Lezione 2ª- Statistica descrittiva per variabili quantitative

Esempio 5. Nella tabella seguente sono riportati i valori del tasso glicemico rilevati su 10 pazienti:

Paziente	Glicemia (mg/100cc)
1	x_1=103
2	x_2=97
3	x_3=90
4	x_4=119
5	x_5=107
6	x_6=71
7	x_7=94
8	x_8=81
9	x_9=92
10	x_{10}=96
Totale	**950**

- Numerosità del campione: **n=10**
- Variabile rilevata: **glicemia**
- Unità di misura: **mg/100cc**

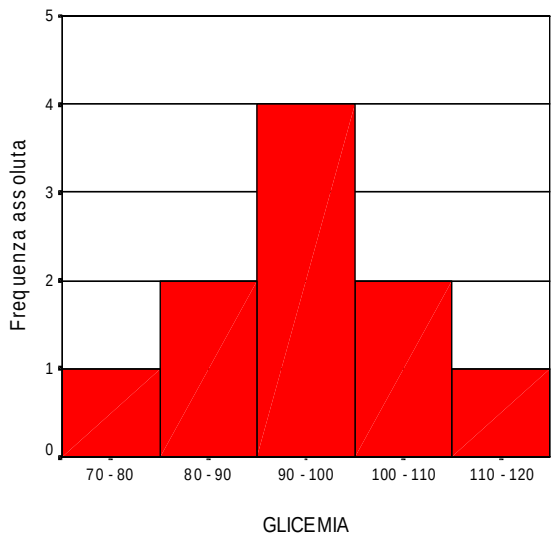
I° QUESITO: Come si distribuiscono questi dati?

Una prima analisi descrittiva dei dati può essere di tipo grafico, attraverso la costruzione di un istogramma o un poligono di frequenza. Essendo la variabile X quantitativa e continua, la si può suddividere in classi di valori di data ampiezza. Si può scegliere, ad esempio, una suddivisione in 5 classi di ampiezza = (valore massimo – valore minimo)/5 = (119 – 71)/5 ≈ 10 mg/100 cc, come in tabella seguente (in ogni classe il primo estremo è escluso, il secondo è incluso). Si fa osservare, comunque, che la scelta del numero di classi non è sempre agevole, può essere anche arbitraria e dipende dalla numerosità campionaria.

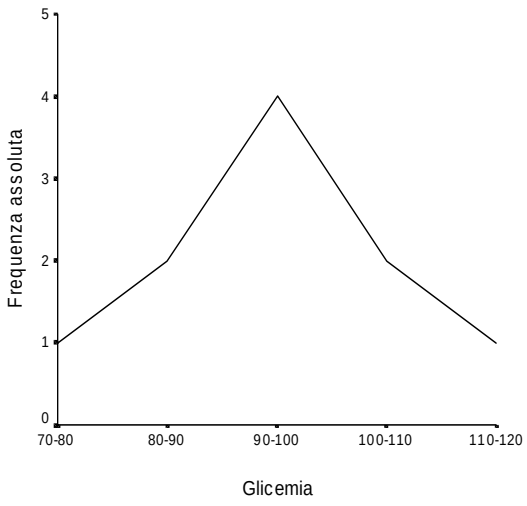
Calcolo delle frequenze di ogni classe: assolute e relative percentuali

Classi di valori di glicemia	Frequenza assoluta	Frequenza relativa
70 — | 80	1	1 / 10 ·100% = 10 %
80 — | 90	2	2 / 10 ·100% = 20 %
90 — | 100	4	4 / 10 ·100% = 40 %
100 — | 110	2	2 / 10 ·100% = 20 %
110 — | 120	1	1 / 10 ·100% = 10 %
Totale	**10**	**100 %**

Costruzione dell'istogramma e del poligono di frequenza



Istogramma



Si tratta ora di scegliere una misura di posizione (o di tendenza centrale) più appropriata per “sintetizzare” la distribuzione in esame.

Le misure di posizione più usate sono:

1. media aritmetica;
2. mediana;
3. moda;
4. media armonica;
5. media geometrica.

Per i dati quantitativi (variabili statistiche quantitative) si possono utilizzare le medie algebriche (aritmetica, geometrica, armonica).

La **<u>media aritmetica</u>** è quel valore che avrebbero tutte le osservazioni se non ci fosse la variabilità (casuale o sistematica). Più precisamente, è quel valore $\bar{x}$ che sostituito a ciascun degli **n** dato ne fa rimanere costante la somma:

$$x_1 + x_2 + x_3 + \ldots + x_n = \sum_{i=1}^{n} x_i = n \cdot \bar{x} \quad \Rightarrow \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Nell’Esempio 5 si ha:

$$\sum_{i=1}^{n} x_i = 950 = 10 \cdot \bar{x} \quad \Rightarrow \quad \bar{x} = 950/10 = 95 \text{ mg/100 cc.}$$

95 x 10=950

Esempio 6. Nella tabella seguente ci sono i voti riportati da uno studente universitario in 19 esame sostenuti

Voto (x_i)	Frequenza (f_i)	$x_i \cdot f_i$
18	2	36
20	4	80
22	8	176
24	2	48
27	2	54
30	1	30
Totale	19	424

Dalla tabella la media aritmetica (**ponderata o pesata**) è data da:

$$\bar{x} = \frac{\sum_i x_i \cdot f_i}{\sum_i f_i} = \frac{424}{19} = 22,32$$

Proprietà della media aritmetica:

a) minimo dei dati < $\bar{x}$ < massimo dei dati;

b) $\sum_i (x_i - \bar{x}) = 0$ (cioè: la somma degli scarti dalla media è zero);

c) $\sum_i (x_i - z)^2$ assume valore **minimo** per $z = \bar{x}$;

d) la media dei valori: $k \cdot x_i$ è pari a: $k \cdot \bar{x}$ (dove k è un numero reale qualsiasi);

e) la media dei valori: $x_i \pm h$ è pari a: $\bar{x} \pm h$ (dove h è un numero reale qualsiasi).

Limite della media aritmetica: è notevolmente influenzata dai valori estremi della distribuzione.

Si consideri infatti il seguente esempio.

Esempio 7: Età alla morte di 5 soggetti:

$x_1 = 34$ anni; $x_2 = 70$ anni; $x_3 = 74$ anni; $x_4 = 64$ anni; $x_5 = 68$ anni.

La media aritmetica è pari a:

$$\bar{x} = (34+70+74+64+68)/5 = 62 \text{ anni}$$

e tale valore è seriamente influenzato dall'osservazione di una morte avvenuta all'età di 34 anni; in realtà 4 delle 5 osservazioni sono superiori alla media.

Esempio 7': Peso in grammi di 5 topini

x_1=19gr, x_2=17gr, x_3=25gr, x_4=18, x_5=15gr

La media aritmetica è pari a:

$$\bar{x} = (19+17+25+18+15)/5 = 18.8gr$$

La **mediana (Me)** è quell'osservazione che bipartisce la distribuzione in modo tale da lasciare al "di sotto" lo stesso numero di termini che lascia al "di sopra".

Ritornando all'Esempio 5, per il calcolo della mediana è necessario disporre i dati in ordine crescente:

71, 81, 90, 92, **94, 96,** 97, 103, 107, 119

la mediana è quel dato che cade a metà della distribuzione ordinata. Se il numero di osservazioni è pari (come nel caso dell'esempio della glicemia) la mediana è la media aritmetica delle due osservazioni centrali:

Me = (94+96)/2 = 95 mg/100 cc.

Esempio 7': 15 17 **18** 19 25 n=5

Il fatto che mediana e media aritmetica in questo caso coincidano non è casuale in quanto la distribuzione è **simmetrica**. Ma, in generale, ciò non avviene.

Vantaggio nell'uso della mediana: non è influenzata dalle osservazioni aberranti o estreme.

Così nell'Esempio 7, disposti i dati in ordine crescente:

34 anni; 64 anni; 68 anni; 70 anni; 74 anni;

si ottiene il valore: Me = 68 anni, misura "più attendibile" di sintesi dei (pochi) dati a disposizione.

In realtà, in presenza di una distribuzione non simmetrica di dati è più appropriato far ricorso alla mediana che non alla media aritmetica.

Le **fasi operative** per il calcolo della mediana sono le seguenti:

a) ordinamento crescente dei dati;

b) – se il numero di dati **n** è dispari, la mediana corrisponde al dato che occupa la (n+1)/2 esima posizione

– se il numero di dati **n** è pari, la mediana è data dalla media aritmetica dei due dati che occupano la posizione n/2 e quella n/2+1.

In presenza di una distribuzione di frequenze è necessario considerare le frequenze cumulate, come illustrato nell'Esempio 6 di seguito ripreso in esame.

Voti ordinati (x_i)	Frequenze (f_i)	Freq. Cumulate (F_i)
18	2	2
20	4	2+4 = 6
22	8	4+8 = 14
24	2	14+2 = 16
27	2	16+2 = 18
30	1	18+1 = 19
Totale	19	19

n/2 = 19/2 = 9,5 $\Rightarrow$ la più piccola frequenza cumulata maggiore o uguale a **n/2** è pari a 14, dunque la mediana è data da Me = 22 (voto corrispondente alla frequenza cumulata 14).

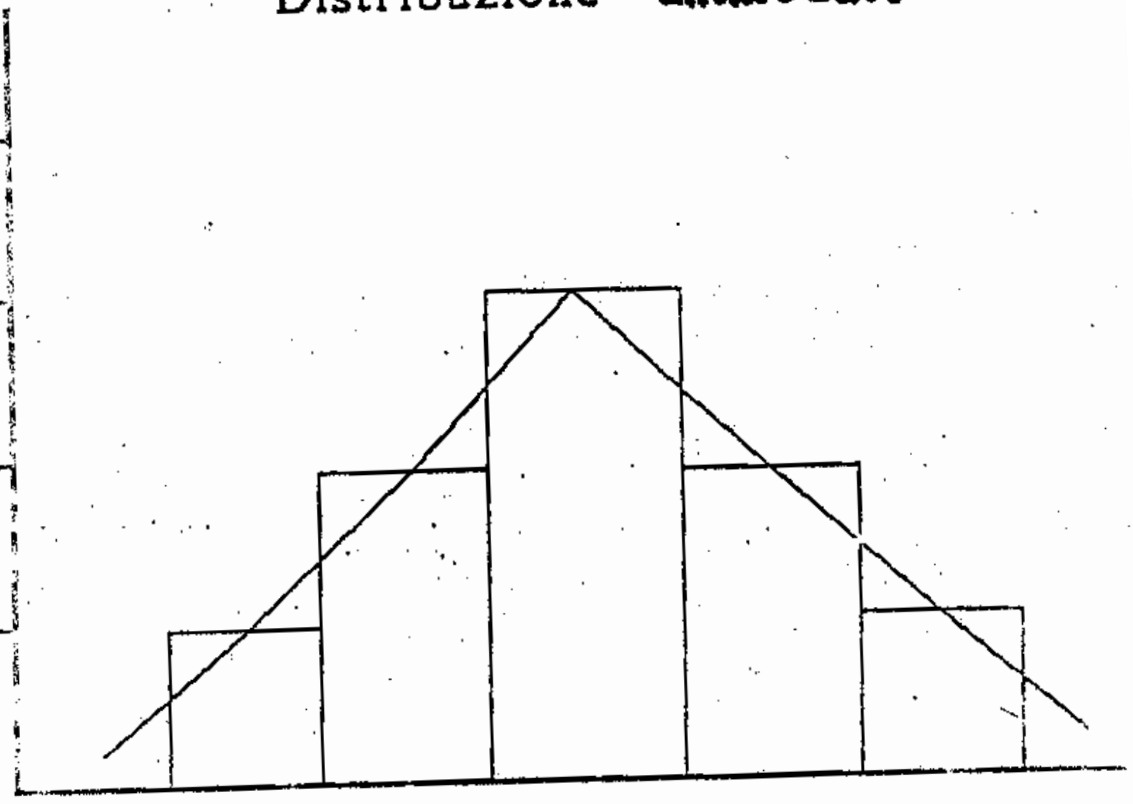
Se, infine, i dati sono raggruppati in classi, per il calcolo della mediana **si può** far riferimento al valore centrale di ciascuna classe (dato dalla semisomma dei valori estremi di classe) o, più in genere, alla "classe mediana".

La **Moda (Mo)** è l'osservazione che si verifica con maggiore frequenza in una data distribuzione. Si possono avere anche più valori modali.

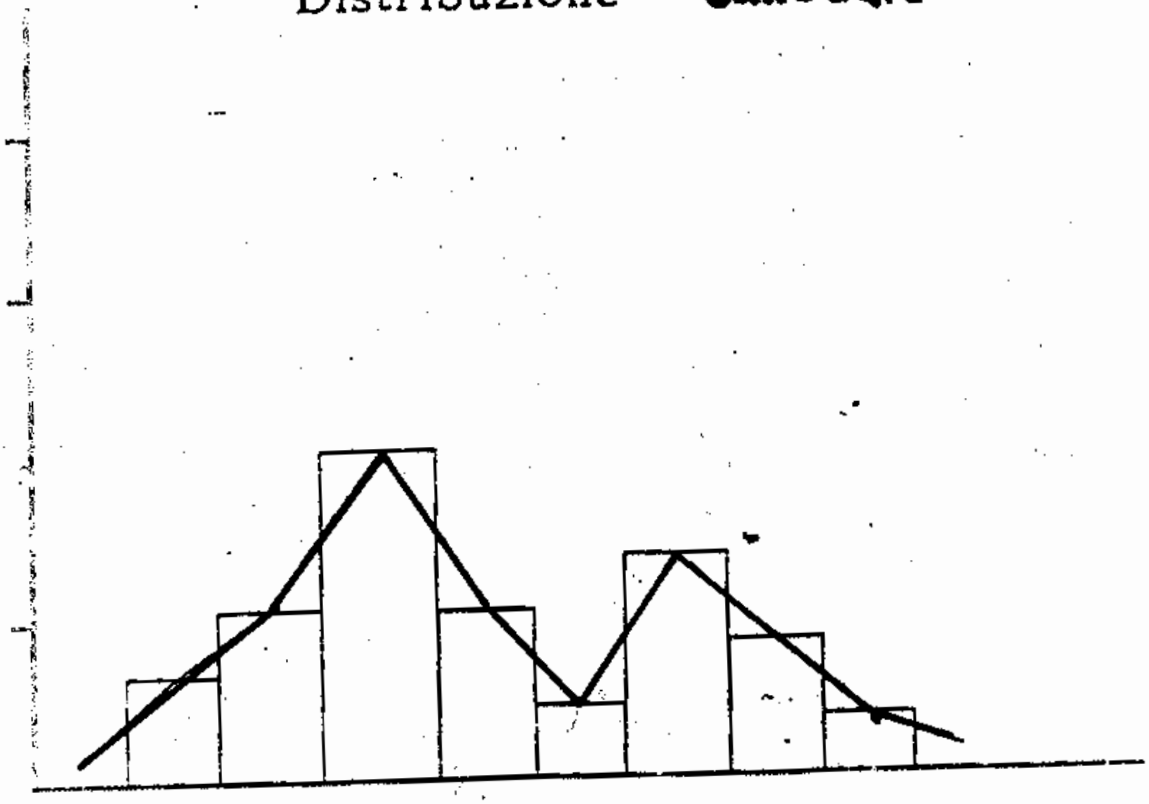
Ad esempio, la moda della distribuzione di voti (Esempio 6) è pari a Mo = 22; nel caso della glicemia si può considerare la "classe modale" pari all'intervallo: 90 —| 100.

Altre misure di tendenza centrale sono la media armonica e quella geometrica.

Distribuzione **unimodale**



Distribuzione **bimodale**



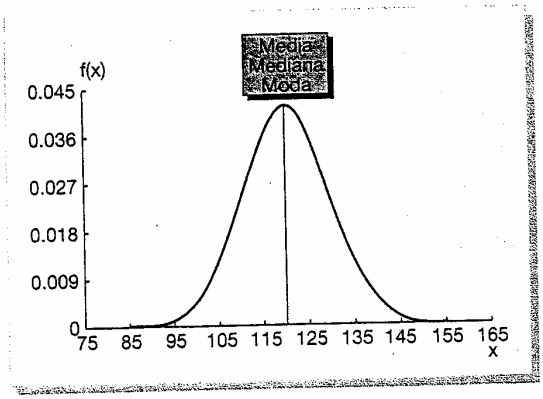


Grafico 4.7 Media, moda e mediana in una distribuzione normale.

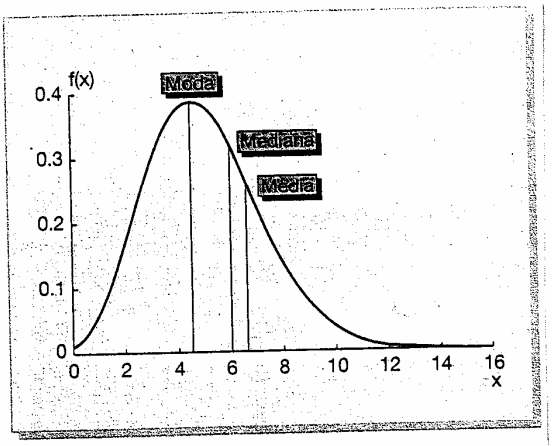


Grafico 4.8 Media, moda e mediana in una distribuzione asimmetrica positiva.

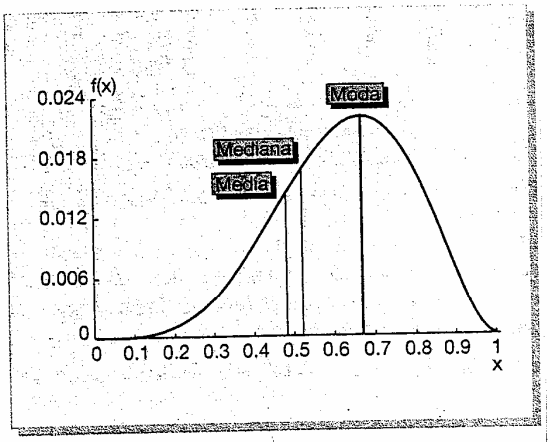


Grafico 4.9 Media, moda e mediana in una distribuzione asimmetrica negativa.

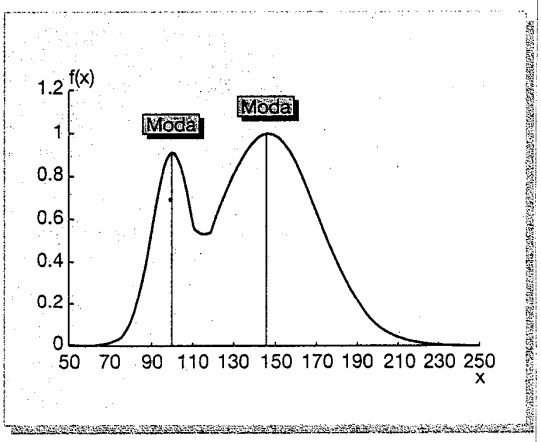


Grafico 4.10 Distribuzione ad andamento bimodale.

RIEPILOGO SULLE MISURE DI POSIZIONE

	simbolo	formula	tipo di misura	osservazioni/ applicazioni
1. media aritmetica	$\bar{x}$ μ	$= \frac{\sum_{i=1}^{n}}{n}$	media analitica	è la media più usata soprattutto quando la distribuzione è simmetrica
2. mediana	$\tilde{x}$ μ_e	/	media lasca	si usa per distribuzioni asimmetriche o in presenza di valori aberranti
3. moda	M_o	/	media lasca	si usa solo per definire la classe (modale) in cui cade il maggior numero di osservazioni
4. media armonica	M_a	$= \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	media analitica	trova applicazione nel caso in cui la variabile in esame è rappresentata dai tempi di reazione
5. media geometrica	M_g	$= \sqrt[n]{\prod_{i=1}^{n} x_i}$	media analitica	si usa per la descrizione di fenomeni che seguono una legge esponenziale