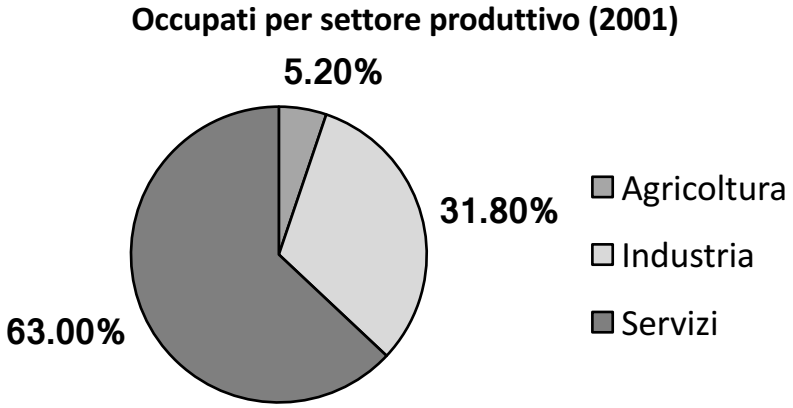
01 – Rappresentazioni statistiche e grafiche — Unità n° 03

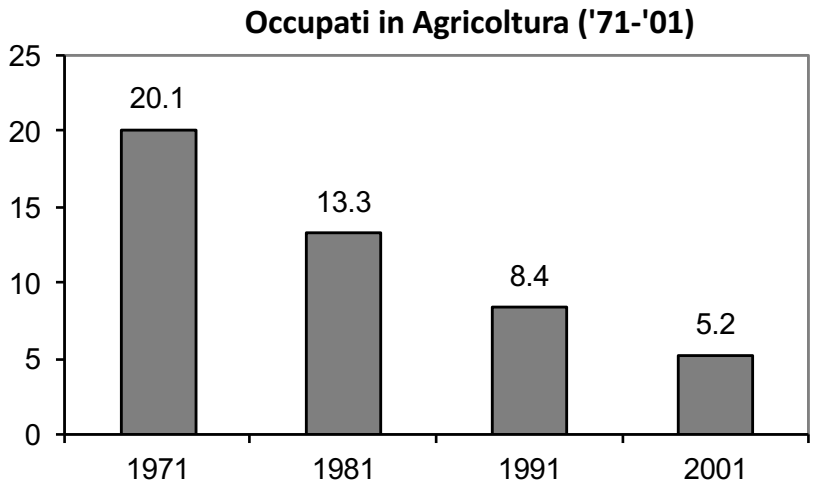
Per descrivere l'insieme delle modalità di un carattere osservato su un collettivo è possibile ricorrere alla distribuzione di frequenza o ad una sua opportuna rappresentazione grafica

Occupati per settore produttivo - 1971-2001

	Anni			
Settore	1971	1981	1991	2001
Agricoltura	20,1	13,3	8,4	5,2
Industria	39,5	37,2	32,0	31,8
Servizi	40,4	49,5	59,6	63,0

Vogliamo studiare la distribuzione per settore produttivo in Italia tra il 1971 e il 2001





02 – Intensità (totale) di un fenomeno

Unità n° 03

Solitamente vogliamo studiare la manifestazione del carattere oggetto di studio in ciascuna delle unità che compongono il collettivo

Allo stesso tempo spesso è utile conoscere anche l'**intensità totale** del fenomeno nel collettivo studiato, cioè l'ammontare di carattere posseduto complessivamente da tutte le unità

Intuitivamente siamo portati ad intendere l'ammontare totale di carattere come la somma delle misurazioni/osservazioni effettuate su ciascuna unità: questo però è solo uno dei possibili modi, quindi è necessario valutare caso per caso come calcolare l'intensità sul collettivo

In generale possiamo dire che i due operatori matematici più utilizzati sono la somma e il prodotto. E' possibile considerare quindi tre diversi metodi di calcolo dell'intensità totale:

- $f(x_1, x_2, .., x_k)$ = somma
- $f(x_1, x_2, .., x_k)$ = prodotto
- $f(x_1, x_2, .., x_k)$ = somma di potenze r^{me} dei valori del carattere

03 – Come calcolare l'intensità totale

Unità n° 03

Il calcolo dell'intensità totale differisce a seconda del tipo di distribuzione che andiamo a considerare e del tipo di fenomeno studiato:

DISTRIBUZIONE UNITARIA

x_1	x_2	x_3	x_4	...	x_N

$$x_1 + x_2 + x_3 + x_4 + \ldots + x_N = \sum x_i$$

$$x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot \ldots \cdot x_N = \prod x_i$$

DISTRIBUZIONE DI FREQUENZA

x_1	x_2	x_3	x_4	...	x_k
n_1	n_2	n_3	n_4	...	n_k

$$x_1 \cdot n_1 + x_2 \cdot n_2 + \ldots + x_k \cdot n_k = \sum x_i \cdot n_i$$

$$x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \cdot \ldots \cdot x_k^{n_k} = \prod x_i^{n_i}$$

(nelle distribuzioni in classi sostituiamo alle modalità x_i i valori centrali delle classi c_i)

04 – Esempio (1)

Consideriamo le commesse ottenute da un'industria manifatturiera nel corso di un anno (in ML di €):

2 0,5 0,72 1 0,375 1 0,5 3,25

Per ottenere il fatturato complessivo dell'industria nell'anno è sufficiente sommare gli importi delle differenti commesse, ottenendo 9,345 ML di €

Da un punto di vista statistico le diverse commesse rappresentano un collettivo e gli importi sono le osservazioni su ciascuna unità statistica del carattere "valore della commessa"

Se trasformiamo la distribuzione unitaria in una distribuzione di frequenza otteniamo

x_i	0,375	0,5	0,72	1	2	3,25
n_i	1	2	1	2	1	1

Moltiplicando ogni modalità per il numero di unità statistiche su cui sono state osservate si ottiene allo stesso modo l'intensità totale (cioè 9,345 ML di €)

05 – Esempio (2)

Consideriamo di depositare il 1 gennaio una certa quantità di denaro S in banca e di ricevere su questo ammontare un interesse annuo pari a i_1: alla fine dell'anno avremo un ammontare pari alla quantità iniziale più gli interessi maturati

$$S + S \cdot i_1 = S \cdot (1 + i_1)$$

Il secondo anno, la quantità in banca è differente. Supponiamo di non prelevare alcuna quantità e di avere questa volta un interesse annuo pari a i_2:

$$S \cdot (1 + i_1) + S \cdot (1 + i_1) \cdot i_2 = S \cdot (1 + i_1) \cdot (1 + i_2)$$

Dopo k anni, considerando un interesse annuo pari a (i_1, i_2, i_3, …, i_k) l'ammontare complessivo (in questo caso l'intensità totale del nostro fenomeno) risulta essere pari a:

$$S \cdot (1 + i_1) \cdot (1 + i_2) \cdot \ldots \cdot (1 + i_k)$$

Come si vede la quantità ottenuta di anno in anno è proporzionale a quella degli anni precedenti

06 – La statistica mente?

Unità n° 03

Sai ched'è la statistica? E' 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.

Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre eguale
puro co' la persona bisognosa.

Me spiego, da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:

e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perchè c'è un antro che se ne magna due.

Spesso è utile sintetizzare i dati ma occorre farlo nel modo opportuno!

07 – "Redistribuire" l'intensità totale… Unità n° 03

Il passaggio da un elenco di modalità alle distribuzioni di frequenza con modalità distinte o con classi di modalità, consente una prima sintesi dei dati: il processo di sintesi non può limitarsi solo a questa diversa rappresentazione dei dati ma deve spingersi oltre fino a sintetizzare in un unico dato numerico una caratteristica d'interesse

L'idea è quella di sostituire tutte le modalità del carattere in esame con un'unica modalità che le rappresenti: ottenuta l'intensità complessiva del fenomeno è necessario quindi procedere ad una "redistribuzione" dello stesso su tutte le unità statistiche

Questa finalità può essere perseguita attraverso la determinazione di opportuni indici sintetici del fenomeno considerato, dette **misure o indici di centralità**

Alcuni indici sono adatti a sintetizzare tutti i tipi di carattere, altri invece sono utilizzabili solo se si studiano caratteri quantitativi

08 – Misure di centralità di una distribuzione — Unità n° 03

Le *misure di centralità* (o tendenza centrale) esprimono sinteticamente il centro ideale della distribuzione, ossia quel valore intorno al quale tendono a gravitare i dati

Per questo occorrono misure sintetiche che "centrino" la distribuzione di un certo fenomeno e consentano il passaggio da una pluralità di informazioni ad un solo valore numerico

La scelta di un indice sintetico è legato essenzialmente a tre fattori:

1) **tipologia del carattere in esame (qualitativo o quantitativo)**
2) **la sua rappresentazione statistica (distribuzione unitaria, di frequenza o in classi)**
3) **le motivazioni che inducono a "riassumere" la distribuzione in un unico valore rappresentativo**

Per le ragioni riportate non è possibile definire una sola misura di centralità

Inoltre è bene tenere presente che indipendentemente dall'indice adoperato, il valore di sintesi ottenuto non è detto che coincida esattamente con una delle modalità osservate

09 – Misure di centralità di una distribuzione: le medie — Unità n° 03

Le **medie** sono utili perché sono espresse nella stessa unità di misura del carattere osservato e perché danno una idea immediata della manifestazione del fenomeno nel collettivo

Medie analitiche → Calcolate con operazioni algebriche sulle modalità

Medie di posizione → Non richiedono operazioni algebriche sulle modalità

Le medie, data la loro funzione di sintesi, possono essere impiegate per confrontare:
- uno stesso fenomeno rilevato su collettivi diversi
- uno stesso fenomeno rilevato in tempi diversi e/o luoghi diversi
- due o più fenomeni diversi tra di loro

10 – La media

Unità n° 03

La media è un concetto primitivo per gli esseri umani, percepito con immediatezza, tuttavia la sua misura è arbitraria perché, come visto, il criterio utilizzato dipende strettamente dalle informazioni ritenute rilevanti e dagli obiettivi per i quali l'indice è calcolato

Una media di una variabile X secondo *Cauchy* (1821) è qualunque valore reale M intermedio tra il valore più piccolo della distribuzione ordinata in senso crescente $x_{(1)}$ (minimo) e il valore più grande $x_{(N)}$ (massimo)

$$\mathbf{X_{(1)} \leq M \leq x_{(N)}}$$

Per quanto ovvio e convincente, tale requisito costituisce in effetti più un aspetto importante da considerare che una soluzione, essendo generalmente infiniti i numeri reali che soddisfano tale criterio, detto di **internalità**

Una media di una variabile X secondo *Chisini* (1929) è invece quel valore (interno) che rispetto ad una funzione sintetica lascia inalterato il valore

$$\mathbf{f(x_1, x_2, \ldots, x_n) = f(M, M, \ldots, M)}$$

11 – Media aritmetica

Unità n° 03

Supponiamo di volere effettuare uno studio per capire qual è il mezzo più conveniente per raggiungere l'università. Registriamo i tempi impiegati da un piccolo collettivo di studenti che solitamente utilizza l'auto: qual è il tempo medio impiegato per raggiungere il campus?

studente	Tempo (min.)	studente	Tempo (min.)
1	23	7	28
2	32	8	33
3	44	9	45
4	21	10	34
5	36	11	29
6	30	12	31

$$\bar{x}_a = (23+32+44+21+36+30+28+33+45+34+29+31)/12 = 386/12 = 32{,}17 \text{ minuti}$$

12 – Notazione

La media aritmetica di un insieme di n osservazioni $x_1, x_2, \ldots x_N$ di un carattere quantitativo X è data da:

$$\bar{x}_a = \frac{1}{N}\left(x_1 + x_2 + \ldots + x_N\right) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Se il carattere X è quantitativo discreto e conosciamo la sua distribuzione di frequenza:

con frequenze assolute $\bar{x}_a = \frac{1}{N}\sum_{i=1}^{K} x_i n_i$

Perché ?

con frequenze relative $\bar{x}_a = \sum_{i=1}^{K} x_i f_i$

13 – Media aritmetica per distribuzioni di frequenza

Consideriamo la distribuzione unitaria relativa al n° di impianti di 12 imprese industriali:

3, 6, 6, 5, 3, 2, 5, 3, 2, 4, 2, 4

Vogliamo calcolare quanti impianti in media possiede ciascuna impresa

$$\overline{x}_a = \frac{3+6+6+5+3+2+5+3+2+4+2+4}{12} = 3{,}75 \text{ (impianti)}$$

N° impianti	imprese
2	3
3	3
4	2
5	2
6	2
Totale	12

$$\left(\frac{2\times(3)+3\times(3)+4\times(2)+5\times(2)+6\times(2)}{12} \right)$$

Dal confronto delle due formule si comprende come nel caso di distribuzioni di frequenza si tenga conto dell'*importanza* di ciascuna modalità in termini di manifestazione nel collettivo

14 – Media aritmetica per distribuzioni in classi

Nel caso di una distribuzione di frequenza per un carattere X suddiviso in classi, possiamo calcolare approssimativamente la media utilizzando il valore centrale di ciascuna classe

Prezzi e quantità delle specialità vendute da una farmacia giornalmente

Prezzo (a confezione)	Valore centrale	N° confezioni	Ammontare carattere
20 –| 30	25	112	25*112 = 2800
30 –| 35	32.5	53	32.5*53 = 1722.5
35 –| 40	37.5	27	37.5*27 = 1012.5
40 –| 50	45	9	45*9 = 405
Totale		201	5940 euro

$$\bar{x}_a = \frac{1}{N}\sum_{i=1}^{k} c_i n_i$$ = 5940/201 = 29.55 € *(a confezione)*

15 – Notazione

Per le distribuzioni in classi vale quanto appena detto: se calcoliamo la media a partire dalla rappresentazione in frequenze assolute o in frequenze relative abbiamo lo stesso risultato

N.B.: in entrambi i casi non otteniamo la vera media ma una sua approssimazione

X	v.c.	n	f
x_1-x_2	c_1	n_1	f_1
x_2-x_3	c_2	n_2	f_2
…	…	…	…
x_{i-1}-x_i	c_i	n_i	f_i
…	…	…	…
x_{k-1}-x_k	c_k	n_k	f_k
totale		N	1

$$c_i = \frac{\text{estr. inferiore} + \text{estr. superiore}}{2}$$

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

16 – Esempio

Vogliamo calcolare l'altezza media di un collettivo di 200 individui a partire dalla distribuzione in classi del carattere altezza (in centimetri)

$x_i - \mid x_{i+1}$	n_i	f_i	c_i	$c_i\, n_i$	$c_i\, f_i$
70 -| 100	20	0,1	85	1700	8,5
100 -| 120	7	0,035	110	770	3,85
120 -| 140	18	0,09	130	2340	11,7
140 -| 170	65	0,325	155	10075	50,375
170 -| 180	21	0,105	175	3675	18,375
180 -| 200	45	0,225	190	8550	42,75
200 -| 220	24	0,12	210	5040	25,2
Totale	**200**			**32150**	**160,75**

Una volta individuati i v. centrali per ogni classe si calcola la media: si ottiene uguale risultato sia con le frequenze assolute che relative

$$\bar{x}_a = \frac{1}{200}[(85\times 20)+(110\times 7)+\ldots+(210\times 24)] = \frac{32150}{200} = 160{,}75 \text{ cm}$$

$$\bar{x}_a = (85\times 0{,}1)+(110\times 0{,}035)+\ldots+(210\times 0{,}12) = 160{,}75 \text{ cm}$$

17 – Quali proprietà ha la media aritmetica? Unità n° 03

La media aritmetica gode di cinque proprietà fondamentali:

1) **La media è sempre un valore interno alla distribuzione ordinata dei dati (INTERNALITA')**

2) **La somma di tutte le differenze tra i valori della distribuzione e il loro valore medio è sempre pari a zero**

3) **La somma di tutte le differenze al quadrato tra i valori della distribuzione e il loro valore medio è sempre un minimo**

4) **La media è invariante per trasformazioni affini (LINEARITA')**

5) **La media di un carattere osservato su una popolazione divisa in sottogruppi è pari alla media delle medie di tutti i sottogruppi (ASSOCIATIVITA')**

18 – Linearità della media aritmetica

Consideriamo due valori costanti **a** e **b**: se i valori $\mathbf{x_i}$ vengono trasformati nei valori $\mathbf{y_i = a + bx_i}$, allora tra le media aritmetica delle $\mathbf{y_i}$ e quella delle $\mathbf{x_i}$ esiste la stessa relazione (lineare) che esiste tra le $\mathbf{y_i}$ e le $\mathbf{x_i}$, cioè

$$\bar{y}_a = a + b\bar{x}_a$$

La relazione ci dice che la media è invariante per trasformazioni affini: questo vuol dire che se aggiungiamo a tutti i valori di una distribuzione una costante **a** allora la media della nuova distribuzione sarà pari a quella della distribuzione originaria maggiorata della quantità **a**; allo stesso modo se moltiplichiamo a tutti i valori di una distribuzione per una costante **b** allora la media della nuova distribuzione sarà proporzionale a quella della distribuzione originaria di una quantità **b**

PROVA: calcolare la media della distribuzione (23, 24, 22, 20, 26, 23), poi la media della distribuzione (23+3, 24+3, 22+3, 20+3, 26+3, 23+3) e la media della distribuzione (23x3, 24x3, 22x3, 20x3, 26x3, 23x3)

19 – Associatività della media aritmetica

La media aritmetica complessiva di più gruppi è uguale alla media aritmetica delle medie di ciascun gruppo, considerando quante unità del collettivo appartengono a ciascun gruppo

Supponiamo che la popolazione sia suddivisa in H gruppi, ognuno di numerosità N_i (i=1,2,…,H)

Sia $\bar{x}_i$ la media aritmetica del carattere in esame nell'i-esimo gruppo (i=1,2,…,H)

La proprietà afferma che la media complessiva del carattere oggetto di studio può essere calcolata nel modo seguente

$$\bar{x}_a = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \ldots + \bar{x}_H N_H}{N_1 + N_2 + \ldots + N_H} = \frac{\sum_{i=1}^{H} \bar{x}_i N_i}{N}$$

20 – Esercizio

La giornata lavorativa di un'azienda metallurgica è organizzata su tre turni diversi. Il 1° turno di lavoro è composto da 100 operai che hanno un'età media di 50 anni; il 2° turno è composto da 75 operai che hanno un'età media di 46 anni, mentre l'ultimo turno è composto da 45 operai con una età media di 37 anni. Determinare l'età media degli operai che lavorano nell'azienda

Turni	Numerosità Turno	Età media (in anni)
Primo Turno	100	50
Secondo Turno	75	46
Terzo Turno	45	37

Il numero complessivo N di operai è $100+75+45=220$. L'età media dei 220 operai è data da:

$$\overline{x}_a = \frac{(50\times 100)+(46\times 75)+(37\times 45)}{220} = \frac{10115}{220} = 45{,}98 \text{ anni}$$

Quale proprietà abbiamo utilizzato ?

21 – Esercizio

Supponiamo che il reddito lordo annuo (in €) relativo al 2003 di 5 dirigenti di una amministrazione pubblica si quello sotto riportato:

Nome Dirigente	*Dirigente 1*	*Dirigente 2*	*Dirigente 3*	*Dirigente 4*	*Dirigente 6*
Reddito	85000	130000	210000	150000	97000

Si supponga che nel 2004 tutti i dirigenti abbiamo ricevuto un "bonus di produttività" pari a 10000 euro e che inoltre il loro reddito lordo, a causa degli adeguamenti salariali, sia aumentato dell'1,35% rispetto al 2003. Determinare il reddito medio lordo dei 5 dirigenti nel 2004

$$Y = 10000 + X + 0{,}0135X = 10000 + (1 + 0{,}0135)X$$

$$= 10000 + 1{,}0135X$$

$$\bar{y}_a = 10000 + 1{,}0135\ (\bar{x}_a) = 146214{,}4\ €$$

$$\bar{x}_a = \frac{85000 + 130000 + 210000 + 150000 + 97000}{5}$$

$$= \frac{672000}{5} = 134400\ €$$

Quale proprietà abbiamo utilizzato ?

22 – Sintesi di variabili doppie miste e quantitative

Unità n° 03

	y_1	...	y_j	...	y_c	TOT
x_1	n_{11}	...	n_{1j}	...	n_{1c}	$n_{1.}$
...	...	...	...	...	...	...
x_i	n_{i1}	...	n_{ij}	...	n_{ic}	$n_{i.}$
...	...	...	...	...	...	...
x_r	n_{r1}	...	n_{rj}	...	n_{rc}	$n_{r.}$
TOT	$n_{.1}$	...	$n_{.j}$	...	$n_{.c}$	$n_{..}$

Consideriamo una variabile doppia **(X,Y)** e supponiamo che sia stata organizzata in una tabella che contiene sulle righe le r modalità di X e sulle colonne le c modalità di Y

Supponiamo che la variabile **(X,Y)** sia mista o quantitativa (cioè che almeno una delle due variabili in gioco sia quantitativa)

Se entrambe le variabili sono quantitative è possibile calcolarne la media aritmetica

$$\bar{x}_a = \frac{\sum_{i=1}^{r} x_i n_{i.}}{n}$$ ← media generale di X

$$\bar{y}_a = \frac{\sum_{j=1}^{c} y_j n_{.j}}{n}$$ ← media generale di Y

Supponiamo ora che **X** sia una variabile qualitativa e che **Y** sia invece una variabile quantitativa: non è più possibile calcolare la media di **X** mentre invece possiamo calcolare la media di **Y**

23 – Esempio

Consideriamo la distribuzione doppia di frequenza del ***tipo di birra*** preferito e dell'***età*** di un collettivo di consumatori: ovviamente possiamo calcolare la media solo della variabile in colonna

		Età			
		18 - 22	***23 - 26***	***27 - 30***	
Tipo di Birra	*Bionda*	12	22	11	**45**
	Rossa	5	9	14	**28**
	Scura	3	15	18	**36**
		20	**46**	**43**	**109**

Individuiamo i valori centrali delle diverse classi di età e quindi calcoliamo la media considerando quante unità statistiche appartengono alle diverse classi

v.c.	20	24,5	28,5	
$n_{.j}$	20	46	43	109

$$\bar{y}_a = \frac{20 \cdot 20 + 24{,}5 \cdot 46 + 28{,}5 \cdot 43}{109} = 25{,}25 \text{ anni}$$

I consumatori del collettivo che stiamo esaminando hanno una età media di 25,25 anni

24 – Medie condizionate

Supponiamo di voler studiare la distribuzione condizionata **Y|x** : siamo interessati a vedere come si distribuisce nel collettivo il carattere misurato dalla variabile Y fissato un certo valore della variabile X

$$f(Y|x_1)=\frac{f(X=x_1,Y=y_j)}{f(X=x_1)}=\frac{f_{1j}}{f_{1.}} \qquad j=1,2,\ldots,c$$

al variare di j abbiamo la distribuzione di Y rispetto a x_1

Possiamo allora calcolare anche la media di Y fissato un certo valore della variabile X, ottenendo così una ***media condizionata***: è come se calcolassimo la media della sottopopolazione che presenta x_1

	Voto					*Totale*
Sesso	18-20	21-23	24-25	26-28	29-30	
maschi	11	23	38	116	75	263
femmine	55	163	107	48	64	437
Totale	66	186	145	164	139	700

$$M(Y|x_i)=\frac{\sum_{j=1}^{c} y_j n_{ij}}{n_{i.}}$$

M(voto|m) $=\frac{(19\cdot 11)+(22\cdot 23)+(24,5\cdot 38)+(27\cdot 116)+(29,5\cdot 75)}{263}=$
$=26,58$

M(voto|f) $=\frac{(19\cdot 55)+(22\cdot 163)+(24,5\cdot 107)+(27\cdot 48)+(29,5\cdot 64)}{437}=$
$=23,882$

M(voto) $=\frac{(19\cdot 66)+(22\cdot 186)+(24,5\cdot 145)+(27\cdot 164)+(29,5\cdot 139)}{700}=$
$=24,896$

25 – Esercizio

Nella tavola che segue è riportata la distribuzione di 250 individui che abitualmente trascorrono le vacanze in località di mare della provincia di Cosenza secondo il luogo prescelto (X) e l'età (Y):

	0 -| 20	20 -| 40	40 -| 60	60 -| 80
Fuscaldo	14	25	21	34
Diamante	15	23	11	15
Sibari	12	11	19	5
Cariati	5	23	7	10

(1) Tra tutti gli individui con età superiore ai 40 anni, quanti preferiscono la costa ionica?

(2) Calcolare l'età media dei turisti e l'età media dei turisti per luogo di villeggiatura prescelto

26 – Un diverso punto di vista

Nella realtà bisogna spesso considerare l'*importanza* delle diverse modalità di un carattere secondo un criterio diverso dalla sua frequenza…

	VOTO	CF
1	27	5
2	23	10
3	20	10
4	30	5
5	25	5

Se calcoliamo la media considerando solo il voto degli esami otteniamo:

$$\overline{x}_a = 125/5 = 25$$

Se però assegniamo ad ogni esame un diverso peso, ad esempio i crediti, cosa accade al valore medio?

27 – Media ponderata

Unità n° 03

La media aritmetica ponderata di un insieme di N valori osservati di un carattere quantitativo X con pesi non negativi, è data da:

$$\bar{x}_a = \frac{x_1 p_1 + x_2 p_2 + \ldots + x_k p_k}{p_1 + p_2 + \ldots + p_k} = \frac{\sum_{i=1}^{k} x_i p_i}{\sum_{i=1}^{k} p_i} = \sum_{i=1}^{k} x_i \left(\frac{p_i}{\sum_{i=1}^{k} p_i} \right)$$

la somma dei pesi relativi deve essere sempre pari a 1

Attraverso la media ponderata è possibile valutare sinteticamente un fenomeno, espresso in termini di carattere quantitativo discreto o continuo, inserendo anche un sistema di pesi che tenga conto dell'importanza che le diverse manifestazioni dello stesso hanno all'interno della popolazione oggetto di studio

28 – Esempio

	VOTO	CF
1	27	5
2	23	10
3	20	10
4	30	5
5	25	5

unità — osservazioni — pesi

Calcoliamo il voto medio con la media ponderata:

$$\bar{x}_a = \frac{(27 \cdot 5) + (23 \cdot 10) + (20 \cdot 10) + (30 \cdot 5) + (25 \cdot 5)}{5+10+10+5+5} = \mathbf{24}$$

Il voto medio è 24, più basso di quello calcolato con la media aritmetica (pari a 25)

Attenzione! Nel calcolo della media ponderata l'intensità totale del fenomeno deve essere rapportata non al numero di unità statistiche (nell'esempio 5) ma al totale dei pesi utilizzati

29 – La media geometrica

Nello studio di alcuni fenomeni, ad esempio quelli economici, si osservano dati distribuiti secondo un andamento di tipo geometrico: ciò vuol dire che il carattere studiato si modifica mediante proporzioni

In questi casi ha senso da un punto di vista statistico moltiplicare i dati piuttosto che sommarli

$$\overline{x}_g = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N} = \left(\prod_{i=1}^{N} x_N \right)^{\frac{1}{N}}$$

Per poter calcolare la media geometrica è necessario che i valori siano tutti positivi

30 – Calcoliamo l'intensità totale

Riprendiamo l'esempio di un capitale S depositato in banca per 3 anni: supponiamo di avere ogni anno un tasso d'interesse diverso (tasso variabile)

Il **tasso medio** è quel tasso che sostituito ai tre diversi tassi applicati dalla banca ci consente di ottenere lo stesso ammontare (intensità totale del fenomeno)

ANNO	TASSO VARIABILE	TASSO (MEDIO) FISSO i^*
Fine 1° anno	$S+i_1S=S(1+i_1)$	$S+i^*S=S(1+i^*)$
Fine 2° anno	$S(1+i_1)+i_2S(1+i_1)=S(1+i_1)(1+i_2)$	$S(1+i^*)+i^*S(1+i^*)=S(1+i^*)(1+i^*)=S(1+i^*)^2$
Fine 3° anno	$S(1+i_1)(1+i_2)+i_3S(1+i_1)(1+i_2)= S(1+i_1)(1+i_2)(1+i_3)$	$S(1+i^*)^2+i^*S(1+i^*)^2= S(1+i^*)^3$
TOTALE	$S(1+i_1)(1+i_2)(1+i_3)$	$S(1+i^*)^3$

31 – Dall'intensità totale a quella media

Per comodità assumiamo S=1: per calcolare il tasso medio dobbiamo soddisfare la relazione

$$(1 + i_1) \cdot (1 + i_2) \cdot \ldots \cdot (1 + i_N) = (1 + i^*)^N$$

La nostra incognita è quindi il tasso i^*, eleviamo i due elementi della relazione per il reciproco di N così da ottenere

$$[(1 + i_1) \cdot (1 + i_2) \cdot \ldots \cdot (1 + i_N)]^{1/N} = (1 + i^*)$$

Quindi il tasso medio sarà uguale a

$$[(1 + i_1) \cdot (1 + i_2) \cdot \ldots \cdot (1 + i_N)]^{1/N} - 1 = i^*$$

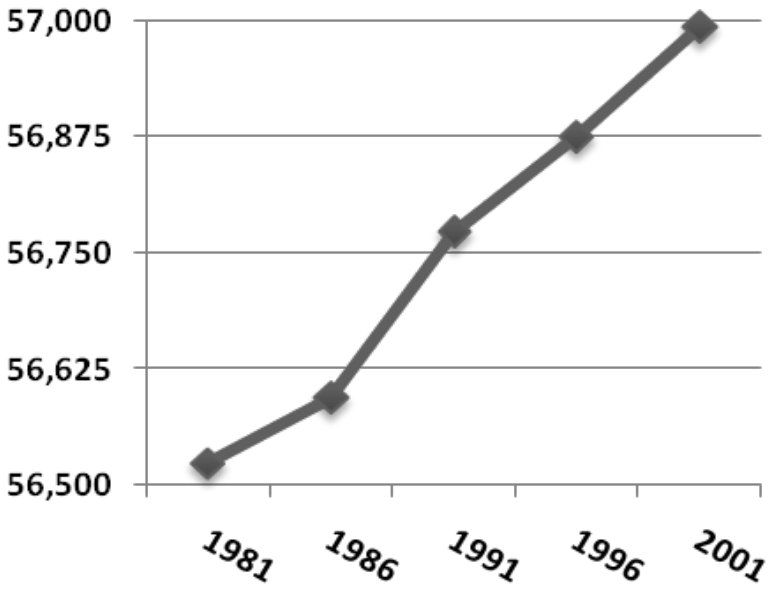
$$[(i_1) \cdot (i_2) \cdot \ldots \cdot (i_N)]^{1/N} = i^*$$

Abbiamo utilizzato la media geometrica, moltiplicando tutte le modalità e "redistribuendole" tra le diverse unità elevando l'intensità totale per 1/N, reciproco della numerosità del collettivo

32 – Esempio

Nella tabella seguente sono riportati i dati sulla popolazione italiana residente dal 1981 al 2001

	Popolazione (in mil)	Variazione relativa	Tasso sviluppo
1981	56.52	-	-
1986	56.59	1.0012	1.2‰
1991	56.77	1.0032	3.2‰
1996	56.88	1.0018	1.8‰
2001	56.99	1.0021	2.1‰



$$\overline{x}_g = \sqrt[4]{0.0012 \times 0.0032 \times 0.0018 \times 0.0021} = 0.0020 \;:\; 2{,}0‰$$

Dal calcolo della media geometrica possiamo dedurre che la popolazione italiana è cresciuta nel periodo 1981-2001 con un tasso medio di variazione pari allo 2,0 ‰

33 – Media geometrica per distribuzioni di frequenza

Nel caso in cui dobbiamo calcolare la media geometrica di dati organizzati in una distribuzione di frequenza dobbiamo tener conto di quante unità statistiche hanno manifestato le differenti modalità del carattere

$$\overline{x}_g = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$$

Se abbiamo delle frequenze relative dobbiamo ovviamente modificare la formula ottenendo

$$\overline{x}_g = x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_k^{f_k}$$

Nel caso di distribuzioni in classi si sostituisce il valore centrale delle classi alle modalità x_i

34 – Esercizio

Si consideri la seguente distribuzione che riporta il tasso praticato da alcune banche sui mutui per l'acquisto della prima casa:

Tasso %	N. banche
5,3	2
5,5	3
6,1	1
6,3	1

Tramite un indice opportuno, si calcoli il tasso medio praticato dalle banche in questione

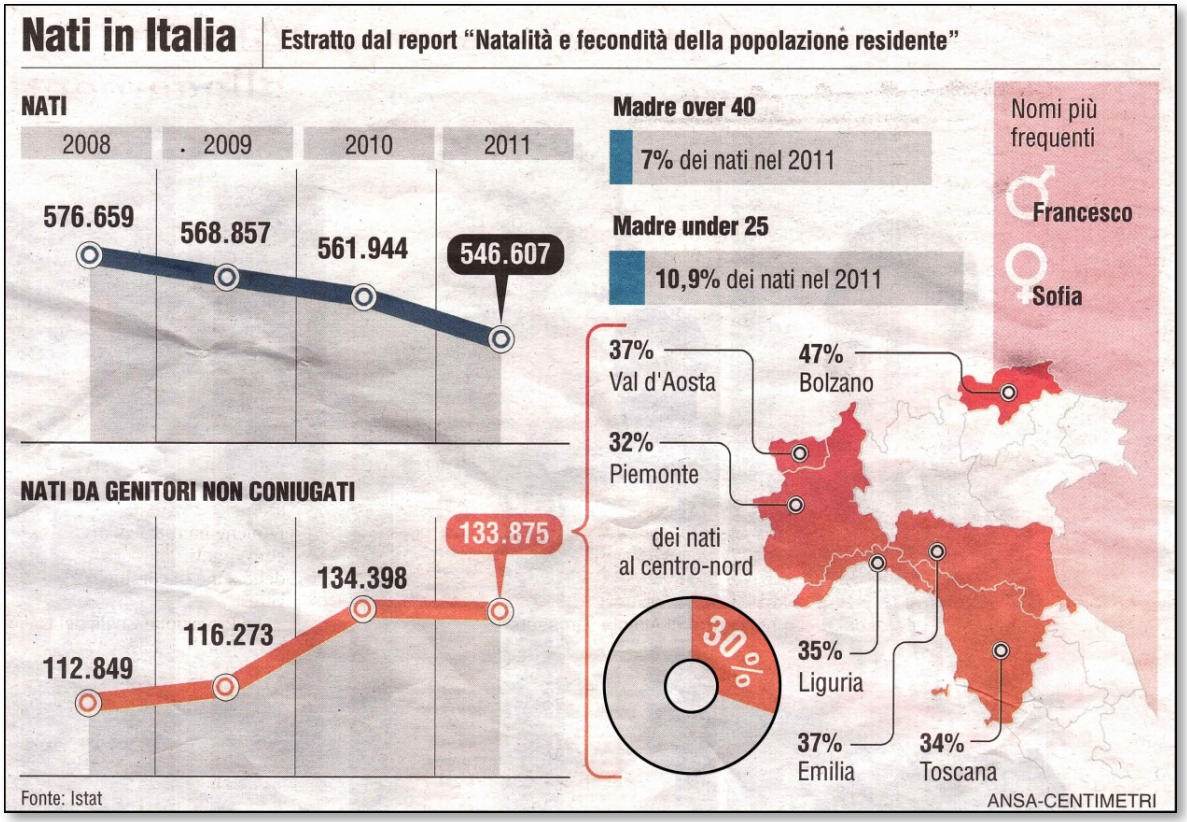
---

Soluzione

Tasso (t)	n_i	$x_i = t/100$	$x_i^{n_i}$
5,3	2	0,053	0,0028
5,5	3	0,055	0,0002
6,1	1	0,061	0,0610
6,3	1	0,063	0,0630

➡ **il tasso medio è pari al 5,63%**

35 – Esercizio



Il 15 novembre 2012 l'Istat ha pubblicato il rapporto sulla **Natalità e fecondità della popolazione residente 2011**

Sulla base dei dati riportati sull'infografica qui presente:

1) calcolare e commentare la variazione relativa media dei *nati* e dei *nati da genitori non coniugati* tra il 2008 e il 2011

2) calcolare e commentare i tassi di variazione per i nati e i nati da genitori non coniugati, motivando le scelte effettuate

rapporto completo

36 – Medie di dati percentuali (1)

Le percentuali possono talvolta ingannare perché consentono di rappresentare fattispecie molto diverse da un punto di vista concettuale e interpretativo

(a) ➡ il 16% degli impiegati dell'azienda Alfa è di genere femminile

(b) ➡ il n° di impiegate di Alfa è cresciuto del 16% tra il 2008 e il 2012

Nel caso (a) si rapporta una parte del collettivo all'ammontare totale del collettivo stesso, nel caso (b) si confronta una parte del collettivo osservata in un dato momento con la stessa osservata in un momento diverso

➡ qual è la percentuale media di donne impiegate nelle aziende del settore in cui opera l'azienda Alfa?

➡ qual è la percentuale media di crescita dell'occupazione femminile in Alfa dalla sua fondazione a oggi?

Che tipo di media si deve utilizzare nei due casi in esame?

37 – Medie di dati percentuali (2)

La percentuale media è sempre uguale alla media delle percentuali?

	n_f	N	n_f/N	%
Alfa	8	50	0.160	16.0
Beta	24	83	0.289	29.0
Gamma	13	26	0.500	50.0
SETTORE	45	159	0.283	28.3
media aritmetica %			0.316	31.6
media ponderata %			0.283	28.3

se la proporzione di donne impiegate, rispetto alle corrispondenti popolazioni di riferimento, è molto dissimile nelle diverse aziende, la media aritmetica non è un buon indice di sintesi e va invece preferita una media ponderata con pesi pari al n° di impiegati per ciascuna azienda

Solo se le popolazioni sono pressoché uguali è possibile utilizzare la media semplice

	n_f	N	n_f/N	%
Alfa	18	50	0.360	36.0
Beta	22	83	0.265	26.5
Gamma	20	26	0.769	76.9
SETTORE	60	159	0.377	37.7
media aritmetica %			0.465	46.5
media ponderata %			0.377	37.7

	n_f	N	n_f/N	%
Alfa	8	68	0.118	11.8
Beta	24	80	0.300	30.0
Gamma	13	66	0.197	19.7
SETTORE	45	214	0.210	21.0
media aritmetica %			0.205	20.5
media ponderata %			0.210	21.0

38 – Robustezza della media aritmetica

Abbiamo visto come con le medie sia possibile sintetizzare i dati rilevati in un collettivo, allo scopo di descrivere il fenomeno che ci interessa studiare

Con le medie analitiche si opera una trasformazione "matematica" dei dati:

1) non è possibile utilizzare questi indici su dati qualitativi (non esiste la media del colore degli occhi o del titolo di studio…)
2) i valori osservati molto piccoli o molto grandi rispetto alla distribuzione del carattere nel collettivo (detti **valori anomali**) influenzano il valore dell'indice e portano ad una lettura del fenomeno fuorviante

Consideriamo la statura delle giocatrici di una squadra di pallavolo:

cm	173	169	173	175	170	175	209	172

ordiniamo i dati in senso crescente

cm	169	170	172	173	173	175	175	209

L'altezza media è pari a 177 cm…

Come si vede dai dati in tabella, la media rispetta la proprietà dell'internalità ma di fatto non è rappresentativa perché superiore a 7 modalità su 8: questo è dovuto al fatto che il valore 209 risulta essere molto più grande rispetto agli altri, e quindi tende ad "attrarre" il valore medio

39 – Le medie di posizione

Le medie di posizione consentono di sintetizzare il fenomeno con criteri differenti rispetto a quelli visti per le medie analitiche

MODA → *si tiene conto della frequenza delle modalità del carattere studiato (può essere calcolata su qualsiasi tipo di carattere)*

MEDIANA → *si tiene conto della posizione delle modalità del carattere studiato (può essere calcolata su qualsiasi tipo di carattere ordinabile)*

Analogamente alle medie analitiche sono espresse nella stessa unità di misura dei dati

40 – La moda

Unità n° 03

La moda **Mo** è la modalità più frequente della distribuzione del carattere

Consumi ml.(€)	Numero reparti
1.0	20
1.5	80
2.0	90
2.5	**140**
3.0	70
Totale	400

Consideriamo la distribuzione dei consumi (in ML di €) dei reparti ospedalieri delle strutture della ASL di una grande città

La moda è in questo caso 2.5, perché a questa modalità corrisponde la frequenza più alta (140)

Mo = 2.5 ML di € (è il **consumo modale**)

(il consumo medio è invece $\overline{x}_a$ = 2.2 ML di €)

I reparti spendono più frequentemente tale ammontare

Per determinare la moda possiamo utilizzare le frequenze assolute, le frequenze relative o le frequenze percentuali: la moda è sempre la modalità prevalente

41 – Moda per distribuzioni in classi

Nel caso di distribuzioni in classi ovviamente non possiamo individuare una modalità prevalente ma dobbiamo invece fare riferimento alla classe di modalità più frequenti

Consumi ml.(€)	N° reparti	Ampiezza classe	Densità frequenza
5 -| 2.5	100	2	100/2 = 50
2.5 -| 3.5	**90**	**1**	**90/1 = 90**
3.5 -| 6.0	210	2.5	210/2.5 = 84
Totale	400		

Così come per le distribuzioni di frequenza possiamo considerare le frequenze assolute, relative o percentuali

Se però le classi non sono equiampie bisogna fare riferimento alla densità di frequenza

La **classe modale** è la classe della distribuzione corrispondente alla frequenza più alta (se le classi sono equiampie) o alla densità di frequenza più alta (se le classi non sono equiampie)

Nell'esempio in tabella la moda Mo, o meglio la classe modale, è pari a 2.5-|3.5 (ML di €)

42 – Caratteristiche della Moda

La moda di una distribuzione è la modalità a cui è associata la frequenza più elevata, quindi si evince facilmente che la moda è un indice di posizione che può essere determinato per qualsiasi tipo di carattere, quantitativo o qualitativo

È comunque necessario puntualizzare alcune aspetti fondamentali

1) la moda può ritenersi un buon criterio di sintesi quando si presenta con una frequenza "nettamente maggiore" di tutte le altre modalità (almeno il 50% delle osservazioni). In tal caso è ragionevole assumerla come valore tipico del fenomeno, cioè come quel valore più idoneo a rappresentarlo sinteticamente

2) la moda potrebbe non essere unica: se si individuano due modalità con frequenza maggiore si parla di distribuzione *bimodale* (due mode)

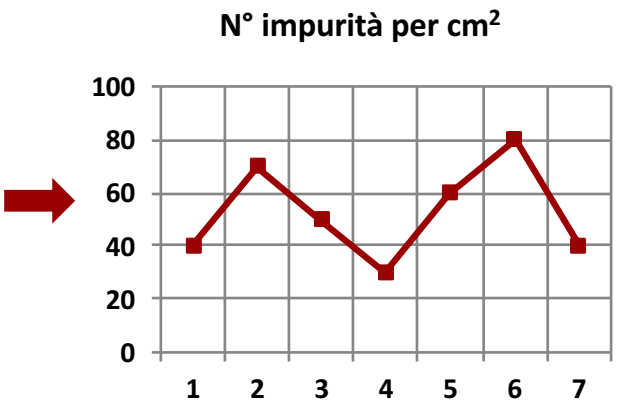
3) se tutte le modalità del carattere presentano all'incirca le stesse frequenze, allora non ha senso determinare la moda: per alcuni studiosi già non ha senso parlare di moda se nella distribuzione si individuano più di due valori maggiormente ricorrenti

43 – Distribuzioni bimodali

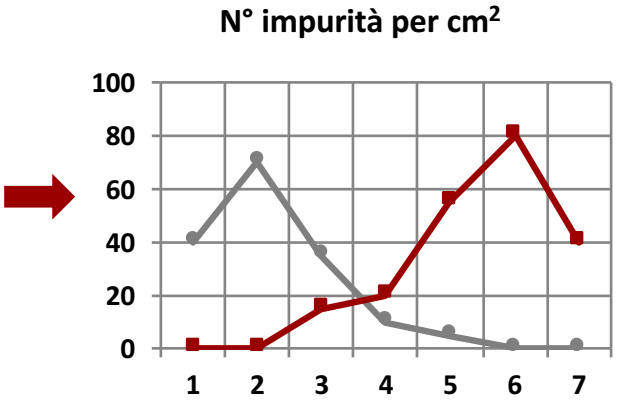
È possibile che nel collettivo ci siano 2 gruppi omogenei rispetto ad un'altra caratteristica

Un'industria di vasellame vuole controllare la qualità della creta utilizzata nella lavorazione

N° impurità per cm^2	Campioni
1	40
2	70
3	50
4	30
5	60
6	80
7	40
Totale	370



N° impurità per cm^2	Cava 1	Cava 2	Campioni
1	40	0	40
2	70	0	70
3	35	15	50
4	10	20	30
5	5	55	60
6	0	80	80
7	0	40	40
Totale	160	210	370



Dall'analisi della distribuzione si evince che le modalità prevalenti sono due (70 e 80)

Se teniamo conto del fatto che i campioni di creta sono prelevati da due diversi siti allora possiamo vedere come in effetti il collettivo esaminato può essere suddiviso in due diversi sotto-collettivi e di conseguenza sfruttare questa informazione per meglio studiare il fenomeno

44 – La mediana

La mediana è il centro di un insieme di valori ordinati, è cioè il valore che bipartisce il collettivo statistico in due gruppi di uguale numerosità

La determinazione della mediana richiede quindi, come prerequisito, che il carattere in esame sia almeno ordinale. Pertanto potrà essere determinata per tutti i tipi di caratteri quantitativi o qualitativi, tranne quelli sconnessi (es. colore degli occhi -> **no** | reddito, titolo di studio -> **si**)

A seconda della numerosità e di come sono organizzati i dati, cambia il modo di determinare l'indice: in generale per le distribuzioni unitarie si guarda alla numerosità, cioè se le unità del collettivo sono pari o dispari; per le distribuzioni di frequenza si guarda invece alle frequenze cumulate, meglio se relative

Considerando una successione di N valori ordinati in senso non decrescente:

$$X_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(i)} \leq \ldots \leq x_{(N)}$$

la mediana **Me** è definita come il valore centrale della successione, cioè come quel valore che è preceduto e seguito dallo stesso numero di dati della distribuzione (50%-50%)

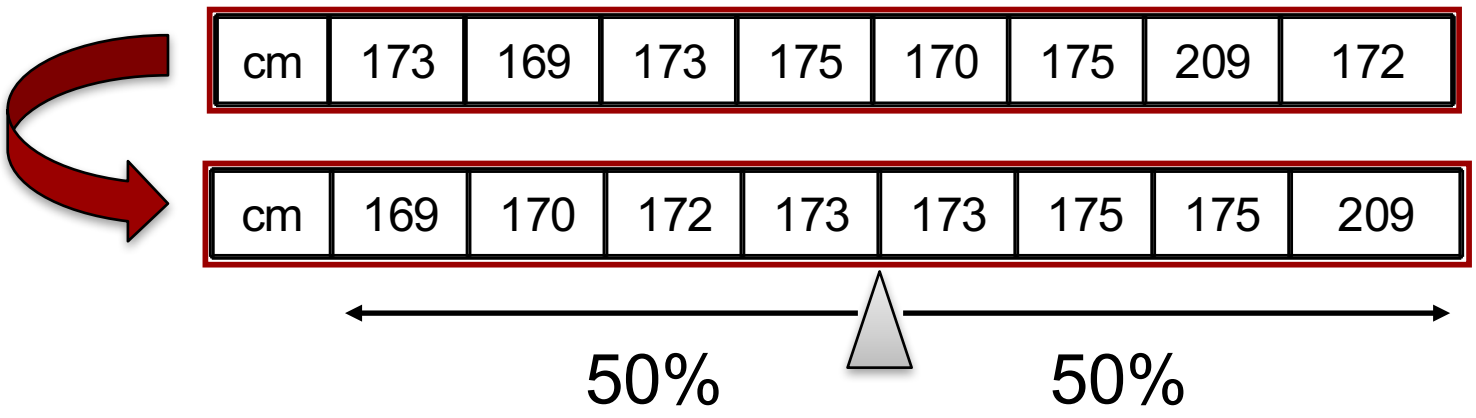
45 – La mediana per distribuzioni unitarie

E' la modalità presentata dall'unità centrale del collettivo

Divide il collettivo in due sottoinsiemi di uguale numerosità: uno con modalità di ordine più basso e l'altro con modalità di ordine più alto

Statura delle giocatrici di una squadra di pallavolo

Ordiniamo i dati



Il calcolo della mediana è possibile solo per caratteri quantitativi o qualitativi ordinabili

46 – Come si calcola la mediana

1) Ordinare le unità in senso crescente

2) Individuare la posizione in graduatoria dell'unità centrale:

   - se *n* è *dispari*, la posizione è (n+1)/2

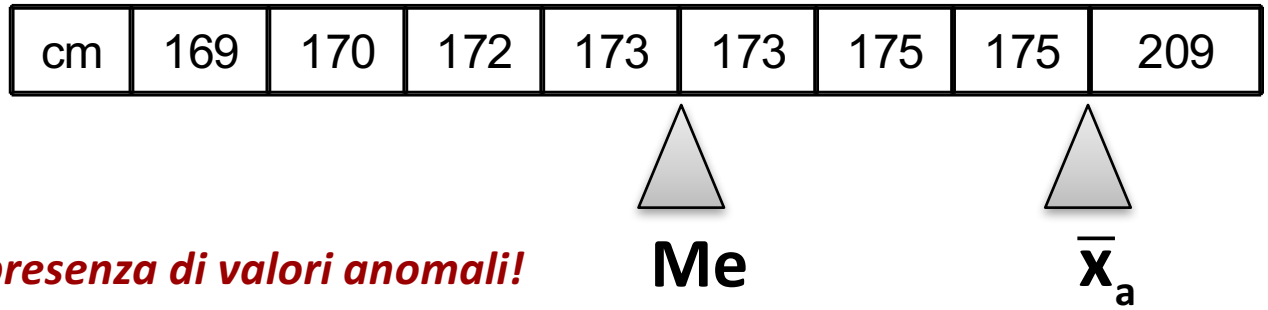
     (la mediana è la modalità presentata dall'unità centrale)

   - se *n* è *pari* si hanno due unità centrali con posizione *n*/2 e *n*/2 +1

     (si hanno due mediane date dalle modalità delle due unità centrali:

     se il carattere è quantitativo, possiamo considerare come mediana

     la semisomma dei valori delle due unità centrali)



La mediana non è sensibile alla presenza di valori anomali!

47 – Esempio (dati quantitativi)

Consideriamo i seguenti valori:

10 -5 0,8 -2 3 2 5 0,515 3

e ordiniamoli in senso non decrescente:

mediana

Valori	**-5**	**-2**	**0,515**	**0,8**	**2**	**3**	**3**	**5**	**10**
Rango	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*

posizione centrale

Poiché $n=9$, esiste un'unica posizione centrale: la posizione $(9+1)/2=5$. Pertanto:

$$Me = x_{(5)} = 2$$

Consideriamo la stessa successione dell'esempio precedente a cui aggiungiamo il valore 7. La successione ordinata, composta dunque da $n=10$ termini, sarà:

2,5 Modalità centrali

Valori	**-5**	**-2**	**0,515**	**0,8**	**2**	**3**	**3**	**5**	**7**	**10**
Rango	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

Posizioni centrali: $P_1=5$, $P_2=6$

$$Me = \frac{x_{(5)} + x_{(6)}}{2} = \frac{2+3}{2} = 2{,}5$$

La mediana è la modalità corrispondente alla posizione centrale della distribuzione se la distribuzione ha una numerosità dispari o alla semisomma delle modalità corrispondenti alle posizioni centrali della distribuzione se la distribuzione ha numerosità dispari

Nell'esempio proposto la mediana **Me** non è pari a 5 ma a $\mathbf{x_{(5)}}$, quindi alla modalità 2

Analogamente nel secondo caso la mediana **Me** non è pari a 5 e 6 ma a $\mathbf{x_{(5)}}$ e $\mathbf{x_{(6)}}$, quindi per convenzione si considera il valore

$$\mathbf{Me} = 0{,}5 \cdot (\mathbf{x_{(5)}} + \mathbf{x_{(6)}}) = 2.5$$

48 – Esempio (dati qualitativi)

Consideriamo i giudizi ricevuti dagli scolari di una classe elementare alla fine dell'anno

Ordiniamo le modalità in senso crescente poiché il giudizio è espresso sulla base di un carattere qualitativo ordinato

Studente	Giudizio
1	Insufficiente
2	Sufficiente
3	Buono
4	Insufficiente
5	Sufficiente
6	Discreto
7	Ottimo
8	Ottimo
9	Buono

Studente	Giudizio
1	Insufficiente
4	Insufficiente
2	Sufficiente
5	Sufficiente
6	Discreto
3	Buono
9	Buono
7	Ottimo
8	Ottimo

La modalità DISCRETO rappresenta la mediana della distribuzione

49 – La mediana per distribuzioni di frequenza

Unità n° 03

Per le distribuzioni di frequenza è necessario osservare le frequenze cumulate relative:

X	n	f	F
x_1	n_1	f_1	F_1
…	…	…	…
x_i	n_i	f_i	F_i
…	…	…	…
x_k	n_k	f_k	1
Totale	N	1	

La mediana Me è la modalità x_i se la corrispondente frequenza cumulata relativa F_i è maggiore o uguale a 0,50 (50%):
in tal caso l'unità che lascia a destra e a sinistra lo stesso numero di osservazioni si trova tra le unità che presentano x_i

Consideriamo ad esempio la distribuzione del numero di figli in un collettivo di 50 famiglie:

Mediana →

Me=2

X	n_i	f_i	F_i
0	5	0,10	0,10
1	12	0,24	0,34
2	19	0,38	0,72
3	9	0,18	0,90
4	4	0,08	0,98
5	1	0,02	1,00
Totale	**50**	1,00	

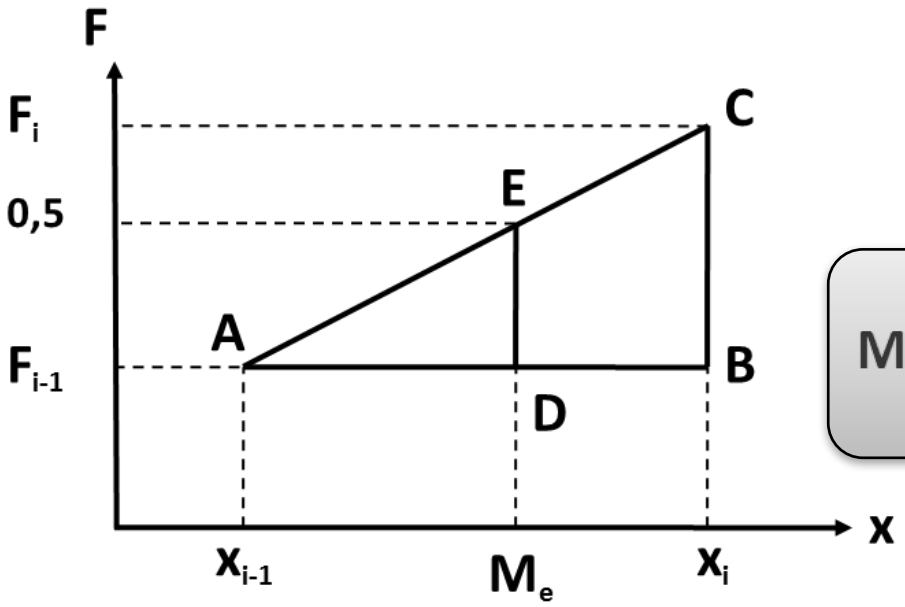
La Me non può essere 1 perché le famiglie con meno di un figlio sono il 34%: in corrispondenza di 2 abbiamo 0,72 e cioè il 72%, quindi le famiglie con meno di 2 figli sono quelle che bipartiscono la distribuzione

50 – Mediana per distribuzioni in classi

Se il carattere è suddiviso in classi, si può ottenere un valore ben approssimato assumendo implicitamente l'ipotesi che nella classe mediana le unità siano distribuite uniformemente

X	n	f	F
x_1-|x_2	n_1	f_1	F_1
x_2-|x_3	n_2	f_2	F_2
...	...	...	...
x_{i-1}-|x_i	n_i	f_i	F_i
...	...	...	...
x_{k-1}-|x_k	n_k	f_k	1
totale	N	1	

La classe mediana è la classe x_{i-1} -| x_i se la corrispondente frequenza cumulata relativa F_i è maggiore o uguale a 0,50: una volta individuata la classe mediana, si può ottenere la mediana per **approssimazione lineare**



$$AB:AD = BC:DE$$

$$AD = \frac{AB \times DE}{BC}$$

$$Me \approx x_{i-1} + \left(\frac{0,5 - F_{i-1}}{F_i - F_{i-1}} \right) \cdot \omega_i$$

51 – Esempio

Consideriamo la distribuzione del numero di impiegati per anni di servizio in una industria

Anni di servizio	n. impiegati	N	F
0 -| 1	7	7	0,06
1 -| 5	18	25	0,22
5 -| 10	45	70	0,61
10 -| 20	25	95	0,83
20 -| 30	20	115	1,00
Totale	115		

Dalle fr. cumulate individuiamo la classe mediana: in corrispondenza di 5-|10 anni abbiamo 0,61 -> il 61% degli impiegati ha meno di 10 anni di servizio, quindi in questa classe si trova il valore che lascia a destra e sinistra lo stesso numero di osservazioni

A questo punto applichiamo la formula

La mediana, per quanto detto, sarà uguale a:

$$M_e \approx 5 + \left(\frac{0{,}5 - 0{,}22}{0{,}61 - 0{,}22} \right) \cdot 5 = 5 + 3{,}59 = 8{,}59 \text{ anni}$$

- 5: estremo inferiore classe mediana
- 0,22: fr. cumulata alla classe precedente la classe mediana
- 0,61: fr. cumulata alla classe mediana
- 5: ampiezza classe mediana

52 – I percentili

Come abbiamo visto, la mediana è quel valore che divide il collettivo statistico in due parti uguali ognuna contenente il 50% delle unità

Estendendo il discorso possiamo immaginare di suddividere il collettivo in 100 parti, ognuna delle quali contenente lo stesso numero di unità. I valori che suddividono la distribuzione in 100 parti di uguale numerosità sono detti **percentili** (o quantili)

Si definisce p-mo **percentile**, corrispondente alla frazione **p/100** del collettivo, la modalità x_i del carattere che suddivide il collettivo in due gruppi tali che:

1) il primo gruppo ha numerosità N(p/100) e le sue unità hanno una modalità al più (inferiore o uguale) pari a x_i
2) il secondo gruppo ha numerosità N(1-p/100) e le sue unità hanno una modalità almeno (superiore o uguale) pari a x_i

In generale il p-esimo percentile è quello che lascia alla sua sinistra il p% della distribuzione ordinata dei valori osservati e alla sua destra il (100-p)%

53 – Percentili delle distribuzioni unitarie

In generale il p-esimo percentile è la modalità che (in senso crescente) si trova nella posizione

$$i = [p / 100 \cdot \text{Num. Collettivo}]$$

Se i è intero si considera la media tra x_i e x_{i+1}, se invece i non è intero si considera x_{i+1}

Che cosa significa essere al 5° percentile di statura?

→

Significa che su una popolazione di 100 individui abbiamo 5 soggetti più bassi di noi e 95 più alti

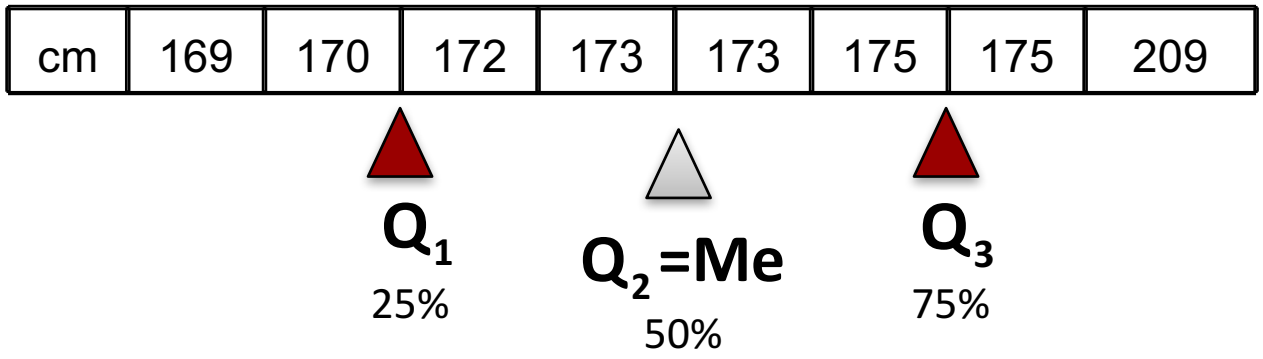
5° percentile

54 – Il 25° e 75° percentile: i quartili

Alcuni percentili in particolare sono di interesse per la descrizione dei fenomeni

I **quartili** sono dei percentili che consentono di dividere la distribuzione in quattro parti uguali:

- se p=25 allora abbiamo il 25° percentile (detto **primo quartile**, Q_1), cioè la modalità che lascia a sinistra il 25% delle unità
- se p=75 allora abbiamo il 75° percentile (detto **terzo quartile**, Q_3), cioè la modalità che lascia a sinistra il 75% delle unità



Il **secondo quartile**, per p=50, coincide con la mediana della distribuzione e rappresenta quella modalità che lascia a sinistra (e a destra) il 50% delle unità statistiche

55 – I percentili delle distribuzioni unitarie e di frequenza — Unità n° 03

Per determinare i percentili nelle distribuzioni unitarie e di frequenza si può utilizzare lo stesso procedimento visto per la mediana

Per le distribuzioni unitarie è necessario ordinare i dati in senso crescente e quindi individuare il percentile come quella modalità corrispondente all'unità statistica che divide la distribuzione in due gruppi, in base a quanto detto precedentemente

Per le distribuzioni di frequenza è sempre utile riferirsi alle frequenze cumulate relative per individuare la modalità che ci interessa

ATTENZIONE

Vale la pena sottolineare come i percentili siano sempre le modalità e non le posizioni!

→ Per non confondersi è possibile riferirsi a quanto visto per la media: anche in questo caso il valore che ci interessa è compreso tra il valore più piccolo e più grande della distribuzione, ed espresso nella stessa unità di misura con cui si stanno leggendo i dati

56 – Percentili delle distribuzioni di frequenza

Unità n° 03

Per le distribuzioni di frequenza è necessario osservare le frequenze cumulate relative:

X	n	f	F
x_1	n_1	f_1	F_1
…	…	…	…
x_i	n_i	f_i	F_i
…	…	…	…
x_k	n_k	f_k	1
totale	N	1	

Il percentile p è la modalità x_i se la corrispondente frequenza cumulata relativa F_i maggiore o uguale a p/100: in tal caso tra le unità che presentano x_i si trova l'unità statistica che lascia a destra il p% delle osservazioni

Consideriamo ad esempio la distribuzione del numero di figli in un collettivo di 50 famiglie:

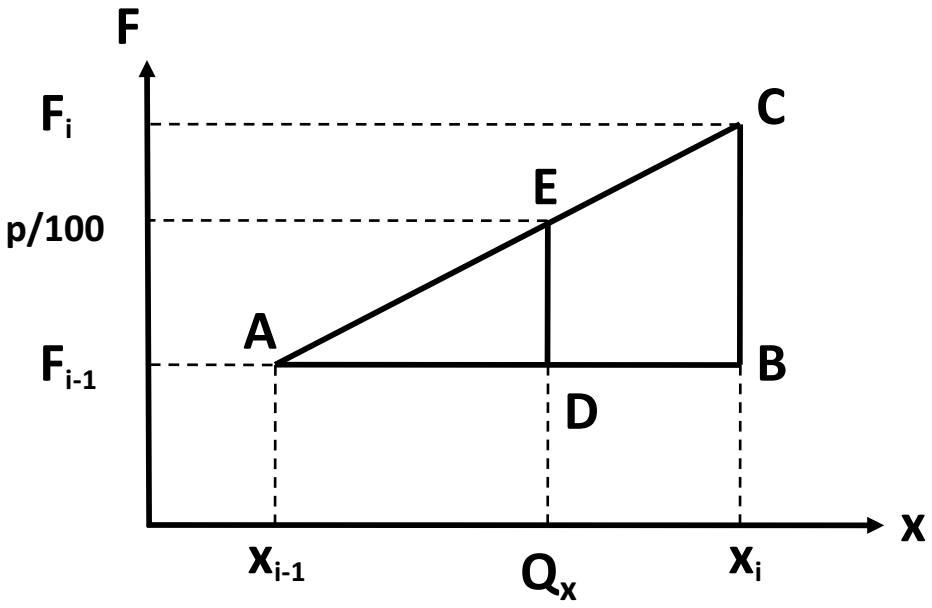
X	n	f	F
0	5	0.10	0.10
1	12	0.24	0.34
2	19	0.38	0.72
3	9	0.18	0.90
4	4	0.08	0.98
5	1	0.02	1
Totale	**50**	1	-

Possiamo individuare il 10° percentile in corrispondenza della modalità 0:

vuol dire che è al più pari al 10% il numero di famiglie che non hanno figli, o alternativamente che più del 10% delle famiglie ha almeno un figlio

57 – Percentili delle distribuzioni in classi

Per le distribuzioni in classi vale il principio dell'approssimazione lineare: si individua la classe per la quale la funzione di ripartizione è al più pari a p/100 e quindi si determina il valore del percentile per approssimazione lineare attraverso la stessa proporzione già utilizzata per la mediana



$$AB:AD = BC:DE \qquad AD = \frac{AB \times DE}{BC}$$

$$Q_1 \approx x_{i-1} + \left(\frac{0,25 - F_{i-1}}{F_i - F_{i-1}} \right) \cdot \omega_i$$

$$Q_3 \approx x_{i-1} + \left(\frac{0,75 - F_{i-1}}{F_i - F_{i-1}} \right) \cdot \omega_i$$

58 – Esempio

Consideriamo la distribuzione del numero di impiegati per anni di servizio in una industria

Anni di servizio	n. impiegati	N	F
0 -| 1	7	7	0,06
1 -| 5	18	25	0,22
5 -| 10	45	70	0,61
10 -| 20	25	95	0,83
20 -| 30	20	115	1,00
Totale	115		

Per calcolare ad esempio il terzo quartile si deve individuare la classe che lascia a destra il 75% delle osservazioni

La classe 10-20 lascia a destra l'83% delle modalità e quindi anche il 75%

Per quanto detto, Q_3 sarà uguale a:

$$Q_3 \approx 10 + \left(\frac{0,75 - 0,61}{0,83 - 0,61} \right) \cdot 10 = 10 + 6,36 = 16,36 \text{ anni}$$

59 – Rappresentazione grafica

Da un punto di vista grafico possiamo individuare e commentare i percentili dal poligono delle frequenze, rappresentazione della funzione di ripartizione empirica della distribuzione

"Tagliando" il poligono delle fr. in corrispondenza di 0,25 (25%) si ottiene il valore del primo quartile, a 0,50 (50%) si ha la mediana e a 0,75 (75%) si ha il terzo quartile

In corrispondenza degli altri valori (compresi tra 0 e 1) si hanno tutti gli altri percentili

60 – Valori positivi e valori negativi

Quando nella distribuzione di un carattere osserviamo sia valori positivi sia valori negativi abbiamo alcuni problemi ad utilizzare le medie analitiche

Esempio:
Analizzando la distribuzione dei prezzi della benzina negli ultimi 6 anni si sono osservati i seguenti tassi di variazione

2% -5% 2,5% 1% -0,75%

Qual è il tasso medio di variazione?

Per calcolare il tasso medio dovremmo utilizzare la media geometrica: il problema è che la media geometrica può essere calcolata solo in presenza di valori positivi, quindi nell'esempio proposto siamo "costretti" ad utilizzare la media aritmetica o la mediana per sintetizzare la distribuzione dei prezzi e avere una idea della intensità media del fenomeno oggetto di studio

61 – Riepilogo

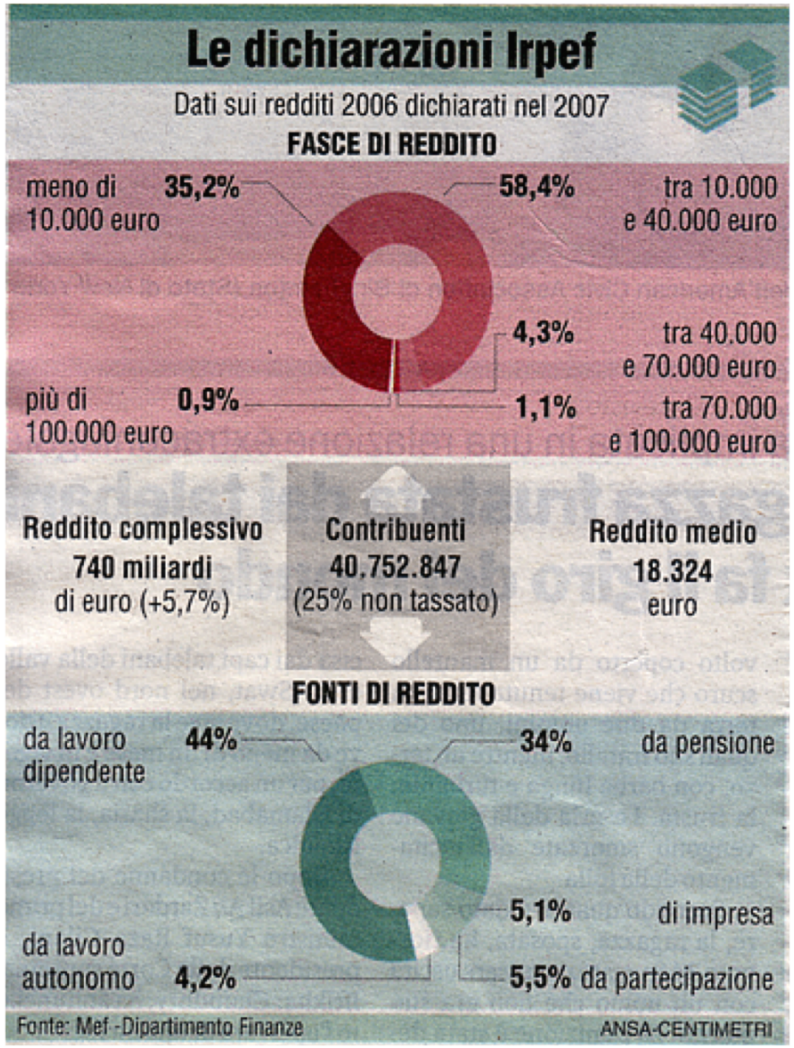
La scelta dell'indice più opportuno per descrivere sinteticamente la distribuzione di un certo fenomeno nel collettivo oggetto di studio dipende da diversi fattori

1) dobbiamo innanzi tutto considerare la natura del carattere, perché come visto non tutti gli indici sono idonei, ad esempio, a descrivere dei caratteri qualitativi. Se il carattere è qualitativo ordinato allora è possibile utilizzare sia la moda sia la mediana, se il carattere è invece qualitativo sconnesso allora è possibile utilizzare solo la moda

2) nel caso di caratteri quantitativi è possibile utilizzare le medie di posizione e le medie analitiche. In relazione a queste ultime la scelta dipende dal modo di calcolare l'intensità totale del fenomeno, cioè se il carattere è additivo o moltiplicativo

3) un altro elemento da prendere in considerazione è, nel caso dei caratteri quantitativi, la presenza o meno dei cosiddetti valori anomali. È infatti dimostrato che ad esempio la media aritmetica è sensibile a valori molto più piccoli o molto più grandi rispetto a quelli presenti nella distribuzione: potrebbe essere comunque conveniente in questi casi usare la mediana

62 – Un caso reale…



Il 4 aprile 2009 i maggiori quotidiani italiani hanno pubblicato i dati del Ministero dell'Economia relativi alle dichiarazioni dei redditi delle persone fisiche nel 2007. In questa elaborazione grafica dell'ANSA sono riportate delle informazioni che possono essere lette con alcuni degli strumenti della Statistica Descrittiva studiati in questo corso

Esercizio. Sulla base delle informazioni riportate:

1) Individuare il collettivo, il fenomeno e i caratteri studiati, specificandone la natura
2) Individuare intensità totale e intensità media del fenomeno nel collettivo
3) Determinare le distribuzioni di frequenza a partire dai dati delle rappresentazioni grafiche