

L'importanza della statistica

La Statistica è la scienza che analizza i fenomeni quantitativi, cercando di evidenziarne le caratteristiche salienti, le regolarità, le eccezioni.

Esempi:

- Rilevazione del numero di fratelli/sorelle per ogni corsista presente in aula.
- Mese di nascita di ogni corsista presente in aula.
- Rilevazione della temperatura a intervalli orari in una stazione meteorologica.
- Rilevazione del primo numero estratto sulla ruota di Palermo nelle ultime 1000 estrazioni del lotto.
- Mezzo di trasporto utilizzato da ciascun corsista per raggiungere la sede del corso.
- Tempo di spostamento impiegato da ciascuno corsista per raggiungere la sede del corso, congiuntamente al mezzo di trasporto usato.

Affinché le informazioni deducibili da tali rilevazioni siano proficuamente utilizzate è necessario imparare a conoscere:

- Gli obiettivi primari della statistica;
- Il linguaggio statistico;
- Come si organizza una indagine statistica;
- Le principali tecniche di analisi e sintesi;
- Il metodo statistico.

Obiettivi della statistica

- Separare il “segnale” dal “rumore”;
- Valutare la complessità dei fenomeni;
- Saper “prevedere”.

In base al tipo di obiettivi è necessario:

- A. saper predisporre, logicamente e praticamente, il tipo di indagine più adatta al conseguimento dei nostri obiettivi.
- B. Definire con precisione qual è la popolazione di riferimento della nostra indagine.
- C. Stabilire quali sono le caratteristiche della popolazione che “interessano”.

Le fasi di una indagine statistica

1. Definizione degli obiettivi;
2. Pianificazione della raccolta dei dati;
3. Rilevazione dei dati;
4. Elaborazione metodologica;
5. Presentazione dei risultati;
6. Utilizzazione dei risultati della ricerca.

Un buon corso di Statistica deve soffermarsi:
(senza indugiare nel dettaglio) sui punti 1-3, approfondire il punto 4, sorvolare sul punto 5 (serve a vendere meglio la ricerca) lasciare all'esperto del settore il punto 6.

Nota importante:

Pur avendo tracciato un percorso logico lineare, nella realtà le cose sono più contorte.

Spesso si parte con un obiettivo che sottintende una teoria, ma i dati a disposizione sembrano confutare la teoria, suggerendo al tempo stesso una interpretazione differente.

La rilevazione (o indagine) statistica

E' il complesso di operazioni rivolte ad acquisire informazioni su un insieme di elementi, oggetto dello studio.

Le rilevazioni possono essere:

- **semplici;**
- **composte.**

La rilevazione fornisce:

- **risposte;**
- **misure.**

Esistono rilevazioni:

- **totali** (censimenti), relative a tutte le unità della popolazione;
- **parziali** (campionarie), relative a un sottoinsieme della popolazione (il campione).

Un po' di terminologia statistica:

Popolazione

Qualsiasi insieme di elementi, reale o virtuale, che forma oggetto di studio.

Esempi:

- Tutti i residenti nel comune di Messina il 30/09/2002;
- Tutte le possibili sestine giocabili al Superenalotto;
- Tutti i malati di *Sindrome della Statistica*.

E' di fondamentale importanza (nonché indicatore di serietà della ricerca) definire esattamente la popolazione di riferimento della nostra indagine.

Unità statistica

Elemento di base della popolazione sulla quale viene effettuata l'indagine. E' *indivisibile* nell'ambito della ricerca ma non in senso assoluto (es.: *famiglie*).

Carattere (o Variabile)

Fenomeno oggetto di studio che è rilevato sulle unità statistiche. Esso si manifesta attraverso diverse **modalità**.

Esempi:

- Il carattere **Sesso** si manifesta attraverso le modalità **M** e **F**.
- Il carattere **Tempo di spostamento** si manifesta attraverso infinite modalità.
- Il carattere **Numero di fratelli** ha come modalità i numeri interi positivi e lo zero.

Nota: di una popolazione interessano soltanto le modalità del carattere che è oggetto di studio.

Frequenza

Il numero di volte che una data modalità si manifesta nel collettivo di riferimento.

La distribuzione delle frequenze descrive come il fenomeno in esame si manifesta nella popolazione (o campione). Distingueremo tra:

- *Frequenze assolute;*
- *Frequenze relative o percentuali;*
- *Frequenze cumulate;*
- *Frequenze retrocumulate.*

Esempio:

Nel collettivo dei corsisti frequentanti questo corso di Statistica il carattere *Numero di fratelli/sorelle* è così distribuito:

Modalità	Freq. Assol.	Freq. Relat.	Freq. Cum.	Freq. Retr.
0	28	0.140	0.140	1.000
1	59	0.295	0.435	0.860
2	51	0.255	0.690	0.565
3	30	0.150	0.840	0.310
4	13	0.065	0.905	0.160
5	7	0.035	0.940	0.095
6	5	0.025	0.965	0.060
7	3	0.015	0.980	0.035
>7	4	0.020	1.000	0.020
Totale	200	1.000		

Tipologia dei caratteri

I Caratteri possono essere:

- **Quantitativi** (le modalità sono numeri)
si chiamano anche **variabili**
- **Qualitativi** (le modalità sono attributi non numerici)
si chiamano anche **mutabili**

I caratteri quantitativi si possono poi suddividere in:

- quantitativi **discreti**
le modalità sono solo alcuni numeri
- quantitativi **continui**
le modalità sono tutti i numeri reali, almeno in teoria

Tale distinzione è più teorica che pratica perché, nelle misurazioni concrete, tutte i caratteri sono discreti.

I caratteri possono essere classificati anche in base alla scala di misurazione o, in altre parole, alle operazioni che possiamo fare con le loro modalità.

- scala *nominale*:

Le modalità non hanno un ordinamento (è il caso dei caratteri qualitativi)

Esempi:

- giudizio su un film: bello / brutto;
- sì/no;
- colore degli occhi.

- scala *ordinale*:

Le modalità sono attributi non numerici ma logicamente *ordinabili*

Esempi di caratteri misurabili su scala ordinale:

- titolo di studio;
- livello di soddisfazione per un prodotto
(per niente, poco, abbastanza, etc...).

- scala *per intervallo*:

Sono i caratteri quantitativi che consentono confronti solo per differenza ma non per rapporto.

Esempio: la temperatura.

Se misuriamo un giorno la temperatura minima e massima a Messina e a New York, in gradi C o F $[C=5(F-32)/9]$ otteniamo:*

	<i>Massime C</i>	<i>Minime C</i>	<i>Massime F</i>	<i>Minime F</i>
<i>Messina</i>	25	16	77	60.8
<i>New York</i>	17	8	62.6	46.4

Ha senso dire che l'escursione termica è la stessa nelle due città, ma non ha senso dire che la minima a New York è la metà della minima a Messina.

In altri termini si tratta di caratteri numerici in cui non esiste uno *zero* assoluto.

- *scala per rapporto:*

caratteri numerici per i quali è intrinseca la definizione dello *zero*.

Es. *peso, altezza, durata temporale.*

Operazioni statistiche elementari

Differenza tra due modalità x e y di un carattere

$$x-y$$

(espressa nella stessa unità di misura del carattere)

Differenza relativa tra due modalità x e y di un carattere

$$(x-y)/x$$

(è un numero puro)

Esempi:

Tasso di variazione del cambio euro/dollaro

Al tempo (a) 1 euro = 0,97 dollari

Al tempo (b) 1 euro = 0,99 dollari

Tasso di variazione

$$(0,99-0,97)/0,97= 0,02 \quad (\text{oppure } 2 \%)$$

Rapporti Statistici

Sono importanti indicatori descrittivi di un fenomeno.
Ne esistono diversi tipi:

- **Rapporti di composizione**

Valore rilevato in una certa circostanza rapportato al totale della popolazione. Esempio: Percentuale di giovani (<14 anni) nella popolazione.

- **Rapporti di derivazione**

Rapporto tra la modalità di un carattere e la corrispondente modalità di un altro carattere che ne costituisce presupposto logico.

Esempio: Numero di nati rapportato al totale della popolazione femminile in età fertile.

- **Rapporti di densità**

Esempio: Numero di abitanti per unità di misura spaziale.

- **Rapporti di coesistenza**

Rapporto tra frequenze (o quantità) di una modalità rispetto a un'altra

Es. Indice di vecchiaia = N. di anziani (>64 anni) / Numero di giovani (<15 anni)

- **Numeri Indice**

Rapporto tra due valori di uno stesso fenomeno misurato in due diverse occasioni o in due località differenti.

Es. se il pane costa € 1,50/Kg. a Roma e € 1,20/Kg a Messina il numero indice di Messina con base Roma è:

$$\text{Prezzo a Me} / \text{Prezzo a Rm} * 100$$

cioè

$$1,20/1,50 * 100 = 80$$

I numeri indice più frequentemente usati riguardano le variazioni dei prezzi, delle produzioni a diversi livelli di aggregazione.

Rappresentazione delle rilevazioni statistiche

- **Distribuzioni per unità**

Mero elenco delle modalità di un carattere osservati sulle unità statistiche del campione o della popolazione.

Esempio: Ad un collettivo di 40 individui viene richiesto il numero di fratelli/sorelle

0 2 1 4 0 3 2 3 2 1

Poco adatto a grandi numerosità...

- **Distribuzioni di frequenza**

Tabella in cui vengono elencate le modalità e le rispettive frequenze (ass. e/o rel. e/o cum.).

Consideriamo lo stesso esempio:

<i>Modalità</i>	<i>Freq. Ass.</i>
<i>0</i>	<i>6</i>
<i>1</i>	<i>8</i>
<i>2</i>	<i>10</i>
<i>3</i>	<i>8</i>
<i>4</i>	<i>4</i>
<i>5</i>	<i>2</i>
<i>>5</i>	<i>2</i>
<i>Totale</i>	<i>40</i>

• Distribuzioni per classi

Un carattere continuo può, in teoria, assumere infinite modalità. Per questo può essere conveniente organizzare i risultati in una tabella in cui le modalità sono in realtà *intervalli*.

Esempio:

Su un collettivo di 40 individui si registra il peso e le modalità vengono raggruppate in classi di ampiezza 5kg.

Classe	Frequenza
<60	7
[60,65)	9
[65,70)	10
[70, 75)	8
≥ 75	6

• Serie storiche e territoriali

Esprimono la dinamica temporale o spaziale di un certo fenomeno (registrato istantaneamente (var. di stato) o in relazione a un certo periodo (var. di flusso)).

Esempi.

- *Numero di nati vivi a Messina mese per mese;*
- *Cambio euro - dollaro registrato giornalmente;*
- *Il numero di assist di un giocatore di basket ad ogni partita durante una stagione.*

- **Matrice di dati**

Un modo generale di rappresentare i risultati di una rilevazione statistica, soprattutto quando i caratteri rilevati sono più di uno è la cosiddetta

matrice unità - variabili

composta da tante righe quante sono le unità osservate e su ogni riga vengono riportate le modalità specifiche per i diversi caratteri.

Esempio: Su un collettivo di 10 città si rilevano:

1. *Popolazione residente;*
2. *Numero di ospedali pubblici;*
3. *Numero di cinema/multisale;*
4. *Numero di centri commerciali.*

Città	Popolaz.	Ospedali	Cinema	C.Comm.
Roma	3.824.000	18	72	14
Milano	2.726.000	16	53	18
Napoli	2.121.000	15	50	11
Torino	895.000	12	27	8
Genova	598.000	13	24	7
Palermo	680.000	10	21	6
Firenze	728.000	14	26	5
Bologna	568.000	18	30	8
Venezia	389.000	12	12	7
Bari	450.000	8	19	2

La particolare tipologia di rappresentazione dei dati è strettamente legata agli obiettivi dell'analisi

- **Rappresentazioni grafiche**

Istogramma;

Torta;

Scatter Plot;

Diagramma cartesiano.

Le distribuzioni statistiche

- Distribuzioni per unità
- Distribuzioni di quantità
- Distribuzioni di frequenza

Le distribuzioni di quantità sono simili a quelle di frequenza ma ad ogni modalità associano un "ammontare" (Kg, dollari, litri) piuttosto che una frequenza.

Esempio:

Redditi da lavoro dipendente in Italia secondo i rami dell'economia (in miliardi di lire)

Rami	Amm. Assoluto	Amm. relativo
Agricoltura	15.043	2,2
Industria	227.765	33,0
Servizi	235.016	34,3
Pubb. Ammin.	209.339	30,5
Totale	687.136	100,0

Si utilizzano in presenza di caratteri **trasferibili**

Distribuzioni di frequenza

Rappresentazione tabellare in cui, accanto ad ogni modalità del carattere viene riportato "quante unità"

(frequenza assoluta) assumono quella specifica modalità.

Oltre alla frequenza assoluta si possono considerare le:

1. Frequenze relative
2. Frequenze cumulate
3. Frequenze retrocumulate

Es. (*Titolo di studio del padre di un collettivo di 326 studenti*)

Modalità	F. Ass.	F. Rel.	F. Cum.	F. Retr.
Lic.Elem.	88	0.270	0.270	1.000
Lic. Med.	77	0.236	0.506	0.730
Maturità	130	0.399	0.905	0.494
Laurea	31	0.095	1,000	0.095
Totale	326	1.000		

La distribuzione di frequenza può essere determinata per qualunque carattere statistico. Però le frequenze cumulate e quelle retrocumulate hanno senso soltanto per caratteri almeno ordinabili.

Le frequenze assolute e quelle relative forniscono la stessa informazione ma quelle relative

- Consentono confronti tra due diverse distribuzioni
- Fanno perdere l'informazione relativa al totale

Un po' di formalizzazione

Caratteri discreti (numero finito o numerabile di modalità)

Assumiamo di avere un collettivo di n individui e un carattere X con k modalità. La generica distribuzione di frequenza sarà

Modalità	Freq. Ass.	Freq. Rel.
x_1	n_1	n_1 / n
x_2	n_2	n_2 / n
...
x_i	n_i	n_i / n
...
x_k	n_k	n_k / n
<i>Totale</i>	N	1

Le quantità n_i per $i=1, \dots, k$ sono le frequenze assolute delle modalità, rappresentate genericamente con il simbolo x_i per $i=1, \dots, k$

Note:

- *Ogni n è un numero intero compreso tra 0 e n*
- $n_1 + n_2 + \dots + n_i + \dots + n_k = n$
- le frequenze relative si indicano spesso con il simbolo

$$f_i = n_i / n \quad i=1, \dots, k$$

Ovviamente, per ogni $i=1, \dots, k$

$$0 < f_i < 1$$

$$f_1 + f_2 + \dots + f_i + \dots + f_k = 1$$

- Rappresentazione grafica
Diagramma a barre separate

Caratteri continui

Non è possibile elencare tutte le modalità
Esigenza di raggruppare in classi

In questo caso la generica modalità del carattere verrà indicata col termine **classe**
e rappresentata dai simboli

$(x_i, x_{i+1}]$ per $i=1, \dots, k$
chiuso a destra e aperto a sinistra

oppure

$[x_i, x_{i+1})$ per $i=1, \dots, k$
chiuso a sinistra e aperto a destra

Convenzione: $x_{k+1} = \text{infinito} \dots$

Esempio:

Distribuzione delle durate di 1192 brani musicali.

Poiché la durata minima osservata è 30 secondi e quella massima osservata è 1022 secondi ma i dati sono perlopiù concentrati su valori medio-bassi, adottiamo il seguente raggruppamento

Criteri per la scelta delle classi

- Non ne esistono di univoci
- Quando ha senso, si considerano classi della stessa ampiezza e tali da avere una frequenza "significativa" (classi **equiampie**)
- In alternativa si costruiscono classi in modo da avere approssimativamente, frequenze simili.

Questo secondo criterio è più complesso perché va elaborato dopo aver osservato tutti i dati

Rappresentazioni grafiche

Istogramma.

L'istogramma è formato da k rettangoli adiacenti (uno per ogni classe) che indicano la frequenza delle classi.

La costruzione dell'istogramma nasconde qualche insidia, soprattutto quando le classi non hanno la stessa ampiezza.

In questo caso la base dei rettangoli sarà proporzionale alla ampiezza della classe e la frequenza verrà indicata dall'**area** del rettangolo e non dalla sua altezza.

Frequenze cumulate e Funzione di ripartizione empirica. (solo per caratteri qualitativi ordinabili o quantitativi)

Consideriamo un collettivo di n unità su cui osserviamo un carattere X ordinabile ed elenchiamo in ordine crescente le n modalità assunte

$$x_1 < x_2 < x_3 < \dots < x_i < \dots < x_n$$

Attribuiamo ad ogni unità un peso pari a $1/n$; possiamo dire che la frazione di unità che assume valori minori o uguali a x_1 è pari a $1/n$, che la frazione di unità che assume valori minori o uguali a x_i è pari a i/n .

Quando X è quantitativo si ha la seguente

Definizione: Si chiama Funzione di *ripartizione empirica della variabile X* , la funzione che associa ad ogni valore x di X il valore

$$F(x) = (\# \text{ di osservazioni minori o uguali di } x) / n$$

Note:

- La funzione $F(x)$ è non decrescente
- La funzione $F(x)$ è compresa tra 0 e 1
- $\lim_{[x \rightarrow \infty]} F(x) = 1$, $\lim_{[x \rightarrow -\infty]} F(x) = 0$,
- Nel caso di distribuzione di frequenze la funzione di ripartizione corrisponde (logicamente) alle frequenze cumulate

Gli indicatori di posizione

E' spesso necessario sintetizzare le informazioni fornite da una distribuzione attraverso un semplice indicatore.

Diversi indicatori forniscono diversi tipi di sintesi

Gli indicatori di posizione (*location index*) rappresentano un valore “rappresentativo” di tutti i valori della distribuzione.

Per forza di cose, essi sacrificano delle informazioni.

Indicatori per caratteri quantitativi

Cominciamo a considerare il caso di un carattere quantitativo discreto X . In questo caso è possibile parlare di **medie** di una distribuzione.

Consideriamo il seguente esempio:

Vogliamo controllare la qualità dei televisori prodotti nelle varie catene di montaggio di una data fabbrica, che ha un ritmo produttivo di 1600 televisori giornalieri, cioè di 200 televisori per ciascuna delle 8 ore lavorative. E' risultato che i televisori difettosi prodotti sono distribuiti , nelle 8 ore come esposto nella tabella seguente:

Ora	Televisori difettosi
1 ^a	7
2 ^a	5
3 ^a	5
4 ^a	8
5 ^a	6
6 ^a	9
7 ^a	7
8 ^a	9
Totale	56

L'intera distribuzione di frequenza dei pezzi difettosi nelle 8 ore, dà un'idea confusa del fenomeno che studiamo. In un caso come questo, per esprimere sinteticamente il fenomeno si fa ricorso a un procedimento di questo tipo: *si sommano tutti i pezzi difettosi costruiti nelle 8 ore lavorative e si divide tale somma per il numero delle ore*; si ottiene così:

$$\frac{7 + 5 + 5 + 8 + 6 + 9 + 7 + 9}{8} = \frac{56}{8} = 7$$

Si vede facilmente che il numero 7, sostituito ai singoli valori di pezzi difettosi, dati dalla tabella, *ne lascia inalterata la somma*:

$$7 \times 8 = 56.$$

Per questo motivo, si dice che 7 è la “*media aritmetica*” di tali valori. Possiamo così affermare che “in media” si sono presentati 7 televisori difettosi per ognuna delle 8 ore di lavoro.

Si chiama quindi media aritmetica (semplice) di n numeri: x_1, x_2, \dots, x_n , il numero M che si ottiene dividendo la loro somma per il numero n. Cioè:

$$M = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Per vedere a quale concetto intuitivo corrisponde il numero M , consideriamo una successione di valori x_1, x_2, \dots, x_n , e insieme alla precedente, una seconda successione di ugual numero di termini M, M, \dots, M . E' intuitivo come la somma dei termini x_1, x_2, \dots, x_n della prima successione è uguale alla somma dei termini della seconda, cioè risulterà:

$$\sum_i x_i = n M$$

M rappresenta il valore che, sostituito ai singoli valori, mantiene inalterato il totale. Ne segue che

$$M = \sum_i x_i / n$$

*ovvero la **Media Aritmetica** (che si indica con M o con μ).*

In generale, data una tabella

$$x_1, x_2, \dots, x_i, \dots, x_k$$

$$n_1, n_2, \dots, n_i, \dots, n_k$$

la media aritmetica (ponderata) è

$$M = \frac{\sum_i x_i \cdot n_i}{\sum_i n_i} = \sum_i x_i f_i$$

Proprietà della media aritmetica

- Internalità $x_1 \leq \mu \leq x_k$
- La somma degli scarti della media è sempre uguale a 0

$$\sum_i (x_i - \mu) = 0$$

- La media aritmetica è quel valore che minimizza la somma dei quadrati degli scarti, ovvero per qualsiasi valore δ , si ha

$$\sum_i (x_i - \mu)^2 \leq \sum_i (x_i - \delta)^2$$

Caratteri continui

In questo caso, per poter calcolare una media relativamente ad un collettivo di unità, occorre diminuire il numero di modalità possibili attraverso una riduzione in classi e poi considerando il valore centrale della classe come valore rappresentativo.

La media aritmetica per variabili quantitative

Se le modalità sono classi non è possibile applicare direttamente le formula nota perché le modalità non sono numeri bensì intervalli. In questo caso occorre determinare un valore centrale per ogni classe.

In genere si considera la semisomma degli estremi dell'intervallo.

Esempio: Classi di statura.

Classi	<i>Val. centr. x_i</i>	f_i	$x_i f_i$
[155-164)	159.5	0.093	14.833
[164-169)	166.5	0.194	32.398
[169-174)	171.5	0.290	49.880
[174-179)	176.5	0.248	43.896
[179-184)	181.5	0.126	22.932
[184-194)	189	0.049	9.285
Totali		1.000	173.225

Quando le classi estreme sono aperte, il valore centrale viene scelto in altro modo, tenendo conto del problema specifico.

Moda, Mediana e Quantili

Esistono indicatori di posizione che, per essere calcolati, non utilizzano tutti i dati a disposizione
(a differenza di quanto avviene per la media).

Il più semplice di tali indicatori è la **MODA**, definito come
la modalità con frequenza più elevata

Es. Per un collettivo di 28 studenti registriamo il comune di residenza: risulta

Comune	Freq. Ass.
Messina	9
Milazzo	6
Barcellona	3
Patti	6
Villafranca	4
Totale	28

Qui il carattere è il ***Comune di residenza*** e la modalità modale è **Messina**.

- La moda non è necessariamente unica: se nell'esempio precedente anche Milazzo avesse avuto una frequenza pari a 9 avremmo avuto **due mode**.
- La moda è un indice molto *rozzo* perché non tiene conto di quello che avviene “dietro”: distribuzioni molto diverse potrebbero avere la stessa moda pur essendo sostanzialmente diverse
- Può avere un comportamento contro intuitivo come dimostra l'esempio seguente

Es.

X	1	2	3	4	5
Freq.	13	30	35	17	5

La moda è la modalità **3**. Se ora spostiamo 20 unità dalla modalità **3** e le mettiamo alla modalità **5**, otteniamo

X	1	2	3	4	5
Freq.	13	30	15	17	25

E' innegabile che la distribuzione si è spostata verso valori più grandi ma la moda ora è **2** !!

Il caso di distribuzioni continue

In questo caso sappiamo che le modalità vengono raggruppate per classi. Occorre allora determinare la *classe modale* che non necessariamente corrisponde a quella di maggiore frequenza: infatti, bisognerà tenere conto dell'*ampiezza delle classi*.

Es.

X	[0,1)	[1,3)	[3,6)	[6,10)	[10,15]
Freq.	10	22	24	36	25
Fr./Am.	10	11	8	9	5

Qui la classe modale **non è** [6,10) bensì è [1,3).

In generale la classe modale è quella con maggiore *densità di frequenza*.

La mediana

Un altro indicatore molto utile e molto usato è la mediana (**Me**). Essa può essere calcolata quando

- Il carattere è quantitativo
- Il carattere è qualitativo ordinabile

Occorre prima ordinare le osservazioni in modo che le modalità osservate risultino in ordine crescente e poi definire la mediana come

La modalità assunta dalla/e unità che occupano la posizione centrale

Es. Si osserva il carattere *numero di fratelli* su un collettivo di 9 studenti e, dopo aver ordinato i valori si ha

1, 2, 3, 4, 4, 5, 5, 6, 7

In questo caso n è dispari e la mediana è la modalità relativa all'unità che occupa la posizione $(n+1)/2$, ovvero, in questo caso la quinta: la mediana è quindi la modalità **Me=4**.

Quando n è pari non esiste una sola unità mediana, bensì 2. Infatti, se nell'esempio precedente aggiungiamo un'osservazione pari, ad esempio a 5, la distribuzione diventa

1, 2, 3, 4, 4, 5, 5, 5, 6, 7

e le mediane sono ora le modalità osservate sulle unità che occupano le posizioni $n/2$ e $n/2+1$.

Questo problema è importante quando n è piccolo ma perde importanza per grandi valori di n .

• Il caso delle distribuzioni di frequenze

Quando la distribuzione è organizzata per frequenze, la mediana si calcola utilizzando la Funzione di ripartizione (o le Frequenze cumulate). La mediana è quella modalità x_i per la quale risulta

- $F(x_{i-1}) < 0.5$
- $F(x_i) \geq 0.5$

Esempio

$X = \text{Tit. di studio}$	n_i	F_i
Lic. Elementare	6	0.3
Lic. Media	2	0.4
Maturità	6	0.7
Laurea	6	1.0
<i>Totale</i>	20	

La mediana è dunque **Me = Maturità** perché la decima e la undicesima unità assumono entrambe questa modalità.

• Il caso delle variabili continue

Quando le modalità sono raggruppate per classi occorre

- Individuare la classe mediana
- Assumere una uniforme distribuzione all'interno delle classi
- Individuare esattamente la mediana all'interno della classe

In parole povere è sufficiente disegnare l'istogramma relativo alla distribuzione e tracciare una linea che divide in due l'area sottesa all'istogramma.

Es.: Classi di statura

Classi	<i>Val. centrale</i>	f_i	F_i
[155-164)	159.5	0.093	0.093
[164-169)	166.5	0.194	0.287
[169-174)	171.5	0.290	0.577
[174-179)	176.5	0.248	0.825
[179-184)	181.5	0.126	0.951
[184-194)	189	0.049	1.000
Totali		1.000	

La classe mediana è la classe [169-174). Assumendo una distribuzione uniforme nella classe occorre allora individuare il valore che lascia alla sua sinistra esattamente il 50% delle unità. Il problema si risolve partendo dalla classe mediana $(x_{i+1}, x_i]$ attraverso la formula

$$Me = x_{i-1} + (x_i - x_{i-1}) * (0.5 - F(x_{i-1})) / (F(x_i) - F(x_{i-1}))$$

difficile da ricordare ma semplice da ricavare. Nell'esempio si ha

$$Me = 169 + 5 * (0.5 - 0.287) / (0.577 - 0.287) = 172.67$$

Caratteristiche della mediana

- **Me** minimizza la somma degli scarti in valore assoluto. Per ogni valore reale δ si ha

$$\sum_i |x_i - Me| \leq \sum_i |x_i - \delta|$$

- **Me** è più robusta della media aritmetica μ rispetto a valori anomali della distribuzione

Es.

Caso A:

1, 2, 2, 4, 6, 6, 7 **Me** = 4 μ = 4.

Cambiando l'ultimo valore da 7 a 700 si ha

Caso B:

1, 2, 2, 4, 6, 6, 700 **Me** = 4 μ = 180.25.

- Può essere più facile ed economico calcolare **Me** piuttosto che μ .

Es.: *Dati di sopravvivenza.*

Si osserva il tempo di durata di un insieme di 21 lampadine. Per calcolare il tempo medio occorre attendere che le 21 lampadine si rompano e calcolare il tempo medio. Per calcolare la mediana, invece, è sufficiente osservare le prime 11 “morti”.

I Quantili

Si può reinterpretare la mediana come *la più piccola modalità* che lascia alla sua sinistra il 50% delle unità statistiche. Si può effettuare lo stesso ragionamento cercando di individuare la modalità che lascia alla sua sinistra una percentuale di unità statistiche pari ad una frequenza relativa p . In questo senso la mediana diventa il quantile di ordine $p=1/2$.

Più in generale si definisce quantile di ordine p la modalità x_i tale che

- $F(x_{i-1}) < p$
- $F(x_i) \geq p$

I quantili più utilizzati sono i **percentili**, soprattutto il 25-esimo, 50-esimo (mediana) e il 75-esimo

Tutte le proprietà della mediana si estendono naturalmente ai quantili.

La variabilità

Introduciamo l'argomento della variabilità mediante un opportuno esempio:

Tre studenti, nel primo quadrimestre, hanno riportato le seguenti successioni di voti nelle prove scritte di matematica:

<i>Studente</i>	<i>Voti</i>			
Anna	5	6	6	7
Giovanni	4	5	7	8
Giuseppe	3	4	8	9

Se si calcolano le medie aritmetiche dei voti di ogni studente si vedrà che esse coincidono ($M' = M'' = M''' = 6$).

Tuttavia, a parità di media, è facile rendersi conto che le tre successioni di voti presentano una misura diversa della variabilità, definita come *l'attitudine che la grandezza in oggetto ha di assumere valori più o meno diversi tra loro*.

Se i dati sono vicini al loro valore medio, allora, intuitivamente, la variabilità è bassa. Se i dati si discostano fortemente dal loro valore medio allora, intuitivamente la variabilità è alta.

Poiché nessuno dei valori medi è in grado di darci informazioni sulla misura della variabilità dei dati, occorre introdurre nuovi indici, detti **indici di variabilità**, capaci di misurare questa grandezza.

Campo di variazione

E' il più semplice degli indici di variabilità. Esso è dato dalla *differenza tra il dato massimo e il dato minimo*. Ossia:

$$x_{max} - x_{min}.$$

Tale indice equivale all'ampiezza del minimo intervallo che contiene tutti i dati. Con riferimento all'esempio precedente si ha che:

Il campo di variazione dei voti di Anna è: $7 - 5 = 2$

Il campo di variazione dei voti di Giovanni è: $8 - 4 = 4$

Il campo di variazione dei voti di Giuseppe è: $9 - 3 = 6$

Il campo di variazione, dà soltanto una misura grossolana e insufficiente della variabilità, in quanto è influenzato soltanto dai dati minimi e massimi (i dati intermedi, infatti, non intervengono nel calcolo del suddetto indice).

Per tale motivo, può accadere che due successioni di dati abbiano stesso campo di variazione ma variabilità diverse, a causa dei dati intermedi. E' dunque indispensabile introdurre un nuovo indice di variabilità, più sensibile del precedente.

Scarto semplice medio

Consideriamo la seguente successione di dati statistici:

$$x_1, x_2, \dots, x_n$$

aventi la seguente media aritmetica:

$$M = \sum_i x_i / n$$

Le differenze sotto indicate:

$$x_1 - M, x_2 - M, \dots, x_n - M$$

tra ciascun dato e la media aritmetica si chiamano **scarti** semplici dei dati statistici dalla loro media aritmetica M .

Si verifica facilmente che la sommatoria di tutti gli scarti è uguale a zero, ossia che:

$$\sum_i (x_i - M) = 0$$

Ciò è dovuto al fatto che gli scarti semplici positivi e quelli negativi si neutralizzano a vicenda.

Per rendere utili gli scarti semplici al fine della misura della variabilità, occorre considerarli in valore assoluto.

Così facendo si ottiene la formula di un nuovo indice di variabilità, detto **scarto semplice** medio:

$$S = \sum_i |x_i - M| / n$$

Esso è così definito:

Lo scarto semplice medio è uguale alla media aritmetica degli valori assoluti degli scarti semplici di ciascun dato x dalla media aritmetica M .

Ci proponiamo di calcolare lo scarto semplice medio dei voti di Anna riferito all'esempio precedente:

Voti	Scarti dalla media	Valori assoluti scarti
$x_1=5$	$x_1-M=5-6=-1$	$ x_1-M =1$
$x_2=6$	$x_2-M=6-6=0$	$ x_2-M =0$
$x_3=6$	$x_3-M=6-6=0$	$ x_3-M =0$
$x_4=7$	$x_4-M=7-6=+1$	$ x_4-M =1$
$M=24/4=6$	$\sum_i (x_i - M) = 0$	$\sum_i x_i - M = 2$

Quindi lo scarto semplice medio risulta:

$$S = \sum_i |x_i - M| / n = 2/4 = 0.5 \text{ (cioè } \frac{1}{2} \text{ voto)}$$

Il risultato si interpreta così:

Mediamente i voti di Anna si discostano dalla loro media di una frazione pari a mezzo voto.

Tuttavia, noi ci proponiamo di dimostrare come, *a parità di media e di campo di variazione, due successioni di dati possono avere misure diverse della variabilità.*

A tal fine consideriamo la successione di voti di Paolo, un quarto studente:

5, 5½, 6½, 7.

Se calcoliamo la media e il campo di variazione di questi dati ci accorgiamo che sono identici a quelli di Anna (ossia $M=6$, $C.v.=2$).

Tuttavia se calcoliamo lo scarto semplice medio dei voti di Paolo, si ha:

$$S = \sum_i |x_i - M| / n = 3/4 = 0.75 \text{ (cioè } \frac{3}{4} \text{ voto)}$$

Pertanto i voti di questo studente, discostandosi dalla media del 6 mediamente di $\frac{3}{4}$ voto, presentano una variabilità maggiore dei voti di Anna.

Lo scarto semplice medio, si dimostra inadeguato in alcuni casi, per cui si introduce un nuovo indice, detto **scarto quadratico medio**.

Scarto quadratico medio

Consideriamo la seguente successione di dati statistici:

$$x_1, x_2, \dots, x_n$$

e sia:

$$M = \sum_i x_i / n$$

la loro media aritmetica.

Calcoliamo poi gli scarti semplici dei dati dalla media:

$$x_1 - M, x_2 - M, \dots, x_n - M$$

ed elevando al quadrato ciascuno di essi:

$$(x_1 - M)^2, (x_2 - M)^2, \dots, (x_n - M)^2$$

otteniamo lo scopo di renderli tutti non negativi, evitando così che la loro somma dia zero; il risultato ottenuto costituisce la successione degli **scarti quadratici**.

Se calcoliamo la media aritmetica di questi scarti quadratici:

$$\sum_i (x_i - M)^2 / n$$

ricaviamo un indice di variabilità, detto **varianza**.

Infine, per tornare alla stessa unità di misura dei dati iniziali, eseguiamo la radice quadrata della varianza, ottenendo per risultato quell'importante indice di variabilità che si chiama **scarto quadratico medio**:

$$\sigma = \sqrt{\frac{\sum_i (x_i - M)^2}{n}}$$

Calcolando i due indici S e σ su una identica successione di dati, si noterà che essi conducono a risultati quantitativamente diversi. Ciò è irrilevante, in quanto ciò che interessa è il confronto con indici analoghi di un'altra successione di dati.

Indici di variabilità per distribuzioni di frequenze

Negli esempi precedenti, abbiamo visto esempi di calcolo degli indici di variabilità per successioni di dati.

Vediamo ora un esempio di calcolo degli indici di variabilità σ^2 (varianza) e σ (scarto quadratico medio) con dati organizzati per frequenza.

Le formule per il calcolo degli indici suddetti vengono così modificate:

$$\sigma^2 = \frac{\sum (x_i - M)^2 f_i}{\sum f_i}$$

e

$$\sigma = \sqrt{\frac{\sum (x_i - M)^2 f_i}{\sum f_i}}$$

Esempio: *Distribuzione delle partite di calcio dello scorso campionato di calcio di serie A per numero di gol segnati:*

N. gol (x_i)	f_i	$(x_i - M)$	$(x_i - M)^2$	$(x_i - M)^2 f_i$
0	36	-2.65	7.0225	252.81
1	51	-1.65	2.7225	138.8475
2	80	-0.65	0.4225	33.8
3	52	+0.35	0.1225	6.37
4	36	+1.35	1.8225	65.61
5	22	+2.35	5.5225	121.495
6	18	+3.35	11.2225	202.005
7	7	+4.35	18.9225	132.4575
8	4	+5.35	28.6225	114.49
Totale	$\Sigma f_i = 306$			$\Sigma (x_i - M)^2 f_i = 1067.885$

La media M , vale:

$$M = \frac{\sum x_i f_i}{\sum f_i}$$

per cui:

$$(0*36+1+51+2*80+3*52+4*36+5*22+6*18+7*7+8*4)/306 = 2.65 \text{ (M)}$$

la varianza è:

$$\sigma^2 = \frac{\sum (x_i - M)^2 f_i}{\sum f_i} = \frac{1067.885}{306} = 3.4898$$

e lo scarto quadratico medio:

$$\sigma = \sqrt{\frac{\sum (x_i - M)^2 f_i}{\sum f_i}} = \sqrt{\frac{1067.885}{306}} = 1.8681$$

Indici di variabilità relativi

Gli indici E (campo di variazione), S (scarto semplice medio), σ (scarto quadratico medio) e σ^2 (varianza) che abbiamo introdotto nel capitolo precedente, sono indici **assoluti**, ossia espressi nella stessa unità di misura dei dati elaborati (nell'esempio considerato erano voti o frazioni di voto).

Gli indici assoluti servono solo per confrontare la variabilità di due successioni di dati omogenei, cioè misurabili con la stessa unità di misura. Non avrebbe senso, infatti, confrontare direttamente, per esempio, voti e temperature.

Per poter effettuare il confronto tra due successioni di dati non omogenei, occorre svincolarsi dalle rispettive unità di misura.

Tale obiettivo si raggiunge introducendo nuovi indici, detti **indici di variabilità relativi**.

Essi sono numeri puri che si ottengono, in generale, dai *rapporti degli indici assoluti con la media aritmetica dei dati*.

Quindi, in simboli, si ha:

1) Il campo di variazione relativo E_r :

$$E_r = \frac{E}{M} ;$$

2) Lo scarto semplice medio relativo S_r :

$$S_r = \frac{S}{M}$$

3) Lo scarto quadratico medio relativo σ_r :

$$\sigma_r = \frac{\sigma}{M}$$

4) La varianza relativa σ_r^2 :

$$\sigma_r^2 = \frac{\sigma^2}{M}$$

Per fissare le idee, utilizzando i dati dell'ultimo esempio, otterremmo, relativamente allo scarto quadratico medio e alla varianza:

$$\sigma_r = \frac{\sigma}{M} = \frac{1.8681}{2.65} = 0.7049$$

e

$$\sigma^2_r = \frac{\sigma^2}{M} = \frac{3.4898}{2.65} = 1.3169$$

Per confrontare la variabilità di due fenomeni distinti, riconsideriamo i dati della tabella dei voti riportati da Giuseppe nelle prove scritte di matematica (vedi capitolo precedente):

Voti (x_i)	$(x_i - M)$	$(x_i - M)^2$
3	-3	9
4	-2	4
8	+2	4
9	+3	9
$M = 24/4 = 6$	$\Sigma(x_i - M) = 0$	$\Sigma(x_i - M)^2 = 26$

Da essi si ottiene:

$$\sigma = \sqrt{\frac{\sum (x_i - M)^2}{n}} = \sqrt{\frac{26}{4}} = 2.5495$$

e

$$\sigma^2 = \frac{\sum (x_i - M)^2}{n} = \frac{26}{4} = 6.5$$

Se adesso calcolo i rispettivi indici relativi, ottengo:

$$\sigma_r = \frac{\sigma}{M} = \frac{2.5495}{6} = 0.4249$$

e

$$\sigma^2_r = \frac{\sigma^2}{M} = \frac{6.5}{6} = 1.0833$$

per cui, dal confronto con gli indici dell'esempio precedente, essendo in questo caso sia σ_r che σ^2_r numericamente inferiori, concludiamo affermando che la *distribuzione delle partite di calcio dello scorso campionato per numero di gol segnati* presenta una variabilità superiore rispetto alla *distribuzione dei voti di Giuseppe nelle prove scritte di matematica*.

Importanza dello scarto quadratico medio

Lo scarto quadratico medio, detto anche scarto tipico, fra tutti gli indici di variabilità è il più importante.

Esso è ovviamente preferibile a quel grossolano indice che è il cmapo di variazione E ed è inoltre preferibile anche allo scarto semplice medio S , nonostante quest'ultimo sia molto più semplice da calcolare.

I motivi che ci inducono ad affermare che σ è più significativo di S sono i seguenti:

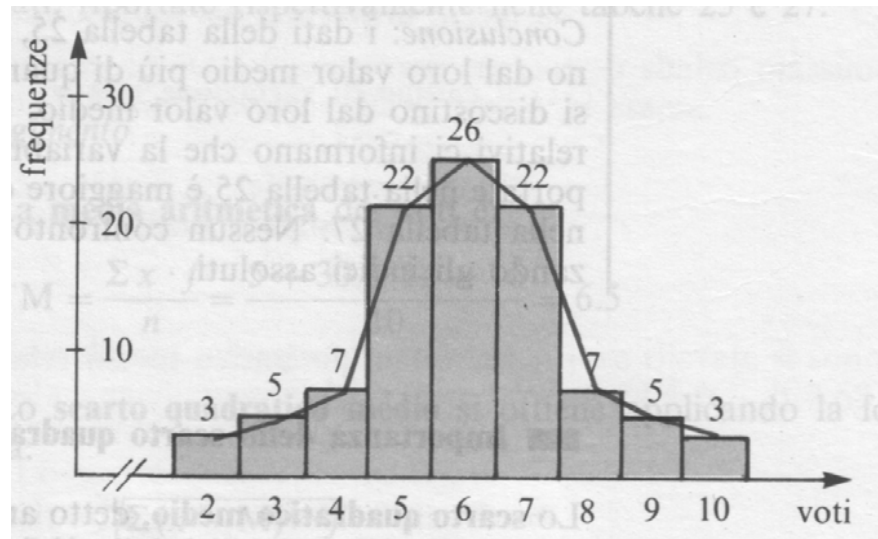
- 1) σ risulta sempre maggiore o uguale a S , cosicché attraverso di esso si possono meglio rilevare anche le più piccole differenze della variabilità, allorché si confrontano due insiemi di dati statistici.
- 2) σ è, in generale, un indice più sensibile di S , dimostrandosi capace di percepire più intensamente di quest'ultimo anche lievissimi mutamenti della variabilità.
- 3) σ è molto importante per lo studio di una notevolissima distribuzione di frequenze, chiamata **distribuzione normale**.

La distribuzione normale

Consideriamo adesso questo tipo di distribuzione e il ruolo che in essa svolge lo scarto quadratico medio σ . Osserviamo, la seguente tabella:

voti	2	3	4	5	6	7	8	9	10
frequenze	3	5	7	22	26	22	7	5	3

Essa riporta i risultati di un test di ingresso di matematica, assegnato a un campione casuale di $n=100$ studenti iscritti al 1° anno dell'ITIS “Copernico”:



Il corrispondente istogramma, ci fornisce le seguenti indicazioni:

1. La **media aritmetica**, la **mediana** e la **moda** sono coincidenti nel voto 6;
2. Il grafico è a forma di **campana**; tale forma risulta maggiormente evidente se si congiungono i punti medi delle basi superiori dei rettangoli, ottenendo così una spezzata detta poligono delle frequenze;
3. La maggior parte dei voti è **addensata** nelle vicinanze della **media** (ben 70 studenti su 100 hanno riportato voti appartenenti all'intervallo $[5, 7]$ il cui centro è la media 6). Inoltre, via via che i voti si discostano dalla media aritmetica, sino a diventare molto bassi o molto alti, notiamo che le frequenze corrispondenti decrescono velocemente, tendendo ad assumere valori prossimi allo zero (ciò significa che in un campione casuale piuttosto numeroso, normalmente sono pochi gli studenti che si discostano fortemente dalla media nei due sensi, ossia sono pochi gli studenti completamente impreparati e altrettanto pochi quelli eccezionalmente preparati).

Osservazione: Tutte le volte che una distribuzione di frequenze porta a un risultato simile a quello visto nell'esempio, si dice che essa è più o meno analoga alla *distribuzione normale*.

Quest'ultima è una distribuzione ottenibile in circostanze ideali, avente come rappresentazione grafica una curva perfettamente a *campana*, denominata *curva normale* o di *Gauss*.

La curva normale è il grafico di una notevole funzione matematica del tipo:

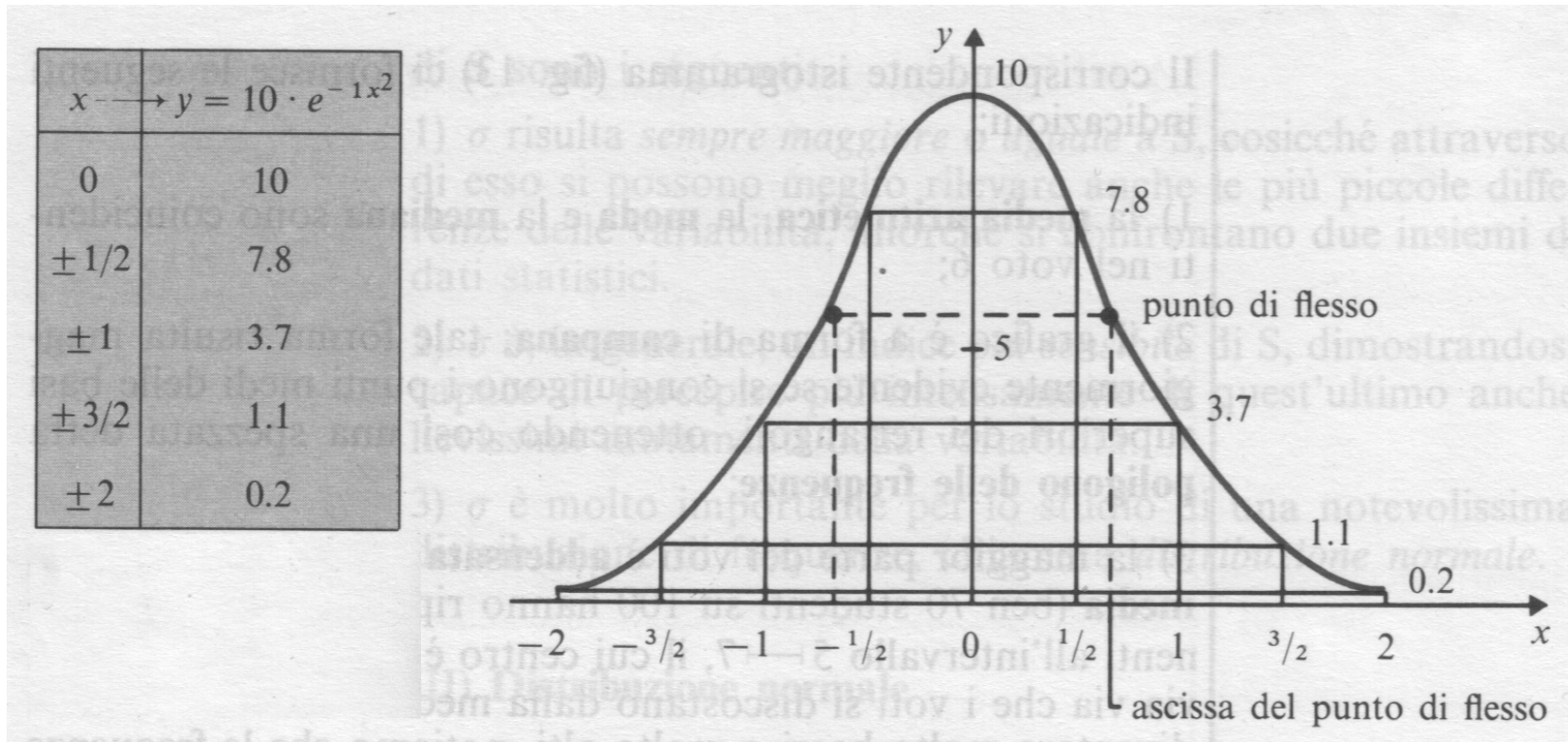
$$y = k \cdot e^{-h \cdot x^2}$$

detta *funzione di Gauss*, in cui i numeri h e k sono due costanti reali positive, mentre la base e è il famoso numero di Nepero, conosciuto anche come la base dei logaritmi cosiddetti neperiani o naturali, uguale a 2,7182....

Se per chiarirci le idee, proviamo a tracciare il grafico delle seguente funzione di Gauss

$$y = 10 \cdot e^{-1 \cdot x^2}$$

ottenuta per $k=10$ e $h=1$, risulta una curva perfettamente a campana.



Come si osserva facilmente, **la curva di Gauss**:

1. Presenta il **massimo** valore della y per $x=0$ (ossia nell'origine);
2. E' **simmetrica** rispetto all'asse delle ordinate;
3. Ha come **asintoto** *l'asse delle ascisse*;
4. Presenta **due** punti notevoli detti **flessi** (la curva quando li attraversa cambia concavità).

Il ruolo di σ in una distribuzione normale

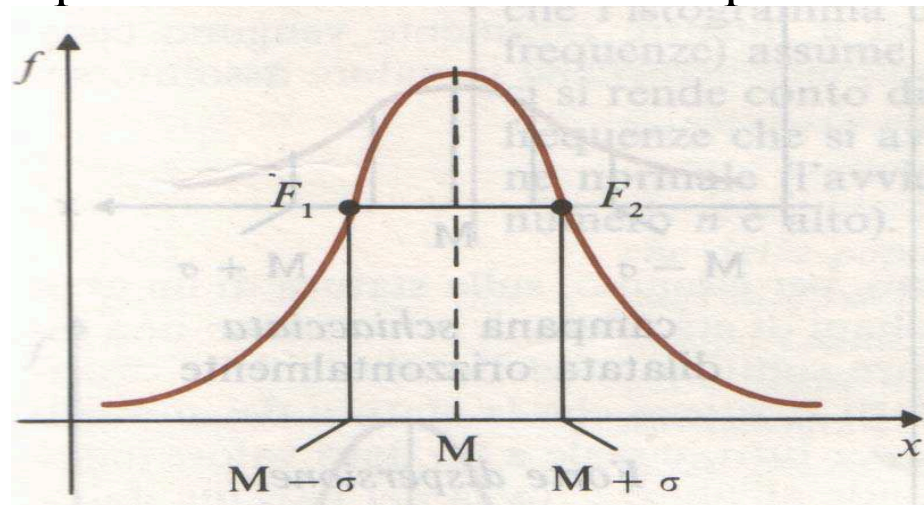
Abbiamo visto che la curva di Gauss presenta *due punti di flesso*, simmetrici rispetto all'asse della campana. In corrispondenza a tale asse di simmetria, in una distribuzione normale, si trova la media aritmetica **M** dei dati.

Si dimostra che i suddetti punti di flesso hanno ascisse funzioni di σ , rispettivamente uguali a:

flesso sinistro: $M - \sigma$

flesso destro: $M + \sigma$

ossia, graficamente si presenta una curva simile a quella illustrata in figura:



Si dimostra ancora che:

1. In una distribuzione normale, **68.27 unità statistiche su 100** (il 68.27% della *popolazione*) hanno intensità del carattere, cioè valori della x , appartenenti all'intervallo: $(M-\sigma, M+\sigma)$. In altre parole, 68.27 unità statistiche su 100 hanno intensità del carattere che si discostano dalla media **M** , di uno scarto semplice assoluto minore di **σ** .
2. Nella stessa distribuzione, **95.45 unità statistiche su 100** hanno intensità del carattere appartenenti all'intervallo: $(M-2\sigma, M+2\sigma)$, ossia quasi tutte le unità statistiche della popolazione hanno intensità che differiscono da **M** , in valore assoluto, meno di **2σ** .

Ad esempio: Se tra una popolazione di 1000 persone si osserva un peso medio $M=70$ kg, con scarto quadratico medio $\sigma=5$ kg, poiché i pesi delle persone, come è noto, hanno frequenze che seguono una legge normale, si può affermare che:

- Circa 683 persone su 1000 hanno un peso compreso tra 65 e 75 kg;
- Circa 955 persone, su un totale di 1000, hanno un peso compreso tra 60 e 80 Kg.

Il valore di σ influisce in maniera determinante sulla forma della campana. Si dimostra che a seconda del valore di σ , si possono avere le seguenti situazioni:

