Esercitazioni di statistica

Gli indici statistici di sintesi: Gli indici di centralità

Stefania Spina

Universitá di Napoli Federico II

stefania.spina@unina.it

7 Ottobre 2014

Introduzione

Per poter analizzare un fenomeno, del quale possediamo una o più **distribuzioni di frequenza** e che sono state rilevate in tempi, luoghi o circostanze diverse, è opportuno individuare delle **misure sintetiche** che ordinano la diversità tra le distribuzioni.

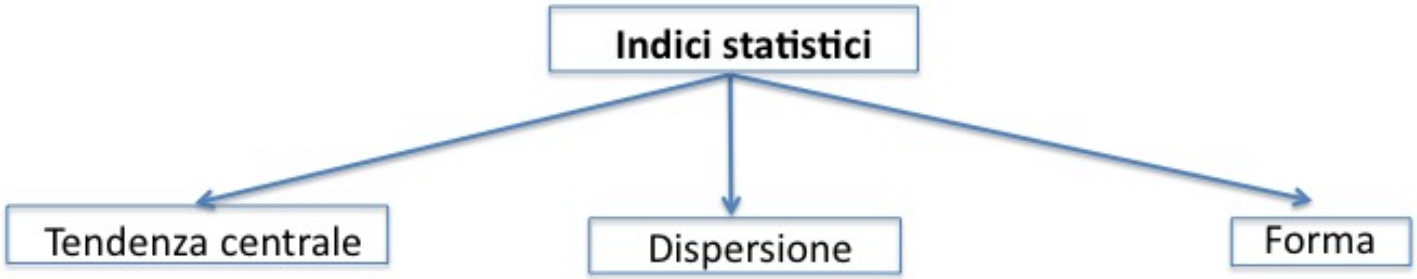
Per fare ciò è necessario esplicitare quali **aspetti di una distribuzione di frequenza** si vogliono esaminare e successivamente individuare le misure più idonee.

Introduzione

Gli aspetti più importanti di una distribuzione di frequenza riguardano:

- **la posizione**, cioè la misura della sua centralità complessiva in rapporto alle modalità e alle rispettive frequenze. La sintesi dovrà essere un valore rappresentativo della variabile nella sua globalità, espresso nella stessa unità di misura del fenomeno e capace di sostituire in qualche modo tutte le osservazioni;
- **la variabilità**, cioè la mutevolezza dei dati della popolazione, ossia la capacità della variabile ad assumere modalità molto diverse tra loro;
- **la forma**, ossia l'aspetto generale della distribuzione di frequenza.

Gli indici di sintesi



- Media
- Mediana
- Moda
- Quartili
- Percentili

- Campo di variazione
- Scarto medio assoluto
- Varianza
- Deviazione standard
- Coefficiente di variazione

- Coefficiente di asimmetria
- Coefficiente di curtosi

Gli indici centralità o di posizione

La tendenza centrale o posizione di un insieme di dati indica dove, numericamente, i dati sono posizionati o concentrati.

Livello di misurazione della variabile	**Moda**	**Mediana /Quartili**	**Media**
Variabili qualitative nominali	■	❑	❑
Variabili qualitative ordinali	■	■	❑
Variabili quantitative discrete	■	■	■
Variabili quantitative continue	■	■	■

Indici di centralità: La moda

La **moda** è la **modalità** con la **frequenza più alta**.
Si può calcolare per tutti i tipi di dati.

Telespettatori (in migliaia) delle emittenti televisive

	Telespettatori
Raiuno	7873
Raidue	2377
Raitre	2664
Canale 5	7665
Rete 4	2007
Italia 1	3162
La 7	910
Altre emittenti	1857

Esempio sulla moda

	Telespettatori
Raiuno	7873
Raidue	2377
Raitre	2664
Canale 5	7665
Rete 4	2007
Italia 1	3162
La 7	910
Altre emittenti	1857

→ MODA

Esercizio sulla moda

La seguente tabella riporta il numero di componenti per famiglie, individuare la moda

	Frequenza
1	153
2	225
3	335
4	564
5	346
6	133
7	75
8	49
Totale	**1880**

Esercizio sulla moda

	Frequenza
1	153
2	225
3	335
4	564
5	346
6	133
7	75
8	49
Totale	1880

MODA

Esercizio sulla moda per dati raggruppati in classi

La tabella seguente riporta la distribuzione di 100 fumatori per classi di età, determinare l'età modale.

	Frequenza
30 - 33	2
34 - 37	3
38 - 41	9
42 - 45	19
46 - 49	29
50 - 53	17
54 - 57	10
58 - 61	7
62 - 65	4
Totale	**100**

Esercizio sulla moda per dati raggruppati in classi

Per i dati raggruppati in classi è necessaria la distinzione che per:

- le classi di modalità che hanno **uguale ampiezza**, la moda cade in quella con maggiore frequenza;

- le classi di modalità che hanno **diversa ampiezza**, la moda cade nella classe con maggiore densità di frequenza che si calcola con la seguente formula:

$$d_i = \frac{n_i}{x_i - x_{(i-1)}}$$

Esercizio sulla moda per dati raggruppati in classi

	Frequenza
30 - 33	2
34 - 37	3
38 - 41	9
42 - 45	19
46 - 49	29
50 - 53	17
54 - 57	10
58 - 61	7
62 - 65	4
Totale	100

Classe modale

Indici di centralità: La mediana

La **mediana** è il valore dell'osservazione centrale di una **distribuzione ordinata** di dati ed è quel valore che si lascia a **destra e a sinistra** un **numero uguale di dati**.

La **mediana** può essere calcolata su caratteri **quantitativi e qualitativi ordinali**, ma non su qualitativi semplici.

Il calcolo della posizione della mediana varia a seconda della numerosità del collettivo esaminato.

Mediana per n dispari

Se la **numerosità del collettivo** (n) è **dispari**, la mediana è il valore o la modalità che occupa la posizione $(\mathbf{n}+\mathbf{1})/\mathbf{2}$, ovvero la mediana è pari a:

$$Me = x_{\frac{n+1}{2}}$$

Supponiamo di aver rilevato, su di un campione di 9 individui, il numero di scarpe:

$$X = \{38, 40, 41, 36, 38, 45, 43, 44, 42\}$$

La prima operazione da compiere è **ordinare il carattere**

$$X = \{36, 38, 38, 40, 41, 42, 43, 44, 45\}$$

La mediana è quel valore che si trova nella posizione $(9+1)/2 = 5$, quindi, il valore **41** che si trova nella **posizione 5**.

Mediana per n pari

Se la **numerosità del collettivo** (n) è **pari**, la mediana è il valore o la modalità che occupa la posizione $(\mathbf{n}/\mathbf{2}) + \mathbf{1}$, ovvero la mediana è pari a:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

Supponiamo di aver rilevato, su di un campione di 10 individui, il numero di scarpe:

$$X = \{38, 40, 41, 36, 38, 45, 43, 44, 42, 36\}$$

La prima operazione da compiere è **ordinare il carattere**

$$X = \{36, 36, 38, 38, 40, 41, 42, 43, 44, 45\}$$

La mediana è quel valore che si trova tra la posizione 5 (con valore pari a 40) e 6 (con valore pari a 41), quindi, la mediana si individua calcolando la semisomma dei valori che si trovano in tali posizioni $(\mathbf{40} + \mathbf{41})/\mathbf{2} = \mathbf{40}, \mathbf{5}$

Mediana per dati raggruppati in classi

Nella seguente tabella sono raccolti i dati relativi al peso dei giocatori di una rosa di una squadra di calcio (n=23).

Peso in *Kg*	n_i	N_i
60 - 65	1	1
65 - 70	2	3
70 - 75	4	7
75 - 80	6	13
80 - 85	6	19
85 - 90	2	21
90 - 95	1	22
95 - 100	1	23

Calcolare la mediana della distribuzione data. Nota: gli intervalli di frequenza si intendono del tipo “primo valore incluso – secondo valore escluso”.

Mediana per dati raggruppati in classi

- Per prima cosa, identifichiamo la **classe mediana**.
Pos $Me = \frac{n+1}{2} = \frac{23+1}{2} = 12 \rightarrow 4$
Quindi la classe che contiene la **mediana è la quarta**.

- Si applica la formula per il calcolo della mediana:

$$Me = L_{inf} + \left[\frac{(n/2) - N_{inf}}{n_{mediana}}\right] c$$

dove

- L_{inf} è il limite inferiore della classe mediana
- $n_{mediana}$ è la frequenza della classe mediana
- c è l'ampiezza della classe mediana
- n è la numerosità dei casi
- N_{inf} è la frequenza cumulata fino al limite inferiore della classe.

Mediana per dati raggruppati in classi

Nello specifico

- L_{inf} è il limite inferiore della classe mediana, cioè 75;
- $n_{mediana}$ è la frequenza della classe mediana, cioè 6;
- c è l'ampiezza della classe mediana, cioè 5;
- n è la numerosità dei casi, cioè 23;
- N_{inf} è la frequenza cumulata fino al limite inferiore della classe, cioè 7.

$$Me = 75 + \left[\frac{(\frac{23}{2}) - 7}{6}\right] * 5 = 78.75$$

Si può quindi affermare che il 50% dei calciatori della squadra pesa meno di 78.75 Kg.

Indici di centralità: I percentili

Il **p -esimo percentile** di un insieme di dati è il valore per cui una percentuale pari a p delle osservazioni è inferiore o uguale a esso.

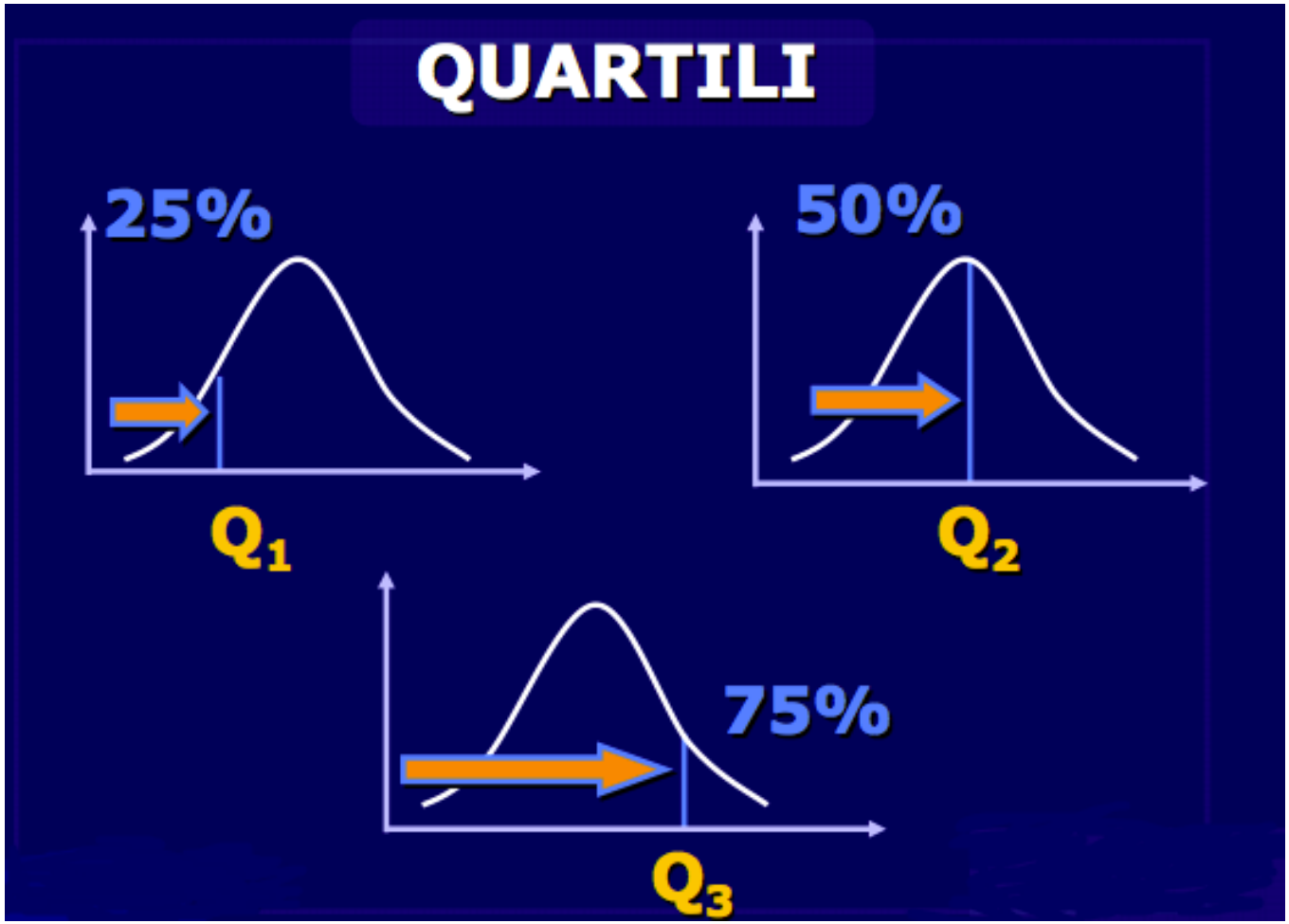
Percentili speciali

25-esimo percentile: quartile inferiore Q_i. Valore tale che il 25% dei dati è inferiore o uguale a esso.

50-esimo percentile: coincide con la Mediana.

75-esimo percentile: quartile superiore Q_s. Valore tale che il 75% dei dati è inferiore o uguale a esso.

Indici di centralità: I quartili



Un esempio sui quartili

Studio che esamina i tempi di attesa al ristorante in un campione di 10 clienti

Dati ordinati

58.6 59 59.3 59.4 62.7 62.8 63.7 65.4 67.3 68.1

Si calcoli la mediana, il primo quartile e il terzo quartile.

Un esempio sui quartili

Formule per il calcolo della posizione dei **quartili**

- Per il 1° quartile Pos $Q_1 = \frac{\frac{N}{4} + \left(\frac{N}{4} + 1\right)}{2}$
- Per il 2° quartile Pos $Q_2 = \frac{\frac{N}{2} + \left(\frac{N}{2} + 1\right)}{2}$
- Per il 3° quartile Pos $Q_3 = \frac{\left(\frac{N}{4} * 3\right) + \left(\frac{N}{4} * 3\right) + 1}{2}$

Un esempio sui quartili

Calcolo della mediana

MEDIANA

58.6 59 59.3 59.4 **62.7** 62.8 63.7 65.4 67.3 68.1

Calcolo del 1° Quartile

1° QUARTILE

58.6 59 **59.3** 59.4 62.7 62.8 63.7 65.4 67.3 68.1

Calcolo del 3° Quartile

3° QUARTILE

58.6 59 59.3 59.4 62.7 62.8 63.7 **65.4** 67.3 68.1

Esercizio

Nella tabella seguente è riportata la distribuzione di frequenze dei tassi di criminalità per i 50 stati degli Stati Uniti registrati in ciascuno di essi nel 2005 (il tasso di criminalità misura il numero di crimini registrati in un dato Stato ogni 10000 abitanti residenti nello stesso nel 2005).

Tasso di criminalità	n
0 – 11	3
12 - 23	3
24 – 35	18
36 – 47	11
48 – 59	7
60 – 71	6
72 - 83	2
Totale	50

Calcolare i quartili, il primo decile, il novantesimo percentile

Esercizio

Tasso di criminalità	n	N
0 – 11	3	3
12 - 23	3	6
24 – 35	18	24
36 – 47	11	35
48 – 59	7	42
60 – 71	6	48
72 - 83	2	50
Totale	50	

$$Pos(Q_1) = \frac{N}{4} = \frac{50}{4} = 12{,}5$$

$$Pos(Q_2) = \frac{N}{2} = \frac{50}{2} = 25$$

$$Pos(Q_3) = \frac{3N}{4} = \frac{3*50}{4} = 37{,}5$$

$$Pos(D_1) = \frac{N}{10} = \frac{50}{10} = 5$$

$$Pos(P_{90}) = \frac{90N}{100} = \frac{90*50}{100} = 45$$

Esercizio

$$Q_1 = L_{\text{inf}} + \frac{N/4 - N_{Cl_{(prec)}}}{n_{Cl_{(Q)}}} * c = 24 + \frac{12{,}5 - 6}{18} * 12 = 28$$

L_{inf} = *limite inferiore della classe in cui cade il primo quartile*

$\frac{N}{4}$ = *Frequenza cumulata corrispondente alla classe del primo quartile*

$N_{Classe.prec}$ = *frequenza cumulata della classe precedente*

$n_{Cl.Q}$ = *frequenza della classe in cui cade il primo quartile*

c = *ampiezza della classe*

Esercizio

Tasso di criminalità	n	N
0 – 11	3	3
12 - 23	3	6
24 – 35	18	24
36 – 47	11	35
48 – 59	7	42
60 – 71	6	48
72 - 83	2	50
Totale	50	

$$Pos(Q_1) = \frac{N}{4} = \frac{50}{4} = 12{,}5 \qquad Q_1 = L_{inf} + \frac{N/4 - N_{Cl_{(prec)}}}{n_{Cl_{(Q)}}} * c = 24 + \frac{12{,}5 - 6}{18} * 12 = 28$$

$$Pos(Q_2) = \frac{N}{2} = \frac{50}{2} = 25 \qquad Q_2 = L_{inf} + \frac{N/2 - N_{Cl_{(prec)}}}{n_{Cl_{(Q)}}} * c = 36 + \frac{25 - 24}{11} * 12 = 37$$

$$Pos(Q_3) = \frac{3N}{4} = \frac{3*50}{4} = 37{,}5 \qquad Q_1 = L_{inf} + \frac{3N/4 - N_{Cl_{(prec)}}}{n_{Cl_{(Q)}}} * c = 48 + \frac{37{,}5 - 35}{7} * 12 = 52$$

$$Pos(D_1) = \frac{N}{10} = \frac{50}{10} = 5 \qquad D_1 = L_{inf} + \frac{N/10 - N_{Cl_{(prec)}}}{n_{Cl_{(D)}}} * c = 12 + \frac{5 - 3}{3} * 12 = 20$$

$$Pos(P_{90}) = \frac{90N}{100} = \frac{90*50}{100} = 45 \qquad P_{90} = L_{inf} + \frac{90N/100 - N_{Cl_{(prec)}}}{n_{Cl_{(P)}}} * c = 60 + \frac{45 - 42}{6} * 12 = 66$$

Approssimativamente un quarto degli stati ha tassi di criminalità al di sotto di 28, un quarto tra 28 e 37, un quarto tra 37 e 52, un quarto oltre 52.

La media aritmetica

La media aritmetica ($\bar{x}$) di un insieme di k valori $x_1, x_2, \ldots, x_k$ di un carattere **quantitativo** X è pari alla somma dei valori divisa per il loro numero.
In simboli, la media è:

$$\bar{x} = \frac{1}{N}(x_1 + x_2, \ldots + x_k) = \frac{1}{N}\sum_{i=1}^{k} x_i$$

Ovviamente, se si conoscono i dati mediante una distribuzione di frequenza, la precedente formula è modificata per tenere conto dei raggruppamenti delle modalità nel modo seguente:

$$\bar{x} = \frac{1}{N}(x_1 n_1 + x_2 n_2, \ldots + x_k n_k) = \frac{1}{N}\sum_{i=1}^{k} x_i n_i$$

Esempio sulla media

Supponiamo di aver rilevato il peso (in kg) di cinque studenti:

	Peso in kg
Maria	50
Simonetta	65
Giovanni	78
Franco	70
Carlo	90

Esempio sulla media

Supponiamo di aver rilevato il peso (in kg) di cinque studenti:

	Peso in kg
Maria	50
Simonetta	65
Giovanni	78
Franco	70
Carlo	90

La media è : $\bar{x} = \frac{1}{5}(50 + 65 + 78 + 70 + 90) = \frac{353}{5} = 70.6$

La media ponderata

Nel caso di **distribuzioni di frequenza**, calcoleremo la *media ponderata* che è data dalla somma dei prodotti delle singole modalità $(x_j, per\ j = 1, \ldots, k)$ e le rispettive frequenze $(n_j, per\ j = 1, \ldots, k)$

$$\bar{x} = \frac{1}{n}(x_1 n_1 + x_2 n_2, \ldots + x_k n_k) = \frac{1}{n}\sum_{j=1}^{k} x_j n_j$$

Se si utilizzano le frequenze relative $(f_j, per\ j = 1, \ldots, k)$ l'espressione utilizzata per il calcolo della media è la seguente:

$$\bar{x} = \sum_{j=1}^{k} x_j f_j$$

Esercizio sulla media ponderata

Nella tabella seguente è riportata la distribuzione delle famiglie per numero di componenti in un dato comune, calcolare il numero medio di componenti:

	Frequenze
1	153
2	225
3	335
4	564
5	346
6	133
7	75
8	49
Totale	**1880**

Esercizio sulla media ponderata

	n_i	$x_i n_i$
1	153	153
2	225	450
3	335	1005
4	564	2256
5	346	1730
6	133	798
7	75	525
8	49	392
Totale	**1880**	**7309**

La media ponderata è : $\bar{x} = \frac{7309}{1880} = 3.88$

Esercizio sulla media ponderata

Il corso di Matematica è frequentato da 95 studenti, con un'età compresa tra i 19 e i 24 anni.

	Frequenza
19	10
20	20
21	35
22	20
23	5
24	5

Esercizio sulla media ponderata

Il corso di Matematica è frequentato da 95 studenti, con un'età compresa tra i 19 e i 24 anni.

	Frequenza
19	10
20	20
21	35
22	20
23	5
24	5

La media del carattere età è:

$$\bar{x} = \frac{1}{95}(19x10) + (20x20) + (21x35) + (22x20) + (23x5) + (24x5) = \frac{2000}{95} = 21,05$$

La classe di matematica è composta da studenti con un'età media di circa 21 anni.

La media ponderata per dati raggruppati in classi

Se il carattere quantitativo X è suddiviso in k classi, si può approssimare la media aritmetica con la seguente espressione:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} c_j n_j$$

oppure

$$\bar{x} = \sum_{j=1}^{k} c_j f_j$$

dove

- k è il numero delle classi nella distribuzione di frequenza;
- c_j è il valore centrale della classe j-esima;
- n_j è la corrispondente frequenza assoluta;
- f_j è la corrispondente frequenza relativa.

Esempio per la media ponderata per dati raggruppati in classi

	Frequenza assoluta
19 - 20	30
21 - 22	55
23 - 24	10
Totale	95

Esempio per la media ponderata per dati raggruppati in classi

	Frequenza assoluta
19 - 20	30
21 - 22	55
23 - 24	10
Totale	95

- **Valore centrale nella classe i-*esima***

$$c_i = \frac{(x_{i-1} + x_i)}{2} \text{ oppure } c_i = x_{i-1} + \frac{(x_i - x_{i-1})}{2}$$

Esempio per la media ponderata per dati raggruppati in classi

In questo caso bisogna individuare il **valore centrale della classe**, ossia occorre calcolare il punto medio di ogni classe.

	Frequenza assoluta	**Frequenza relativa**	**Valore centrale della classe**	$c_j * f_j$
19 - 20	30	0.32	19.5	6.24
21 - 22	55	0.58	21.5	12.47
23 - 24	10	0.11	23.5	2.59
Totale	95	1		21.30

$$\bar{x} = \frac{1}{95}\left[(19,5x30) + (21,5x55) + (23,5x10)\right] = \frac{2003}{95} = 21,01$$

oppure

$$\bar{x} = \left[(19,5x0.32) + (21,5x0.58) + (23,5x0.11)\right] = 21,01$$

L'età media del campione esaminato è 21 anni.