Statistica

Statistica Descrittiva

2° parte

Isabella Corazziari

corazzia@istat.it

Le medie

Obiettivo: sintesi della distribuzione di un carattere mediante una modalità "tipica", altrimenti detto valore medio o caratteristico.

Strumenti:

- **Indici di posizione**: moda (tutte le scale); mediana, quantili (scala almeno ordinale)
- **Medie analitiche**: media aritmetica, media troncata (caratteri quantitativi); media geometrica, media armonica (scala di rapporti).

Le medie analitiche sono una funzione esplicita, analitica, delle modalità del carattere.

La moda:

La modalità (valore; classe; categoria) della variabile a cui corrisponde la **frequenza** (intensità) maggiore.
Nel caso di variabile quantitativa (discreta con tanti valori, oppure continua) raggruppata in classi di diversa ampiezza, la moda corrisponde a quella classe con **intensità** maggiore.

La media

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Media Aritmetica
Esempio: distribuzione unitaria

Apartment Rent Data

445	615	430	590	435	600	460	600	440	615
440	440	440	525	425	445	575	445	450	450
465	450	525	450	450	460	435	460	465	480
450	470	490	472	475	475	500	480	570	465
600	485	580	470	490	500	549	500	500	480
570	515	450	445	525	535	475	550	480	510
510	575	490	435	600	435	445	435	430	440

$$\bar{x} = \frac{\sum_{i=1}^{70} x_i}{70} = \frac{34356}{70} = 490.80$$

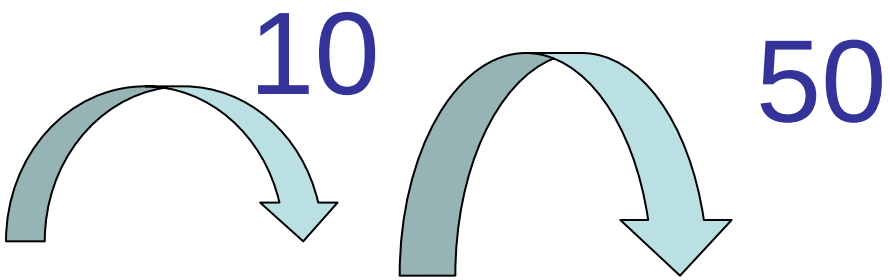
Esempio: dati raggruppati, distribuzione di frequenze (X=numero di cellulari)

Data la seguente distribuzione del numero di cellulari per famiglia, ottenuta dall'intervista di 100 famiglie. Calcolare la media aritmetica

Numero di cellulari	Numero di famiglie
0	10
1	50
2	10
3	30

Identificare: variabile, modalità, collettivo, unità statistiche, frequenze

Esempio: dati raggruppati (numero di cellulari)



$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0+\cdots+0+1+\cdots+1+2+\cdots+\ldots}{100}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0\times 10 + 1\times 50 + 2\times 10 + 3\times 30}{100} = \frac{160}{100} = 1.6$$

Esempio: dati raggruppati

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j^* \times n_j}{n} = \sum_{j=1}^{m} x_j^* \times f_j$$

n numero di unità (i=1,..,n)
m numero di modalità (j=1,…,m)

x_j^*	n_j	$n_j \times x_j^*$
0	10	0
1	50	50
2	10	20
3	30	90
	100	**160**

$\bar{x} = 1.6$

Esempio: dati raggruppati

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j^* \times n_j}{n} = \sum_{j=1}^{m} x_j^* \times f_j$$

n numero di unità (i=1,..,n)
m numero di modalità (j=1,…,m)

x_j^*	f_j	$f_j \times x_j^*$
0	0.10	0
1	0.50	0.5
2	0.10	0.2
3	0.30	0.9
		1.6

$\bar{x} = 1.6$

QUANDO IL CARATTERE è QUANTITATIVO CONTINUO (o assimilabile a un carattere continuo, vd reddito)

La suddivisione in classi

Arbitrarietà della suddivisione in classi.
Linee guida:

- al fine di facilitare l'interpretrazione della distribuzione, qualora possibile, le classi dovrebbero avere la stessa ampiezza
- evitare di costruire classi caratterizzate da un numero di frequenze molto basso
- equilibrio tra due esigenze in conflitto: sintesi e grado di risoluzione

Attenzione: la suddivisione in classi comporta una perdita di informazioni (le differenze presenti entro la classe).
Tale operazione ha senso soltanto se l'obiettivo finale è produrre una tabella di sintesi o un istogramma.
Per tutti gli altri scopi occorre lavorare con la distribuzione unitaria di partenza.

Esempio: Dati in classe

Considerata la seguente distribuzione del reddito di 100 famiglie

Stipendio	n_j
[0,10)	10
[10,20)	30
[20,30)	20
[30,50]	40

Chiusura delle classi: quando la variabile è discreta e quando è continua

Valore centrale della classe: (estremo inferiore + estremo superiore) /2

Stipendio	n_j	c_j	$n_j x c_j$
[0,10)	10	5	50
[10,20)	30	15	450
[20,30)	20	25	500
[30,50]	40	40	1600
	100		**2600**

Proprietà della Media Aritmetica

- La m.a. equiripartisce il totale di un carattere tra le unità

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

- Internalità

$$x_{(\min)} \leq \bar{x} \leq x_{(\max)}$$

- La somma degli scarti dalla media aritmetica e nulla:

$$\sum_{i=1}^{n} \left(x_i - \bar{x} \right) = 0$$

Proprietà media aritmetica

- La m.a. rende minima la somma dei quadrati degli scarti da una costante:

$$\bar{x} = \arg\min_{c} \sum_{i=1}^{n} (x_i - c)^2$$

- Linearità $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$ tale che $y_i = a + bx_i$ allora

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

Dimostrazione della linearità della media aritmetica

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} (a + bx_i)}{n} = \frac{\sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i}{n}$$

$$\frac{na + b\sum_{i=1}^{n} x_i}{n} = a + b\frac{\sum_{i=1}^{n} x_i}{n} = a + b\bar{x}$$

Dimostrazione della proprietà di minimizzazione degli scarti al quadrato: derivata prima =0 e derivata seconda >0

$$\sum_{i=1}^{n}\left(x_i - c\right)^2$$

$$\frac{d\sum_{i=1}^{n}\left(x_i - c\right)^2}{dc} =$$

$$\sum_{i=1}^{n}\frac{d\left(x_i - c\right)^2}{dc} = \sum_{i=1}^{n} -2\left(x_i - c\right) = -2\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n} c$$

Continua dimostrazione della proprietà di minimizzazione degli scarti al quadrato: derivata prima =0 e derivata seconda >0

$$\sum_{i=1}^{n} \frac{d\left(x_i - c\right)^2}{dc} = -2\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n} c$$

$$\sum_{i=1}^{n} \frac{d\left(x_i - c\right)^2}{dc} = 0$$

$$\sum_{i=1}^{n} x_i = nc$$

$$c = \bar{x}$$

$$\frac{d^2 \sum_{i=1}^{n} \left(x_i - c\right)^2}{dc} = 2n$$

Proprietà media aritmetica: associativa

- La media di un collettivo è la media aritmetica delle medie dei sottogruppi in cui può essere ripartito il medesimo, ponderata per le numerosità relative dei sottogruppi.
  - *Se* $\bar{x}_1$ *e* $\bar{x}_2$ *sono le medie di due campioni di ampiezza rispettivamente* n_1 *e* n_2 *la media può essere calcolata come*

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$\bar{x} = \frac{\sum_{j=1}^{m} n_j\bar{x}_j}{\sum_{j=1}^{m} n_j}$$

Proprietà media aritmetica: associativa

– *Se* $\bar{x}_1$ *e* $\bar{x}_2$ *sono le medie di due sottogruppi di ampiezza rispettivamente* n_1 *e* n_2 *, per la prima proprietà della media aritmetica possiamo scrivere*

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{\sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n} x_i}{n_1 + n_2} = \frac{\sum_{i=1}^{n} x_i}{n}$$

dove n_2=n-n_1

Il procedimento può essere generalizzato a un numero generico m di sottogruppi del collettivo in esame.

Esercizio

In un’azienda gli stipendi annui, in migliaia di euro, sono così distribuiti:

1	direttore	30
3	capi ufficio	20
10	impiegati	16
25	operai	12
30	manovali	10

Calcolare la media aritmetica, la mediana e la moda degli stipendi.

Esempio (stipendio annuo)

- Qual è l’unità statistica?
- Qual è il carattere?
- Cosa rappresenta la parola «impiegati»?

Rappresentazione tabellare

Impiego	**stipendio**	**frequenze**	**$x_j * n_j$**
Laborer	10	30	300
Worker	12	25	300
Employee	16	10	160
Head office	20	3	60
Director	30	1	30
total		**69**	**850**

$$\bar{x} = \frac{\sum_{i=1}^{69} x_i}{69} = \frac{\sum_{j=1}^{5} x_j^* n_j}{69} = \frac{850}{69} = 12.319$$

Esercizio

In una stanza ci sono 12 persone con un peso medio pari a 75kg. Se arriva un’altra persona che pesa 60 kg, qual è il peso medio delle 13 persone?

Esercizio

In una stanza ci sono 12 persone con un peso medio pari a 75kg. Se arriva un'altra persona che pesa 60 kg, qual è il peso medio delle 13 persone?

$$\bar{x}_{13} = \frac{\sum_{i=1}^{13} x_i}{13} = \frac{\sum_{i=1}^{12} x_i + x_{13}}{13} = \frac{12\bar{x}_{12} + x_{13}}{13} = \frac{12 \times 75 + 60}{13} = 73.84$$

associativa

Esercizio

Le temperature della neve in gradi Celsius di una nota località sciistica nel mese di gennaio sono state le seguenti

t_j	-4	-3	-2	-1	0	1
g_j	6	5	8	6	4	2

dove t_j è la temperatura rilevata in gradi Celsius e g_j è il numero di giorni in cui si è registrata la temperatura t_j

Esercizio

- Si calcoli la temperatura media: in gradi Celsius, e in gradi Fahrenheit, sapendo che

$$T_{Fahr} = 32 + 1.8\ T_{Cels}$$

t_j	g_j	$t_j \cdot g_j$
-4	6	-24
-3	5	-15
-2	8	-16
-1	6	-6
0	4	0
1	2	2
tot	31	-59

$$\frac{\sum_{j=1}^{6} t_j \cdot g_j}{\sum_{j=1}^{6} g_j}$$

$$\frac{-59}{31} = -1.903$$

Nel periodo di osservazione, la temperatura media della neve nella nota località sciistica è stata pari $-1.903\ C°$. Più precisamente, $-1.903\ C°$ indica la temperatura che si sarebbe dovuta osservare nell'intero mese di gennaio nel caso in cui si fosse avuta la stessa temperatura in ogni giorno.
Si osservi che le relazioni che ci permettono di passare dalle temperature in gradi Celsius a quelle in gradi Fahrenheit e assoluti, sono lineari. In forza della proprietà di linearità della media aritmetica[1] le medie ricercate risultano:

$$\begin{aligned} M_1(T_{Fahrenheit}) &= 32 + 1.8 \cdot M_1(T_{Celsius}) \\ &= 32 + 1.8 \cdot (-1.903) = 28.574 \end{aligned}$$

linearità

Esercizio

In una scuola elementare si è misurata l'altezza di 100 bambini di quarta e si è trovato un valore medio di 126 cm. Ci si è accorti però che lo strumento era stato erroneamente posizionato e ciascun bambino è risultato 4 cm più basso, qual è il vero valore dell'altezza media dei bambini?

Esercizio

In una scuola elementare si è misurata l'altezza di 100 bambini di quarta e si è trovato un valore medio di 126 cm. Ci si è accorti però che lo strumento era stato erroneamente posizionato e ciascun bambino è risultato 4 cm più basso, qual è il vero valore dell'altezza media dei bambini?

$$y_i = x_i + 4$$

$$\bar{y} = \bar{x} + 4 = 126 + 4 = 130cm$$

Punti deboli della media aritmetica

- Robustezza. Sensibilità ai valori estremi.
- Rappresentatività nei confronti di distribuzioni asimmetriche. Più avanti argomenteremo che la media aritmetica è un valore di sintesi rappresentativo nei confronti di distribuzioni simmetriche.

Mediana

- La mediana è la modalità pertinente all'unità statistica che occupa la posizione centrale nella distribuzione **ordinata** delle osservazioni.
- Divide la distribuzione ordinata in due parti ciascuna contenente la meta delle osservazioni.
- Può essere calcolata per i caratteri misurati su scala ordinale e per quelli quantitativi.

Esempio Mediana: numero dispari osservazioni

Dati osservati:	24.1	22.6	21.5	23.7	22.6
Dati ordinati:	21.5	22.6	**22.6**	23.7	24.1
Posizione:	1	2	**3**	4	5

$$\text{Positioning Point} = \frac{n+1}{2} = \frac{5+1}{2} = 3.0$$

$$\text{Mediana} = 22.6$$

Esempio Mediana
numero pari osservazioni

Dati grezzi:	10.3	4.9	8.9	11.7	6.3	7.7
Ordinati:	4.9	6.3	**7.7**	**8.9**	10.3	11.7
Posizione:	1	2	**3**	**4**	5	6

$$\frac{n}{2} = 3 \quad \text{e} \quad \frac{n}{2} + 1 = 4$$

$$\text{Mediana} = \frac{7.7 + 8.9}{2} = 8.30$$

Mediana

1. Ordinare le osservazioni in ordine crescente
   n = numero di osservazioni.

2a. Se n è dispari, la mediana è l'osservazione che occupa la posizione (n + 1) / 2

2b. Se n è pari, la mediana è la media delle due osservazioni centrali (n/2 e (n/2+1)

Order	Data
1	78
2	91
3	94
4	98
5	99
6	101
7	103
8	105
9	114

← $n = 9$
$(n+1)/2 = 10/2 = 5$
Mediana = 99

$n = 10$
$n/2=5$ e $n/2+1 = 6$
Mediana = (99+101) /2 = 100

Order	Data
1	78
2	91
3	94
4	98
5	99
6	101
7	103
8	105
9	114
10	121

Median Apartment Data

Averaging the 35th and 36th data values:

Median = (475 + 475)/2 = 475

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Note: Data is in ascending order.

Mediana dati raggruppati

Data un collettivo di 100 famiglie, calcola la mediana

Numero di cellulari	Numero di famiglie
0	10
1	50
2	10
3	30

Mediana dati raggruppati

Numero di cellulari	Numero di famiglie	freq	Cumulative freq
0	10	0.1	0.1
1	50	0.5	0.6
2	10	0.1	0.7
3	30	0.3	1

La mediana è il primo valore la cui frequenza cumulata supera (**o è uguale**) 0.5

Mediana = 1

Mediana: dati ordinali qualitativi

Rating	Relative Frequency	Percent Frequency
Poor	0.10	10%
Below Average	0.15	15%
Average	0.25	25%
Above Average	0.45	45%
Excellent	0.05	5%
Total	**1.00**	**100%**

Ordinal qualitative data (Maradann Inn example)

Rating	Relative Frequency	Cumulative Frequency
Poor	0.10	0.10
Below Average	0.15	0.25
Average	0.25	0.50
Above Average	0.45	0.95
Excellent	0.05	1
Total	**1.00**	

Median Rating= Average

Caratteri quantitativi suddivisi in classi

Occorre fare riferimento alla distribuzione unitaria di partenza. Altrimenti, non è possibile calcolare la mediana se non in modo approssimativo, sotto l'ipotesi di *equidistribuzione* del carattere all'interno di ciascuna classe.

Ai fini dell'individuazione della classe entro cui cade la mediana si procede come sopra, facendo riferimento alle frequenze cumulate

Esempio: Dati in classe

Considerata la seguente distribuzione del reddito di 100 famiglie

Stipendio	n_j
[0,10)	10
[10,20)	30
[20,30)	20
[30,50]	40

Stipendio	n_j	f_j	F_j
[0,10)	10	0.10	0.10
[10,20)	30	0.30	0.40
[20,30)	20	0.20	0.60
[30,50]	40	0.40	1
	100	**1**	

Classe mediana: facoltativo calcolo valore approssimato della mediana

Proprià della mediana

La mediana minimizza la somma degli scarti in valore assoluto dei valori di un carattere quantitativo da una costante: definendo

$$S^*(c) = \sum_{i=1}^{n} |x_i - c|,$$

$$\arg\min_c \{S^*(c)\} = M_e.$$

Robustezza. La mediana è resistente, cioè insensibile alla presenza di valori anomali.

Moda

La moda è la modalità della distribuzione che si presenta con la massima frequenza.

Se la distribuzione di un carattere quantitativo e ripartita in classi di diversa ampiezza è necessario eliminare questa diversità, calcolando la densità di ogni classe per poi individuare la classe modale. La classe modale è quella a cui corrisponde la **densità di frequenza** più elevata.

Una distribuzione può essere unimodale o plurimodale.

Moda

Apartment Data

450 occurred most frequently (7 times)

Mode = 450

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Note: Data is in ascending order.

Moda dati raggruppati

- La moda del numero di cellulari per famiglia= 1

Number of cellular	Number of families
0	10
1	50
2	10
3	30

Esempio di applicazione dei tre indici di sintesi

I seguenti valori si riferiscono ai valori di un titolo rilevati mensilmente:

1.4; 1.7; 2.3; 2.5; 3.2; 3.8

Se il valore 3.8 fosse erroneamente trascritto come 38, quale sarebbe l'effetto sulle misure di posizione calcolate a partire da questi dati?

a) Un incremento della mediana.

b) Un incremento della moda.

c) Un incremento della media aritmetica.

d) Un incremento sia della mediana sia della moda.

e) Un incremento della mediana, della moda e della media aritmetica.

I quantili

I *quantili* sono modalità del carattere che suddividono la distribuzione in q distribuzioni parziali ciascuna contenente $1/q$ della numerosità totale (distribuzioni di frequenza) o della quantità totale (distribuzione di quantità).
Se $q = 10$ si parla di decili; se $q = 5$ di quintili (vedi foto); se $q = 4$ di quartili; se $q = 100$ di percentili.
Ad esempio, i *quartili* ripartiscono la distribuzioni in quattro parti caratterizzate dalla stessa numerosità, pari al 25% della numerosità totale.
La mediana è il 5^o decile, il 2^o quartile e il 50^o percentile.

80th Percentile

"At least 80% of the items take on a value of 542 or less."

56/70 = .8 or 80%

"At least 20% of the items take on a value of 542 or more."

14/70 = .2 or 20%

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Retro-cumulata è il complemento a n, 1, 100 rispettivamente di Ni, Fi, Pi

Quartiles

- Quartiles are specific percentiles
- First Quartile = 25th Percentile (median first half (50%) of distribution)
- Second Quartile = 50th Percentile = Median
- Third Quartile = 75th Percentile (median second half (50%) of distribution)

Remember:

- *Il primo e il terzo quartile possono essere considerati e calcolati come la mediana rispettivamente della prima metà del collettivo ordinato in base ai valori di X, e della seconda metà del collettivo.*
- *Come visto per la mediana: se la numerosità del collettivo è divisibile per 4, avremo due unità per il calcolo del primo quartile e due unità per il calcolo del terzo quartile, come anche due unità per la mediana o secondo quartile (se è divisibile per 4 è sicuramente pari); se non è divisibile per 4, avremo un'unica unità rispettivamente per Q1 e una per Q3.*

Third Quartile

Third quartile = 75th percentile

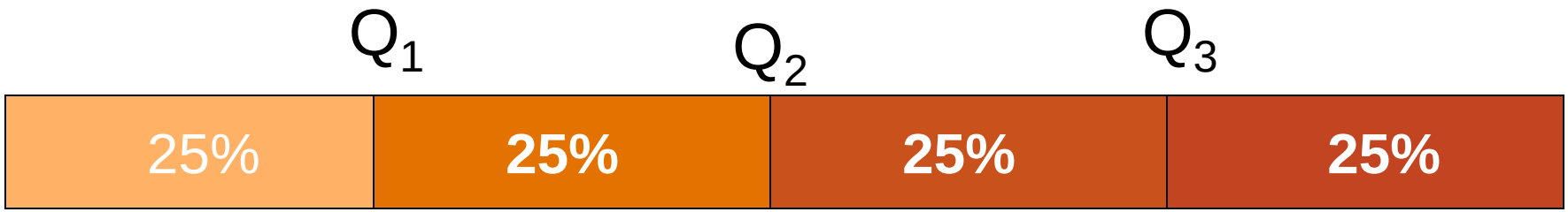
$i = (p/100)n = (75/100)70 = 52.5 = 53$

Third quartile = 525

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Note: Data is in ascending order.

Quartiles



Q_i element in position $i \times (n+1)/4$

Q_2 median

Regola per distribuzioni unitarie

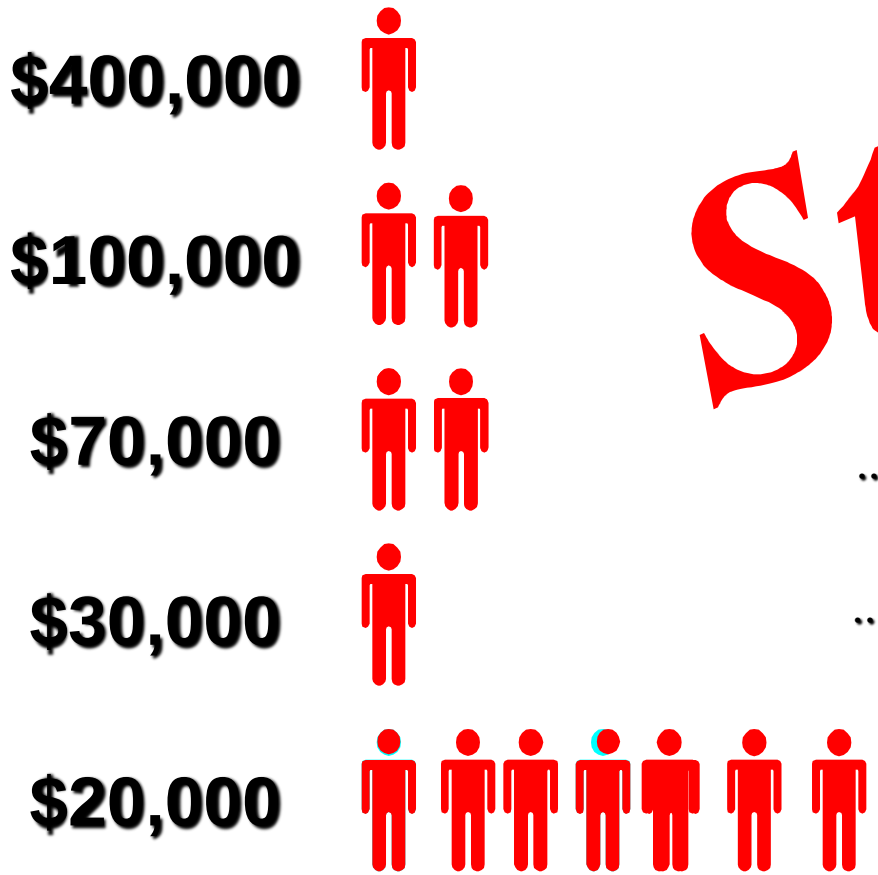
Quartiles grouped data

Number of cellular	Number of families	freq	Cumulative freq
0	10	0.1	0.1
1	50	0.5	0.6
2	10	0.1	0.7
3	30	0.3	1

$Q_1=Q_2=1$

$Q_3=3$

Thinking Challenge



Strike!

... employees cite low pay -- most workers earn only $20,000.

... President claims average pay is $70,000!

Calcolare media, moda, mediana