i$$

$x(i)$: valore della modalità (o classe) i -esima
 $n(i)$: frequenze assolute modalità i -esima
 k : numero di modalità distinte o di classi
 N : numero totale di osservazioni

- Vi sono altre situazioni, in cui i dati possono essere ponderati con pesi diversi dalle frequenze: al denominatore avremo sempre la somma dei pesi, ma non sarà più uguale a N

Età	Frequenze Assolute	
	Scienze Formazione	Filosofia
19	350	80
20	300	70
21	250	60
22	200	55
23	150	70
24	180	60
25	200	30
26	80	20
27	130	10
28	60	0
29	25	5
30	50	0
31	10	0
32	5	0
33	0	15
34	5	25
35	5	0
Totale	2000	500

ANALISI DI UNA DISTRIBUZIONE

- La media ponderata può essere calcolata anche ponderando le osservazioni direttamente con le frequenze relative:

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

- Per esserne convinti, esplicitiamo le somme:

$$\begin{aligned} \bar{x} &= \frac{1}{N} (x_1 n_1 + \dots + x_i n_i + \dots + x_k n_k) = \\ &= (x_1 \frac{n_1}{N} + \dots + x_i \frac{n_i}{N} + \dots + x_k \frac{n_k}{N}) = \\ &= (x_1 f_1 + \dots + x_i f_i + \dots + x_k f_k) = \\ &= \sum_{i=1}^k x_i f_i \end{aligned}$$

- Nel caso della ponderazione con le frequenze relative, la somma dei pesi è uguale a 1

Età	Frequenze Relative	
	Scienze Formazione	Filosofia
19	0,18	0,16
20	0,15	0,14
21	0,13	0,12
22	0,10	0,11
23	0,08	0,14
24	0,09	0,12
25	0,10	0,06
26	0,04	0,04
27	0,07	0,02
28	0,03	0,00
29	0,01	0,01
30	0,03	0,00
31	0,01	0,00
32	0,00	0,00
33	0,00	0,03
34	0,00	0,05
35	0,00	0,00
Totale	1,00	1,00

ANALISI DI UNA DISTRIBUZIONE

- Esercizio.
Calcoliamo l'età media ponderata dei nostri iscritti:

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

Avendo una tabella di frequenze assolute applichiamo la prima formulazione e otteniamo:

$$\begin{aligned} M(\text{Età Scienze Formazione}) &= \\ &= (19 \cdot 350 + 20 \cdot 300 + 21 \cdot 250 + 22 \cdot 200 + \dots + \\ &+ 34 \cdot 5 + 35 \cdot 5) / 2000 = \\ &= (6650 + 6000 + 5250 + \dots + 170 + 175) / 2000 = \\ &= 45380 / 2000 = 22,69 \end{aligned}$$

- Per esercizio calcolare $M(\text{Età Filosofia})$
[= 22,94]

x(i)	n(i)	x(i) n(i)
19	350	6650
20	300	6000
21	250	5250
22	200	4400
23	150	3450
24	180	4320
25	200	5000
26	80	2080
27	130	3510
28	60	1680
29	25	725
30	50	1500
31	10	310
32	5	160
33	0	0
34	5	170
35	5	175
Totale	2000	45380
Media		22,69

ANALISI DI UNA DISTRIBUZIONE

- Nel caso di tabella di frequenze con dati raggruppati in classi di valori, non si conoscono i valori di tutte le osservazioni della classe, ma solo gli estremi dell'intervallo
- Come si calcola allora la media ?
Bisogna scegliere che valore adottare per le classi, come $x(i)$, per applicare la formula
- Ipotizzando anche in questo caso che i dati siano distribuiti uniformemente nell'intervallo, come valore rappresentativo di tutte le unità della classe, si utilizza il valore *centrale* dell'intervallo
- Nel determinare il valore centrale dell'intervallo, occorre prestare attenzione a considerarne correttamente gli estremi
- Nel nostro esempio, se consideriamo che l'età è in realtà una variabile continua, è più corretto considerare la classe indicata con 19-21 (anni compiuti) come intervallo [19,22) : quindi il valore centrale della classe da utilizzare nel calcolo della media sarà:
 $(22 + 19) / 2 = 20,5$
e non 20.

Età	n(i)	x(i) n(i)
19-21	900	?
22-24	530	?
25-27	410	?
28-30	135	?
31-33	15	?
34-36	10	?
Totale	2000	
Età	Intervallo	x(i)
19-21	[19-22)	20,5
22-24	[22-25)	23,5
25-27	[25-28)	26,5
28-30	[28-31)	29,5
31-33	[31-34)	32,5
34-36	[34-37)	35,5

ANALISI DI UNA DISTRIBUZIONE

- Dunque risulterà:

$$M(\text{Età Scienze Formazione}) =$$

$$= (20,5 \cdot 900 + 23,5 \cdot 530 +$$

$$+ 26,5 \cdot 410 + \dots + 32,5 \cdot 15 +$$

$$+ 35,5 \cdot 10) / 2000 = 46595 / 2000 =$$

$$= 23,3$$
- Per esercizio, calcolare $M(\text{Età Filosofia})$
[= 23,53]

Età	Frequenze Assolute		
	x(i)	n(i)	x(i) n(i)
[19-22)	20,5	900	18450,0
[22-25)	23,5	530	12455,0
[25-28)	26,5	410	10865,0
[28-31)	29,5	135	3982,5
[31-34)	32,5	15	487,5
[34-37)	35,5	10	355,0
Totale	-	2000	46595,0
Media	-	-	23,3
Età	Frequenze Relative		
	x(i)	f(i)	x(i) f(i)
[19-22)	20,5	0,4500	9,2250
[22-25)	23,5	0,2650	6,2275
[25-28)	26,5	0,2050	5,4325
[28-31)	29,5	0,0675	1,9913
[31-34)	32,5	0,0075	0,2438
[34-37)	35,5	0,0050	0,1775
Totale	-	1,0000	23,2975
Media	-	-	23,2975

ANALISI DI UNA DISTRIBUZIONE



- **Proprietà della Media Aritmetica**
- È la media *algebrica* di gran lunga più utilizzata, tanto che quando si parla di media senza specificare, si intende quella aritmetica, perché gode di importanti proprietà.
- Per definizione, la media aritmetica conserva la somma dei valori, cioè può essere sostituita ai singoli valori e il totale resta lo stesso.
- Tiene conto dei valori di *tutti* i dati osservati: questa caratteristica può rappresentare un pregio ma anche un difetto in relazione alla situazione in cui si applica (sensibilità vs. stabilità)
- È sempre compresa tra il minore e il maggiore dei dati.
- La media di una costante è la costante stessa.
- **La somma degli scarti dei valori osservati dalla media aritmetica è sempre uguale a zero.**
- **La somma dei quadrati degli scarti dalla media aritmetica è "minima", cioè è la minore possibile.**

ANALISI DI UNA DISTRIBUZIONE

- Proprietà della Media Aritmetica
- La media di una costante è la costante stessa.
Ovvero, se i dati sono tutti uguali a una costante c , la media è uguale a c .

se

$$x_1 = x_2 = \dots = x_n = c$$

allora :

$$M(x) = M(c) = \frac{1}{N} \sum_{i=1}^N c = \frac{1}{N} (c + c + \dots + c) = \frac{1}{N} (N \cdot c) = c$$

- Osserviamo che la media di una serie di valori, una volta calcolata, è una costante: in effetti è un numero (es. 23,3) e quindi una quantità costante
- Un ruolo importante hanno i cosiddetti **scarti** dei valori osservati dalla media.
- Gli scarti da una costante (a) sono definiti come la differenza tra i valori osservati e la costante:

$$(x_i - a) \quad \forall i = 1, \dots, N$$

ANALISI DI UNA DISTRIBUZIONE

- Proprietà della Media Aritmetica
- La somma degli scarti dalla media è sempre uguale a 0

$$\sum_{i=1}^N (x_i - \bar{x}) = 0 \quad \text{e nel caso della media ponderata:} \quad \sum_{i=1}^k (x_i - \bar{x}) n_i = 0$$

Dimostrazione per il caso semplice:

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

Dimostrazione per il caso ponderato:

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

ANALISI DI UNA DISTRIBUZIONE

- Proprietà della Media Aritmetica
- La somma dei quadrati degli scarti dalla media è minima
- Si dice che la media *minimizza* la somma dei quadrati degli scarti

$$(a + b)^2 = a^2 + b^2 + 2ab$$

$$\sum_{i=1}^N (x_i - \bar{x})^2 = \min \quad \text{ovvero} \quad \sum_{i=1}^N (x_i - \bar{x})^2 < \sum_{i=1}^N (x_i - a)^2 \quad \forall a \neq \bar{x}$$

Infatti, se prendiamo un qualunque altro numero a diverso da \bar{x} :

$$\begin{aligned} \sum (x_i - a)^2 &= \sum (x_i - a + \bar{x} - \bar{x})^2 = \sum [(x_i - \bar{x}) + (\bar{x} - a)]^2 = \\ &= \sum [(x_i - \bar{x})^2 + (\bar{x} - a)^2 + 2(x_i - \bar{x})(\bar{x} - a)] = \\ &= \sum (x_i - \bar{x})^2 + \sum (\bar{x} - a)^2 + 2 \sum (x_i - \bar{x})(\bar{x} - a) = \\ &= \sum (x_i - \bar{x})^2 + N(\bar{x} - a)^2 + 2(\bar{x} - a) \underbrace{\sum (x_i - \bar{x})}_{=0} = \\ &= \sum (x_i - \bar{x})^2 + N(\bar{x} - a)^2 + \underbrace{2(\bar{x} - a) \cdot 0}_{=0} = \sum (x_i - \bar{x})^2 + \underbrace{N(\bar{x} - a)^2}_{\text{sempre } \geq 0} \geq \\ &\geq \sum (x_i - \bar{x})^2 \end{aligned}$$

ANALISI DI UNA DISTRIBUZIONE



Esercizio: Topi che nuotano

Un ricercatore vuole verificare se i topi hanno la stessa capacità di nuotare in un liquido opaco (e quindi alla cieca) rispetto ad uno trasparente. Come liquido opaco decide di usare il latte. L'esperimento viene condotto misurando il tempo impiegato da 5 topi per percorrere a nuoto una stessa distanza, nei due tipi di liquido.

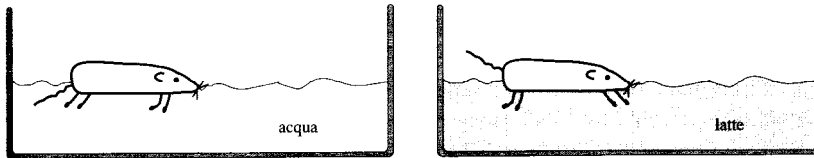


Tabella - Tempo impiegato dai topi per nuotare nel latte o nell'acqua (secondi).

Liquido	Identificativo del topo				
	1	2	3	4	5
Acqua	10	12	13	15	11
Latte	12	14	17	126	13

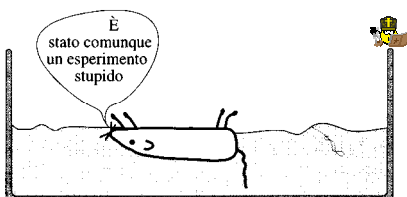
- Calcolare media e mediana. Qual è l'indicatore migliore in questo caso e cosa suggerisce il risultato? (... e cosa è successo al topo n. 4?)

ANALISI DI UNA DISTRIBUZIONE

Risposte:

Acqua: Mediana = 12 Media=12,2

Latte: Mediana = 14 Media=36,4



- I topi sembrano un po' più a loro agio a nuotare in acqua che non alla cieca in un liquido opaco
- La media aritmetica risente molto più della mediana del valore estremo (126)

Un difetto della media aritmetica

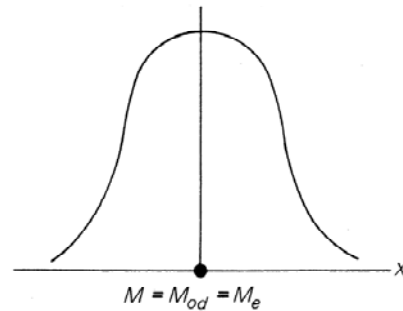
Non è del tutto infrequente trovare degli insiemi di dati contenenti una piccola frazione di osservazioni anomale o atipiche, ovvero, osservazioni che assumono valori lontani da quelli assunti dalla maggior parte delle altre osservazioni e che, quindi, sembrano provenire da una popolazione diversa o essere state generate da un meccanismo differente.

In una situazione del tipo descritto, bisogna tenere presente che la media aritmetica può essere molto sensibile alla presenza delle osservazioni anomale potendo anche, a volte, fornire risultati non molto sensati.

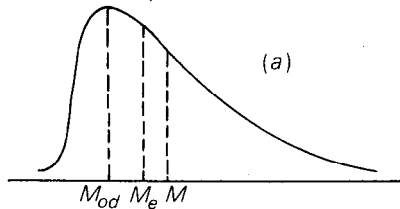
Infatti, come è facile capire dalla definizione stessa, una sola osservazione molto grande o molto piccola può *dominare* il valore assunto dalla media.

ANALISI DI UNA DISTRIBUZIONE

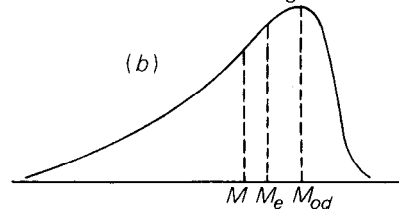
- **Relazioni tra Moda, Mediana e Media**
- La moda è il valore con la frequenza maggiore, quindi è quello in corrispondenza del massimo della curva della distribuzione di frequenze
- La mediana divide l'area sottesa alla curva in due metà uguali (50% a destra e 50% a sinistra)
- La media tiene conto dei valori di tutte le osservazioni, quindi risente maggiormente dei valori estremi (molto piccoli o molto grandi)
- In una distribuzione unimodale simmetrica: media, moda e mediana coincidono
- Al crescere dell'asimmetria, i tre indicatori si allontanano progressivamente



asimmetria positiva



asimmetria negativa

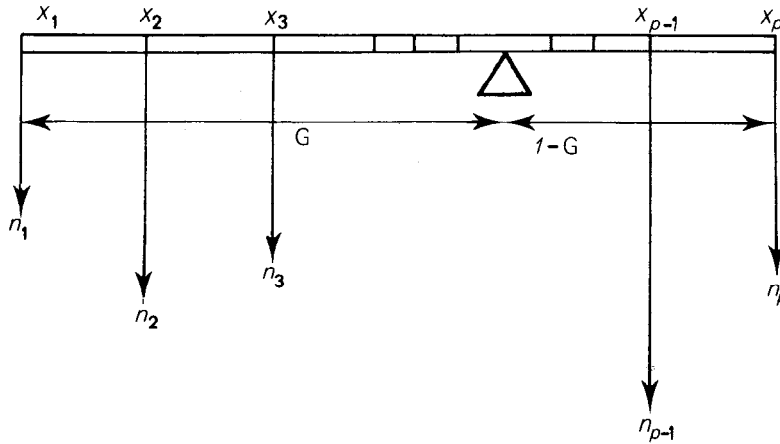


ANALISI DI UNA DISTRIBUZIONE

- La media aritmetica è fondamentale nei processi di misura, e nel campionamento, perché permette di "controllare" la precisione delle stime: per aumentare la precisione di una misura, si effettuano più misurazioni e se ne fa la media
- È semplice da calcolare: è la più semplice delle medie algebriche, e per questo tende ad essere usata anche quando non è appropriata.
- Calcolare la media aritmetica ha senso quando le quantità espresse dalla variabile sono additive, cioè ha significato sommarle. Le grandezze additive sono le più diffuse.
- Non tutte le grandezze però sono additive, ovvero non sempre ha senso sommare determinate quantità in tutti i contesti.
- La media aritmetica è il valore che conserva la somma dei dati da mediare. Quando il risultato dell'operazione che ha significato conservare non è la somma, la media aritmetica non è appropriata, cioè conduce a conclusioni non corrette.
- Ad es. non è corretto fare la media aritmetica di tassi di crescita, o di tassi di interesse (grandezze moltiplicative), o di velocità (rapporto spazio/tempo), ...
- Per *mediare* correttamente grandezze non additive è necessario introdurre altri tipi di medie algebriche: media *geometrica*, media *armonica*, media *quadratica*, ...

ANALISI DI UNA DISTRIBUZIONE

- Proprietà della Media Aritmetica Ponderata
- Una interessante proprietà fisica della media è quella di essere il *baricentro* (centro di gravità) cioè il punto di equilibrio del sistema rappresentato dai dati



Medie Algebriche

un ultimo sforzo ...



MEDIE ALGEBRICHE

- **Media dei Quadrati:** è la media aritmetica del quadrato dei valori osservati

$$M(x^2) = \frac{1}{N} \sum_{i=1}^N x_i^2$$

- Per dati in tabella di frequenze diventa:

$$M(x^2) = \frac{1}{N} \sum_{i=1}^k x_i^2 n_i$$

$x(i)$: valore della modalità i -esima

$n(i)$: frequenze assolute modalità i -esima

N : numero totale di osservazioni

k : numero di modalità distinte della variabile (ovvero di classi della tabella di frequenza)

- La media dei quadrati è una quantità molto importante in statistica, che vedremo tornare spesso nei nostri discorsi

$x(i)$	$n(i)$	$x(i)^2$	$x(i)^2 n(i)$
19	350	361	126350
20	300	400	120000
21	250	441	110250
22	200	484	96800
23	150	529	79350
24	180	576	103680
25	200	625	125000
26	80	676	54080
27	130	729	94770
28	60	784	47040
29	25	841	21025
30	50	900	45000
31	10	961	9610
32	5	1024	5120
33	0	1089	0
34	5	1156	5780
35	5	1225	6125
Totale	2000		1049980
Media			524,99

MEDIE ALGEBRICHE

- **Media Quadratica:** è la radice quadrata della media dei quadrati dei valori

$$\sqrt{M(x^2)} = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$$

- Per dati in tabella di frequenze:

$$\sqrt{\frac{1}{N} \sum_{i=1}^k x_i^2 n_i}$$

- **Proprietà:**
La media quadratica è sempre maggiore della media aritmetica, calcolata sugli stessi dati

$$\sqrt{M(x^2)} > M(x)$$

- **Esercizio.** Verifichiamolo sui nostri dati:

$$\sqrt{M(x^2)} = \sqrt{524,99} = 22,91$$

$$\hat{e} > M(x) = 22,69$$

$x(i)$	$n(i)$	$x(i)^2$	$x(i)^2 n(i)$
19	350	361	126350
20	300	400	120000
21	250	441	110250
22	200	484	96800
23	150	529	79350
24	180	576	103680
25	200	625	125000
26	80	676	54080
27	130	729	94770
28	60	784	47040
29	25	841	21025
30	50	900	45000
31	10	961	9610
32	5	1024	5120
33	0	1089	0
34	5	1156	5780
35	5	1225	6125
Totale	2000		1049980
Media			524,99

MEDIE ALGEBRICHE

- **Media Geometrica:** è costruita come il valore che può essere sostituito ai dati osservati senza farne variare il prodotto

$$M_{geom}(x) = \sqrt[N]{x_1 \cdot x_2 \cdot \dots \cdot x_N} = \sqrt[N]{\prod_{i=1}^N x_i} = \left(\prod_{i=1}^N x_i \right)^{1/N}$$

- Per dati in tabella di frequenze:

$$M_{geom}(x) = \sqrt[N]{\prod_{i=1}^k (x_i)^{n_i}} = \left(\prod_{i=1}^k x_i^{n_i} \right)^{1/N}$$

indica il prodotto di N termini

- **Proprietà:**
La media geometrica è sempre minore della media aritmetica (calcolata sugli stessi dati)

$$M_{geom}(x) < M(x)$$

- La media geometrica è appropriata per calcolare la media di grandezze moltiplicative, come tassi di crescita o tassi di interesse

MEDIE ALGEBRICHE

- **Media Armonica:** è il reciproco della media aritmetica dei reciproci dei valori

$$M_{arm}(x) = \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{1}{x_i}} = \frac{N}{\sum_{i=1}^N \frac{1}{x_i}}$$

- **Proprietà:**
La media armonica è sempre minore della media geometrica e quindi anche della media aritmetica (calcolata sugli stessi dati)

$$M_{arm}(x) < M_{geom}(x) < M(x)$$

- La media armonica è appropriata per calcolare la media di grandezze che derivano da rapporti, come la velocità (rapporto spazio/tempo)
- **Esempio:**

Percorriamo il tragitto Verona-Trento alla velocità di 90km/h e il ritorno da Trento a Verona a 10km/h. Qual'è stata la nostra velocità media ?

$$M_{arm} = \frac{2}{\frac{1}{90} + \frac{1}{10}} = \frac{2}{\frac{1+9}{90}} = \frac{2}{\frac{10}{90}} = \frac{2}{\frac{1}{9}} = 18$$

- **Esercizio:** La fermata dell'autobus
- Lungo una strada rettilinea sono collocati cinque condomini: A, B, C, D ed E. L'AMT deve decidere dove posizionare la fermata dell'autobus, in modo che risulti più comoda possibile per i potenziali utenti che abitano nella strada.
- I dati rilevati per prendere la decisione sono i seguenti:
 - nei 5 stabili abitano rispettivamente il seguente numero di inquilini:
6, 6, 20, 12, 8
 - le distanze tra gli edifici sono le seguenti:
distanza di A da B = 1000 m
distanza di B da C = 1000 m
distanza di C da D = 100 m
distanza di D da E = 50 m
- Si vuole determinare la posizione della fermata in modo da minimizzare il disagio complessivo dei residenti nella strada per raggiungere la fermata, considerando due differenti ipotesi:
 - il disagio cresce linearmente con la distanza
 - il disagio cresce con il quadrato della distanza