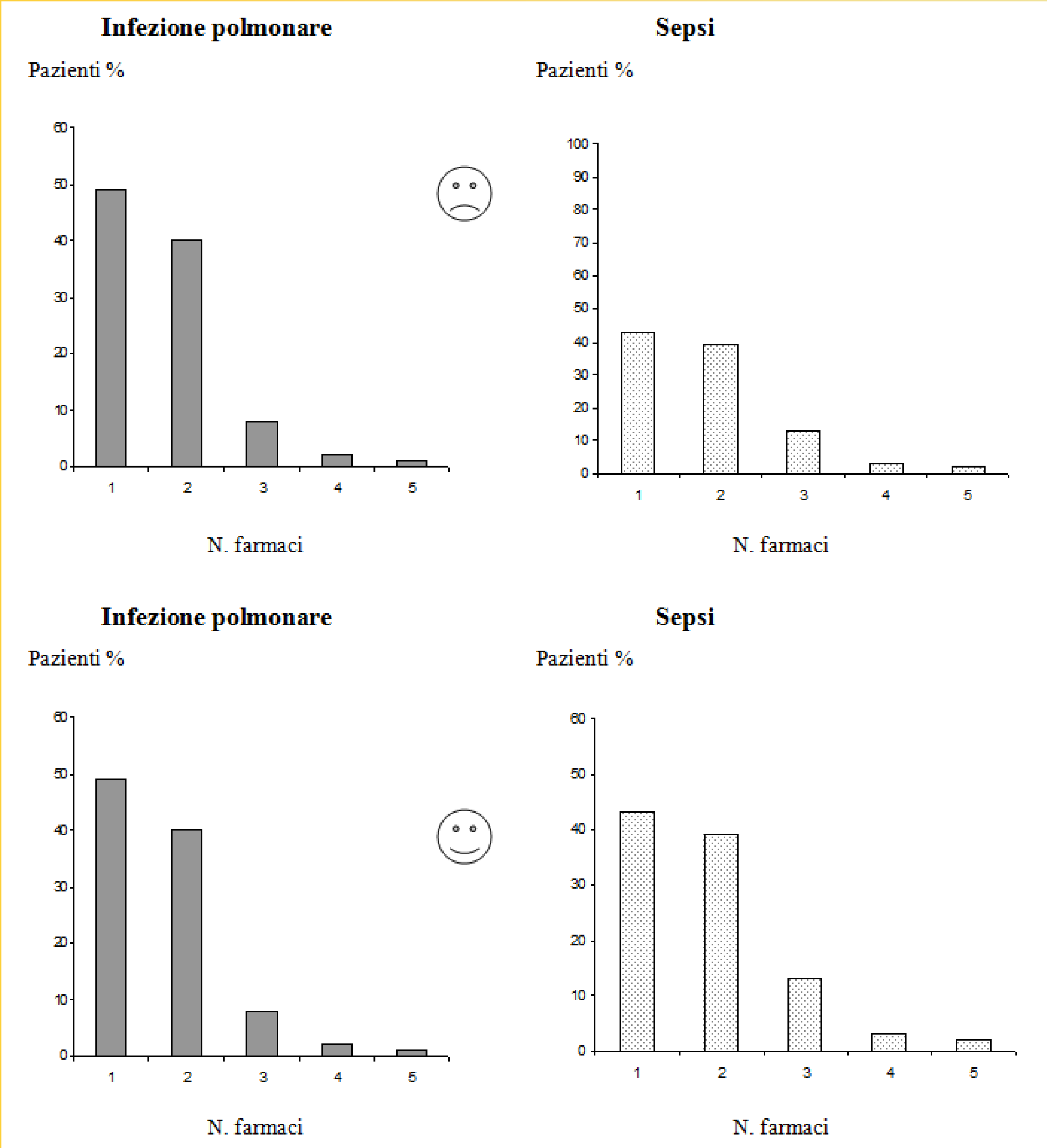
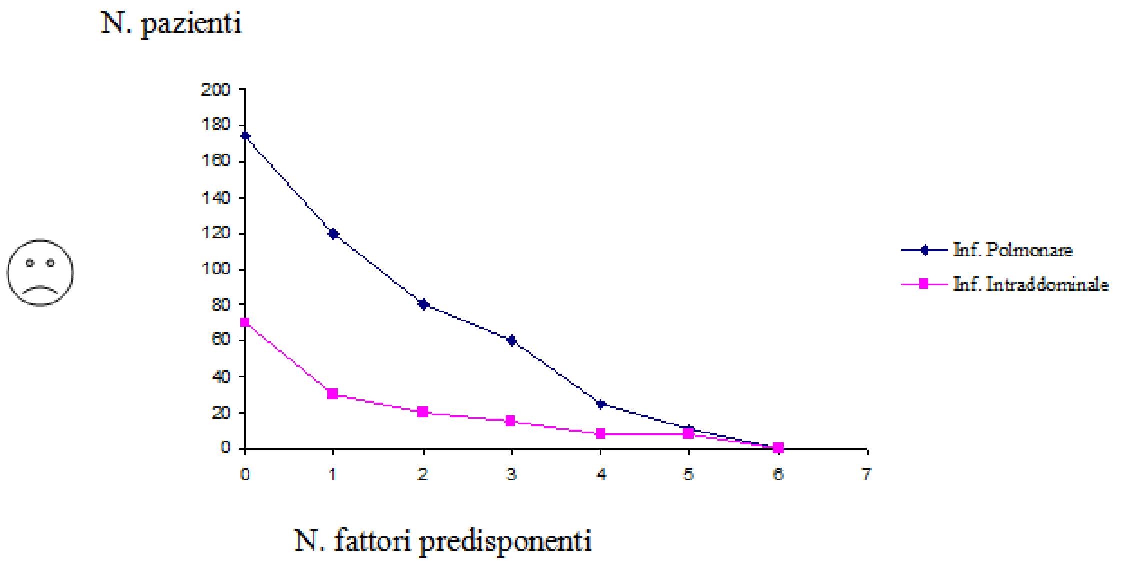
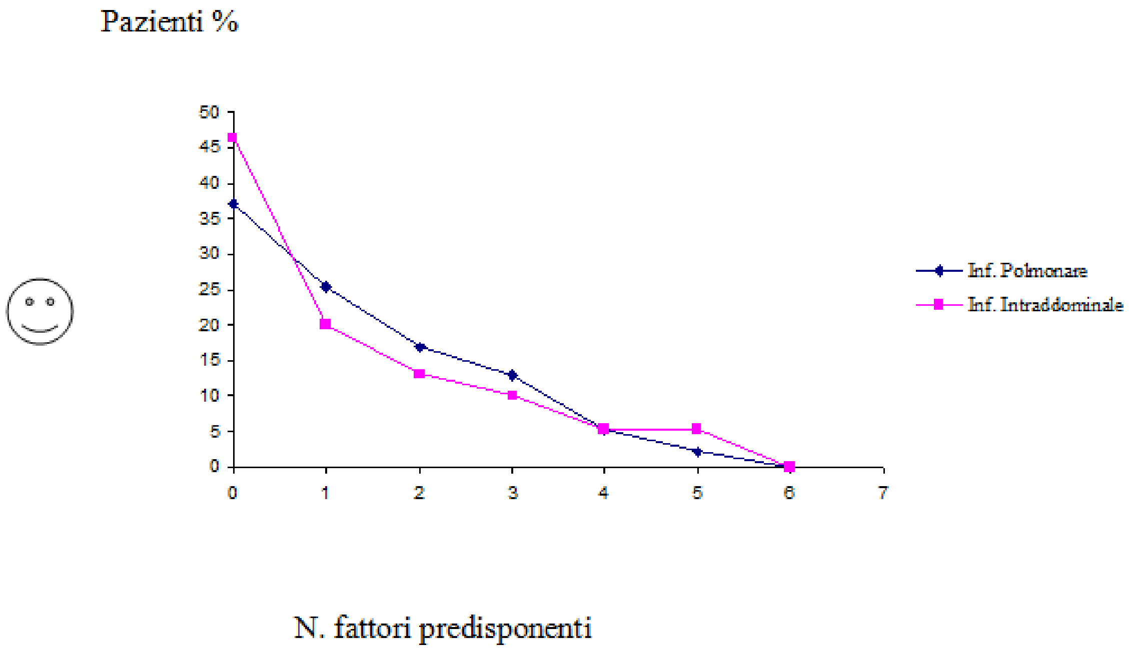
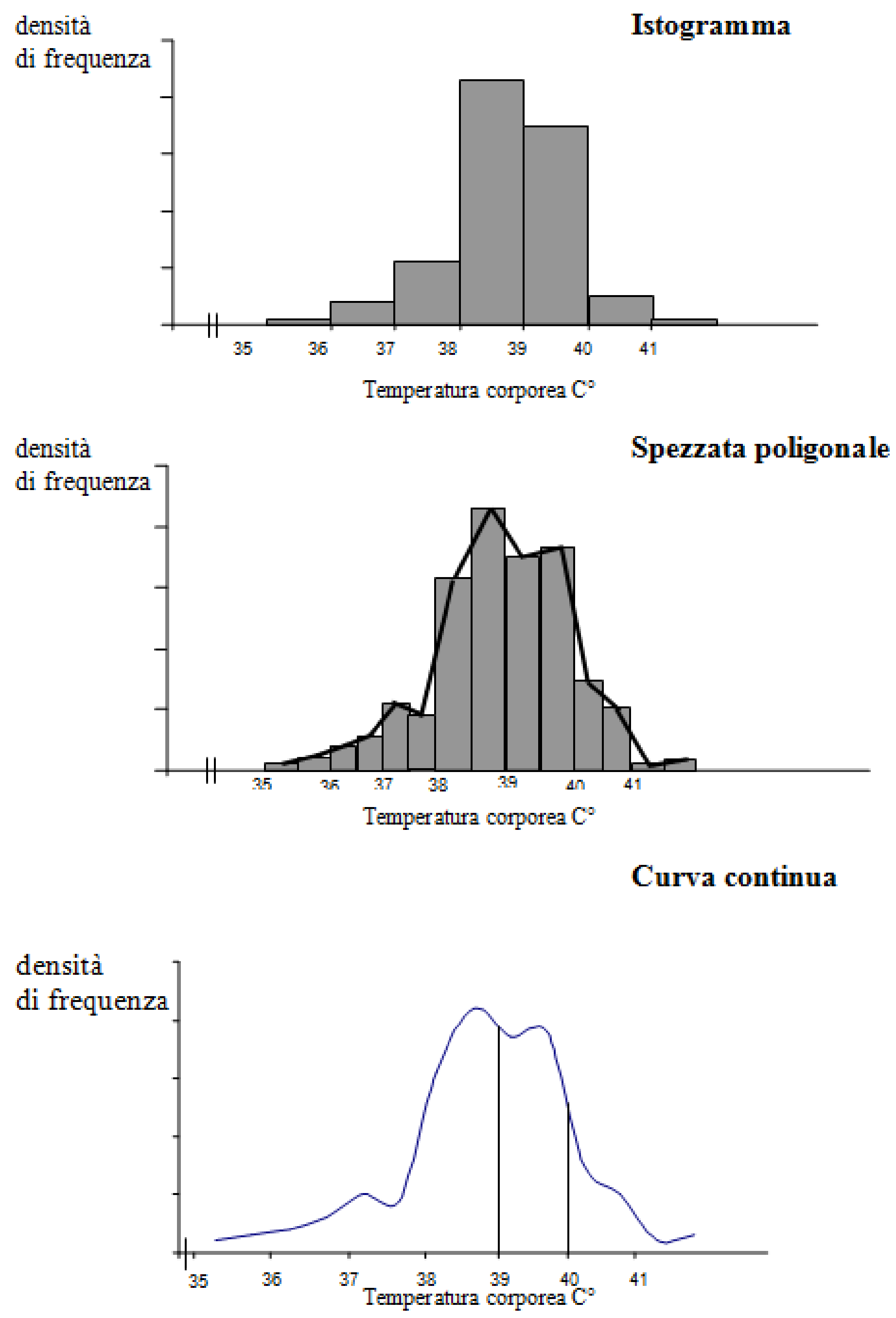
Esempi di confronti grafici

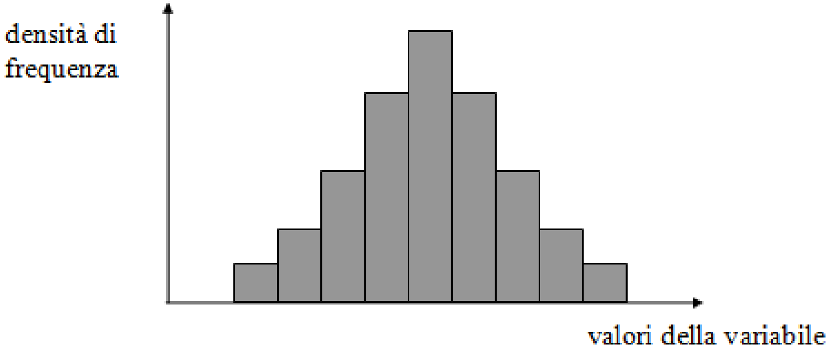


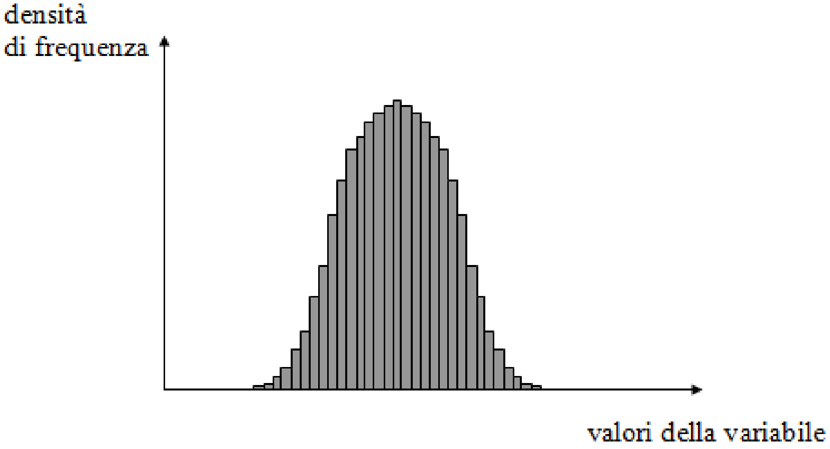
Esempi di confronti grafici

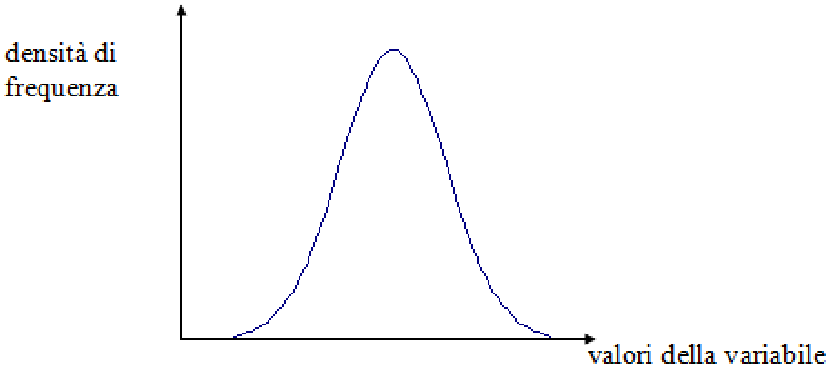





Istogramma
densità
di frequenza
35
36
37
38
39
40
41
Temperatura corporea C°
Spezzata poligonale
densità
di frequenza
35
36
37
38
39
40
41
Temperatura corporea C°
Curva continua
densità
di frequenza
35
36
37
38
39
40
41
Temperatura corporea C°

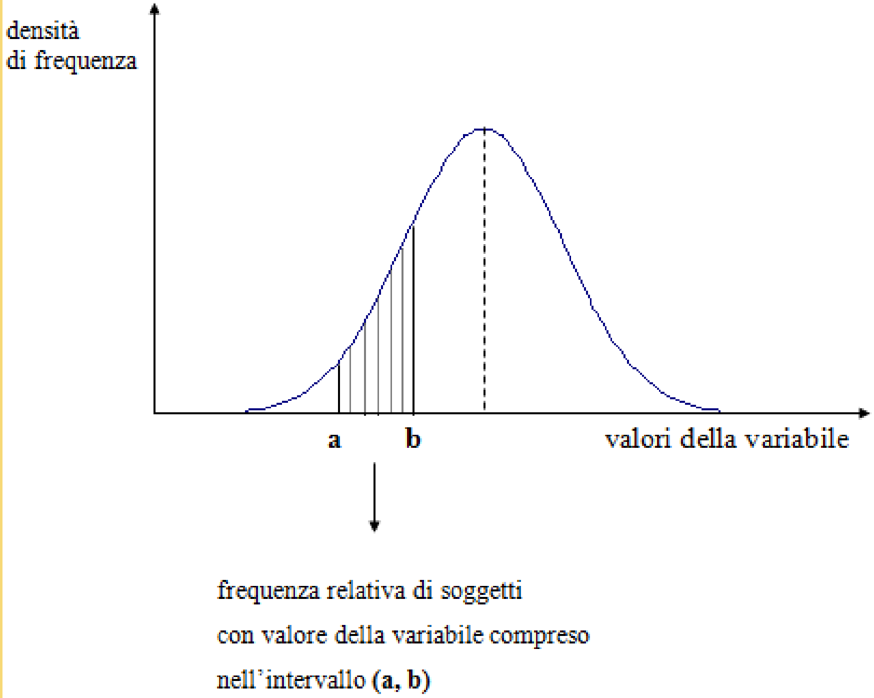




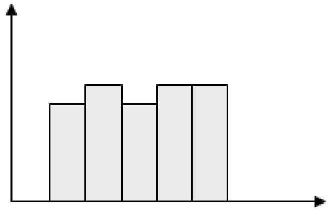


Distribuzione teorica normale

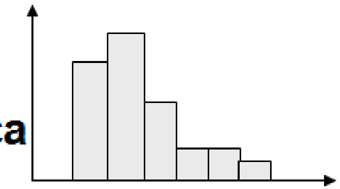
La normale è un modello teorico di distribuzione dei valori di una variabile. Ha un profilo campanulare e simmetrico:



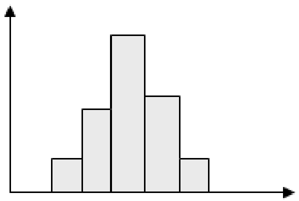
Distribuzione uniforme



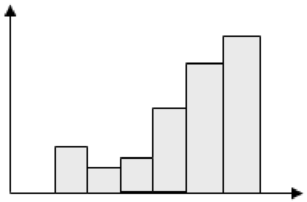
“ “ asimmetrica



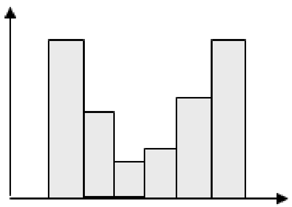
“ “ simmetrica



“ “ a forma di J



“ “ a forma di U



Le misure descrittive per dati univariati

Sintetizziamo ulteriormente i dati raccolti su ciascuna variabile mediante il calcolo di un numero ridotto di quantità in grado di riassumere alcuni aspetti peculiari della distribuzione delle variabili rilevate.

Possono riferirsi in particolare ai seguenti aspetti:

- dimensione media della variabile
- variabilità dei valori osservati
- forma della distribuzione dei valori osservati

Possono essere ricondotte a tre classi principali:

- medie
- misure di variabilità
- indicatori della forma distributiva

Tali misure costituiscono un SISTEMA di indicatori in grado di qualificare l'insieme dei dati.

Capitolo 3

LE MEDIE

Le medie sono quantità che descrivono le seguenti proprietà di una variabile:

- la dimensione media
- la centralità
- la configurazione tipica

- La media aritmetica
- La media geometrica
- La trimmed mean
- La mediana
- La moda
- I percentili

Introduzione

Medie di posizione → *non richiedono operazioni algebriche sulle modalità*

Medie analitiche → *calcolate con operazioni algebriche sulle modalità, richiedono variabili quantitative*

La Media Aritmetica

Tempo impiegato per raggiungere il posto di lavoro

tempo impiegato (min.)			tempo impiegato (min.)		
giorno	auto	metro	giorno	auto	metro
1	23	22	**7**	28	24
2	32	24	**8**	33	28
3	44	22	**9**	45	32
4	21	33	**10**	34	31
5	36	26	**11**	29	37
6	30	31	**12**	31	24

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

$$\bar{x}_a(metro) = \frac{(22+24+22+33+26+31+24+28+32+31+37+24)}{12} = \frac{334}{12} = 27{,}83$$

Media aritmetica

La media aritmetica di un insieme di n valori $x_1, x_2, \ldots x_n$ di una variabile quantitativa X è data da:

$$\bar{x}_a = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Se la variabile X è quantitativa discreta e conosciamo la sua distribuzione di frequenza:

$$\bar{x}_a = \frac{1}{n}\sum_{j=1}^{K} x_j n_j \qquad \bar{x}_a = \sum_{j=1}^{K} x_j f_j$$

Esempio (variabile non suddivisa in classi):

E' stato rilevato il numero di antibiotici prescritti a 82 pazienti. Dai dati raccolti è stata costruita la seguente distribuzione di frequenza:

N. antibiotici	N. pazienti	%
0	15	13.8
1	40	36.7
2	37	33.9
3	17	15.6
Totale	**109**	**100**

Calcoliamo il consumo medio di farmaci:

$$\textbf{\textit{media aritmetica}} = \frac{(0\times 15)+(1\times 40)+(2\times 37)+(3\times 17)}{109}$$
$$= 1.5\ \textit{farmaci}$$

<u>Alternativamente</u>, utilizzando le frequenze relative percentuali:

$$\textbf{\textit{media aritmetica}} = [(0\times 13.8)+(1\times 36.7)+(2\times 33.9)+(3\times 15.6)]\ /\ 100$$
$$= 1.5\ \textit{farmaci}$$

Esempio:

Tabella 3 - Famiglie per numero di componenti e classe di età della persona di riferimento della famiglia. Italia - Anno 2001

CLASSI DI ETÀ	Componenti						
	1 persona	2 persone	3 persone	4 persone	5 persone	6 o più persone	Totale
Fino a 25 anni	109.056	52.884	32.296	8.786	2.010	714	205.746
25-34	629.651	713.703	629.988	341.563	61.638	13.633	2.390.176
35-44	631.802	615.667	1.029.719	1.382.206	369.819	83.575	4.112.788
45-54	528.504	569.001	1.007.634	1.346.055	465.740	125.829	4.042.763
55-64	679.247	1.119.754	1.077.170	724.642	234.117	77.271	3.912.201
65-74	1.146.219	1.605.949	665.285	251.448	91.950	45.607	3.806.458
75-84	1.191.231	987.616	218.503	66.380	32.753	18.610	2.515.093
85 e oltre	511.911	240.837	45.611	15.126	7.799	4.167	825.451
Totale	**5.427.621**	**5.905.411**	**4.706.206**	**4.136.206**	**1.265.826**	**369.406**	**21.810.676**

Fonte: Istat. 14º Censimento generale della popolazione e delle abitazioni

Tabella 18 - Calcolo del numero di componenti delle famiglie residenti. Italia - Anno 2001

NUMERO DI COMPONENTI	Famiglie (frequenze assolute)	Numero di componenti*numero di famiglie
1 componente	5.427.621	1*5.427.621= 5.427.621
2 componenti	5.905.411	2*5.905.411=11.810.822
3 componenti	4.706.206	3*4.706.206=14.118.618
4 componenti	4.136.206	4*4.136.206=16.544.824
5 componenti	1.265.826	5*1.265.826=6.329.130
6 componenti e oltre	369.406	7*369.406=2.585.842
Totale	**21.810.676**	Somma=56.816.857

Fonte: Elaborazione su dati Istat. 14º Censimento della popolazione e delle abitazioni

numero medio di componenti per famiglia: 56.816.857/21.810.676=2,6.

http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_1.htm

Valore centrale della classe

Nel caso di una distribuzione di frequenze per una variabile X suddivisa in classi, possiamo approssimare la media utilizzando il valore centrale della classe c_j

$$\bar{x}_a \cong \frac{1}{n}\sum_{j=1}^{K} c_j n_j$$

Esempio

Prezzi di farmaci e quantità acquistate da un ospedale

Valore centrale	Prezzo a confezione (€)	Numero Confezioni *(migliaia)*	Ammontare variabile (costo) ml. (€)
25	20 ⊢ 30	11	25*11= 275.0
32.5	30 ⊢ 35	5	32.5*5= 162.5
37.5	35 ⊢ 40	15	37.5*15= 562.5
45	40 ⊢ 50	9	45*9 = 405.0
	Totale	40	1405

$\bar{x}_a$ = **1405/40 = 35.72 €** (a confezione) *(approssimato)*

Esempio:

Tabella 3 - Famiglie per numero di componenti e classe di età della persona di riferimento della famiglia. Italia - Anno 2001

CLASSI DI ETÀ	Componenti						
	1 persona	2 persone	3 persone	4 persone	5 persone	6 o più persone	Totale
Fino a 25 anni	109.056	52.884	32.296	8.786	2.010	714	205.746
25-34	629.651	713.703	629.988	341.563	61.638	13.633	2.390.176
35-44	631.802	615.667	1.029.719	1.382.206	369.819	83.575	4.112.788
45-54	528.504	569.001	1.007.634	1.346.055	465.740	125.829	4.042.763
55-64	679.247	1.119.754	1.077.170	724.642	234.117	77.271	3.912.201
65-74	1.146.219	1.605.949	665.285	251.448	91.950	45.607	3.806.458
75-84	1.191.231	987.616	218.503	66.380	32.753	18.610	2.515.093
85 e oltre	511.911	240.837	45.611	15.126	7.799	4.167	825.451
Totale	**5.427.621**	**5.905.411**	**4.706.206**	**4.136.206**	**1.265.826**	**369.406**	**21.810.676**

Fonte: Istat. 14° Censimento generale della popolazione e delle abitazioni

[18,25)

[85,100)

Tabella 19 - Calcolo dell'età media della persona di riferimento della famiglia al Censimento 2001

CLASSI DI ETÀ	Numero di famiglie (1)	Valore centrale classe di età (2)	Valore centrale* numero di famiglie (1)*(2)
Fino a 25 anni	205.746	21,5	4.423.539
25-34	2.390.176	30,0	71.705.280
35-44	4.112.788	40,0	164.511.520
45-54	4.042.763	50,0	202.138.150
55-64	3.912.201	60,0	234.732.060
65-74	3.806.458	70,0	266.452.060
75-84	2.515.093	80,0	201.207.440
85 e oltre	825.451	92,5	76.354.218
Totale	**21.810.676**		**1.221.524.267**

Fonte: Elaborazione su dati Istat. 14° Censimento generale della popolazione e delle abitazioni

Età media della persona di riferimento = 1.221.730.013/21.810.676=56 anni.

http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_1.htm

Considerazioni

- La Media aritmetica ***sintetizza*** la distribuzione di un carattere con un solo valore;

> ***La media aritmetica rappresenta la dimensione o intensità media di una variabile.***

- La Media aritmetica dipende da tutti i valori osservati e quindi risente dei valori estremi (valori anomali);

Proprietà della media aritmetica

1) La somma dei valori osservati è uguale al valore medio moltiplicato per il numero di unità;

2) La somma delle differenze tra i valori e la loro media aritmetica, è pari a zero;

3) La somma degli scarti al quadrato dei valori da una costante *c* è minima quando *c* è uguale alla media aritmetica;

4) Se un collettivo viene suddiviso in *L* sottoinsiemi disgiunti, allora la media aritmetica generale si può ottenere come media ponderata delle medie dei sottoinsiemi con pesi uguali alle loro numerosità.

Il calcolo della media aritmetica combinatamente per più gruppi o campioni

All'interno di un campione di 35 pazienti sono state calcolate le medie aritmetiche separatamente per i maschi (15 pazienti) e per le femmine (20 pazienti) per la variabile **'numero patogeni isolati'**:

	Maschi	**Femmine**
N. medio di patogeni	**2.3**	**1.8**

Il numero medio di patogeni isolati per il totale dei 35 pazienti sarà:

$$\textit{media aritmetica} = \frac{(15 \times 2.3) + (20 \times 1.8)}{35}$$
$$= 2.0 \text{ patogeni}$$

Media aritmetica ponderata

La **media aritmetica ponderata** di un insieme di *n* valori osservati di una variabile quantitativa X con pesi non negativi, è data da:

$$\bar{x}_a = \frac{x_1 p_1 + x_2 p_2 + x_k p_k}{p_1 + p_2 + p_k} = \frac{\sum_{j=1}^{k} x_j p_j}{\sum_{j=1}^{k} p_j}$$

La trimmed mean

La **trimmed mean** è la media aritmetica calcolata su una fissata percentuale di valori centrali di un insieme di dati.

- Elimina l'influenza dei valori anomali

Ad esempio nella trimmed mean al 50% si escludono il 25% dei valori più piccoli e il 25% dei valori più grandi.

Esempio

Con valori della variabile (**3, 5, 5, 6, 8, 8, 9, 150**) la **trimmed mean** al 50% sarà ottenuta escludendo i due valori più piccoli e i due più grandi

La moda

Tipologia di farmaco	Numero reparti	Frequenze %
Antidolorifico	100	25
Antibiotico	200	50
Antiblastico	80	20
Altro	20	5
Totale	400	100

La moda è la modalità prevalente della variabile, a cui corrisponde la massima frequenza

Essendo definita come modalità prevalente, la moda individua la **caratteristica tipica** dell'insieme dei dati.

Consumi ml.(€)	N. reparti	Ampiezza classe	Densità frequenza
5 ⊢ 25	100	20	100/20 = 5
25 ⊢ 35	90	10	90/10 = **9**
35 ⊢ 60	210	25	210/25 = 8.4
Totale	400		

Considerazioni sulla moda

- La moda acquista validità solo se vi è una netta prevalenza di una modalità;
- La moda si calcola su tutti i tipi di variabili;
- La distribuzione di una variabile può avere più mode (es. distribuzioni bi-modali o tendenzialmente bi-modali)

La Mediana

E' la modalità presentata dall'unità centrale del collettivo ordinato in senso crescente o decrescente rispetto ai valori della variabile. Essa divide il collettivo in due sottoinsiemi di uguale numerosità: uno con modalità di ordine più basso rispetto alla mediana e l'altro con modalità di ordine più alto.

La MEDIANA è quel valore della caratteristica rilevata che divide l'insieme dei dati in modo tale che il 50% dei pazienti ha un valore inferiore alla mediana e il 50% ha un valore superiore.

Il calcolo della mediana è possibile solo per variabili quantitative o qualitative ordinabili.

Calcolo della 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
   - se *n* è *pari* si hanno due unità centrali con posizione *n*/2 e *n*/2 +1;
3. Se n è dispari, la mediana è la modalità presentata dall'unità centrale
4. Se *n* è *pari* si hanno due mediane date dalle modalità delle due unità centrali. (se la variabile è quantitativa, possiamo considerare come mediana la semisomma dei valori delle due unità centrali)

Esempio:

E' stato rilevato il numero di farmaci prescritti a 10 pazienti:

4 1 3 2 4 1 6 5 4 3

Per calcolare la mediana ordiniamo i valori in senso crescente:

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

Essendo il numero delle osservazioni pari, vi sono due valori centrali che occupano il 5° (n/2) e il 6° (n/2+1) posto. La mediana è per convenzione la media aritmetica dei due valori: $\boldsymbol{mediana} = \frac{\mathbf{3+4}}{\mathbf{2}} = \mathbf{3.5\,farmaci}$

Esempio

Calcolo della mediana delle stature (in metri) di 5 studenti:
1,72 1,68 1,75 1,80 1,74.

Prima di tutto ordiniamo la statura dei ragazzi in senso crescente:
1,68 1,72 1,74 1,75 1,80.

Il valore mediano è quello presentato dallo studente che si trova al 3° posto di questa distribuzione e che ha altezza pari a 1,74. In modo formale procediamo così:
se n è il numero di unità del collettivo (5 nel nostro esempio), individuiamo il posto occupato nella distribuzione dal valore mediano con:

$$\frac{n+1}{2}$$

nel nostro caso:

(5+1): 2=3

Il valore mediano è quello presentato dalla terza unità del collettivo allorchè ordinato, cioè 1,74.

http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_1.htm

Esempio

Calcolo del tempo mediano realizzato da 6 atleti nei 200 m
I tempi (in secondi) ordinati dei 6 atleti sono:
24,7 25,1 25,2 25,6 25,7 26,1.

Applicando la formula $\frac{n+1}{2}$ con n=6 si ha:

(6+1):2=3,5, che può leggersi come il valore compreso tra il 3° e il 4° posto.
Il valore mediano si situa pertanto tra 25,2 e 25,6, approssimabile con:
(25,2 + 25,6):2=25,4.

http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_1.htm

Calcolo della mediana per una distribuzione di frequenza

Per il calcolo della mediana per una distribuzione di frequenza, è necessario passare alla distribuzione delle *frequenze cumulate*.

La mediana è la prima modalità o classe a cui corrisponde una frequenza percentuale cumulativa maggiore o uguale al 50%.

Esempio:

Si consideri la seguente distribuzione di frequenza:

Età (in anni)	**N. pazienti**	**%**	**% cumulate**
6-14	15	25.4	25.4
15-19	24	40.7	66.1
20-24	13	22.0	88.1
≥ 25	7	11.9	100
Totale	**59**	**100**	

Classe di età mediana: **15-19 anni**

Considerazioni

- La mediana è meno sensibile alla presenza di valori anomali rispetto alla media aritmetica;

Esempio

Consideriamo i seguenti punteggi realizzati da tre giocatori in 11 partite:
Giocatore a: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3
Media = 22 ÷ 11 = 2
Mediana = 2
Giocatore b: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4
Media = 23 ÷ 11 = 2.1
Mediana = 2
Giocatore c: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 14
Media = 33 ÷ 11 = 3
Mediana = 2.

http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_1.htm

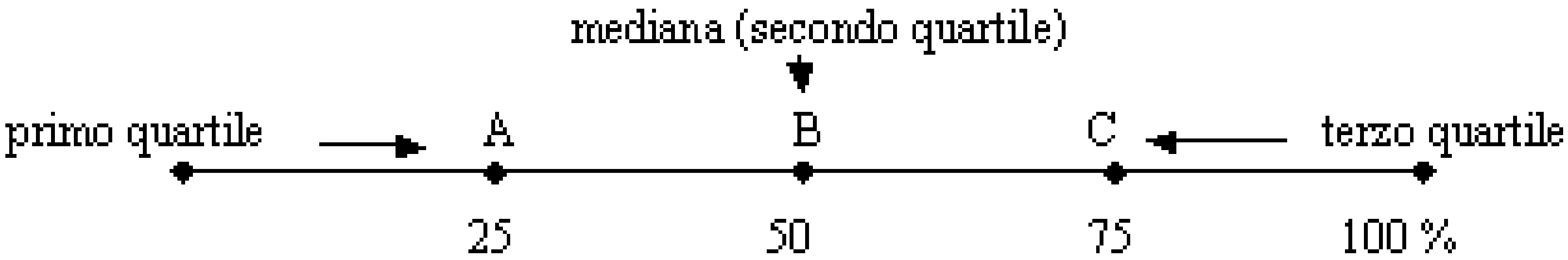
Quantili

- Definiamo **quantili** quei valori che dividono la distribuzione ordinata di una variabile in un certo numero di parti di uguale numerosità.

Terzili	**2 valori (T_1, T_2) che dividono la distribuzione ordinata in 3 parti**
Quartili	**3 valori (Q_1, Q_2, Q_3) che dividono la distribuzione ordinata in 4 parti**
…	
Decili	**9 valori (D_1, D_2, …, D_9) che dividono la distribuzione ordinata in 10 parti**
…	

I percentili usati più frequentemente sono i *quartili* che dividono in quattro parti di uguale ampiezza (quarti) la distribuzione ordinata:

- il primo quartile lascia prima di sé il 25 per cento delle unità che hanno modalità inferiori e dopo di sé il 75 per cento di unità con modalità superiori. Si indica con Q_1;
- il secondo quartile corrisponde alla mediana e divide la distribuzione in due parti uguali, lasciando prima di sé il 50 per cento di unità con modalità più piccole. Può essere indicato, oltre che con Me, anche con Q_1;
- il terzo quartile, Q_3, è quella modalità che divide in due parti la distribuzione ordinata, lasciando prima di sé il 75 per cento delle unità che presentano modalità inferiori.



http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_1.htm

Esempio:

E' stata valutata la richiesta di visite specialistiche da parte di 10 medici di base di una data ASL. Dall'osservazione sono risultati i seguenti dati: 50, 35, 28, 37, 33, 35, 48, 39, 47, 27. Supponiamo di voler calcolare i quartili. Ordiniamo la lista dei valori in senso crescente:

1°	**2°**	**3°**	**4°**	**5°**	**6°**	**7°**	**8°**	**9°**	**10°**
27	28	33	35	35	37	39	47	48	50

Q_1
(primo quartile)
33 richieste

Mediana
(secondo quartile)
(35+37)/2=36 richieste

Q_3
(terzo quartile)
47 richieste

25%

50%

75%

Il **primo quartile** è il valore che occupa il posto centrale della prima metà della successione ordinata delle osservazioni.

Il **secondo quartile** coincide con la mediana per definizione.

Il **terzo quartile** è il valore centrale della seconda metà della graduatoria.

Esempio:

Se abbiamo la seguente distribuzione di pesi (in Kg) di 20 studenti:

66 57 65 54 84 48 56 76 73 72 75 76 78 68 69 70 70 71 85 68

li mettiamo in ordine

48 54 56 57 65 66 68 68 69 70 70 71 72 73 75 76 76 78 84 85

1° quartile

3° quartile

il primo quartile si troverà tra il 5° e 6° posto e assumerà la modalità kg 65,5.
Il terzo quartile si troverà tra il 15° e il 16° posto e assumerà la modalità kg 75,5.

http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_1.htm

I percentili

- Definiamo **percentili** quei valori che dividono la distribuzione in cento parti di uguale numerosità.

I percentili di uso più frequente sono il 25-esimo e il 75-esimo percentile, che coincidono con il **primo** (Q_1) e **terzo quartile** (Q_3) che insieme alla mediana (**secondo quartile,** Q_2) dividono la distribuzione in quattro parti uguali.

Calcolo dei quantili per una lista di valori individuali

Ricordando che il primo quartile è quel valore al di sotto del quale c'è il 25%, cioè una *proporzione* pari a *0.25* di casi, e così via, possiamo utilizzare una semplice regola per il calcolo dei quantili:

- si ordina l'insieme dei dati dal più piccolo al più grande o viceversa;
- si determina il prodotto *np* dove *n* è il numero totale delle osservazioni e *p* la proporzione di casi inferiore al quantile; tale prodotto indica il posto nella graduatoria in corrispondenza del quale troviamo il valore del quantile; se il prodotto *np* non è un numero intero, si arrotonda all'intero successivo; se è intero si considera tale numero e quello successivo e il quantile sarà la media aritmetica dei due valori corrispondenti.

Applicando tale regola ai dati precedenti abbiamo:

I quartile:

$n \times p = 10 \times 0.25 = 2.5$ arrotondiamo all'intero successivo: 3° posto

⇨ **33 richieste**

II quartile:

$n \times p = 10 \times 0.50 = 5$ prendiamo 5 e l'intero successivo 6: 5° e 6° posto

⇨ **(35+37)/2=36 richieste**

III quartile:

$n \times p = 10 \times 0.75 = 7.5$ arrotondiamo all'intero successivo: 8° posto

⇨ **47 richieste**

Calcolo dei quantili per una distribuzione di frequenza

Esempio:

Consideriamo la seguente distribuzione dei costi giornalieri di una data terapia in 195 pazienti:

Costo (in migliaia di lire)	N. pazienti	%	Frequenze cumulate	% cumulate
10-29	11	5.7	11	5.7
30-49	33	16.9	44	22.6
50-69	23	11.8	67	34.4
70-89	15	7.7	82	42.1
90-109	35	17.9	117	60.0
110-129	22	11.3	139	71.3
130-149	17	8.7	156	80.0
150-169	21	10.8	177	90.8
170-189	18	9.2	195	100
totale	**195**	**100**		

I quartile $\longrightarrow$ classe **50-69**

Mediana (II quartile) $\longrightarrow$ " " **90-109**

III quartile $\longrightarrow$ " " **130-149**.

La mediana, i quantili e la definizione delle classi di una variabile

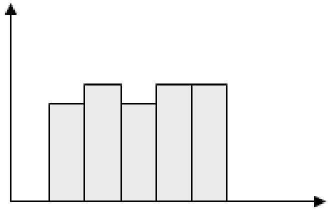
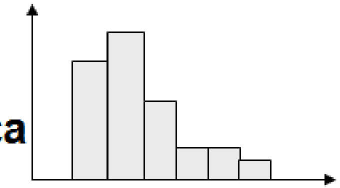
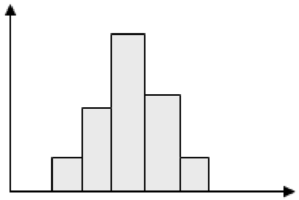
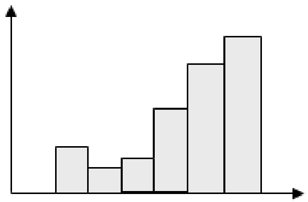
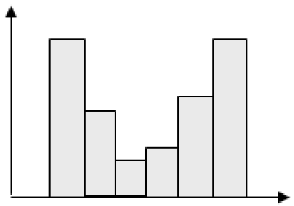
Esempio

In uno studio coorte sul ricorso a rimedi naturali in un campione di 6545 soggetti di età compresa tra i 45 e i 65 *anni (Pharmacoepidemiology and Drug Safety, 1996, 5:303-314*), è stata valutata l'associazione tra alcune variabili con l'uso o il non uso di rimedi naturali. Nel prospetto seguente sono riportate le classi di alcune di queste variabili definite coll'ausilio della mediana e dei quantili:

Percentuale di grasso corporeo (quartili)	**Percezione dello stato di salute (terzili)**	**Grado di autogestione (mediana)**
Meno di 17.7	1-4	alto
17.7-20.4	5	basso
20.5-23.6	6-7	
Più di 23.6		

Aspetti che orientano nella scelta del tipo di media

- scala di misura in cui è espressa la variabile da sintetizzare
- la forma della distribuzione della variabile
- le finalità conoscitive per le quali interessa effettuare la sintesi
- la variabilità interna ai dati

Distribuzione uniforme		quantili
" " asimmetrica		mediana quantili
" " simmetrica		media aritmetica, moda, mediana, quantili
" " a forma di J		mediana, quantili
" " a forma di U		quantili