UNIVERSITA' DEGLI STUDI DI PERUGIA
DIPARTIMENTO DI FILOSOFIA SCIENZE SOCIALI UMANE E DELLA FORMAZIONE
Corso di Laurea in Scienze per l'Investigazione e la Sicurezza

4. I VALORI CENTRALI

Prof. Maurizio Pertichetti

Statistica sociale

4. I VALORI CENTRALI

Come si è avuto modo di dire una distribuzione di frequenze viene rappresentata in forma tabellare o grafica (od anche analitica). Sovente, però, quando si devono operare raffronti con situazioni analoghe in condizioni di contesto diverse è necessario invece fornire una efficace rappresentazione del fenomeno nella sua globalità riassumendone in modo sintetico le caratteristiche più importanti attraverso un solo "numero" o valore. Questi numeri o valori, definiti caratteristici, sono i **valori centrali** (o **indici di posizione** o **medie)** e gli **indici di variabilità**. I primi sintetizzano la distribuzione di frequenza con un unico "numero", che di fatto viene a rappresentare il valore intorno al quale si addensa la distribuzione ed è in grado di dare una descrizione dell'ordine di grandezza del fenomeno e confrontarlo con altri, i secondi danno invece una informazione su quanto diversi possono essere i valori della distribuzione in riferimento alle modalità.

I valori centrali si articolano in:

- **valori centrali analitici** (o sintetici), utilizzati solo per variabili quantitative, che si ottengono effettuando operazioni algebriche su tutti i valori della distribuzione. Tra questi vi sono la media aritmetica, la media geometrica, la media armonica, la media quadratica;
- **valori centrali non analitici** (o di posizione), la cui determinazione coinvolge le frequenze o solo alcuni valori della distribuzione. Tra questi vi sono la mediana e la moda.

In questo corso saranno utilizzate prevalentemente la media aritmetica, la mediana e la moda.

La media aritmetica

La **media aritmetica**, spesso indicata semplicemente come valore medio, può essere usata unicamente per le distribuzioni di caratteri quantitativi. Per la sua immediata intuizione, per le proprietà di cui gode e per la facilità del calcolo, è la più utilizzata fra le espressioni sintetiche della grandezza di una serie di dati.

Il concetto di media aritmetica è assai semplice. E' infatti il valore medio risultante dal quoziente tra la somma di tutti i valori di una distribuzione e il numero di questi valori.
Dati i valori tra loro diversi, $x_1 , x_2 , \dots , x_i , \dots , x_k$ di una distribuzione, la media μ si ottiene sommando questi valori e dividendo per il loro numero, che in tal caso corrisponde allo stesso numero delle modalità. In formula:

$$\mu = \frac{\sum_{i=1}^{k} x_i}{k}$$

Se i valori della distribuzione si ripetono, ovvero se le modalità si presentano avendo associate le frequenze $n_1 , n_2 , \dots , n_i , \dots , n_k$, la media μ si ottiene sommando i prodotti di ciascuna modalità per la corrispondente frequenza e dividendo per la somma di tutte le frequenze o, se vogliamo, per il totale dei casi:

$$\mu = \frac{\sum_{i=1}^{k} x_i\, n_i}{\sum_{i=1}^{k} n_i = n}$$

La media aritmetica **μ** è quel valore che sostituito a ciascuno dei valori della distribuzione, non ne altera la somma, ovvero produce lo stesso risultato:

$$\sum_{i=1}^{k} x_i = \mu k$$

La media aritmetica costituisce il baricentro della distribuzione di frequenza ed è sempre compresa tra i valori minimo e massimo della distribuzione, ovvero $x_1 \leq \mu \leq x_k$, e può coincidere con uno o più di essi o con nessuno di essi.

Alcune proprietà della media aritmetica:

- la somma algebrica degli scarti di ciascun termine di una distribuzione dalla propria media aritmetica è nulla:

$$\sum_{i=1}^{k} (x_i - \mu) = 0 \qquad\qquad \sum_{i=1}^{k} (x_i - \mu) n_i = 0$$

- la somma dei quadrati degli scarti tra ciascun termine della successione e la propria media aritmetica è un minimo, ovvero è inferiore alla somma dei quadrati degli scarti da qualsiasi altro valore diverso dalla media aritmetica:

$$\sum_{i=1}^{k} (x_i - \mu)^2 = \text{minimo}$$

In presenza di una variabile statistica divisa in intervalli, il metodo più semplice di calcolo della media aritmetica, anche se evidentemente introduce un errore di approssimazione, consiste nel sostituire le classi di intensità con il valore centrale di ciascuna di esse.
Inconveniente della media aritmetica è la sua scarsissima resistenza al valori eccezionali, nel senso che, anche un valore atipico può farla variare in misura elevatissima, sino a farle perdere significato. In pratica, la media aritmetica fornisce una utile sintesi quando la variabilità delle osservazioni non è elevata.

x_1	12	12	12	12
x_2	15	15	15	15
x_3	19	19	19	19
x_4	23	23	23	23
x_5	28	50	100	200
tot	**97**	**119**	**169**	**269**
k	**5**	**5**	**5**	**5**
μ	**19,4**	**23,8**	**33,8**	**53,8**

Esempi di calcolo della media aritmetica

Data la seguente distribuzione di valori:

12 15 19 23 28

$$\mu = \frac{12 + 15 + 19 + 23 + 28}{5} = \frac{97}{5} = 19,4$$

In questo caso si dimostra che la media aritmetica è quella quantità μ che sostituita a ciascuna modalità, non ne altera la somma. In formula:

$$\sum_{i=1}^{k} x_i = \mu k \qquad \underbrace{19,4 + 19,4 + 19,4 + 19,4 + 19,4}_{\text{5 volte}} = 19,4 * 5 = 97$$

Data la seguente distribuzione di frequenza:

n. componenti	n. famiglie
1	153
2	225
3	335
4	564
5	346
6	133
7	75
8	49
Totale	1880

x_i	n_i	$x_i n_i$
1	153	153
2	225	450
3	335	1.005
4	564	2.256
5	346	1.730
6	133	798
7	75	525
8	49	392
Totale	1.880	7.309

$$\mu = \frac{7.309}{1.880} = 3,89$$

Data la seguente distribuzione di frequenza per classi di età:

classi di età	componenti
4 ⊢⊣ 14	115
15 ⊢⊣ 25	156
26 ⊢⊣ 40	130
41 ⊢⊣ 55	110
56 ⊢⊣ 70	90
71 ⊢⊣ 85	38
Totale	639

x_i valore centrale	n_i	$x_i n_i$
9	115	1.035
20	156	3.120
33	130	4.290
48	110	5.280
63	90	5.670
78	38	2.964
Totale	639	22.359

$$\mu = \frac{22.359}{639} = 34,99$$

Costituiscono ulteriori casi particolari di valori centrali analitici **la media armonica**, **la media geometrica**, **la media quadratica**. Di esse, per conoscenza, se ne da un cenno.

Dati i valori tra loro diversi, $x_1 , x_2 , ... , x_i , ... , x_k$ di una distribuzione:

- la **media armonica μ_{ar}** si ottiene come reciproco della media aritmetica dei reciproci dei termini.

  In formula:

$$\mu_{ar} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}} \;;$$

con modalità associate alle frequenze n_i si ha:

$$\mu_{ar} = \frac{\sum_{i=1}^{k} n_i = n}{\sum_{i=1}^{k} \frac{1}{x_i} n_i} \;;$$

- la **media geometrica μ_g** è rappresentata dalla radice k-sima del prodotto di tutti i k termini.

  In formula:

$$\mu_g = \sqrt[k]{x_1 * x_2 * x_i * .... * x_k} = \sqrt[k]{\prod_{i=1}^{k} x_i} \;;$$

con modalità associate alle frequenze n_i :

$$\mu_g = \sqrt[n]{\prod_{i=1}^{k} x_i^{ni}} \;;$$

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

- la **media quadratica μ_q** è data dalla radice quadrata della media aritmetica dei quadrati dei termini.
  In formula:

$$\mu_q = \sqrt{\frac{x_1^2 + x_2^2 + x_i^2 + \ldots + x_k^2}{k}} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2}{k}}$$

; con modalità associate alle frequenze n_i :

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

La mediana

La **mediana** è quell'indice di posizione, ovvero quel valore caratteristico determinato tenendo conto sia delle frequenze sia dell'ordine delle modalità, che risponde al concetto più intuitivo di valore medio, essendo la modalità posseduta dall'unità che occupa il **posto centrale** di una distribuzione ordinata in senso non decrescente. Essa bipartisce la distribuzione delle modalità di un carattere, per cui il numero dei termini che precedono la mediana è uguale al numero dei termini che la seguono.

La **mediana**, che può essere calcolata tanto per i caratteri qualitativi ordinati, quanto per quelli quantitativi, è una misura robusta, in quanto poco influenzata dalla presenza di dati anomali. Altresì non è influenzata dalle osservazioni estreme.
Quando il numero delle osservazioni è dispari, la mediana è un termine effettivo in quanto corrisponde esattamente alla modalità che occupa la posizione centrale nella distribuzione, quando il numero delle osservazioni è pari la mediana è invece un valore di conto in quanto è data dalla semisomma dei due valori centrali.

Si distingue:

- per caratteri discreti con modalità ordinate se:

  n* *dispari la mediana è data dal valore corrispondente al posto centrale C = (n+1)/2 . Per cui **Me** è uguale alla modalità di **x** che corrisponde a C.

  n* *pari la mediana è data dal valore corrispondente alla semisomma dei due posti centrali C1=n/2 e C2=(n/2)+1 . Per cui **Me** corrisponde a (C1+C2)/2

- per caratteri continui con modalità raggruppate in classi:
  - preliminarmente si deve individuare la classe mediana, che è quella che corrisponde alla modalità associata alla prima frequenza cumulata relativa superiore al 50 per cento.
  - successivamente per stabilire la mediana, ovvero quale valore scegliere all'interno dell'intervallo della classe, si applica la formula:

$$Me = L_{me} + \frac{\frac{n}{2} - \sum n_{iMe}}{f_{Me}} c$$

dove:
- L_{Me} è il confine inferiore della classe mediana;
- n è la frequenza totale;
- $\sum n_{iMe}$ è l'accumulo delle frequenze di tutte le classi inferiori alla classe mediana;
- f_{Me} è la frequenza della classe mediana;
- c è l'ampiezza della classe mediana.

Esempi di calcolo della mediana

Date le seguenti distribuzioni d'intensità:

a) 13, 13, 15, 16, 16, 17, 18, 18, 19, 22, 23, 23, 23, 26, 45

le osservazioni della distribuzione sono in numero dispari (15) per cui la mediana è quel valore di x che corrisponde alla posizione (15 + 1)/2 = 8 . All'**8°** posto c'è la modalità **18** e quindi Me=**18**.

13 13 15 16 16 17 18 | 18 | 19 22 23 23 23 26 45
1° 2° 3° 4° 5° 6° 7° | **8°** | 9° 10° 11° 12° 13° 14° 15°
(7 valori) (7 valori)

b) 13, 13, 15, 16, 16, 17, 18, 18, 19, 22, 23, 23, 23, 26, 26, 45

in questo caso le osservazioni della distribuzione sono in numero pari (16) per cui la mediana è quel valore dato dalla semisomma dei valori di x corrispondenti alle posizioni 16/2=8 e (16/2)+1=9 . All'**8°** e **9°** ci sono le modalità **18** e **19**, per cui (18+19)/2 = 18,5 e quindi Me=**18,5**.

13 13 15 16 16 17 18 | 18 19 | 22 23 23 23 26 26 45
1° 2° 3° 4° 5° 6° 7° | **8° 9°** | 10° 11° 12° 13° 14° 15° 16°
(8 valori) (8 valori)

a)

x_i	n_i	n'_i
13	2	2
15	1	3
16	2	5
17	1	6
18	2	**8**
19	1	9
22	1	10
23	3	13
26	1	14
45	1	15
totale	15	

b)

xi	ni	n'_i
13	2	2
15	1	3
16	2	5
17	1	6
18	2	**8**
19	1	**9**
22	1	10
23	3	13
26	2	15
45	1	16
totale	16	

a) **Me** = 18 in quanto è dato dal valore d'intensità in corrispondenza del quale le frequenze cumulate hanno appena superato la semisomma delle frequenze totali n/2=7,5.

b) Poiché i posti centrali sono 8 e 9 e dalle frequenze cumulate si desume che si riferiscono alla modalità 18 e 19, **Me** = 18,5 .

Data la seguente distribuzione:

x_i	n_i
1	2
5	4
10	3
Totale	9

x_i	n_i	n'_i
1	2	2
5	4	**6**
10	3	9
Totale	9	

Poiché n=9 è dispari il posto centrale è C=(9+1)/2=5. La modalità che corrisponde al 5^0 posto è 5 in quanto è la prima modalità a cui corrisponde una frequenza cumulata superiore a n/2=9/2=4,5.

1 1 5 5 | 5 | 5 10 10 10
1° 2° 3° 4° | **5°** | 6° 7° 8° 9°

Data la seguente distribuzione:

x_i	n_i
3	4
5	6
13	3
15	1
Totale	14

x_i	n_i	n'_i
3	4	4
5	6	**10**
13	3	13
15	1	14
Totale	14	

Poiché n=14 è pari, i posti centrali sono $C=14/2=7^0$ e $C=(14/2)+1=8^0$. Tali posizioni si riferiscono entrambi alla modalità 5 in quanto è la prima cui corrisponde una frequenza cumulata superiore a n/2=14/2=7.

3	3	3	3	5	5	5	5	5	5	13	13	13	15
1°	2°	3°	4°	**5°**	6°	**7°**	**8°**	9°	10°	11°	12°	13°	14°

Data la seguente distribuzione di frequenza per classi di età:

Classi di età	Partecipanti
0 ⊢⊣ 13	22
14 ⊢⊣ 44	65
45 ⊢⊣ 64	80
65 ed oltre	40
Totale	207

n_i	n'_i
22	22
65	87
80	167
40	207
207	

La classe mediana è la classe 45-64, in quanto è la prima cui corrisponde una frequenza cumulata superiore a n/2=207/2=103,5.

	22	65	80	40
posti	$1^o \leftrightarrow 22^o$	$23^o \leftrightarrow 87^o$	$88^o \leftrightarrow 167^o$	$168^o \leftrightarrow 207^o$
	c'è la classe 0-13	c'è la classe 14-44	c'è la classe 45-64	c'è la classe 65 e +
			88^o **103,5** 167^o	

$$Me = L_{me} + \frac{\frac{n}{2} - \Sigma n_{iMe}}{f_{Me}} c$$

dove:

- L_{me} il confine inferiore della classe mediana è 45;
- n la frequenza totale è 207;
- Σn_{iMe} l'accumulo delle frequenze di tutte le classi inferiori alla classe mediana è 87;
- f_{Me} la frequenza della classe mediana è 80;
- c l'ampiezza della classe mediana è 20.

$$Me = 45 + \frac{\frac{207}{2} - 87}{80} 20 = 49{,}13$$

La moda

La **moda**, detta anche valore normale, è la modalità del carattere cui corrisponde la frequenza assoluta o relativa più alta. Questo valore centrale può essere calcolato per qualsiasi carattere, inclusi quelli qualitativi con modalità non ordinate.

Ha significato di sintesi quando la distribuzione presenta una sola moda, ossia è **unimodale** . Ma può averne anche due, nel qual caso si dice **bimodale**, tre, **trimodale**, o più. Una distribuzione si dice zeromodale quando non presenta alcuna moda, ovvero tutti i termini hanno la stessa frequenza.

Non è influenzata dalla presenza di nessun valore estremo, tuttavia viene utilizzata prevalentemente a scopi descrittivi.

Per caratteri continui con modalità raggruppate in classi, analogamente a quanto visto per la mediana, prima si individua la classe modale, quindi per stabilire il valore si applica la formula:

$$Mo = L_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} c$$

dove:

- L_{mo} è il confine inferiore della classe modale;
- Δ_1 è l'eccesso della frequenza modale sulla frequenza della classe immediatamente inferiore;
- Δ_2 è l'eccesso della frequenza modale sulla frequenza della classe immediatamente superiore;
- c è l'ampiezza della classe modale.

Esempi di calcolo della moda

Data la seguente distribuzione:

13, 13, 15, 16, 16, 17, 18, 18, 19, 22, 23, 23, 23, 26, 26, 45

x_i	n_i
13	2
15	1
16	2
17	1
18	2
19	1
22	1
23	**3**
26	2
45	1
	16

In questo caso la moda è Mo = 23, in quanto il dato 23 si presenta 3 volte. Ovvero ha la frequenza più alta di tutti gli altri.

Data la seguente distribuzione di frequenza per classi di età:

classi di età	iscritti
34 ⊢⊣ 37	4
38 ⊢⊣ 41	9
42 ⊢⊣ 45	19
46 ⊢⊣ 49	**29**
50 ⊢⊣ 53	17
54 ⊢⊣ 57	10
58 ⊢⊣ 61	7
Totale	95

La **classe modale** è la quinta in quanto ad essa corrisponde la frequenza massima (29 iscritti). La moda è invece data da:

- L_{mo} = 46
- Δ_1 = 10
- Δ_2 = 12
- c = 4

$$Mo = L_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} c = 46 + \frac{10}{10 + 12} 4 = 47{,}82$$

Data la seguente distribuzione di frequenza per classi di età:

Classi addetti	Aziende
50 ⊢ 60	8
60 ⊢ 70	10
70 ⊢ 80	**16**
80 ⊢ 90	14
90 ⊢ 100	10
100 ⊢ 110	5
110 ⊢ 120	2
Totale	65

La classe modale in questo caso è la terza in quanto ha associata la frequenza massima (16 addetti). La moda è invece data da:

- L_{mo} = 70
- Δ_1 = 6
- Δ_2 = 2
- c = 10

$$Mo = L_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} c = 70 + \frac{6}{6 + 2} 10 = 77{,}50$$