ALCUNI ELEMENTI DI STATISTICA DESCRITTIVA

The last step of reason is to acknowledge that there is an infinity of things that go beyond it.
B. Pascal

La Statistica ha come scopo la *conoscenza quantitativa dei fenomeni collettivi.* L'analisi statistica mira ad individuare ***modelli*** di interpretazione della realtà, attraverso canoni e tecniche che sono astrazioni, semplificazioni di una moltitudine di aspetti e di manifestazioni del reale.

Nell'analisi statistica si può individuare una:

– FUNZIONE **DESCRITTIVA**, in quanto offre il metodo per riassumere le informazioni in modo da renderle utilizzabili più facilmente. Riduce i dati in forma maneggevole, sostituendo a molti numeri poche misure.

– FUNZIONE **INFERENZIALE** (o induttiva), in quanto permette di generalizzare le informazioni, ricavando proprietà e leggi generali sulla base di dati rilevati solamente su una parte (campione) della popolazione

OGGETTO DELL'ANALISI STATISTICA

Si tratta di analizzare un ***fenomeno collettivo***, fenomeno che si conosce tramite la sintesi di osservazioni di fenomeni semplici (es. natalità, mortalità, reddito nazionale, statura di un insieme di individui, ecc.).

Ogni elemento su cui si osserva il fenomeno in esame è detta ***unità statistica*** **(u.s.)** e l'insieme delle u.s. oggetto dell'osservazione costituisce il ***collettivo statistico***. L'***unità di rilevazione*** a volte può non coincidere con l'unità statistica che forma il collettivo, ad esempio nei censimenti demografici le unità statistiche sono gli individui mentre le unità di rilevazione sono le famiglie. La definizione corretta dell'unità statistica e di quella di rilevazione è condizione indispensabile per l'esattezza di una ricerca: l'inclusione o esclusione errata di numerose unità può portare a conclusioni fuorvianti.

La ***popolazione*** è un collettivo statistico. Un ***campione*** è un sottoinsieme della popolazione di riferimento, opportunamente definito in funzione degli obiettivi dell'indagine.

Il fenomeno collettivo viene studiato mediante l'osservazione o la misurazione di una o più ***caratteristiche*** delle u.s. Ogni caratteristica è detta ***carattere o variabile statistica*** ed il "modo" in cui il carattere si manifesta si chiama ***modalità***.

Esempio: Fenomeno collettivo: natalità;
Collettivo statistico: collettivo dei nati (ad es. solo i legittimi, solo gli illegittimi, o il totale);
Unità statistica: il singolo nato;
Possibili caratteri da analizzare: ordine di nascita;
età della madre al parto;
vitalità (nati vivi o nati morti);
regione di nascita;
peso;
statura;
sesso; ecc.

Modalità: sesso: Maschio o Femmina;
statura: 50 cm, 49 cm, 52 cm, ecc.;
regione: Abruzzo, Calabria, Lazio, Piemonte, Umbria, ecc.
ordine di nascita: primogenito, secondogenito, terzogenito, ecc.

I caratteri (variabili statistiche) hanno diversa natura. Alcuni si esprimono con sostantivi, aggettivi, ..., sono i ***caratteri qualitativi*** (sesso, ordine di nascita, regione); altri con numeri e sono i ***caratteri quantitativi*** (statura, peso, numero componenti la famiglia)

CLASSIFICAZIONE DEI CARATTERI

Carattere	Tipologia	Esempi
QUALITATIVO	**NOMINALE** Date due qualsiasi modalità, è possibile solo affermare se esse sono uguali o diverse.	sesso; professione; diagnosi medica; …
	ORDINALE O PER RANGHI Esiste un criterio predeterminato per ordinare le modalità.	ordine di nascita; giorni della settimana; indice di severità di una malattia;…
QUANTITATIVO	**DISCRETO** L'insieme delle modalità assumibili può essere messo in "corrispondenza biunivoca" con un sottoinsieme dei numeri naturali.	Num. componenti famiglia; num. di figli; num. di denti; num. colonie batteriche in una piastra;…
	CONTINUO L'insieme delle modalità assumibili può essere messo in "corrispondenza biunivoca" con un sottoinsieme dei numeri reali (la variabile può assumere qualsiasi valore all'interno di intervalli di numeri reali).	statura; peso; glicemia; PAS;…

Stabilito il fenomeno collettivo da studiare occorre individuare il collettivo su cui studiarlo ed i caratteri da rilevare e, nel caso di *indagine campionaria*, un campione "significativo" di unità statistiche.

Esempio 1. Su un campione di pazienti si rilevino le caratteristiche: sesso, età, altezza, peso, PAS, tasso glicemico. Di seguito sono riportate 4 "schede" di rilevazione:

nome: Rossi Amerigo	Nome: Bianchi Paolo
sesso: maschio	Sesso: maschio
età: 32	Età: 47
altezza: 172 cm.	Altezza: 170 cm.
peso: 64 Kg.	Peso: 80 Kg.
PAS: 140 mm Hg.	PAS: 148 mm Hg.
Glicemia: 190 mg/100cc	Glicemia: 180 mg/100cc
nome: Valenziani Alberica	nome: Alinori Alfonso
Sesso: femmina	sesso: maschio
età: 45	età: 27
Altezza: 168 cm.	Altezza: 183 cm.
Peso: 51 Kg.	peso: 85 Kg.
PAS: 125 mm Hg.	PAS: 138 mm Hg.
Glicemia: 150 mg/100cc	glicemia: 170 mg/100cc

Le informazioni raccolte per essere "trattate" da un computer devono essere organizzate in strutture chiamate comunemente ***Base di Dati*** (Data Base o File Dati).
Le informazioni vengono, comunemente, organizzate per riga, cioè su ogni riga, consecutivamente, vengono elencati i dati relativi ad un soggetto.

N.	NOME	SESSO	ETA'	ALTEZZA	PESO	PAS	GLICEMIA
1	Rossi Amerigo	M	32	172	64	140	190
2	Bianchi Paolo	M	47	170	80	148	180
3	Valenziani Alberica	F	45	168	51	125	150
4	Alinori Alfonso	M	27	183	85	130	170
5							
6							

Le unità archiviate (righe) si chiamano ***records*** e le informazioni che costituiscono i record si chiamano ***items*** e lo spazio occupato da ciascuno di essi si dice ***campo***.
Ogni colonna contiene la sequenza di tutti i dati relativi ad una caratteristica esaminata nel campione.
L'insieme dei record costituisce il ***file*** (file dati).

Le operazioni che, in genere, un Base di Dati permette di fare sono:

- inserimento di nuovi records;
- ricerca di dati già introdotti con un sistema di interrogazioni;
- correzione o aggiornamento dei records;
- riordinamento per chiavi dei records;
- stampa dei dati organizzati in cartelle o tabulati;
- semplici o più complesse statistiche (descrittive) sui dati.

Esistono dei pacchetti ***software*** per la gestione dei dati:

- **DBMS** (*Data Base Management System*).

Questi possono «esportare» i dati verso altri software che permettono una elaborazione più evoluta delle informazioni.
Si possono utilizzare allora dei

- ***fogli elettronici*** (Excel, ecc.)

o dei

- ***packages statistici*** (SPSS, SAS, BMDP, ecc.).

Fatta le rilevazione con lo spoglio dei dati (operazione che determina le modalità del carattere in ciascuna unità statistica) si perviene alla Distribuzione del collettivo secondo le modalità del carattere o dei caratteri studiati.

Se la distribuzione è relativa ad un solo carattere si dice distribuzione semplice, se a due caratteri è una distribuzione doppia, se, invece, riguarda più caratteri si dice multipla.

DISTRIBUZIONI SEMPLICI DI FREQUENZE

I dati (cioè le informazioni raccolte) spesso sono di non immediata lettura. La prima esigenza è dunque quella di rendere i dati più facilmente interpretabili. Perciò si procede ad una sistematizzazione e sintesi delle informazioni raccolte, cioè alla loro **tabulazione**. Per ogni variabile si calcolano le **frequenze assolute (f.a.)** che rappresentano il numero di u.s. che presentano una stessa modalità del carattere.

Esenpio 2. Alcune distribuzioni semplici di frequenze.

Sesso	**f.a.**
M	2
F	10
Tot	12

Età	**f.a.**
17	3
18	6
19	12
20	1
Tot	22

Altezza	**f.a.**
150-160	2
160-170	10
170-180	15
180-190	7
>190	1
Tot	35

Spesso ci si trova nella necessità di dover fare confronti ad es. se si vuole stabilire in quale, fra i seguenti gruppi di persone, ci siano più donne:

Esempio 3. Distribuzione doppia di frequenze assolute

	Gruppo A	Gruppo B
Sesso	frequ. assolute	frequ. assolute
M	12	7
F	16	10
Tot	28	17

Ci accorgiamo che il confronto non può essere effettuato solo con le f.a. in quanto esse si riferiscono a collettivi di numerosità diversa. Se vogliamo confrontare le frequenze le dobbiamo "depurare" dalla numerosità del collettivo; ciò lo si fa dividendo le f.a. per la numerosità (N) della popolazione

e moltiplicando per 100 (cioè facendo riferimento ad una ipotetica popolazione di 100 unità). Le frequenze così calcolate sono le **frequenze percentuali** (f.%)

Esempio 3'. Distribuzione doppia di frequenze percentuali

	Gruppo A		Gruppo B	
Sesso	f.a.	f.%	f.a.	f.%
M	12	42.9	7	41.2
F	16	57.1	10	58.8
Tot	28	100	17	100

In molti casi oltre alle f.a. e f.% è utile calcolare le **frequenze cumulate** assolute e %.

Esempio 4. Distribuzione di frequenze assolute, relative e cumulate

Età	**f.a.**	**f.%**	**f.a.cum**	**f%cum**
17	3	13.6	3	13.6
18	6	27.3	9	40.9
19	12	54.6	21	95.5
20	1	4.5	22	100
Tot	22	100		

Le frequenze cumulate indicano quante u.s. si presentano fino a quella modalità**.**

Ha senso calcolare le f.cum solamente per le variabili quantitative o qualitative ordinabili.

I GRAFICI STATISTICI

Scopo dei grafici è quello di rendere l'informazione contenuta in una serie di dati:
- di più facile comprensione;
- di più diretta lettura;

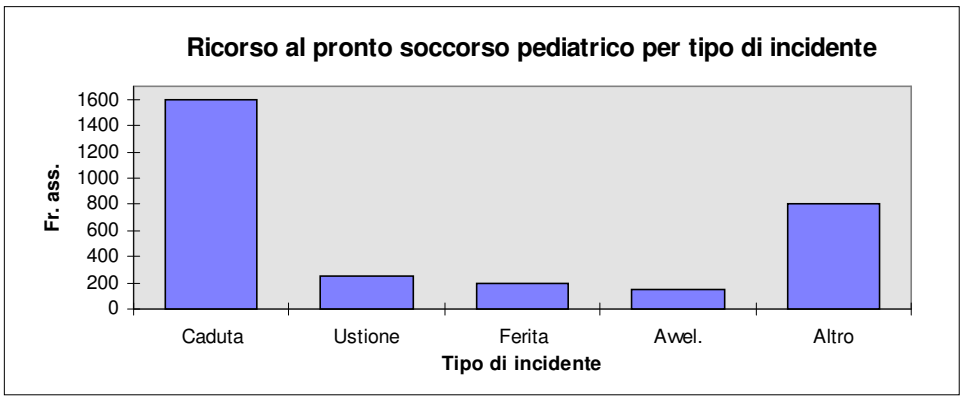
pertanto un grafico deve fornire al lettore una informazione sintetica e facile da interpretarsi.

Una rappresentazione grafica diventa indispensabile nel caso di indagini di elevate dimensioni poiché lunghe serie di dati non sono sempre idonee alla comprensione di fenomeni.

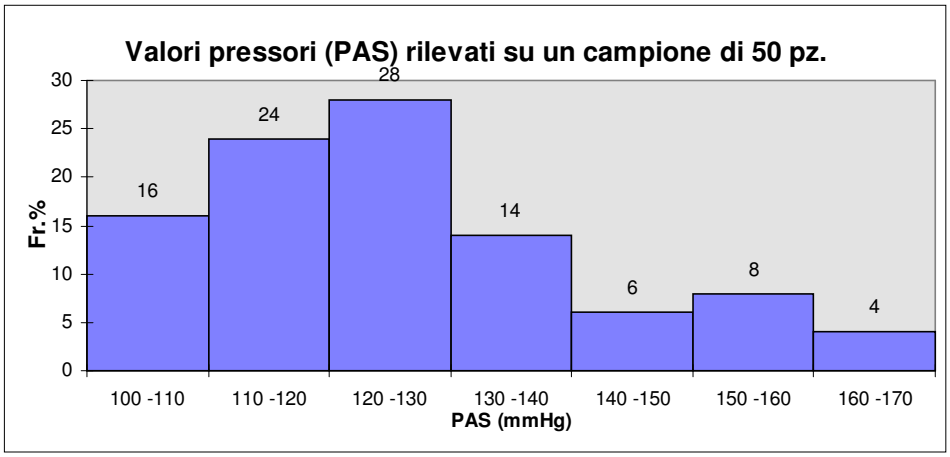
Diagrammi cartesiani: grafici che hanno come riferimento un sistema di assi cartesiani con asse orizzontale x (ascissa) ed asse verticale y (ordinata). Ogni punto viene identificato da una coppia ordinata di valori (x, y).

Diagrammi a bastoncino: indicati per variabili qualitative, evidenziano con la lunghezza del segmento le frequenze delle modalità della variabile.

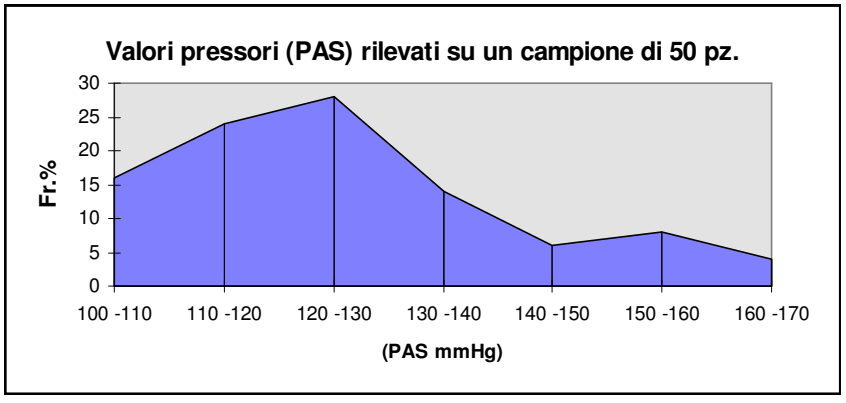
Ortogrammi: usati più frequentemente dei precedenti e si ottengono sostituendo ai bastoncini delle barre.



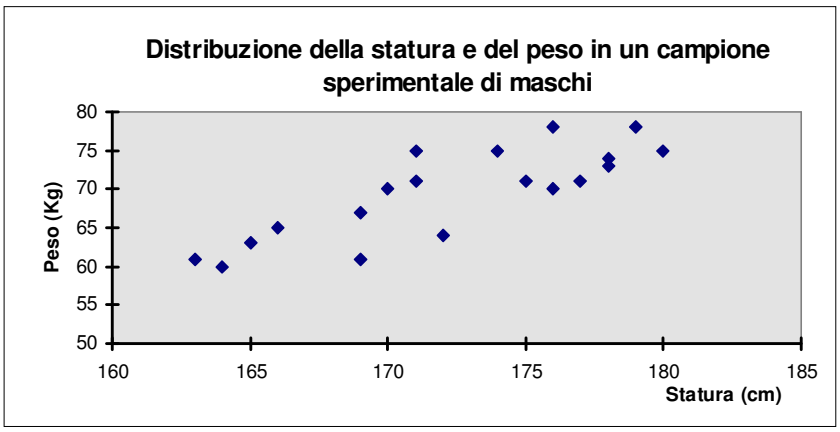
Istogrammi: indicati per rappresentare distribuzioni in classi (variabili quantitative continue). Costituiti da una serie di barre rettangolari contigue ognuna in rappresentanza di una classe e con area proporzionale alla rispettiva frequenza.



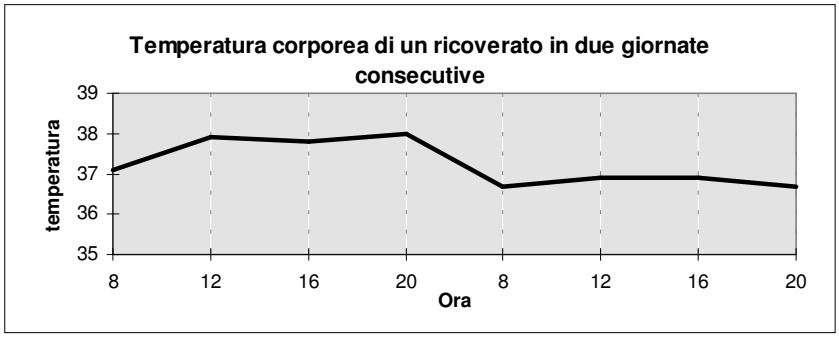
Poligoni e curve di frequenza: si ottengono dai precedenti unendo i valori centrali superiori delle classi.



Grafici per punti (diagrammi a dispersione): costituito dai punti corrispondenti alle diverse coppie di valori rilevati. Indicati per evidenziare le associazioni tra variabili quantitative.

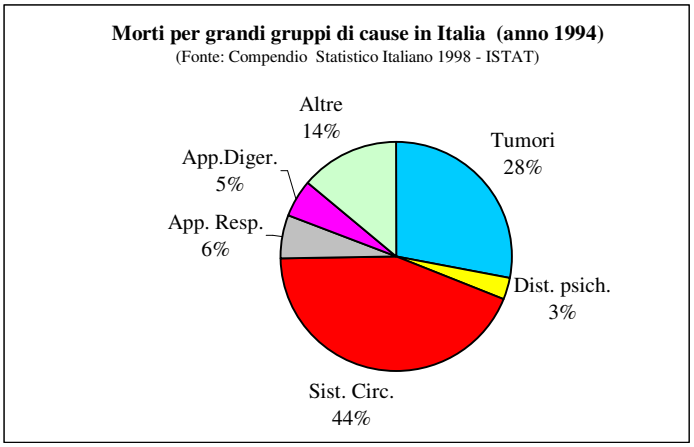


Grafici per spezzate: si ottengono dai grafici per punti congiungendo i vari punti. Indicati per evidenziare una continuità tra valori come ad es. nella rappresentazione delle serie temporali.
Es. : temperatura corporea, pressione sanguigna, tracciato elettrocardiografico.



Diagrammi logaritmici: uno o entrambi gli assi sono in scala logaritmica indicati per rappresentare misure quantitative espresse su ordini di grandezza così differenziati che non possono essere rappresentati su scala decimale.

Diagrammi a settori circolari (torte): indicati per variabili qualitative allo scopo di evidenziare le frequenze % delle singole modalità. L'area di un cerchio viene suddivisa in settori proporzionali alle frequenze %

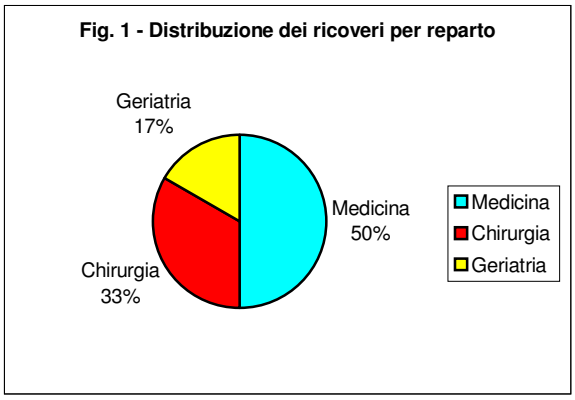


Esempio Grafici. In una Azienda Ospedaliera è stata rilevata le frequenze annua di ricoveri per i reparti indicati in tabella seguente:

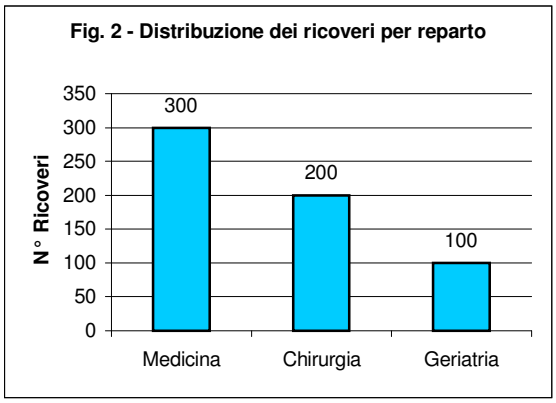
REPARTO	N° RICOVERI
Medicina	300
Chirurgia	200
Geriatria	100

Per visualizzare tali informazione si può far ricorso ad un *diagramma a torta* o ad un *ortogramma:*

Diagramma a torta

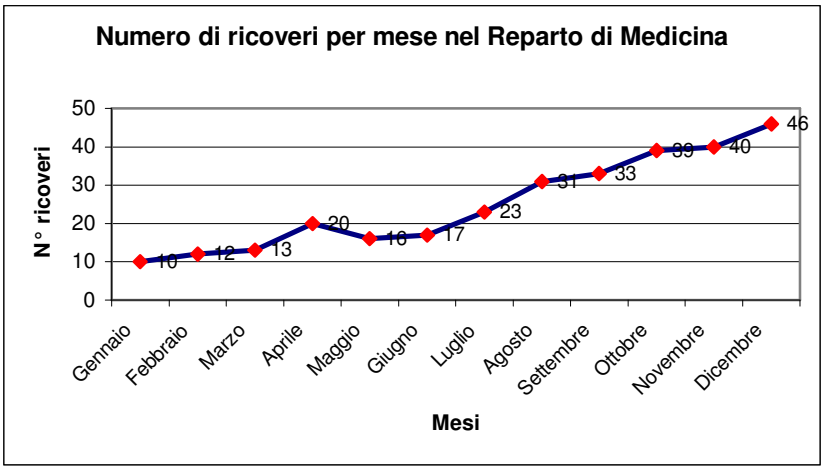


Ortogramma



Avendo a disposizione anche la distribuzione del numero di ricoveri per mese presso il Reparto di Medicina, si può utilizzare *un grafico a spezzata* per visualizzare l'andamento temporale del numero di ricoveri:

Mese	N° Ricov.
Gennaio	10
Febbraio	12
Marzo	13
Aprile	20
Maggio	16
Giugno	17
Luglio	23
Agosto	31
Settembre	33
Ottobre	39
Novembre	40
Dicembre	46
Totale	300



MISURE DI SINTESI DEI DATI

La sintesi numerica di una distribuzione di dati si basa sulla introduzione di ***indici numerici*** che permettono di evidenziare alcuni aspetti essenziali della distribuzione analizzata.

Tali misure si possono classificare in tre famiglie principali:

- indici di tendenza centrale;
- indici di variabilità o di dispersione;
- indici di forma.

INDICI DI TENDENZA CENTRALE

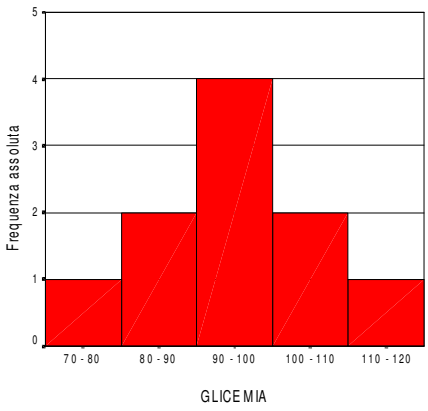
Gli indici di tendenza centrale esprimono dei valori "intorno" ai quali si può ritenere concentrato il carattere statistico di interesse, fornendo un'idea sintetica del fenomeno oggetto di indagine.

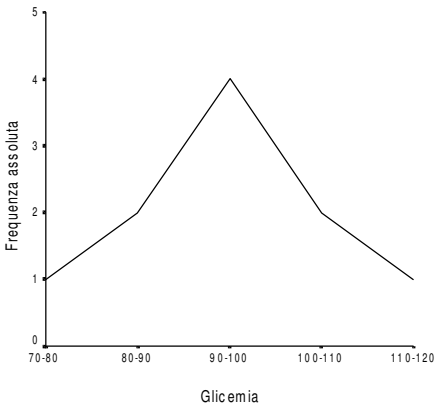
Esempio 5. Nella tabella seguente sono riportati i valori del tasso glicemico rilevati su 10 pazienti:

X	Glicemia (mg/100cc)
x_1	103
x_2	97
x_3	90
x_4	119
x_5	107
x_6	71
x_7	94
x_8	81
x_9	92
x_{10}	96

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 (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 arbitraria e dipende dalla numerosità campionaria.

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 %





Istogramma

Si tratta ora di scegliere una misura di tendenza centrale più appropriata per "sintetizzare" la distribuzione in esame.

Gli indici di tendenza centrale sono:

- i valori medi o medie algebriche (es. media aritmetica, media armonica, media geometrica);
- indici di posizione o medie lasche (es. mediana, moda, quartili,).

Per i dati quantitativi (variabili statistiche quantitative) si possono utilizzare sia le medie algebriche che gli indici di posizione, mentre per i caratteri qualitativi è necessario far ricorso a indici di posizione.

Valori medi

Considerati i valori x_1, x_2, x_3,…,x_n osservati del carattere X e una opportuna funzione $f(x_1,x_2,x_3,\ldots,x_n)$, un valore ***m*** si dice ***valor medio*** di X rispetto alla valutazione di f se risulta:

$$f(m,m,m,\ldots,m) = f(x_1,x_2,x_3,\ldots,x_n).$$

E' questa la definizione di media secondo Chisini.

La **media aritmetica** 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** dati ne fa rimanere costante la somma:

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

Nell'Esempio 5 si ha:

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

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

In questo la media aritmetica (**ponderata**) 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) $\sum_i (x_i - \bar{x}) = 0$: 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 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.

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

Esempio 8. A 5 persone viene effettuato il trapianto di uno stesso organo. Dopo 5 anni si rileva la seguente situazione:
1 persona morta dopo 2 anni dal trapianto;
1 persona morta dopo 3 anni;
1 persona morta dopo 4 anni;
2 persone risultano sopravviventi.
Si vuole valutare la sopravvivenza media dei pazienti.
1^ possibilità: alle persone ancora in vita si attribuisce una sopravvivenza di 5 anni:

$$\bar{x} = \frac{2+3+4+2\cdot 5}{5} = 3.8\ anni$$

2^ possibilità: le persone ancora in vita vengono escluse dal calcolo della media:

$$\bar{x} = \frac{2+3+4}{3} = 3\ anni$$

3^ possibilità: si attribuisce "grande" sopravvivenza alle persone ancora in vita e la media viene determinata attraverso la:

$$\bar{x} = \frac{5}{\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+0+0} = 4.6\ anni$$

Esempio 9. Per illustrare le media armonica si consideri il seguente esempio relativo all'analisi di posti-letto di strutture ospedaliere in alcune regioni italiane nell'anno 1996 (fonte ISTAT).

Regione	Abitanti (Conoscere l'Italia 1997)	Letti (Compendio Stat. Italiano 1998)	Abitanti/Letto
Piemonte	4.294.127	24.850	172,80
Lombardia	8.958.670	58.574	152,95
Emilia Romagna	3.937.924	25.224	156,12
Totale	*17.190.721*	*108.648*	
Toscana	3.524.670	21.217	166,12
Umbria	829.915	4.436	187,09
Lazio	5.217.168	42.660	122,30
Totale	*9.571.753*	*68.313*	
Abruzzo	1.273.665	8.792	144,87
Campania	5.785.352	27.988	206,71
Calabria	2.074.157	11.221	184,85
Totale	*9.133.174*	*48.001*	

Si osservi che l'ultima colonna della tabella precedente riporta l'indice abitanti/letto, ovvero il numero di abitanti che "fanno capo" ad un posto letto.
Supponendo di voler confrontare gli indici *"medi"* abitanti/letto relativi alle regioni geografiche del Nord (Lombardia, Piemonte, Emilia Romagna), del Centro (Toscana, Umbria, Lazio) e del Mezzogiorno (Abruzzo, Campania, Calabria), come vanno calcolati i valori medi ?
Provando con la media aritmetica, si ottiene:

Media Nord = (172,80+152,95+156,12)/3 = 160,62,
Media Centro = (166,12+187,09+122,30)/3 = 158,50,
Media Mezzog. = (144,87+206,71+184,85)/3 = 178,81.

Ma tale modo di calcolo è errato in quanto, se da tali valori medi si risale alla frequenza assoluta di posti letto, si osserva che il numero di letti non è esatto, infatti:

posti letto Nord = 17.190.721/160,62 = 107.027,
posti letto Centro = 9.571.753/158,50 = 60.390,
posti letto Mezzog. = 9.133.174/178,81 = 51.076,

per il Nord c'è una differenza di circa 1621 posti letto in meno, per il Centro di circa 7923 posti letto in meno e per il Mezzogiorno la differenza è di circa 3075 posti letto in più.
In realtà, in questo caso, va effettuata la media utilizzando la ***media armonica***:

$$\text{Media Nord} = \frac{17.190.721}{\frac{4.294.127}{172,80} + \frac{8.958.670}{152,95} + \frac{3.937.924}{156,12}} = 158,22$$ corrispondente a 108.651 posti letto,

$$\text{Media Centro} = \frac{9.571.753}{\frac{3.524.670}{166,12} + \frac{829.915}{187,09} + \frac{5.217.168}{122,30}} = 140,12$$ corrispondente a 68.311 posti letto

$$\text{Media Mezzog.} = \frac{9.133.174}{\frac{1.273.665}{144,87} + \frac{5.785.352}{206,71} + \frac{2.074.157}{184,85}} = 190,27$$ corrispondente a 48.001 posti letto

e in questo modo, come evidentemente risulta, si ottengono valori medi corretti (le differenze riscontrate nel numero di posti letto sono dovute ad approssimazioni numeriche).

In generale, la **Media armonica (Ma)** è quel valore tale che il suo reciproco, sostituito ai dati (che devono essere tutti positivi), fa rimanere invariata la somma dei reciproci dei dati stessi:

$$\text{Ma} = = \frac{n}{f(x_1, x_2, ...x_n)} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

La **Media geometrica (Mg)**, infine, è quel numero che sostituito ai dati (che devono essere tutti positivi) fa rimanere costante il loro prodotto:

$$\text{Mg} = \sqrt[n]{f(x_1, x_2, ...x_n)} = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n}.$$

Esempio 10. Si supponga che in un laboratorio di ricerca si sia rilevato che il numero di batteri in una data cultura è cresciuto da 1500 a 4500 in 4 giorni. Ci si chiede qual è l'incremento percentuale medio giornaliero **r** di batteri.

Osserviamo anzitutto che l'incremento da 1500 a 4500 è pari a:

$$\frac{4500 - 1500}{1500} = 200\,\%,$$

allora si potrebbe dire che l'incremento medio giornaliero è pari a 200/4 % = 50% ?
In realtà è subito visto che un incremento giornaliero del 50% porta ad avere 1500·(1+0.5)=2250 batteri al primo giorno, 3375 al secondo, 5062 al terzo e 7594 al quarto: risultato in evidente contrasto con quanto rilevato!
L'approccio corretto consiste nell'osservare che, se $\mathbf{r_1}$, $\mathbf{r_2}$, $\mathbf{r_3}$ ed $\mathbf{r_4}$ sono gli incrementi percentuali in ognuno dei quattro giorni, deve essere:

$$1500 \cdot (1+\mathbf{r_1}) \cdot (1+\mathbf{r_2}) \cdot (1+\mathbf{r_3}) \cdot (1+\mathbf{r_4}) = 4500$$

e che l'incremento medio giornaliero **r** è dato da:

$$(1+\mathbf{r_1})\cdot(1+\mathbf{r_2})\cdot(1+\mathbf{r_3})\cdot(1+\mathbf{r_4}) = (1+\mathbf{r})^4.$$

Pertanto (1+**r**) risulta la *media geometrica* delle 4 quantità (1+$\mathbf{r_i}$):

$$(1+\mathrm{r}) = \sqrt[4]{(1+\mathrm{r})^4} = \sqrt[4]{\frac{4500}{1500}} = 1.316,$$

da qui la determinazione dell'incremento medio giornaliero:

$$\mathbf{r} = 1.316 - 1 = 31.6\ \%.$$

Si può provare che la media geometrica di un insieme di **dati positivi** è minore o uguale alla loro media aritmetica, ma è maggiore o uguale alla loro media armonica:

$$\mathrm{Ma} \leq \mathrm{Mg} \leq \bar{x}.$$

Indici di posizione

Le medie lasche sono quei valori che si basano solo su alcuni valori caratteristici dell'intera distribuzione dei dati.

La **mediana (Me)** è quell'osservazione che bipartisce la distribuzione di dati, supposta *ordinata* in ordine non decrescente, 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:

$$\mathrm{Me} = (94+96)/2 = 95\ \mathrm{mg/100\ cc}.$$

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	6+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.

Accanto alla mediana vengono considerati anche altri due indici: primo e terzo quartile in quanto presentano caratteristiche molto simili a quelle della mediana.

Il **primo quartile (Q_1)** è un valore della variabile presa in esame tale da lasciare alla sua sinistra ¼ = 25% della frequenza dei dati, mentre il **terzo quartile (Q_3)** lascia alla sua destra ¼ = 25% della frequenza.

Esempio 11. Supponiamo di rilevare il peso di 30 studenti ottenendo la distribuzione di frequenza di Tab. 1

Tab. 1 – Distribuzione di frequenza dei pesi di 30 studenti

Peso (kg)	Frequenza Assoluta
40	1
65	3
55	5
52	9
50	7
73	3
70	2

Ordinando i pesi in modo crescente si ottiene la Tab. 2

Tab. 2 – Distribuzione di frequenza dei pesi di 30 studenti

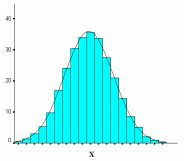
Peso (kg)	Frequ. assoluta	Frequ. %	Frequ. cumulata
40	1	3 %	1
50	7	23 %	8
52	9	30 %	17
55	5	17 %	22
65	3	10 %	25
70	2	7 %	27
73	3	10 %	30
Totale	30	100	

Per il calcolo della mediana e del primo e terzo quartile basta osservare:

$30/2 = 15 \longrightarrow$ mediana = 52 kg; $30/4 = 7.5 \longrightarrow Q_1 = 50$ kg; $3/4 \times 30 = 22.5 \longrightarrow Q_3 = 65$ kg.

Estendendo il concetto di quartile, si possono considerare anche i valori che dividono i dati in dieci parti uguali, ovvero i *decili*, oppure i valori che dividono i dati in cento parti uguali, i *percentili*. Il quinto decile ed il cinquantesimo percentile corrispondono alla media; il venticinquesimo ed il settantacinquesimo percentile corrispondono rispettivamente al primo e terzo quartine. Quartili, decili e percentili sono detti comunemente ***quantili.***

In merito alla scelta dell'indice medio più adeguato alla rappresentazione della distribuzione di dati si può evidenziare quanto segue:



La **media aritmetica** è indicata quando i dati quantitativi X presentano "abbastanza" simmetria (si pensi alla distribuzione normale di figura) e anche quando tali dati sono in progressione aritmetica (la differenza tra un dato e il precedente è costante).

È opportuno ricorrere alla **media geometrica** quando il logaritmo del variabile statistica quantitativa (positiva) presa in esame **ln(X)** presenta una distribuzione "molto simmetrica", più rigorosamente il logaritmo è distribuito normalmente (ad esempio nel caso di risposte alla somministrazione di farmaci), o anche quando i dati sono in progressione geometrica (il rapporto tra un dato e il precedente è costante); esempi di dati sono i pesi degli individui durante la crescita, il numero di microbi in una cultura, i tassi di variazione di una grandezza.

È adeguato far riferimento alla **media armonica** quando sono presenti dati quantitativi (positivi) che si discostano di molto dai limiti di tempo o di luogo stabiliti per l'osservazione; si usa, ad esempio, quando la variabile in esame è rappresentata da "tempi di reazione" (in prove di tossicità, nell'analisi di sopravvivenza).

La **mediana** può essere calcolata per caratteri qualitativi ordinali e caratteri quantitativi; si presta meglio di altri valori medi per esprimere il valore centrale di distribuzioni di caratteri che non possono essere misurati "esattamente" (ad es. i caratteri psicologici graduabili) oppure quando la distribuzione di una carattere presenta valori "anomali" o asimmetria (in sostanza, non si può far riferimento alla distribuzione normale), in quanto essa non risulta influenzata dai valori estremi della distribuzione.

La **moda** può essere calcolata per un carattere statistico qualunque (sia qualitativo che quantitativo) e non risente dei valori estremi (ad esempio è indicata per caratteri per i quali non si conoscono i valori estremi); tuttavia risulta l'indice con minor contenuto informativo.

INDICI DI VARIABILITÀ

Esempio 12. Si considerino inizialmente le seguenti due distribuzioni di valori riferiti all'età di 10 individui:

I serie	II serie
20	10
30	25
40	40
50	55
60	70

La media aritmetica è pari a 40 anni per entrambe le distribuzioni; ma nella seconda i dati sono più "dispersi" attorno alla media.
Pertanto accanto alle medie vanno introdotti anche indici di misura della *variabilità* dei dati.

Le misure di dispersione più usate sono:

1. campo di variazione (range);
2. deviazione standard.
3. differenza interquartile

Campo di Variazione o Range:

$$R = X_{max} - X_{min}$$

Per l'Esempio 5 si ha:

$$R = 119 - 71 = 48 \text{ mg/110 cc;}$$

nel caso dell'Esempio 7:

$$R = 74 - 34 = 40 \text{ anni.}$$

Limiti del campo di variazione
- è troppo influenzato dai valori estremi;
- tiene conto dei due soli valori estremi, trascurando tutti gli altri.
- tende ad aumentare con l'aumento del numero di osservazioni.

Occorre allora un indice di dispersione che consideri tutti i valori (e non solo quelli estremi). Tuttavia va ricordato che: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Si potrebbe calcolare la somma dei valori assoluti: $\sum_{i=1}^{n}|x_i - \bar{x}|$, ma tale quantità è difficile da trattare matematicamente. Un indice alternativo, più agevole da usare, è quello di considerare la **somma dei quadrati degli scarti dalla media aritmetica** $\bar{x}$**.**

Esempio 5'. Valori del tasso glicemico in 10 soggetti

x_i (glicemia mg/100cc)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
103	+8	64
97	+2	4
90	-5	25
119	+24	576
107	+12	144
71	-24	576
94	-1	1
81	-14	196
92	-3	9
96	+1	1
$\bar{x} = 95$	$\sum_{i=1}^{10} \lvert x_i - \bar{x} \rvert = 94$	$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1596$

La quantità $\sum_{i=1}^{n} (x_i - \bar{x})^2$ si chiama ***Devianza*** (**Dev**).

Il limite della Devianza come misura di dispersione è quello di aumentare con il numero di osservazioni. Per ottenere una misura che non dipenda dalla numerosità si può dividere la devianza per il numero n di dati, ottenendo la ***varianza***:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

In pratica il denominatore n è quasi sempre sostituito da (n-1) in modo da ottenere una stima corretta della dispersione della variabile nella popolazione da cui il campione in esame è stato estratto.

$$\mathbf{s^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Nell'Esempio 5' si ha: $s^2 = 1596 / 9 = 177.33$ $(mg/100cc)^2$.

Il limite della Varianza come misura di dispersione è quella di avere una unità di misura espressa al quadrato rispetto all'unità di misura originale, per cui si utilizza la ***Deviazione Standard*** (D.S. o S.D.):

$$\mathbf{s} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}.$$

La Deviazione Standard è l'indice di variabilità più usato e ad esso si farà riferimento nel seguito
Essa indica quanto, in media, ciascun elemento si discosta dalla media.

Nell'esempio 5' : $s = \sqrt{177.33} = 13.32$ mg/100cc.

La Deviazione Standard per distribuzioni di frequenza: assume la seguente forma:

$$s = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1}},$$

dove **k** è il numero di modalità della variabile statistica X o il numero di classi in cui i valori di X sono stati raggruppati.

Indici di variabiltà relativi:

Coefficiente di Variazione $\quad CV = \frac{s}{\bar{x}} = \frac{\text{Deviazione Standard}}{\text{media aritmetica}}$.

Per l'Esempio 5' si ottiene : $CV = 13.32/95 = 14\,\%$. E' interessante anche il confronto tra i coefficienti di variazione delle due serie di dati dell'Esempio 11: per la serie I si ha $CV_I = 15.8/40 = 39.5\%$, mentre per la II: $CV_{II} = 23.7/40 = 59.3\%$, risultati che confermano la maggiore variabilità dei dati della seconda serie rispetto alla prima.
Il Coefficiente di Variazione è un numero "puro", in quanto rapporto di due grandezze omogenee, e perciò consente il confronto anche tra variabili eterogenee.
L'uso del C.V. si rende necessario ogni qualvolta si vogliono confrontare le misure di variabilità relative a distribuzioni le cui modalità sono espresse in unità di misure diverse (confronto tra variabilità dell'altezza e del peso) oppure sono espresse nella stessa unità di misura ma il loro valore medio risulta molto diverso (confronto delle variabilità dei pesi fra un campione di neonati ed uno di adulti).

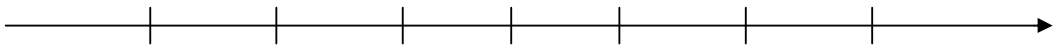
Come indice di variabilità può essere considerato anche la **differenza interquartile**:

$$H = Q_3 - Q_1.$$

Nel caso dell'esempio 10 di pag.15, risulta $H = 65 - 50 = 15$ kg. Si osservi che tra il primo ed il terzo quartile vengono a trovarsi il 50 % delle osservazioni. Se l'intervallo interquartilico è piccolo, tale metà delle osservazioni si trova fortemente concentrata intorno alla mediana; all'aumentare della distanza interquartilica aumenta la dispersione del 50% delle osservazioni centrali intorno alla mediana.

BOX-PLOT

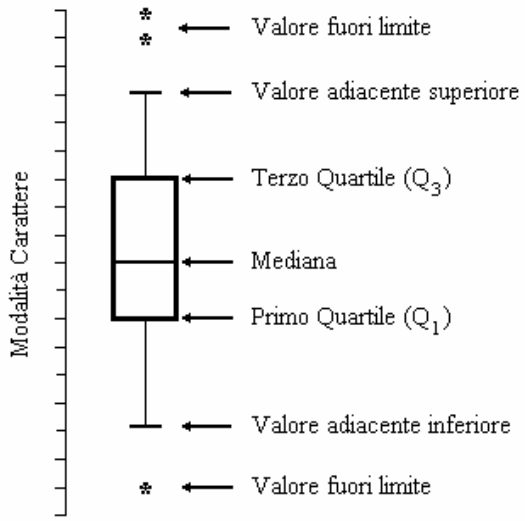
Alcuni metodi statistici esplorativi permettono di esprimere i dati facendo riferimento alla mediana ed ai quartili. Una rappresentazione grafica di notevole utilità per capire come è composto l'insieme delle osservazioni consiste nel rappresentare su di una retta la mediana, i quartili ed altri quattro punti individuati sul grafico in modo da definire 5 intervalli:



Q_1-3H $Q_1-1.5H$ Q_1 Me Q_3 $Q_3+1.5H$ Q_3+3H

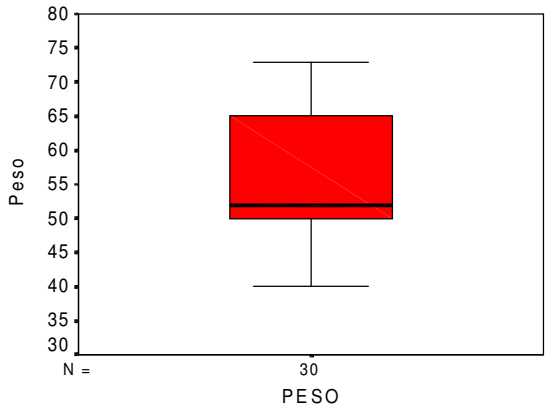
Tra $Q_1-1.5H$ e Q_1 e tra Q_3 e $Q_3+1.5H$ cadono i valori detti *adiacenti* in quanto risultano prossimi al nucleo centrale (tra Q_1 e Q_3 c'è il 50% dei dati) delle osservazioni; tra Q_1-3H e $Q_1-1.5H$ e tra $Q_3+1.5H$ e Q_3+3H cadono i valori detti *lontani* e prima di Q_1-3H e dopo Q_3+3H cadono i valori *molto lontani*. Per i valori molto lontani occorre fare un'analisi attenta e minuziosa per capire se si tratta di valori errati oppure di valori che si staccano effettivamente dal resto dei dati. La

suddivisione posta sopra porta a cogliere intervalli che aiutano a percepire il segnale di qualcosa di anomalo presente fra le osservazioni.



I valori *lontani o molto lontani* (**valori fuori limite**) vengono segnalati *individualmente* nel box-plot per evidenziarne la presenza e la posizione. Questi valori forniscono informazioni ulteriori sulla dispersione e sulla forma della distribuzione. Quando i valori adiacenti, superiore e inferiore, coincidono con gli estremi della distribuzione non comparirà alcun valore fuori limite. I valori adiacenti inferiore e superiore forniscono informazioni sulla dispersione e sulla *forma* della distribuzione ed anche sulle *code* della distribuzione.

Attraverso la costruzione dei "diagrammi a scatola" (**Box-plot**) è possibile effettuare l'esplorazione appena detta. La figura seguente rappresenta tale diagramma per l'esempio 10 di pag. 15. All'interno della "scatola" si posiziona il 50% dei dati (pertanto il lato inferiore indica il primo quartile e quello superiore il terzo); la mediana è indicata dalla linea interna alla scatola; al di sopra e al di sotto della scatola sono anche individuati i punti della zona dei valori *adiacenti* (nell'esempio in esame, poiché $Q_1 - 1.5H = 32.5$ kg < 40 kg e $Q_3 + 1.5H = 84.5$ kg > 73 kg, sono riportati semplicemente il minimo ed il massimo dei valori del peso).



I Box-plot permettono anche di comparare distribuzioni di caratteri diversi.

Esempi su indici di tendenza centrale e di variabilità

Esempio 13. Valori pressori massimi rilevati su 5 pazienti ipertesi

PAS (mmHg)	$\mathbf{f_i}$	$\mathbf{x_i \cdot f_i}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 \cdot \mathbf{f_i}$
170	1	170	-23	529
185	1	185	-8	64
200	1	200	7	49
205	2	410	12	288
Somma	5	965		930

Media Aritmetica: $\bar{x} = 965 / 5$ mmHg $= 193$ mmHg; **Range**: $R = 205 - 170 = 35$ mmHg;
Devianza: $Dev = 930$ (mmHg)^2; **Varianza**: $s^2 = 930 / 4$ (mmHg)^2 $= 232{,}5$ (mmHg)^2;
Dev. St.: $s = \sqrt{232{,}5}$ mmHg $= 15{,}25$ mmHg; **Coeff. Variaz**.: $CV\% = 15{,}25 / 193 \cdot 100 = 7{,}9$ %.

Per il calcolo della mediana (Me) e della Moda (Mo) della distribuzione della pressione si procede come nella tabella:

PAS (mmHg)	f_i	frequ. Cumulate
170	1	1
185	1	2
200	1	3
205	2	5

$5 / 2 = 2{,}5 \longrightarrow$ Me = 200
Mo = 205

I due esempi che seguono illustrano il calcolo di indici medi e di variabilità nel caso di dati raggruppati in classi di frequenze.

Esempio 13. Azoto ureico (mg %) in un gruppo di 50 adolescenti

Azoto	val. centr. $(\mathbf{x_i})$	Frequenze $(\mathbf{f_i})$	Frequ. cum.	$\mathbf{x_i * f_i}$	$(\mathbf{x_i} - \bar{x})^2 * \mathbf{f_i}$
17.1 – 19	18.05	3	3	54.15	82.3728
19.1 – 21	20.05	6	9	120.30	62.9856
21.1 – 23	22.05	11	20	242.55	16.9136
23.1 – 25	24.05	20	40	481.00	11.5520
25.1 – 27	26.05	8	48	208.40	60.9408
27.1 – 29	28.05	1	49	28.05	22.6576
29.1 – 31	30.05	1	50	30.05	45.6976
Totale		50		1164.5	303.12

$\bar{x} = 1164.5/50 = 23.29$; D.S. $= \sqrt{303.12 / 49} = 2.49$;
calcolo mediana: $N/2 = 50 / 2 = 25 \longrightarrow$ la classe mediana (classe che comprende la mediana) è data da: 23.1 - 25, ovvero $23.1 < Me < 25$;
calcolo moda: la frequenza più elevata si ha per la classe 23.1 - 25, dunque: $23.1 < Mo < 25$.

Calcolo della mediana

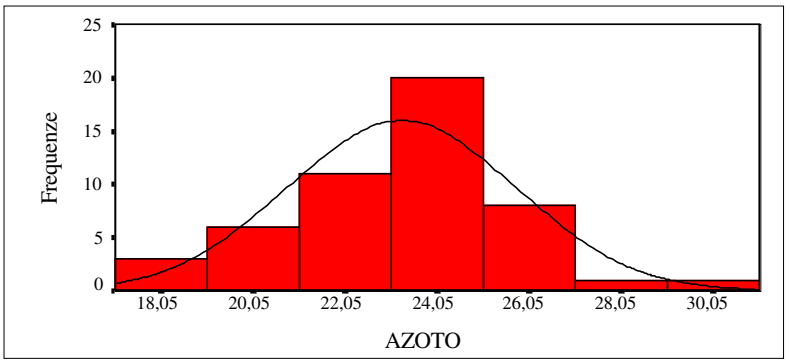
Se è ipotizzabile che all'interno delle classi di frequenze i valori dell'azoto si distribuiscano in modo uniforme, la mediana può essere determinata impostando una relazione di tipo lineare tra dati e le relative frequenze cumulate. Individuata, allora, la classe mediana si può calcolare:

$$Me = x_m + \frac{\frac{N}{2} - \sum f_{prec}}{f_{med}} \cdot c ,$$

dove: x_m, f_{med} e c sono rispettivamente l'estremo inferiore, la frequenza e l'ampiezza della classe mediana, $\sum f_{prec}$ è la somma delle frequenze delle classi inferiori a quella mediana.
Nell'esempio analizzato si ha:

$$Me = 23.1 + \frac{25 - 20}{20} \cdot 2 = 23.6 .$$

Il grafico seguente mostra l'istogramma della distribuzione dell'azoto e, sovrapposta a questo, la curva della distribuzione normale (per lo studio di tale curva si veda in appunti successivi).



Esempio 14. Dosaggio della Fosfatasi Alcalina (UA) in 20 studenti

Fosfatasi Alcalina	Valore centrale (x_i)	Frequenze (f_i)	Frequ. cumul.	$x_i * f_i$	$(x_i - \bar{x})^2 * f_i$
30.1 – 60	45.05	1	1	45.05	7056
60.1 – 90	75.05	3	4	225.15	8748
90.1 - 120	105.05	3	7	315.15	1728
120.1 - 150	135.05	7	14	945.35	252
150.1 - 180	165.05	5	19	825.25	6480
180.1 - 210	195.05	0	19	0	0
210.1 - 240	225.05	1	20	225.05	9216
totale		20		2581	33480

$\bar{x} = 2581 / 20 = 129$; D.S. $= \sqrt{33480 / 19} = 41.98$;
calcolo mediana: $N/2 = 20 / 2 = 10 \rightarrow$ la classe mediana è 120.1 - 150, ovvero $120.1 < Me < 150$;
calcolo moda: la frequenza più elevata si ha per la classe 120.1 - 150, dunque: $120.1 < Mo < 150$.

Con la formula introdotta nell'es. 12 si ottiene anche: $\text{Me} = 120.1 + \frac{10-7}{7} \cdot 30 = 133$.

CURVE DI DISTRIBUZIONI E INDICI DI FORMA

Le osservazioni quantitative continue,se rilevate in gran numero, possono essere suddivise in classi di piccola ampiezza ma anche di numerosità non trascurabile; in tale contesto i poligono di frequenze assolute o di frequenze relative risultano ben essere approssimati (*lisciati*) da curve di frequenze che assumono forme caratteristiche.

Si possono avere distribuzioni con una sola moda (**unimodali**) o con più di una. Sulle prime ci si soffermerà tra breve, mentre per quanto attiene le seconde e, in particolare, le distribuzioni ***bimodali,*** caratterizzate dalla presenza di due mode, si può dire che si è in presenza di dati non omogenei. Ciò può verificarsi, ad esempio, quando i valori si riferiscono ad un campione composto sia da individui sani che malati; in tal caso si avranno **due mode**, una media ed una mediana (Figura IV):

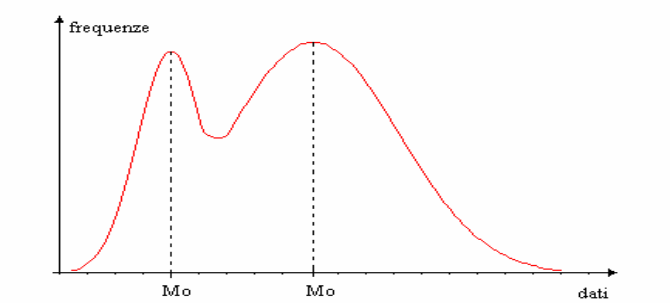


Fig. IV - Curva bimodale

Indici di asimmetria

Per le distribuzioni **unimodali** di frequenze è possibile distinguere le seguenti situazioni.

Distribuzione ***simmetrica***:
le osservazioni equidistanti dalla mediana (coincidente in questo caso col massimo centrale) presentano la stessa frequenza relativa (figura I)
Un esempio importante è fornito dalla curva di ***distribuzione normale***

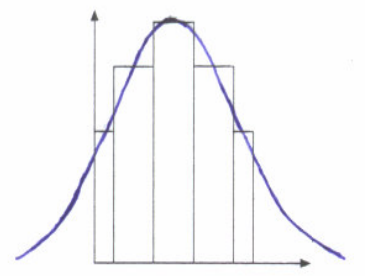


Fig. I - Media = Mediana

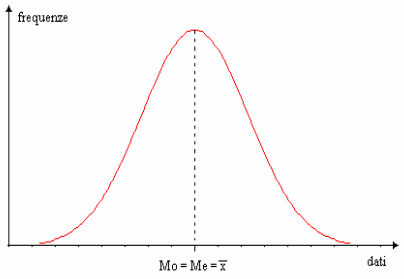


Fig. I - Curva Simmetrica

Distribuzione ***asimmetrica positiva***:
la curva di frequenza ha una *coda* più lunga a destra del massimo centrale (figura II).

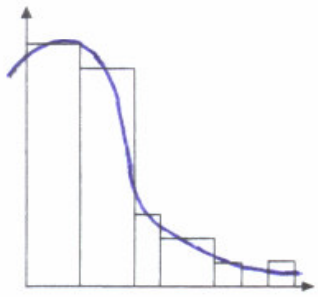


Fig. II - Media > Mediana

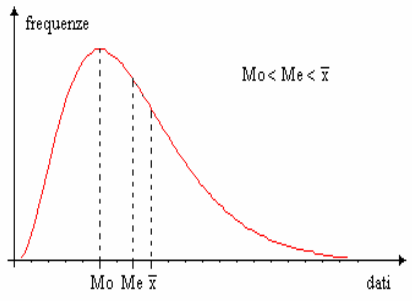


Fig. II - Curva asimmetrica a destra

Distribuzione ***asimmetrica negativa***:
la curva di frequenza ha una *coda* più lunga a sinistra del massimo centrale (figura III).

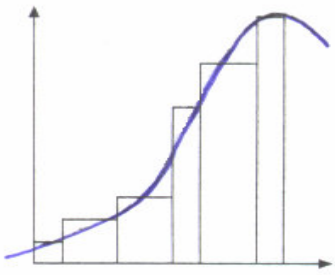


Fig. III - Media < Mediana

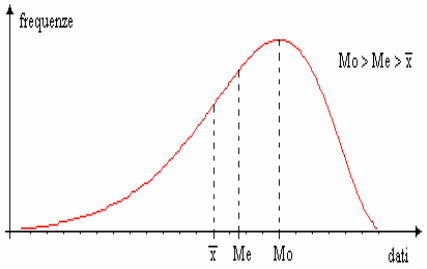


Fig. III - Curva asimmetrica a sinistra

Un indice statistico che permette di informare del grado di simmetria/asimmetria di una distribuzione di n *dati quantitativi* è l'***indice di asimmetria***:

$$\gamma = \frac{\mu_3}{\sigma^3},$$

essendo: $\mu_3 = \dfrac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{n}$ e $\sigma = \dfrac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n}$.

In merito ai valori che tale indice può assumere si può dire che:

- se la distribuzione è simmetrica, $\gamma = 0$;
- se la distribuzione è asimmetrica positiva (a destra), $\gamma > 0$;
- se la distribuzione è asimmetrica negativa (a sinistra), $\gamma < 0$.

Come misura della asimmetria vengono utilizzate anche le seguenti formule *empiriche* (che richiedono la conoscenza di moda, mediana, media e deviazione standard):

$$\frac{\bar{x} - Mo}{DS} \quad \text{oppure} \quad \frac{3(\bar{x} - Me)}{DS},$$

nel caso di simmetria si ha un valore nullo, per l'asimmetria positiva un valore positivo e per quella negativa un valore minore di zero.

Esempio 15. Le seguenti misure caratterizzano le distribuzioni dei voti, in una scala da 1 a 100, riportati in una data Disciplina dagli studenti di due diversi corsi a seguito di una verifica intermedia (formativa):

	Corso 1	Corso 2
Media	78	72
Mediana	65	73
D.S.	16	6

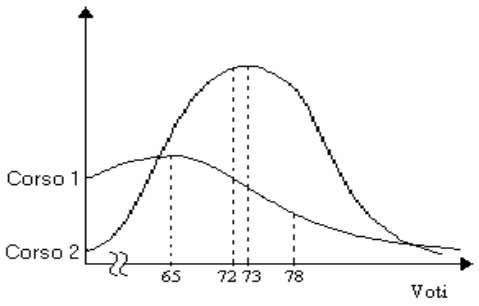
Ci si chiede:

a) In quale di questi due corsi il docente dovrebbe impiegare più tempo nel lavoro individuale con gli studenti?
b) In quale di questi due corsi è più probabile trovare alcuni studenti particolarmente bravi, ovvero molto al di sopra della media?

a) Mentre nel corso 2 circa la metà della classe ha un voto non inferiore a 73, nel corso 1 il 50% degli studenti ha un voto non superiore a 65 ed inoltre c'è maggiore variabilità dei dati (il coeff. di variazione del corso 1 è pari a circa il 21% e quello del corso 2 è di circa l'8%). Pertanto il docente dovrebbe impiegare più tempo nel lavoro individuale con gli studenti del corso 1.

b) La D.S. del corso 1 è decisamente più elevata rispetto a quella del corso 2, pertanto nel primo corso è più probabile trovare studenti molto al di sopra (D.S.=16) della media. D'altra parte, osservato anche i valori delle medie, si può dire anche che nel corso 1 è più probabile trovare studenti "più bravi" di quelli del secondo.

Per illustrare graficamente la situazione, in maniera qualitativa tuttavia, si potrebbe ricorrere ai grafici di distribuzioni di frequenze. La distribuzione relativa al corso 1 potrebbe presentare una ***asimmetrica positiva*** $\left(\frac{3(78-65)}{16}=+2.44\right)$, mentre quella del corso 2 una ***asimmetria negativa*** $\left(\frac{3(72-73)}{6}=-0.5\right)$.



Si osservi anche che informazioni sulla *forma* di una distribuzione possono essere fornite dalle distanze tra ciascun quartile e la mediana. L'indice:

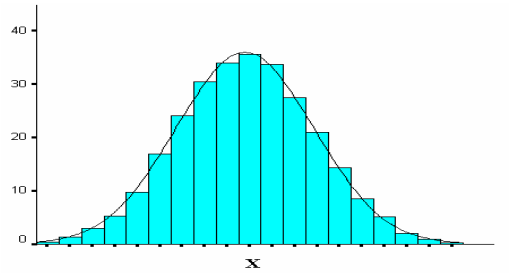
$$\frac{(Q_3-Me)-(Me-Q_1)}{Q_3-Q_1}$$

indica asimmetria positiva se risulta positivo, asimmetria negativa nel caso sia negativo.

Per una distribuzione simmetrica, come quella *normale*, le distanze tra ciascun quartile e la mediana sono uguali, come anche le lunghezze delle linee che si allungano dai bordi della "scatola" utilizzata nei ***box-plot***. In particolare, per una distribuzione normale, il *valore adiacente inferiore* e quello *superiore* risultano pari, rispettivamente, a Me–2.6975·DS e Me+2.6975·DS e l'intervallo individuato da tali valori racchiude il 99.30% dei dati.

Indici di curtosi

Sempre nel caso di distribuzioni unimodali, per interpolare un istogramma con una funzione di densità di tipo continuo spesso nell'analisi statistica si fa riferimento alla **distribuzione normale** come modello teorico di riferimento. L'istogramma di figura a lato, ad esempio, è stato "approssimato" con una normale.



Pertanto vengono costruiti indici statistici che misurano, in qualche modo, lo "scostamento" della distribuzione empirica del carattere quantitativo X da quella relativa alla situazione espressa dalla curva normale. Uno di questi è l'indice di ***curtosi*** (calcolato per n dati quantitativi):

$$\beta = \frac{\mu_4}{\sigma^4}$$

dove: $\mu_4 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n}$ e $\sigma = \dfrac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{n}$.

In merito ai valori assunti si può dire che:

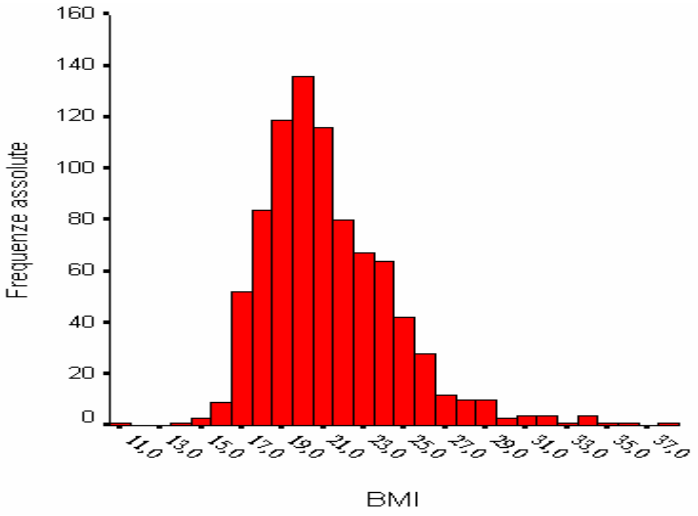
- Se la variabile è perfettamente normale si ha $\beta = 3$;
- Se $\beta > 3$ la distribuzione è detta **ipernormale** (rispetto alla distribuzione normale si ha una maggiore frequenza per valori centrali ed estremi, minore per valori intermedi);
- Se $\beta < 3$ la distribuzione è detta **iponormale** (rispetto alla distribuzione normale si ha una minore frequenza per valori centrali ed estremi, maggiore per valori intermedi).

Va osservato che i **software statistici** spesso calcolano l'indice di curtosi come $\beta - 3$, che pertanto si annulla nel caso di normalità. Va aggiunto che ha senso calcolare tale indice solamente quando la distribuzione empirica è stata determinata da numerose determinazioni del carattere X, in quanto in tal caso si può effettuare un appropriato confronto con la distribuzione normale.

Esempio 16 Su un campione di 853 studenti di età compresa tra 13 e 19 anni sono stati rilevati altezza e peso ed è stato determinato *l'indice di massa corporea* MBI (body mass index) espresso dalla seguente relazione:

$$\text{BMI} = \frac{\text{peso } in\ kg}{(\text{altezza } in\ m)^2}.$$

La distribuzione dei dati del BMI è visualizzata nel grafico seguente ed alcuni indici statistici sono riportati nella tabella laterale:



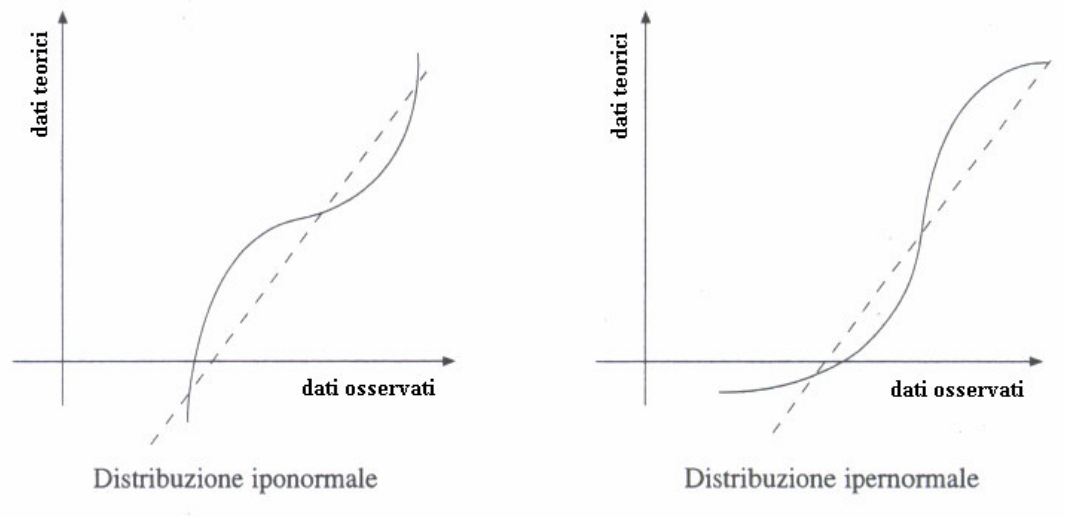
Num. studenti		853
Minimo		11,10
Massimo		38,10
Media		21,29
Mediana		20,72
Dev. Stand.		3,25
Percentili	25	19,03
	50	20,72
	75	23,05
Asimmetria		1,20
Curtosi		2,65

Si evidenzia che la distribuzione è asimmetrica positiva e ipernormale.

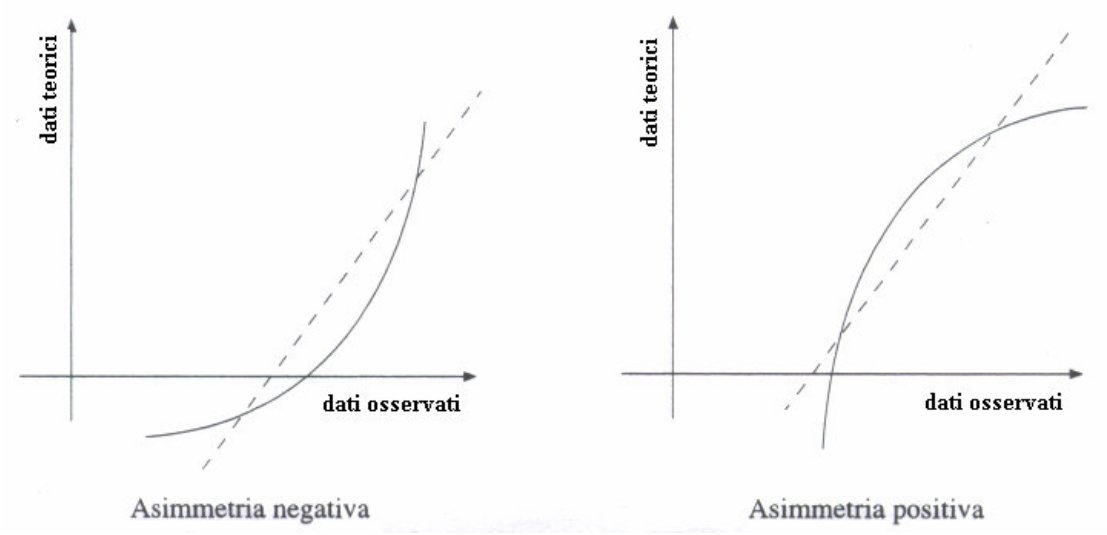
Grafici qq-plot

Un altro modo per vedere se i dati a disposizione possono essere approssimati dalla distribuzione normale è quello di ricorre ai grafici "quantile-quantile" (qq-plot). In tali grafici i quantili ottenuti dai dati numerici "osservati" vengono confrontati con quelli che si avrebbero nel caso che i dati seguissero esattamente una distribuzione normale (dati "teorici"). Si ottengono dei punti su di un piano cartesiano: più tali punti sono allineati e maggiormente la distribuzione di dati osservati è "vicina" a quella normale.

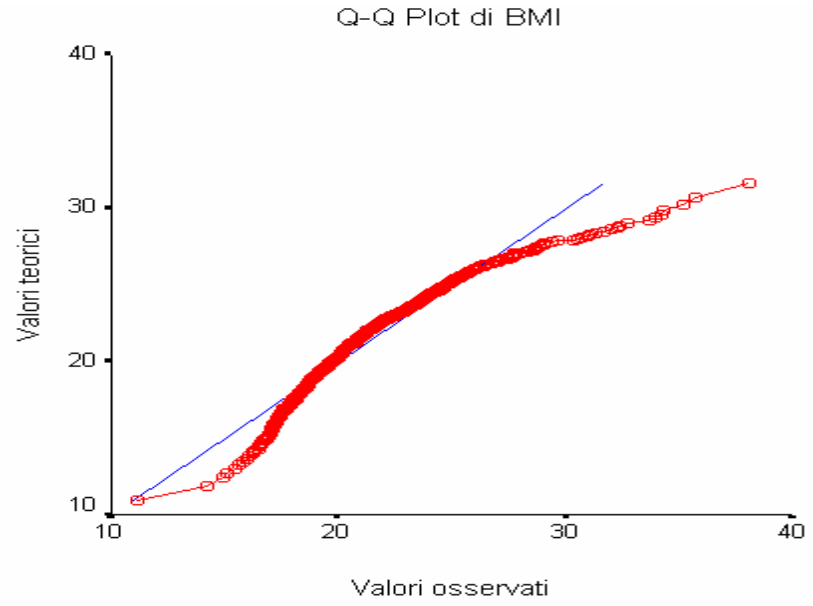
Le figure che seguono schematizzano le situazioni di distribuzioni iponormali ed ipernormali



I qq-plot possono essere utilizzati anche per analizzare la simmetria/asimmetria di una distribuzione, come indicato nelle figure:



In riferimento all'esempio n. 16, il qq-plot è quello sotto riportato:



e si evidenzia l'ipernormalità e l'asimmetria positiva della distribuzione del BMI.