CAPITOLO I

ELEMENTI DI STATISTICA DESCRITTIVA PER DISTRIBUZIONI UNIVARIATE

CAPITOLO I

ELEMENTI DI STATISTICA DESCRITTIVA PER DISTRIBUZIONI UNIVARIATE

1.1. LA STATISTICA NELLA RICERCA AMBIENTALE E BIOLOGICA

Come in tutta la ricerca scientifica sperimentale, anche nelle scienze ambientali e in quelle biologiche è indispensabile la conoscenza dei concetti e dei metodi statistici, sia per i problemi di gestione, sia per quelli di indagine. Per pubblicare i risultati di una ricerca, tutte le riviste scientifiche richiedono che la **presentazione dei dati** e la **loro elaborazione** seguano criteri ritenuti validi universalmente.

Il comportamento nel**la fase di raccolta dei dati**, **la loro descrizione**, **le analisi e infine il riepilogo** sono in buona parte codificati, in modo dettagliato. Inviare ad una rivista uno studio o una relazione che denotino una conoscenza sommaria della statistica comporta generalmente una critica dei metodi seguiti, che può giungere fino al rifiuto delle conclusioni o almeno a una dichiarazione esplicita della loro ridotta attendibilità. Alla ricerca, seppure valida per gli altri aspetti e importante per l'argomento trattato, può essere negata la dignità della pubblicazione.

Una raccolta di dati non corretta, una loro presentazione inadeguata o un'analisi statistica non appropriata rendono impossibile la verifica dei risultati da parte di altri studiosi e il confronto con altre ricerche e analisi del settore. Per il progresso di qualsiasi disciplina sperimentale, una finalità importante di qualsiasi ricerca anche di piccole dimensioni, è la semplice possibilità di sommare le esperienze e confrontare i risultati con altre, effettuate in condizioni simili oppure volutamente differenti. Permette l'accumulo delle conoscenze, la verifica di teorie già proposte, la formulazione di nuove ipotesi.

Al fine di facilitare ai lettori la corretta comprensione dei risultati, per pubblicare una ricerca le riviste internazionali e quelle di maggior prestigio richiedono tassativamente agli autori di seguire uno schema preciso che, in linea di massima, è fondato sullo sviluppo di quattro fasi.

1) Una ***introduzione***, che presenti in modo accurato sia l'**argomento** affrontato, sia le **finalità** della ricerca, mediante citazione dei lavori scientifici pregressi e della **letteratura** specifica.

2) La descrizione di ***materiali e metodi***, nella quale devono essere definiti:

a) il tipo di **scala** utilizzato;

b) le modalità del **campionamento** o di raccolta dei dati;

c) le misure sintetiche delle caratteristiche più importanti della **distribuzione dei dati**, come media e varianza (più raramente simmetria, curtosi e coefficiente di variazione).

Spesso, soprattutto per argomenti nuovi o quando siano stati pubblicati solo pochi dati, è prassi richiedere la distribuzione tabellare completa e dettagliata. Per relazioni scientifiche che non abbiano solo un carattere divulgativo, le tabelle e le distribuzioni di frequenze sono da preferire alle rappresentazioni grafiche. Infatti queste ultime raramente permettono di risalire ai dati originari, che sono indispensabili per la verifica dei calcoli e quindi una valutazione piàù dettagliata delle conclusioni raggiunte.

3) I ***risultati***, che devono comprendere espressamente la citazione dei **test di inferenza** utilizzati, allo scopo di permettere alla comunità scientifica di valutare se la loro scelta è appropriata, cioè se sono in rapporto corretto con

a) le ipotesi che si intendono verificare,
b) il tipo di scala con cui sono state misurate le variabili analizzate,
c) le caratteristiche statistiche della distribuzione dei dati.

4) La ***discussione***, che deve riportare l'**interpretazione** dei risultati ottenuti con i test applicati e infine eventuali confronti con analisi già pubblicate. **L'interpretazione deve non solo comprendere l'analisi statistica, ma essere estesa al significato ecologico, ambientale o biologico dei risultati ottenuti**. Non sempre un risultato statisticamente rilevante assume anche un significato importante nella disciplina specifica. Ne deriva un aspetto di estrema importanza per l'analisi statistica: **per impostare correttamente una ricerca, per formulare ipotesi scientificamente valide, per raccogliere e analizzare i dati, infine per interpretarne i risultati, non è possibile scindere le analisi statistiche dalla loro interpretazione disciplinare**.

1.2. IL DISEGNO SPERIMENTALE, IL CAMPIONAMENTO E L'INFERENZA

Per condurre in modo corretto una ricerca scientifica, cioè per raccogliere un campione con un numero sufficiente di dati, tenendo in considerazione sia le condizioni esistenti nella popolazione, sia la successiva applicazione dei test, occorre seguire alcuni passaggi metodologici, riassumibili in 4 fasi:

- il disegno sperimentale,
- il campionamento,
- la descrizione statistica,
- la scelta dei test per l'inferenza.

1 - Il **disegno sperimentale** è necessario per scegliere e programmare le osservazioni in natura e le ripetizioni in laboratorio, in funzione della ricerca e delle ipotesi esplicative. Già nella prima fase della ricerca, chiamata con termine tecnico appunto "disegno sperimentale" (dall'inglese ***experimental design*** e tradotto più correttamente in italiano con **programmazione dell'esperimento**), occorre avere chiara la formulazione dell'ipotesi che si intende verificare. Raccogliere i dati prima di aver chiaramente espresso le finalità della ricerca conduce spesso ad analisi non adeguate e quindi a risultati poco attendibili.
Con la formulazione dell'ipotesi, si deve rispondere alle domande:
- "*Le eventuali differenze riscontrate tra due o più gruppi di dati, oppure di una serie di osservazioni con quanto è atteso, possono essere imputabili a* ***fattori causali*** *specifici o solamente a* ***fattori casuali*** *ignoti?*
- *Le differenze riscontrate sono generate dalla naturale variabilità delle misure e del materiale utilizzato oppure più probabilmente esiste una causa specifica che le ha determinate?*"

2 - Il **campionamento** permette di raccogliere i dati in funzione dello scopo della ricerca, rispettando le caratteristiche della **popolazione o universo** dei dati.
Uno dei **problemi fondamentali della statistica** è **come raccogliere solamente un numero limitato di dati** (per motivi economici, di tempo, di oggetti effettivamente disponibili, cioè per limiti oggettivi che quasi sempre esistono in qualsiasi ricerca sperimentale), **ma attraverso la loro analisi pervenire ugualmente a conclusioni generali**, che possano essere estese a tutta la popolazione.

3 - La **descrizione** delle caratteristiche statistiche **dell'insieme dei dati raccolti** deve permettere a tutti di verificare sia l'adeguatezza del disegno sperimentale e del campionamento, sia la correttezza delle analisi attuate e dei risultati ottenuti.

4 - I **test** devono essere già programmati nella fase del disegno sperimentale, poiché è da essi che dipende il tipo di campionamento.
Il test è un processo logico-matematico che porta alla conclusione di non poter respingere oppure di poter respingere l'**ipotesi della casualità**, mediante il calcolo di probabilità specifiche di commettere un errore con queste affermazioni.
L'ipotesi che **il risultato ottenuto con i dati sperimentali raccolti sia dovuto solo al caso** è chiamata **ipotesi nulla** e è indicata con $\mathbf{H_0}$. Di norma, con essa si afferma che le differenze tra due o più gruppi, quelle tra un gruppo e il valore atteso oppure le tendenze riscontrate siano imputabili essenzialmente al caso.

Per esempio, confrontando il tempo di guarigione tra due gruppi di ammalati ai quali siano stati somministrati due farmaci differenti, con **l'ipotesi nulla H_0** si sostiene che il risultato ottenuto non dipende da una effettiva differenza tra i due principi attivi, ma che esso è dovuto al caso. Se nell'esperimento il farmaco A è risultato migliore del farmaco B, se si accetta l'ipotesi nulla implicitamente si afferma che, con un nuovo esperimento nelle stesse condizioni, si potrebbe ottenere il risultato opposto.
Per giungere a queste conclusioni si deve ricorrere all'**inferenza**, che può essere definita come **la capacità di trarre conclusioni generali (sulla popolazione od universo) utilizzando solo un numero limitato di dati variabili (campione)**.

Il **disegno sperimentale ed il campionamento sono le due fasi preliminari**
- **sia alla raccolta dei dati in natura,**
- **sia per una corretta impostazione degli esperimenti in laboratorio**.

Tuttavia, la presentazione didattica e la corretta comprensione di questi argomenti richiedono concetti complessi e metodologie sofisticate, non sempre facili né intuitivi. Per questi motivi, il disegno sperimentale e il campionamento sono sempre trattati nella fase finale di un corso di statistica applicata, quando è già stata raggiunta sufficiente familiarità con la terminologia, con i concetti e i metodi fondamentali dell'inferenza.

Nell'apprendimento e nell'uso della statistica, il primo passo è comprendere come solamente una corretta applicazione del campionamento e una scelta appropriata dei test permettano di rispondere alla **domanda inferenziale** di verifica dell'**ipotesi nulla**. Con essa si pone il seguente quesito:
" ***Nell'ipotesi che le differenze fra gruppi di osservazioni empiriche siano dovute a fattori esclusivamente casuali, quale è la probabilità che fra tutte le alternative possibili si presenti proprio la situazione descritta dai dati raccolti o una ancora più estrema?***"
Se tale probabilità risulta alta, convenzionalmente uguale o superiore al 5%, si imputeranno le differenze a fattori puramente casuali.
Al contrario, se la probabilità risulta bassa, inferiore al valore prefissato, si accetta come verosimile che le differenze siano dovute a fattori non casuali, rientranti tra i criteri che distinguono i gruppi di dati.

La procedura dell'inferenza statistica è semplice, nelle linee logiche generali. Tuttavia, le analisi e le conclusioni trovano complicazioni per l'**elevata variabilità dei dati**, **a motivo soprattutto di tre cause che, in ordine crescente d'importanza, sono**:

- gli **errori di misurazione**, generati da strumenti e da differenze nell'abilità dei ricercatori;
- l'**operare su campioni**, per cui i dati utilizzati in una ricerca non sono mai identici a quelli rilevati in qualsiasi altra;
- la presenza di vari **fattori contingenti di disturbo** che, come il tempo e la località, possono incidere diversamente sul fenomeno in osservazione, con intensità e direzioni ignote.

Pure se espressi in modo sintetico, questi concetti definiscono il **contenuto della statistica moderna**: ***la raccolta, la presentazione e la elaborazione numerica delle informazioni, per agevolare l'analisi dei dati ed i processi decisionali***.

In un corso completo di statistica applicata, è importante avere in ogni momento una visione complessiva degli argomenti. Il loro elenco è utile anche per comprendere le diverse parti in cui viene distinta la statistica, nel percorso di apprendimento dei concetti e delle procedure.

La statistica moderna può essere distinta **in tre parti**: descrittiva, matematica, inferenziale.

1 - La **statistica descrittiva** spiega come i dati raccolti devono essere riportati in tabella, rappresentati in grafici e sintetizzati in indici matematici, allo scopo di individuare le caratteristiche fondamentali del campione.

2 - La **statistica matematica** presenta le distribuzioni teoriche sia per misure discrete sia per misure continue, allo scopo di illustrarne le caratteristiche fondamentali, le relazioni che esistono tra esse, gli usi possibili;

3 - L**'inferenza statistica**, la parte nettamente prevalente del corso, serve per la verifica delle ipotesi. Essa può essere distinta in vari capitoli, in rapporto
- alle caratteristiche dei dati (se permettono o meno il ricorso alla distribuzione normale: statistica parametrica e non parametrica
- al numero di variabili (se una, due o più: statistica univariata, bivariata, multivariata).

La prima parte dell'inferenza, di solito affrontata in un corso, è la **statistica univariata parametrica**. Come argomenti, essa comprende il test **t** di **Student** e il test **F** di **Fisher-Snedecor** o analisi della varianza:
- il primo serve sia per confrontare la media di un campione con una media attesa o teorica, sia per confrontare le medie di due campioni;

- il secondo rappresenta la sua generalizzazione e permette il confronto simultaneo tra più medie, considerando uno solo oppure più fattori di variabilità.

Appunto perché fondati sulla distribuzione normale, questi test richiedono condizioni di validità restrittive (discusse nei capitoli seguenti), che non sempre i dati raccolti e la misura utilizzata permettono di rispettare.

E' una situazione che si presenta con frequenza elevata nella ricerca applicata, a causa della estrema variabilità dei dati e della presenza di valori anomali. In queste condizioni si ricorre alla **statistica univariata non parametrica**, che è formata da una serie innumerevole di test.

Di norma essi sono raggruppati sulla base dei campioni ai quali viene applicata: test per un campione, per due campioni dipendenti e indipendenti, test per k campioni dipendenti e indipendenti.

Quando per ogni individuo o situazione si raccolgono informazioni relative a due variabili, è possibile analizzare le relazioni che intercorrono tra esse, mediante sia la regressione e la correlazione parametriche, sia la regressione e la correlazione non parametriche. Si parla allora di **statistica bivariata parametrica** e di **statistica bivariata non parametrica**.

Quando i dati raccolti sono relativi a più variabili, si deve ricorrere alla **statistica multivariata**. Per molte analisi è solamente **parametrica**. Più recentemente sono stati proposti metodi, detti di ricampionamento, che sono definiti test di statistica non parametrica.

In questo corso, verranno presentati i metodi relativi

- alla statistica univariata e bivariata sia parametrica che non parametrica.

Alla fine sono presentati anche il bootstrap e il jackknife, test non parametrici applicabili sia distribuzioni univariate, sia bivariate che multivariate.

La serie completa degli argomenti e il loro ordine sono riportati nell'indice del testo, organizzato in capitoli e entro capitoli in paragrafi.

Questi concetti possono essere schematizzati in una tabella, che offre il vantaggio confrontare le finalità dei tre tipi di statistica

I - STATISTICA DESCRITTIVA

a) **Come si presentano i dati in tabelle e grafici.**
b) **Indici sintetici che descrivono la distribuzione dei dati:**

tendenza centrale, variabilità, forma.

II – STATISTICA MATEMATICA

Calcolo delle probabilità. Distribuzioni teoriche:

binomiale, poissoniana, ipergeometrica, normale, …

III – STATISTICA INFERENZIALE

(dedurre leggi generali, disponendo di un campione di dati variabili)

a) **Ipotesi parametriche (su media, varianza, …) e ipotesi funzionali (su tutta la distribuzione).**
b) **Distribuzioni univariate, bivariate, multivariate.**
c) **Statistica parametrica e non parametrica.**
d) **Test per uno, due e più campioni.**

L**a statistica inferenziale permette di trarre conclusioni su tutti i dati di una popolazione, quando se ne conoscono solamente pochi**, raggruppati in uno o più campioni.
Si supponga di voler conoscere la velocità d'accrescimento somatico di una determinata specie animale o vegetale. E' ovvio che non è possibile rintracciare e misurare tutti gli individui di quella specie, la popolazione od universo; se non altro per il tempo e le risorse che sono richiesti da questa operazione, oltre alla difficoltà teorica rappresentata dal suo continuo rinnovarsi per effetto di nuove nascite e delle morti.

E' possibile utilizzare **non tutta la popolazione** ma solamente **alcune unità**, una frazione limitatissima della popolazione: in termini tecnici, **un campione**.
Quando poi si trattasse di misurare rapporti tra organi di una specie animale, è ovvio che non è possibile sezionare tutti gli individui della specie. Nello stesso modo, per contare i globuli rossi o quelli bianchi di una persona, non è possibile estrarre tutto il sangue per un conteggio totale, ma si effettua un prelievo limitato a pochi centimetri cubici.

LA PROCEDURA PER UN

TEST D'INFERENZA

Dal Campione variabile all'Universo

La logica o filosofia scientifica per la scoperta delle leggi della natura

I – **IPOTESI**

Ipotesi nulla, ipotesi alternativa

II - **RACCOLTA DEI DATI**

Tipo di scala; caratteristiche della distribuzione dei dati

III - **SCELTA DEL TEST**

Sulla base dell'ipotesi, del tipo di scala e delle caratteristiche dei dati

IV - **RISULTATO DEL TEST- PROBABILITA'**

Probabilità di ottenere quel risultato, nella condizione espressa dall'ipotesi nulla

V – **DECISIONE - SCELTA TRA LE DUE IPOTESI**

Probabilità α

VI – **ANALISI DEL TEST E DEI DATI PER UN NUOVO ESPERIMENTO**

Potenza a posteriori e a priori, probabilità β

Tuttavia **le conclusioni non devono essere limitate ai pochi (o anche molti) casi realmente raccolti, misurati ed analizzati; ma devono essere generali,** estese a tutti gli individui della specie o a tutto l'organismo.

Ricoprono effettivo interesse non le conclusioni che restano limitate al caso del campione, ai dati utilizzati, ma quelle che sono estese a tutta la popolazione o universo. Solo in questo modo, la ricerca riveste una importanza generale e contribuisce alla costruzione di teorie scientifiche, di modelli o semplicemente di ipotesi che possono essere universalmente validi.

Una condizione essenziale e preliminare all'uso dei metodi di statistica inferenziale è che **il campione sia corretto, che non riporti in modo distorto od alterato la frequenza delle caratteristiche presenti nella popolazione**.

Lo schema precedente elenca i vari passaggi logici che sono necessari. Nello svolgimento del programma, saranno descritti dettagliatamente in tutte le loro fasi.

1.3. TIPI DI DATI E SCALE DI MISURAZIONE

Nell'analisi statistica, occorre porre sempre molta attenzione alle caratteristiche dei dati. Già la fase dell'esperimento che conduce alla raccolta delle informazioni è un punto fondamentale, poiché da essa dipendono sia i metodi di descrizione, sia i test da applicare.

Schematicamente, esistono **due tipi di variabili casuali,** alle quali sono associati **due tipi di dati:** le **variabili qualitative** e le **variabili quantitative.**

Le variabili qualitative o **categoriali** sono quantificate con **conteggi**, ossia con **numeri interi e discreti**. Ad esempio, per valutare gli effetti di un tossico è possibile contare quante cavie muoiono o sopravvivono; con un farmaco, quanti pazienti guariscono o restano ammalati, entro un tempo prefissato; con esperimenti sulle leggi di Mendel, quante piante hanno fiori rossi o bianchi.

Le variabili quantitative richiedono **risposte numeriche,** espresse su una **scala continua**. Ad esempio, per un'analisi del dimorfismo animale, dopo la separazione in maschi e femmine, si possono misurare il peso e l'altezza di ogni individuo.

I **dati** che si raccolgono per analisi statistiche possono quindi essere **discreti o continui**. Questa suddivisione, ormai storica nella presentazione ed elaborazione dei dati, è stata resa più chiara e funzionale dalla classificazione delle scale di misurazione proposta dallo psicologo S.S. **Stevens** nel 1946, (vedi l'articolo ***On the theory of scales of measurement***, pubblicato su **Science**, vol. 103, pp.:677-680). Tale classificazione è stata aggiornata nel 1951 con **le operazioni statistiche "ammissibili"** e in seguito **divulgata da S. Siegel**, nel suo manuale "***Statistica non parametrica***" del 1956.

Una presentazione ampia e dettagliata può essere trovata pure nell'ultima edizione del testo di **S. Siegel** e **N. J. Castellan** del 1988 (***Nonparametric Statistics for the Behavioral Sciences***, 2nd ed., Mc Graw Hill, New York), tradotto anche in italiano.

Le **misure possono essere raggruppate in 4 tipi di scale,** che godono di proprietà formali differenti; di conseguenza, esse ammettono operazioni differenti. Come per tutte le discipline, **una scala di misurazione** dei fenomeni biologici ed ambientali può essere:

1) **nominale** o **classificatoria**;
2) **ordinale** o per **ranghi**;
3) ad **intervalli**;
4) di **rapporti**.

1.3.1. La **scala nominale** o **classificatoria** è il livello più basso di misurazione. E' utilizzata quando i risultati possono essere classificati o raggruppati in categorie qualitative, dette anche nominali ed eventualmente identificate con simboli. I caratteri nominali, detti anche "sconnessi", costituiscono variabili le cui modalità o attributi non assumono alcun ordine precostituito. In una popolazione animale si possono distinguere gli individui in maschi e femmine, contando quanti appartengono ai due gruppi; con una classificazione a più voci, possono essere suddivisi e contati secondo la loro specie.

Nella scala **nominale o qualitativa**, esiste **una sola relazione**, quella di **identità**: **gli individui attribuiti a classi diverse sono tra loro differenti, mentre tutti quelli della stessa classe sono tra loro equivalenti**, rispetto alla proprietà utilizzata nella classificazione.

Un caso particolare è quello dei caratteri dicotomici che possono assumere solo due modalità, spesso indicate in modo convenzionale con 0 e 1 oppure + (più) e – (meno).

L'attribuzione di numeri per identificare categorie nominali, come avviene per individuare i giocatori nei giochi di squadra, è solamente un artificio che non può certamente autorizzare ad elaborare quei numeri come se fossero reali, ad esempio calcolandone la media. Quando per la classificazione dei gruppi al posto di nomi vengono usati numeri, si utilizza solo la funzione di identificazione degli elementi numerici come se fossero simboli; ma con tale trasformazione non si determina una informazione differente dalla precedente o ad essa aggiuntiva.

L**'operazione ammessa** è il **conteggio** degli individui o dei dati presenti in ogni categoria.

I **quesiti statistici** che possono essere posti correttamente riguardano le fr**equenze, sia assolute che relative**.

Sono possibili confronti tra frequenze osservate (es.: "Una classe è significativamente più numerosa dell'altra? Le varie classi hanno tutte lo stesso numero di individui, escludendo le variazioni casuali?") oppure tra le frequenze osservate e le rispettive frequenze attese sulla base di leggi biologiche, ipotesi od altro (es.: "I risultati ottenuti da un esperimento sulle leggi di Mendel sono in accordo con la sua distribuzione teorica?").

1.3.2. La **scala ordinale** o per **ranghi** rappresenta una misurazione che contiene una quantità di informazione immediatamente superiore a quella nominale; essa assume modalità logicamente sequenziali, non importa se in ordine crescente o decrescente.

Alla proprietà precedente di **equivalenza tra gli individui della stessa classe, si aggiunge una graduazione tra le classi o tra individui di classi differenti**.

Con la precedente scala nominale, si ha la sola informazione che gli individui appartenenti a gruppi differenti sono tra loro diversi, ma non è possibile **stabilire un ordine**.

Con la scala per ranghi, le differenti classi possono essere ordinate sulla base dell'intensità del fenomeno. (es.: Si supponga che il risultato di un reagente sia di colorare in verde una serie di provette, secondo la quantità di sostanza contenuta. E' possibile mettere in ordine le provette secondo l'intensità del colore, per avere una stima approssimata della quantità di sostanza contenuta. Se si confrontano tre o più provette con intensità di colore differente, è facile stabilirne l'ordine; rimane impossibile misurare ad occhio la quantità di colore di ognuna e la differenza esistente tra esse).

Questa misura ha un limite fondamentale. **In una scala ordinale, non è possibile quantificare le differenze di intensità tra le osservazioni**.

Alcune risposte, apparentemente definite a livello qualitativo o nominale, in realtà possono contenere una scala ordinale o di rango, seppure con molte ripetizioni. E' il caso della suddivisione in giovane, adulto ed anziano per l'età; oppure della classificazione in insufficiente, sufficiente, discreto, buono ed ottimo in valutazioni di merito.

Forniscono l'informazione di una scala ordinale anche

- misure che sono rappresentate con simboli, come

--, -, =, +, ++.

- raggruppamenti convenzionali o soggettivi in classi di frequenza variabili come

0, 1-2, 3-10, 11-50, 51-100, 101-1.000, >1.000

Resta l'impossibilità di valutare quanto sia la distanza tra insufficiente e sufficiente; oppure se essa sia inferiore o superiore alla distanza tra buono ed ottimo.

La scala ordinale o per ranghi è pertanto una scala monotonica. Alle variabili così misurate è possibile applicare una serie di **test non parametrici**; ma non quelli parametrici. In questi casi, non sarebbe possibile utilizzare quei test che fanno riferimento alla distribuzione normale, i cui parametri essenziali sono la media e la varianza, poiché non si possono definire le distanze tra i valori.

Tuttavia questa indicazione di massima sulla utilizzazione della statistica non parametrica è spesso superata dall'osservazione che variabili discrete o nominali tendono a distribuirsi in modo approssimativamente normale, quando il numero di dati è sufficientemente elevato. Per coloro che sono nella fase iniziale delle applicazioni statistiche, permane sempre molta incertezza sulla scelta dei test più appropriati; infatti permane un'ampia varietà di opinioni su quando il numero di osservazioni sia sufficientemente elevato, per ottenere una distribuzione normale. Nel seguito del corso, l'argomento sarà discusso in molte situazioni reali, a chiarimento dei criteri di scelta dei test.

1.3.3. La **scala ad intervalli aggiunge la proprietà di misurare le distanze o differenze tra tutte le coppie di valori**. La scala di intervalli si fonda su una misura oggettiva e costante, anche se il punto di origine e l'unità di misura sono arbitrari. Esempi classici di scale ad intervalli sono la **temperatura** (misurata in gradi **Celsius o Fahrenheit**, ma non Kelvin) ed il **tempo** (misurato secondo **calendari** differenti). Valori di temperatura, oltre a poter essere facilmente ordinati secondo l'intensità del fenomeno, godono della proprietà che **le differenze tra loro sono direttamente confrontabili e quantificabili**; le date di qualsiasi calendario, non importa se gregoriano, islamico, ebraico o cinese, possono essere tra loro ordinate dalla più antica a quella più recente e le **differenze** temporali sono **misurate con precisione oggettiva**.

Ma la scala ad intervalli ha un limite, non gode di un'altra proprietà importante nella elaborazione statistica dei dati, quella del rapporto tra coppie di misure

Ad esempio, una temperatura di 80 gradi Celsius non è il doppio di una di 40 gradi. Se una persona ponesse la mano destra in una bacinella con acqua a 80 gradi e la mano sinistra in un'altra con acqua a 10 gradi, non direbbe certamente che la prima scotta 8 volte più della seconda, ma solo che la prima è molto calda e la seconda fredda.

In una **scala ad intervalli, solo le differenze tra i valori sono quantità continue e isomorfe** alla struttura dell'aritmetica. **Solo per le differenze sono permesse tutte le operazioni**: possono essere tra loro sommate, elevate a potenza oppure divise, determinando le quantità che stanno alla base della statistica parametrica.

Da una scala d'intervalli è possibile scendere ad una scala di ranghi (es.: utilizzando solo l'informazione dell'ordine dei valori) oppure ad una scala nominale (es.: suddividendo le misure in alte e basse, sopra o sotto un valore prefissato). Pertanto, la scala d'intervalli gode anche delle proprietà definite per le due scale precedenti.

Nella presentazione dei test non parametrici verranno discusse le situazioni in cui, avendo dati misurati su scale d'intervalli o di rapporti, **è conveniente scendere nel tipo di scala** seppure con una perdita d'informazione.

1.3.4. La **scala di rapporti ha il vantaggio di avere un'origine reale**. Sono tipiche scale di rapporti l'altezza, la distanza, la velocità, l'età, il peso, il reddito, la temperatura in gradi Kelvin; più in generale, tutte quelle misure in cui **0 (zero) significa quantità nulla**.

Non solo le differenze, ma **gli stessi valori possono essere moltiplicati o divisi** per quantità costanti, senza che l'informazione di maggiore importanza, il rapporto tra essi, ne risulti alterata.

Alle variabili misurate con una **scala di rapporti, il tipo di misurazione più sofisticato e completo**, può essere applicato qualsiasi test statistico. Possono essere utilizzati anche la media

geometrica ed il coefficiente di variazione, i quali richiedono che il punto 0 (zero) sia reale e non convenzionale.

Pure con una scala di rapporti è possibile scendere nella scala di misurazione, trasformandola in una scala di rango o addirittura qualitativa. Ovviamente, si ha una perdita ancor più rilevante della quantità d'informazione, che essa fornisce; di conseguenza, rappresenta un'operazione che deve essere evitata, quando non imposta da altre condizioni dell'analisi statistica o dalle caratteristiche della distribuzione dei dati.

Riassumendo i concetti fondamentali esposti,

- nella **scala nominale**, esistono solo relazioni di ***equivalenza***;
- in quella **ordinale**, alla precedente si aggiungono relazioni di ***minore o maggiore di***;
- in quella **ad intervalli** alle due precedenti si aggiunge la relazione di ***rapporto tra ogni coppia d'intervalli***;
- nella **scala di rapporti** si ha anche la quarta relazione di ***rapporto conosciuto tra ogni coppia di valori***.

Come sarà più volte discusso nei prossimi capitoli, anche nella ricerca e nella gestione ambientali occorre **porre estrema attenzione al reale significato da attribuire ai valori numerici che vengono utilizzati**. Si possono avere numeri che apparentemente hanno le stesse caratteristiche, ma che in realtà richiedono elaborazioni diverse ed impongono il ricorso a test differenti, per rispondere ai medesimi quesiti. Per esempio, i grammi di una determinata sostanza inquinante sciolta in un litro d'acqua, la percentuale di questa sostanza sul peso complessivo, il punteggio della qualità dell'acqua determinata dalla presenza di quella sostanza sono misure che utilizzano scale diverse.

- Nel primo caso, si ha una classica scala di rapporti ed è possibile usare test parametrici, se la distribuzione dei dati è normale;
- nel secondo caso, è possibile utilizzare le stesse procedure statistiche e gli stessi test parametrici, solamente dopo apposita trasformazione dei valori;
- nel terzo, si ha una scala di ranghi, poiché la reale informazione fornita da questa serie di punteggi è solo quella di una graduatoria della qualità, nella quale non hanno reale significato né i rapporti né le differenze tra loro.

1.4. CLASSIFICAZIONE IN TABELLE

Un insieme di misure è detto serie statistica o serie dei dati. Quando la serie non è ordinata, si ha un insieme disordinato di numeri che non evidenzia le caratteristiche fondamentali del fenomeno.

Una sua prima ed elementare elaborazione può essere una distribuzione ordinata di tutti i valori, in modo crescente o decrescente, detta **seriazione**.

Il **valore minimo** e il **valore massimo** insieme permettono di individuare immediatamente il **campo (**od **intervallo) di variazione**.

Successivamente, la serie può essere raggruppata in classi, contando quanti valori od **unità statistiche** appartengono ad ogni gruppo o categoria.

Si ottiene una **distribuzione di frequenza o di intensità**, detta anche semplicemente **distribuzione**.

Come prima applicazione, è utile considerare un caso semplice: una **variabile discreta** ottenuta da un conteggio del numero di foglie, germogliate su 45 rami di lunghezza uguale.

Tabella 1. Numero di foglie contate su 45 rami.

5	6	3	4	7	2	3	2	3	2	6	4	3	9	3
2	0	3	3	4	6	5	4	2	3	6	7	3	4	2
5	1	3	4	3	7	0	2	1	3	1	5	0	4	5

Il primo passaggio, quasi intuitivo in una distribuzione discreta, consiste nel **definire le classi**:

- è sufficiente identificare il **valore minimo (0,** nei dati della tabella**) e quello massimo (9)**,
- contando quante volte compare ogni **modalità di espressione** (cioè quanti sono i rami con un numero di foglie uguali).

Queste informazioni di norma sono presentate in una tabella impostata come la seguente:

Tabella 2. Distribuzione di frequenze assolute e relative delle foglie in 45 rami.

Classe	**x**	0	1	2	3	4	5	6	7	8	9
Freq. Assoluta	**n**	3	3	7	12	7	5	4	3	0	1
Freq. Relativa	**f**	0,07	0,07	0,15	0,27	0,15	0,11	0,09	0,07	0,00	0,02
Freq. Cumulata	**---**	0,07	0,14	0,29	0,56	0,71	0,82	0,91	0,98	0,98	1,00

in cui:

- la **classe** è una modalità di espressione (in questo caso un valore o conteggio);
- la **frequenza assoluta della classe** è il numero di volte con la quale compare ogni valore;
- la **frequenza relativa della classe** è la sua frequenza assoluta divisa per il numero totale;
- la **frequenza cumulata di una classe** (che può essere stimata con quelle assolute e/o con quelle relative) è la somma di tutte le frequenze delle classi minori con quella della classe stessa.

La trasformazione da frequenza assoluta a frequenza relativa risulta utile quando si vogliono confrontare due o più distribuzioni, che hanno un differente numero complessivo di osservazioni.
La frequenza cumulata offre informazioni importanti quando si intende stimare il numero totale di osservazioni inferiore (o superiore) ad un valore prefissato (ad es.: il 71% dei rami ha meno di 5 foglie; il 56% ha un massimo di 3 foglie).
La distribuzione dei dati e la distribuzione delle frequenze cumulate forniscono informazioni non dissimili, essendo possibile passare con facilità dall'una all'altra. Sono diverse nella loro forma, come si vedrà con maggiore evidenza nelle rappresentazioni grafiche. **La prima ha una forma a campana, la seconda una forma a S, di tipo asintotico; si prestano ad analisi differenti** e la scelta è fatta sulla base del loro uso statistico.

La distribuzione di frequenza offre una lettura rapida delle caratteristiche più importanti della serie di dati. Nella tabella precedente, il ramo "tipico" ha 3 foglie; se dovessimo sintetizzare con un solo valore il numero di foglie presenti sui rami raccolti diremmo 3, che rappresenta la **tendenza centrale**. Altra caratteristica importante è il numero minimo e il numero massimo, 0 e 9, che insieme forniscono il campo di variazione, una indicazione della **variabilità o dispersione**. La distribuzione del numero di foglie tende ad diminuire in modo simile allontanandosi da 3, seppure mantenga frequenze più alte nelle classi con un numero maggiore di foglie: sono indicazioni sulla **forma** della distribuzione, che in questo esempio non è simmetrica (ma asimmetrica) rispetto alla tendenza centrale, a causa di un eccesso dei valori più alti.

Nella costruzione di tabelle sintetiche (come la tabella 2 rispetto alla 1) uno dei problemi più rilevanti è **quante classi di frequenza costruire**. La scelta dipende strettamente dal numero totale **N** di osservazioni e, in misura minore, dalla variabilità dei dati.
Se, in riferimento alla dimostrazione precedente, i dati fossero stati in numero inferiore ai 45 presentati (ad esempio i 15 valori della prima riga), il campo di variazione sarebbe stato più ridotto (non più da 0 a 9, ma da 2 a 9). Le classi non sarebbero state 10 come prima, ma solamente 8.

Tuttavia, come si può osservare dai dati, 8 classi per 15 osservazioni sarebbero ugualmente un numero troppo alto, per riuscire ad evidenziare e rappresentare in modo corretto le caratteristiche principali e la forma reale della distribuzione.

Le distribuzioni di frequenza tendono a mostrare la distribuzione reale del fenomeno solo quando è possibile utilizzare un numero sufficientemente elevato di osservazioni.
L'esperienza ha insegnato che il numero di classi abitualmente varia da un minimo di 4-5 (con N = 10-15) ad un massimo di 15-20 (con N > 100), in funzione del numero complessivo di osservazioni.
Un numero troppo basso di classi, raggruppando eccessivamente i dati, determina una perdita di informazione sulle caratteristiche della distribuzione e la rende non significativa; è intuitivo che una o due sole classi determinano l'impossibilità di evidenziare qualunque caratteristica della distribuzione. Inversamente, ma con un risultato finale simile, un numero troppo elevato di classi disperde i valori e non rende manifesta la forma della distribuzione.

Per stimare in modo oggettivo il numero di classi, sono stati proposti vari metodi; tra essi è utile ricordarne due:
1 - quello di **H. Sturges** che nel 1926, sulla base del numero di osservazioni **N**, ha indicato il **numero ottimale di classi** **C** con

$$C = 1 + \frac{10}{3} \cdot \log_{10}(N)$$

2 - quello di **D. Scott** che nel 1979 ha determinato l'**ampiezza ottimale h delle classi** (dalla quale ovviamente dipende direttamente anche il numero di classi C), mediante la relazione

$$h = \frac{3{,}5 \cdot S}{\sqrt{N}}$$

dove
- **S** è la **deviazione standard**,

che sarà presentata più avanti tra le misure di variabilità dei dati.
Nella costruzione di distribuzioni di frequenza, non è strettamente obbligatorio utilizzare intervalli uguali, anche se è prassi consolidata per una lettura più semplice. Nel caso di classi di ampiezza diversa, la rappresentazione grafica ed il calcolo dei parametri fondamentali esigono alcune avvertenze, non sempre intuitive (di seguito presentate).

Nel caso di una **variabile continua**, il raggruppamento in classi richiede alcuni accorgimenti ulteriori rispetto a quelli utilizzati per una variabile discreta. Si supponga che sia stata misurata l'altezza in cm. di 40 giovani piante della stessa specie, arrotondata all'unità per semplificazione.

Tabella 3. Altezza in cm. di 40 giovani piante.

107	83	100	128	143	127	117	125	64	119
98	111	119	130	170	143	156	126	113	127
130	120	108	95	192	124	129	143	198	131
163	152	104	119	161	178	135	146	158	176

E' evidente come non sia conveniente fare una classe per ogni cm., in analogia a quanto fatto con i dati della tabella 1. In questo caso, il numero di modalità sarebbe nettamente superiore al numero di osservazioni, anche se il campione avesse un numero di osservazioni doppio o triplo. Di conseguenza, si impone la necessità di un raggruppamento in classi, che comprendano più modalità di espressione.

Una volta individuato il valore minimo e quello massimo (64 e 198), si stabilisce l'intervallo di variazione (198 - 64 = 134). Nella formazione delle classi, il limite inferiore della prima classe ed il limite superiore dell'ultima classe non devono essere necessariamente i valori osservati, ma li devono ovviamente comprendere. E quindi possibile costruire un campo di variazione, ad esempio di 140 cm. (sempre più ampio di quello calcolato), partendo da cm. 60 e arrivando a cm. 199 compresi. Sulla base del numero di dati (40), si decide il numero di classi. Nel caso specifico, potrebbero essere 7 classi, con un'ampiezza di 20 cm. ognuna.

E' necessario **definire con precisione il valore minimo e quello massimo di ogni classe**, onde evitare incertezze nell'attribuzione di un singolo dato tra due classi contigue. Con i dati dell'esempio, le classi possono essere 60-79 la prima, 80-99 la seconda, 100-119 la terza e così via fino a 180-199 per l'ultima.

E' da evitare la suddivisioni in classi come 60-80, 80-100, 100-120, …

Poiché la scala è continua, i cm. riportati devono essere intesi con almeno 2 cifre decimali, per cui nella classe 60-79 il primo numero deve essere inteso come 60,00 cm. e 79 come 79,99; nello stesso modo la classe 180-199 deve essere intesa tra i cm. 180,00 e 199,99.

Nonostante le indicazioni di massima presentate, la determinazione dei valori estremi, del numero di classi e dell'intervallo di ogni classe è ampiamente soggettiva. Nella costruzione di una tabella, **la scelta soggettiva di una particolare serie o di un'altra può tradursi in una rappresentazione completamente diversa degli stessi dati**. Per piccoli campioni, l'alterazione e le differenze possono essere sensibili; ma all'aumentare del numero di osservazioni, gli effetti delle scelte soggettive, quando non siano estreme, incidono sempre meno sulla concentrazione dei valori e sulla forma della distribuzione.

Tra le altre avvertenze importanti, è da ricordare che la classe iniziale e quella terminale non devono essere **classi aperte** (come < 80 quella iniziale e ≥ 180 quella finale). Con classi estreme aperte, si perde l'informazione del loro valore minimo o massimo e quindi del valore centrale di quella classe; la conseguenza è la perdita di un dato indispensabile, per calcolare la media della classe e quella totale, nonché tutti gli altri parametri da essa derivati. Come verrà successivamente chiarito, con tabelle in cui le classi estreme sono aperte viene impedita o resa soggettiva anche la loro rappresentazione grafica, per la quale è indispensabile conoscere con precisione il valore iniziale e quello terminale.

I dati della tabella 3 possono essere riportati in modo più schematico e più comprensibile, come nella seguente tabella 4.

Tabella 4. Distribuzione di frequenza assoluta e relativa (in %) dell'altezza di 40 giovani piante.

Classe	X_i	60-79	80-99	100-19	120-39	140-59	160-79	180-99
Freq. Assoluta	n_i	1	3	10	12	7	5	2
Freq. Relativa %	f_i	2,5	7,5	25,0	30,0	17,5	12,5	5,0
Freq. Cumulata	---	2,5	10,0	35,0	65,0	82,5	95,0	100,0

Rispetto all'elenco grezzo dei dati, la tabella di distribuzione delle frequenze fornisce in modo più chiaro le indicazioni elementari contenute, in particolare la loro

- **posizione o dimensione** (già chiamata anche tendenza centrale**)** e
- la **variabilità o dispersione**.

Per evidenziare sia queste che altre caratteristiche della distribuzione dei dati raccolti, sovente è di aiuto una rappresentazione grafica che mostra in modo sintetico soprattutto

- la **forma**, come la **simmetria** e la **curtosi**, quando si tratti di grandi gruppi di dati.

Ritornando al problema della rappresentazione tabellare dei dati riportati in tabella 3, secondo le indicazioni di **Sturges** il **numero di classi C** avrebbe dovuto essere

$$C = 1 + \frac{10}{3} \cdot \log_{10}(N) = 1 + \frac{10}{3} \cdot \log_{10}(40) = 6{,}34$$

uguale a 6,34

dal quale si deduce anche un'ampiezza **h** = $\frac{140}{6{,}34} \cong 22$

circa 22 centimetri.

Secondo le indicazioni di **Scott**, **l'ampiezza h delle classi** avrebbe dovuto essere

$$h = \frac{3{,}5 \cdot s}{\sqrt{N}} = \frac{3{,}5 \cdot 28{,}618}{6{,}3246} = 15{,}837$$

uguale a circa 16,

dalla quale si deduce un numero di classi C

$$C = \frac{140}{15{,}837} = 8{,}84$$

uguale a 9 (8,84).

Ovviamente, il numero di classi calcolato (C = 8,84) deve essere arrotondato all'unità.

Secondo i due metodi proposti, con i dati della tabella 3 il numero di classi può ragionevolmente variare da 6 a 9; si evidenzia la correttezza della scelta di fare 7 classi, suggerita dalla semplicità di formare classi con un'ampiezza di 20 cm.

La rappresentazione dei dati in una tabella di frequenza offre i vantaggi descritti; ma soffre anche di alcune controindicazioni. Lo **svantaggio maggiore** deriva dal

- **non poter conoscere come sono distribuiti i dati entro ogni classe**.

Per stimare i parametri della distribuzione (media, varianza, simmetria, curtosi), viene quindi usato **il valore centrale** di ogni classe,

- **nell'ipotesi che in quell'intervallo i dati siano distribuiti in modo uniforme**.

Rispetto alla distribuzione delle singole osservazioni, questa procedura comporta un'approssimazione, poiché

- **tale ipotesi operativa implicita non è vera**

(il concetto sarà ripreso e dimostrato in paragrafi successivi).

1.5. RAPPRESENTAZIONI GRAFICHE DI DISTRIBUZIONI UNIVARIATE

Le rappresentazioni grafiche servono per evidenziare in modo semplice, ***a colpo d'occhio***, le quattro caratteristiche fondamentali di una distribuzione di frequenza (tendenza centrale, variabilità, simmetria e curtosi). Insieme con i vantaggi di fornire una visione sintetica e di essere di facile lettura, hanno però **l'inconveniente fondamentale di mancare di precisione e soprattutto di essere soggettive**, quindi di permettere letture diverse degli stessi dati. Pertanto, ai fini di una elaborazione mediante i test e di un confronto dettagliato dei parametri, è sempre preferibile la tabella, che riporta i dati esatti.

Nell'introdurre le rappresentazioni grafiche, seppure nel caso specifico parli di diagrammi (come quello di dispersione che in questo testo è presentato nel capitolo della regressione), Sir **Ronald A. Fisher** nel suo volume del 1958 "***Statistical Methods for Research Workers***, (13^{th} ed. Oliver and Boyd, Edinburgh, 356 p.) espone con chiarezza **i motivi** che devono spingere il ricercatore a **costruire rappresentazioni grafiche** dei suoi dati:

- un **esame preliminare** delle caratteristiche della distribuzione,
- un **suggerimento per il test** da scegliere, adeguato appunto ai dati raccolti,
- un aiuto alla **comprensione delle conclusioni**,
- **senza per questo essere un test**, ma solo una descrizione visiva.

"The preliminary examination of most data is facilited by use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them".

Le **rappresentazioni grafiche** proposte sono numerose. Esse debbono essere scelte in rapporto al tipo di dati e quindi alla scala utilizzata.

Per **dati quantitativi**, riferiti a variabili continue misurate su scale ad intervalli o di rapporti, di norma si ricorre a **istogrammi o poligoni**. Gli istogrammi sono grafici a barre verticali (per questo detti anche diagrammi a rettangoli accostati), nei quali

- le misure della variabile casuale sono riportate lungo l'asse orizzontale,
- mentre l'asse verticale rappresenta il numero assoluto, oppure la frequenza relativa o quella percentuale, con cui compaiono i valori di ogni classe.

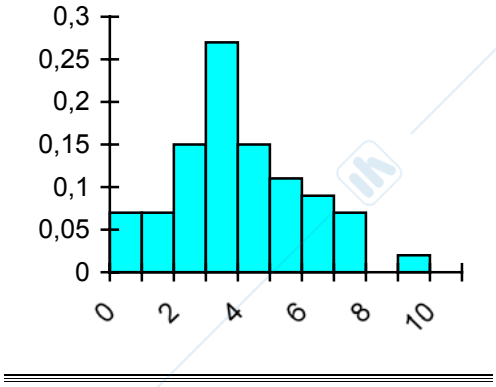


Figura 1. Istogramma dei dati di Tab. 2
(frequenze relative)

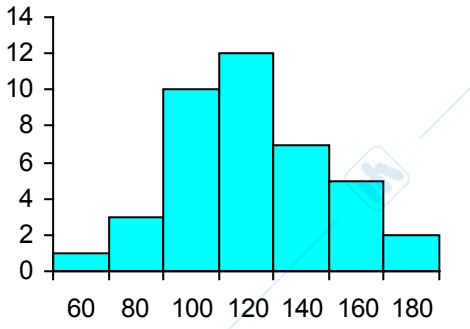


Figura 2. Istogramma dei dati di Tab. 4
(Valore iniz. =60; Valore finale =199; Passo =20; Classi=7)

I lati dei rettangoli sono costruiti in corrispondenza degli estremi di ciascuna classe.

Un istogramma deve essere inteso come una **rappresentazione areale**: sono le superfici dei vari rettangoli che devono essere proporzionali alle frequenze corrispondenti.

Quando **le classi hanno la stessa ampiezza**, le basi dei rettangoli sono uguali; di conseguenza, le loro altezze risultano proporzionali alle frequenze che rappresentano.

Solo quando le basi sono uguali, è indifferente ragionare in termini di altezze o di aree di ogni rettangolo.

Ma se le ampiezze delle classi sono diverse, bisogna ricordare il concetto generale che

- **le frequenze sono rappresentate dalle superfici e quindi è necessario rendere l'altezza proporzionale**.

Tale proporzione è facilmente ottenuta dividendo il numero di osservazioni per il numero di classi contenute nella base, prima di riportare la frequenza sull'asse verticale.

Per esempio, con i dati della precedente figura 2, si supponga di avere raggruppato in una classe sola le frequenze della classe da 80 a 99 e da 100 a 119, per un totale di 13 osservazioni (3 + 10).

Nella successiva figura 3, tale somma è rappresentata

- nel primo caso (istogramma di sinistra) con un grafico errato
- nel secondo caso (istogramma di destra) nella sua versione corretta, che utilizza il valore medio delle classi raggruppate.

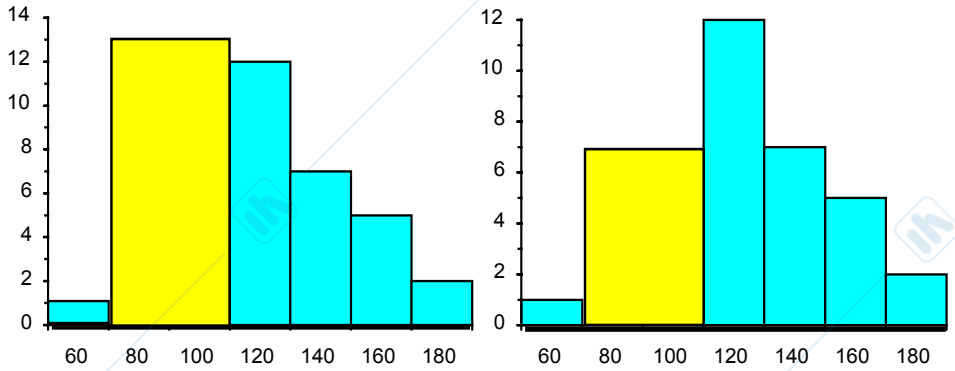


Figura 3. Istogrammi dei dati di Tab. 4

Somma errata di due classi : 2ª e 3ª della figura precedente

Somma corretta di due classi : 2ª e 3ª della figura precedente

Un'altra avvertenza importante nella costruzione degli istogrammi è che l'asse verticale, che riporta le frequenze, deve mostrare lo zero reale od "origine", onde non distorcere o travisare le caratteristiche dei dati ed i rapporti tra essi. In relazione alle caratteristiche della distribuzione dei dati, **la larghezza o base del rettangolo non ha alcun significato** e può essere scelta a piacimento; dipende solamente dal numero di classi che si vogliono rappresentare sull'asse delle ascisse.

Anche **il rapporto tra l'altezza dell'asse delle ordinate e la lunghezza delle ascisse può essere scelto a piacimento e non ha alcun significato statistico**. Tuttavia, sulla prassi di costruire figure graficamente eleganti, le dimensioni utilizzate dai programmi informatici seguono uno schema che è ormai uguale per tutti. E' quasi sempre praticato un accorgimento che ha una finalità esclusivamente **estetica**: per costruire una **relazione armonica tra gli elementi del grafico**, è uso corrente che tutto il disegno dell'istogramma debba essere contenuto in un rettangolo virtuale, in cui **l'altezza sia i 2/3 della base** o, come riportano altri testi per fornire lo stesso concetto, **la base sia 1,5 volte l'altezza.** (Gli esempi riportati sono stati costruiti con programmi informatici a grande diffusione internazionale, che seguono questa regola estetica).

La rappresentazione grafica permette di valutare con immediatezza se il numero di classi costruite è adeguato alle caratteristiche della distribuzione originale dei dati. Con poche eccezioni, le variabili quantitative di fenomeni biologici od ambientali evidenziano una **distribuzione normale**, con caratteristiche specifiche di addensamento verso i valori centrali e di dispersione più o meno simmetrica, ma con declino regolare verso i due estremi.

La rappresentazione grafica deve essere in grado di non alterare od interrompere la regolarità della distribuzione, come può avvenire in particolare quando il numero di classi è troppo alto rispetto al numero di dati. Una dimostrazione di tale caso è riportato nella figura successiva, nella quale sono state fatte 14 classi, invece delle 7 richieste prendendo in considerazione il numero di dati, la loro variabilità e i limiti delle singole classi.

L'istogramma che segue è una chiara dimostrazione di una suddivisione in classi eccessiva: uno o più gruppi di misure (due nell'esempio) comprese entro gli estremi hanno frequenza zero ed alterano la rappresentazione di una distribuzione normale.
La frequenza delle classi e l'altezza dei rettangoli ad essa proporzionali tendono a decrescere in modo relativamente regolare. Quando un grafico presenta una forte alterazione dalla forma normale, che scompare con una suddivisioni in classi meno frammentate, si ha una indicazione di un possibile errore tecnico nella rappresentazione dei dati.

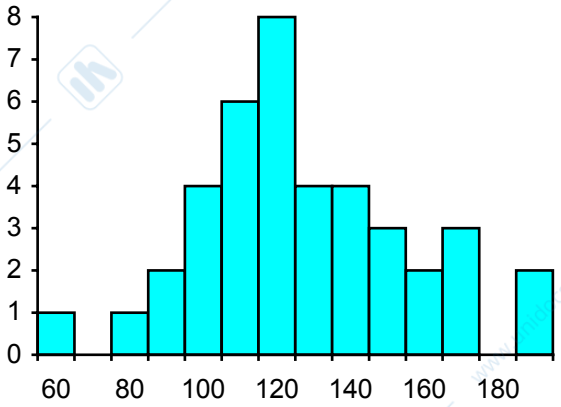


Figura 4. Istogramma dei dati di Tab. 4

(Valore iniziale = 60; Valore finale = 199; Passo = 10; Classi = 14)

(Rappresentazione grafica non adeguata, per eccessiva suddivisione in classi)

I poligoni sono figure simili agli istogrammi e sono utilizzati di norma per la rappresentazione di **valori relativi o di percentuali**, in quanto è implicito che l'area totale sottesa sia uguale a 1 o 100%. Come nel caso degli istogrammi, l'asse orizzontale rappresenta il fenomeno, mentre l'asse verticale rappresenta la proporzione o percentuale di ogni classe.

Un poligono può essere ottenuto a partire dal relativo istogramma, unendo con una linea spezzata i punti centrali di ogni classe.

La linea spezzata deve essere unita all'asse orizzontale, sia all'inizio sia alla fine, per racchiudere l'area della distribuzione. Questo procedimento viene ottenuto con un artificio, simulando la presenza di un istogramma con presenze uguali a 0 (zero) come punto di partenza. Si unisce il valore centrale della prima classe con il valore centrale di questa precedente classe fittizia di valore 0; l'ultimo segmento viene ottenuto unendo il valore centrale dell'ultima classe reale con il valore centrale di una classe successiva, fittizia, di valore 0.

Il poligono rappresentato nella figura 5 corrisponde all'istogramma della figura 2.

E' stato costruito con i dati della tabella 4, spostando le classi sull'asse delle ascisse per comprendere i nuovi estremi della distribuzione.

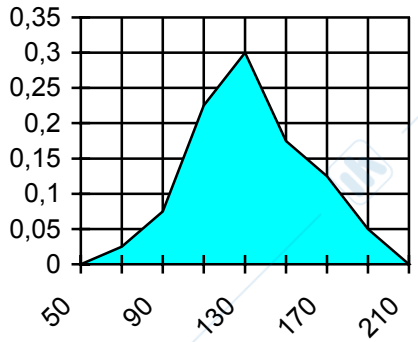


Figura 5. Poligono dei dati di Tab. 4

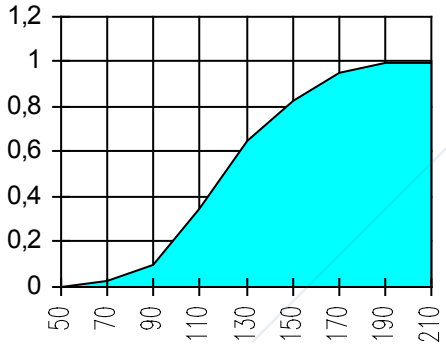


Figura 6. Poligono cumulato di Tab. 4

Le distribuzioni cumulate sono rappresentate sia con istogrammi cumulati sia con **poligoni cumulati**.

Non forniscono informazioni sostanzialmente differenti da quelle dei relativi istogrammi e poligoni già descritti, poiché

- è possibile passare con facilità da una distribuzione di frequenza alla sua cumulata con semplici operazioni di somme o di sottrazioni tra classi.

Sono solamente più convenienti, per meglio evidenziare il concetto di maggiore interesse.

La figura 6 rappresenta il poligono cumulato corrispondente al poligono della figura 5.

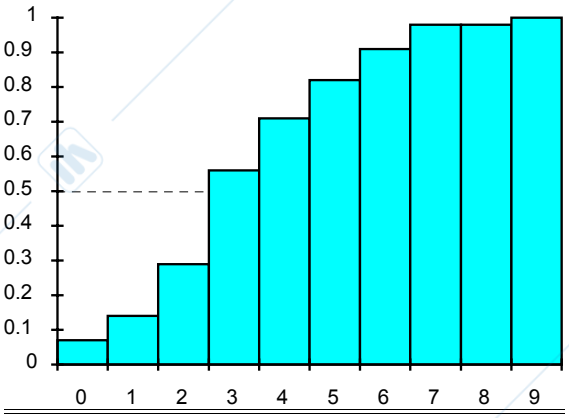


Figura 7. Istogramma cumulato dei dati di Tab. 2

Tuttavia, per la diversa prospettiva che essi offrono a partire dagli stessi dati, gli istogrammi ed i poligoni cumulati sono un altro metodo utile sia per presentare le caratteristiche di dati quantitativi riportati in tabelle, sia per facilitare l'interpretazione e l'analisi. Servono soprattutto per evidenziare, con lettura immediata, quante sono in totale le misure che sono inferiori o superiori ad un certo valore. Il valore dell'asse orizzontale che corrisponde al 50% dei valori identifica la mediana (riportato come linea tratteggiata nella figura 7 che rappresenta un istogramma cumulato); è un parametro di tendenza centrale estremamente importante, quando la distribuzione non è simmetrica (il suo uso e le sue caratteristiche saranno descritte in modo dettagliato nei prossimi paragrafi).

Per le distribuzioni di frequenza di **dati qualitativi**, le rappresentazioni grafiche più frequenti sono

- i diagrammi a rettangoli distanziati,

- gli ortogrammi,

- i diagrammi a punti,

- gli areogrammi (tra cui **i diagrammi circolari**),

- **i diagrammi a figure** (o **diagrammi simbolici**).

I **diagrammi a rettangoli distanziati**, detti anche grafici a colonne, sono formati da rettangoli con basi uguali ed altezze proporzionali alle intensità (o frequenze) dei vari gruppi considerati. A differenza degli istogrammi, i rettangoli non sono tra loro contigui, ma distaccati; di conseguenza, sull'asse delle ascisse non vengono riportati misure ordinate ma nomi, etichette o simboli, propri delle classificazioni qualitative.

Con dati qualitativi o nominali, le basi dei rettangoli sono sempre identiche avendo solo un significato simbolico.

Si può ricorre quindi sia a **diagrammi a punti** o ***line plot***, in cui i punti sono disposti uno sopra l'altro fino ad un'altezza proporzionale alla frequenza della classe, sia a **diagrammi a barre**, che sono un'altra rappresentazione frequente, in cui al posto di rettangoli o colonne di punti vengono usate linee continue più o meno spesse (figura 8).

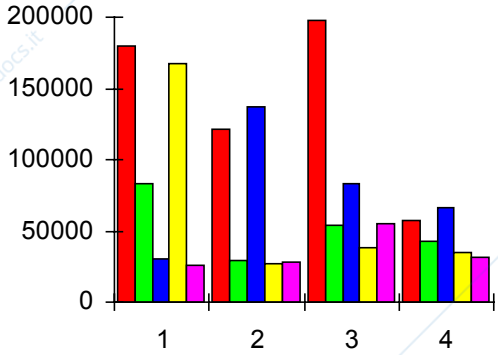


Figura 8. Rettangoli distanziati

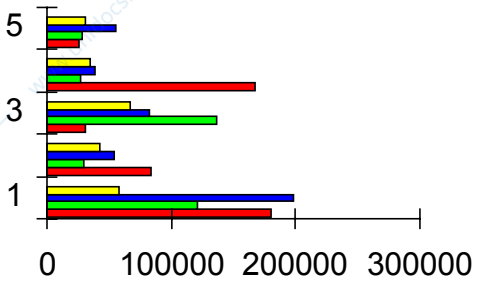


Figura 9. Ortogramma

Nel caso di **dati qualitativi o nominali**, non esiste una logica specifica nell'ordine delle classi.
Per convenzione, i rettangoli o le colonne sovente (ma non obbligatoriamente**) vengono disposti in modo ordinato dal maggiore al minore o viceversa**.
Se le classi qualitative sono composte da **sottoclassi**, è possibile una rappresentazione grafica più articolata, dividendo ogni rettangolo in più parti, con altezze proporzionali alle frequenze delle sottoclassi (figura 10). Avendo basi uguali, le aree sono proporzionali alle altezze; pertanto, anche i diagrammi a rettangoli distanziati sono rappresentazioni areali.

Gli **ortogrammi o grafici a nastri** sono uguali ai rettangoli distanziati; l'unica differenza è che gli assi sono scambiati, per una lettura più facile (figura 9 e figura 11).
Anche in questo caso è possibile sostituire ai rettangoli una linea, eventualmente punteggiata.
Si ottengono **diagrammi a barre o a punti** e l'intensità o frequenza delle varie classi viene letta con una proiezione sull'asse delle ascisse.

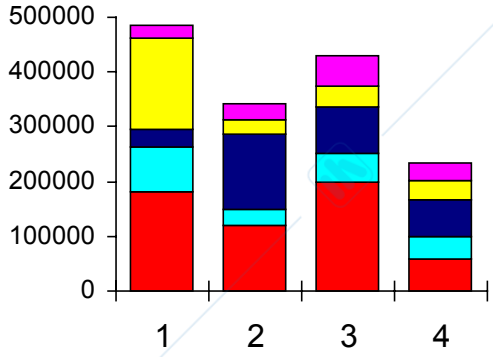


Figura 10. Rettangoli distanziati

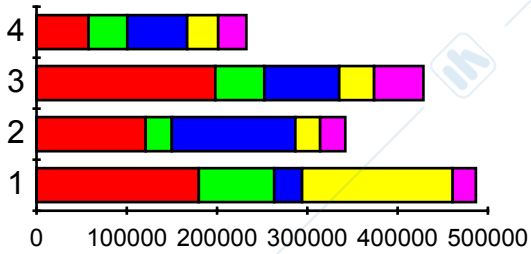


Figura 11. Ortogramma

Secondo alcuni esperti di percezione dei grafici, queste figure vengono lette con maggiore facilità rispetto ai rettangoli distanziati (l'occhio leggerebbe con maggiore facilità la proiezione verticale e di quella orizzontale) e quindi meglio rappresentano le informazioni contenute in distribuzioni di frequenza di dati qualitativi.

Gli areogrammi sono grafici in cui le frequenze o le quantità di una variabile qualitativa sono rappresentate da superfici di figure piane, come quadrati, rettangoli o, più frequentemente, cerchi oppure loro parti. La rappresentazione può essere fatta sia con più figure dello stesso tipo, aventi superfici proporzionali alle frequenze o quantità, sia con un'unica figura suddivisa in parti proporzionali.

Nel caso dei **diagrammi circolari o a torta**, si divide un cerchio in parti proporzionali alle classi di frequenza.

Gli areogrammi vengono usati soprattutto per rappresentare **frequenze percentuali**.

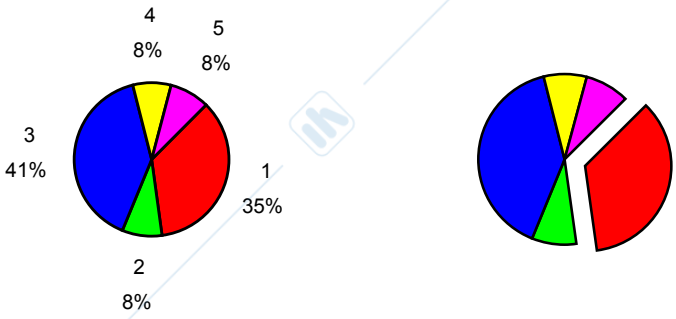


Figura 12. Diagrammi circolari

Hanno il vantaggio di fare capire con immediatezza che la somma di tutte le classi è uguale all'unità (1 o 100%); hanno l'inconveniente che evidenziano con estrema difficoltà le differenze che non sono molto marcate. Per differenze piccole, si dimostrano meno efficaci degli ortogrammi.

I diagrammi circolari sono utilizzati per distribuzioni di variabili nominali, al fine di evitare di stabilire anche involontariamente un ordine, che non esiste tra variabili qualitative. Mettono in evidenza come sono distribuite le singole parti, rispetto all'intero: il cerchio rappresenta l'intero fenomeno ed i componenti sono rappresentati da settori che sono distinti da tratteggi, colori o gradazioni di colore differenti.

Gli angoli (**a,** nella formula successiva) devono essere proporzionali alle percentuali (**Y in %**) che vogliono rappresentare,
in accordo con la relazione

a : 360 = Y in % : 100

Con i **diagrammi a figure**, detti anche **diagrammi simbolici** o **pittogrammi**, la frequenza di ogni carattere qualitativo viene rappresentata da una figura, sovente stilizzata, oppure da simboli che ricordano facilmente l'oggetto. E' una specie di istogramma costruito con figure, dove **l'altezza della figura deve essere proporzionale alla frequenza, quando le basi sono uguali**.

Questi diagrammi a figure **hanno tuttavia il grave inconveniente di prestarsi a trarre in inganno con facilità il lettore inesperto di statistica, quando sono stati costruiti con malizia; è pure facile che un ricercatore non sufficientemente esperto li costruisca in modo sbagliato, generando non volutamente il sospetto che egli abbia voluto fornire una impressione di differenza tra i gruppi a confronto non supportata dai dati**.
Per esempio, una popolazione con un numero triplo di persone rispetto ad un'altra spesso è rappresentata da una figura umana proporzionata, di altezza tripla rispetto all'altra.

L'occhio coglie complessivamente non l'altezza di ogni figura ma la superficie che essa occupa, che è il quadrato del valore che si intende rappresentare: se ne ricava l'impressione distorta di un rapporto di 9 a 1 e non di 3 a 1, come dicono in realtà i dati.
E' possibile ovviare all'inconveniente, costruendo non una figura improbabile di altezza variabile e con base uguale (poiché risulterebbe una figura alterata ed una rappresentazione forse incomprensibile), ma ricorrendo all'artificio di **figure identiche, ripetute tante volte quante**

sono le proporzioni. Per esempio, se l'unità di misura convenuta è 20 individui, 50 persone possono essere rappresentate in modo corretto da due figure umane e mezza e 105 persone da 5 figure intere più un quarto di figura.

In questo modo si rispetta la regola per costruire gli istogrammi in modo corretto: l'altezza delle due figure è proporzionale al numero di dati dei due gruppi

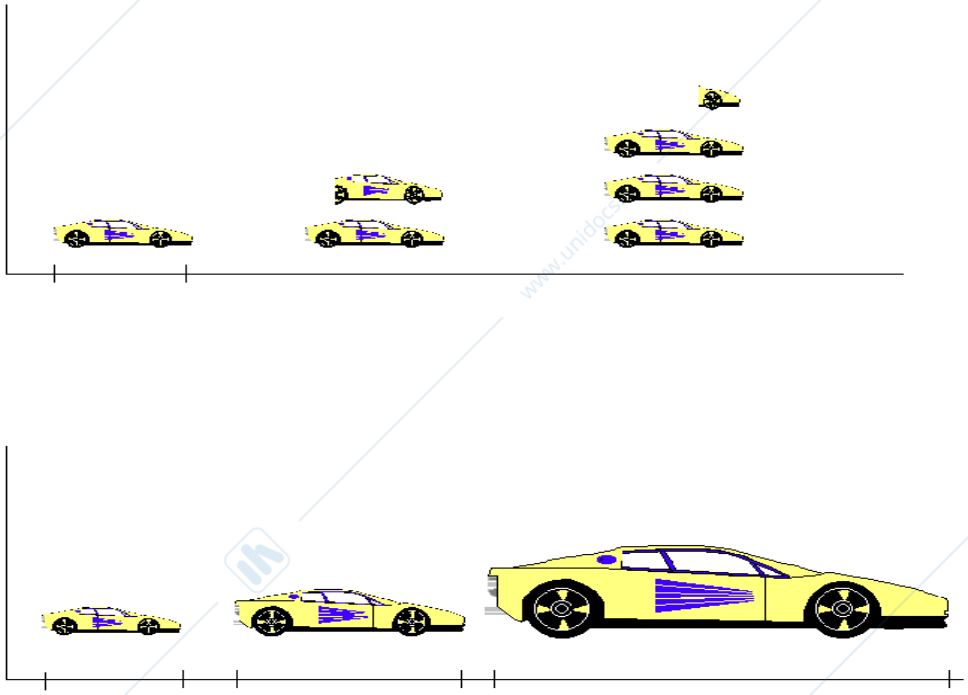


Figura 13. Pittogramma della produzione mensile di auto di 3 case automobilistiche: la prima ha prodotto 100 mila auto, la seconda 180 mila e la terza 320 mila.
La parte superiore della figura fornisce una rappresentazione corretta.
La parte inferiore, fondata sulla proporzione della lunghezza, fornisce una rappresentazione errata: è la superficie coperta dalla figura che deve essere proporzionale, non la lunghezza.

A causa degli inconvenienti, i diagrammi simbolici o a figure sono usati molto raramente nelle pubblicazioni specializzate e mai in quelle scientifiche. Sono riservati a pubblicazioni divulgative, quando è più importante l'impressione della precisione, cioè occorre evidenziare l'importanza del fenomeno a persone che non conoscono esattamente il problema.

Gli specialisti preferiscono i dati, poiché da essi sanno valutare il fenomeno e se un dato è credibile oppure anomalo con elevata probabilità.

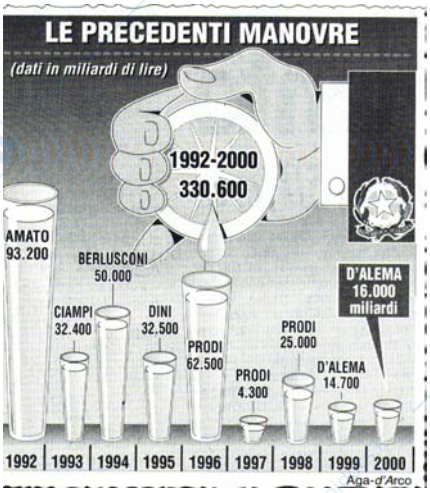


Figura 1**3b**. Esempio tratto da un quotidiano sulle manovre finanziarie in Italia, dal 1992 al 2000
Notare come i numeri diano un'informazione differente dai volumi rappresentati.

Il primo a presentare una trattazione amplia sul problema della rappresentazione grafica dei dati è Edward R. **Tufte** nel 1983 con il volume ***The Visual Display of Quantitative Information*** (Graphics Press, Cheshire, CT).

Egli enuncia chiaramente il concetto, dato per implicito nella tradizione statistica e quindi mai chiaramente discusso, che

- **una rappresentazione grafica di numeri deve essere direttamente proporzionale alla quantità rappresentata:**

The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities representeted.

Come richiesto dalle applicazioni, fornisce anche un metodo per misurare la violazione di tale principio il ***Lie Factor*** (che potrebbe essere tradotto come il Fattore di Imbroglio o il Fattore di Menzogna, ma che non ha ancora avuto una traduzione in Italiano).

Il **Lie Factor** è definito come

- **il rapporto tra la dimensione di un effetto mostrato nel grafico e la dimensione dell'effetto nei dati**:

$$Lie\ \ Factor = \frac{Dimensione\ \ apparente\ \ dell'effetto\ \ mostrato\ \ nel\ \ grafico}{Dimensione\ \ reale\ \ dell'effetto\ \ nei\ \ dati}$$

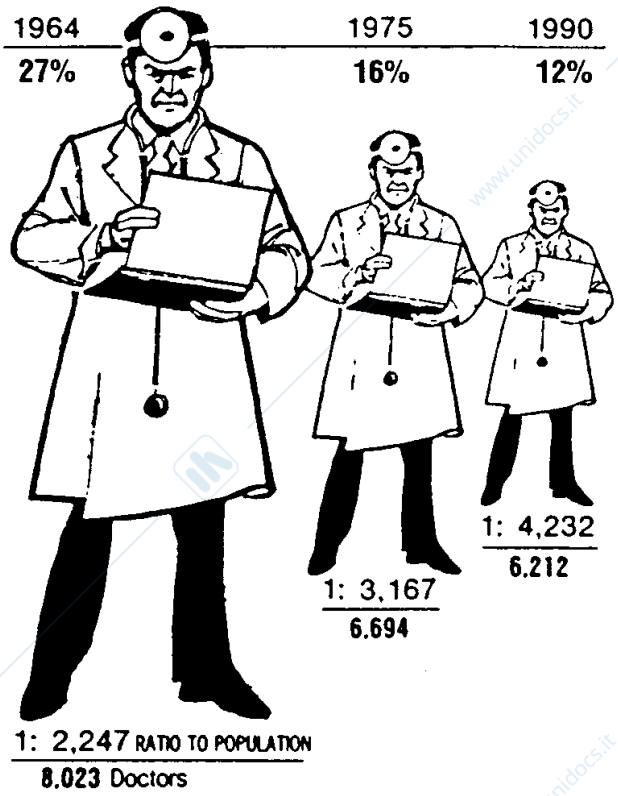
Lie Factor defined as the ratio of the size of an effect shown in the graphic to the size of the effect in the data.

Ovviamente tale rapporto dovrebbe essere uguale a 1.

Se si allontana da 1, è l'indicazione numerica di una rappresentazione non corretta. Fornisce i **limiti di una variazione casuale o accettabile**, come se si trattasse di un test, affermando che il valore del **Lie Factor dovrebbe sempre essere compreso tra 1,05 e 0,95**.

Ma spesso varia da 0,5 a 5. Nel suo testo mostra casi, presi soprattutto dal giornalismo politico, nei quali il **Lie Factor** è maggiore di 10,0.

La figura successiva, anch'essa riportata nel suo testo e ripresa da altri autori come a pag. 218 del volume di B. S. **Everitt** del 2002 ***The Cambridge Dictionary of Statistics*** (2nd edn. Cambridge University Press, UK, IX + 410 p.)



ha un **Lie Factor** uguale a **2,8.**

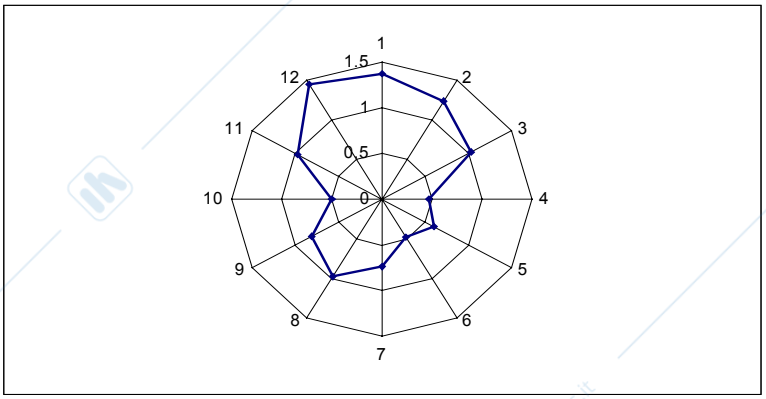
E' la percentuale di medici che svolgono solamente il compito di medici di famiglia: in California dal 1964 al 1990 diminuiscono dal 21% al 12% di tutti i medici dello stato, numericamente di riducono da 8.023 unità a 6.212 e in rapporto alla popolazione passano da 1 su 2.247 residenti a 1 su 4.232.
La figura espone questi concetti, con la costruzione di un pittogramma, che ovviamente viene letto dall'occhio sulla base della superficie mentre la proporzione di rappresentare è riportata solamente sull'asse delle ordinate (l'altezza); inoltre la figura è ulteriormente alterata dal cambiamento dell'asse delle ascisse.
Un metodo diffuso per costruire grafici alterati è l'uso esasperato della prospettiva. Nel grafico riportato il gioco della prospettiva è presente in modo limitato.
Ulteriori approfondimenti di questi argomenti sono riportati in un secondo volume di Edward R. **Tufte**, pubblicato nel 1991 ***Envisioning Information*** (Graphic Press, Cheshire, CT).

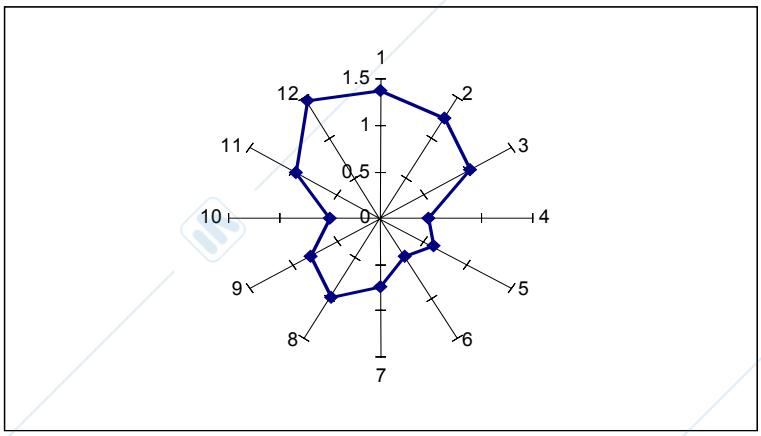
Il cartogramma evidenzia distribuzioni territoriali mediante carte geografiche, in cui nelle località interessate sono riportati **cerchi proporzionali alle frequenze**. E' il caso delle città segnate su carte geografiche con cerchi di dimensioni proporzionali al numero di abitanti. Questi cerchi spesso sono solo simboli, illustrati nelle didascalie: per cui un solo cerchio bianco spesso indica una quantità di base (es.: 1.000 persone), due cerchi concentrici indicano una quantità maggiore, anche in modo non proporzionale (es.: 20.000 persone), tre cerchi o un cerchio annerito una quantità ancora superiore (es.: 100.000 persone) e così via. E' importante ricordare che, quando costruiti in modo proporzionale, anche queste rappresentazioni grafiche sono essenzialmente areogrammi e quindi possono trarre in inganno se ad essere proporzionale alle frequenze è il raggio.

Un'altra rappresentazione grafica che ha un uso specifico per alcuni argomenti è il **diagramma polare o diagramma a coordinate polari**.
Serve per rappresentare le **variabili cicliche** (mensili, settimanali, giornaliere), come la quantità di pioggia e la temperatura media mensile; oppure la quantità di inquinanti presenti nell'aria in un ciclo di 24 ore.
A partire da un **punto centrale**, chiamato **polo**, si traccia una **serie di cerchi concentrici**, la cui **distanza dal centro** misura l'**intensità del fenomeno**.

Figura 14. Valori medi mensili della radioattività beta totale nell'aria a livello del suolo in Italia nell'anno 1993 (mBq per metro cubo).

	Mese	mBq
1	Gennaio	1.37
2	Febbraio	1.24
3	Marzo	1.03
4	Aprile	0.47
5	Maggio	0.60
6	Giugno	0.48
7	Luglio	0.74
8	Agosto	0.98
9	Settembre	0.81
10	Ottobre	0.50
11	Novembre	0.97
12	Dicembre	1.45





Per rappresentare la variabile ciclica, si divide l'angolo giro in tante parti quante sono le modalità (es.: 12 per i mesi, 24 per le ore). Si devono poi collocare punti nei vari cerchi concentrici, per individuare insieme la modalità (es.: il mese o l'ora) e l'intensità del fenomeno (es.: la quantità di pioggia, la temperatura, la misura d'inquinamento atmosferico o di un corso d'acqua). Il diagramma polare è ottenuto congiungendo i vari punti e l'intensità del fenomeno è rappresentata dalla distanza dal centro.
Le figure relative riportano due differenti impostazioni grafiche di costruire un diagramma polare sui valori medi mensili in Italia della radioattività beta totale nell'anno 1993.

Per la rappresentazione di **dati numerici**, è possibile ricorrere anche a **diagrammi cartesiani**. Essi saranno illustrati nel capitolo dedicato ai dati bivariati; sono utilizzati quando per ogni individuo sono rilevati contemporaneamente 2 variabili, come il peso e l'altezza. Ma possono essere usati anche per una sola variabile, collocando i punti di una distribuzione cumulata su un piano cartesiano: la perpendicolare sull'asse delle ascisse coincide con il valore della variabile e quella sull'asse delle ordinate fornisce le corrispondenti quantità o frequenze; i punti sono uniti da segmenti secondo l'ordine stabilito dal valore riportato in ascissa. E' di particolare utilità il **diagramma quantile**, che risulta graficamente simile al diagramma cumulato, soprattutto quando si dispone di poche unità e la variabile è di tipo continuo: vengono eliminate le anomale presenze di classi nulle entro gli estremi.

Per la scelta del metodo grafico con il quale presentare i dati, si deve prendere in considerazione il tipo di dati (qualitativi o quantitativi), la misura (discreta o continua), il dettaglio che si vuole ottenere nella forma della distribuzione. I metodi non aggiungono alcuna informazione che già non sia contenuta nei dati; ma garantiscono una rappresentazione più efficace, in particolare a persone non esperte dell'argomento trattato.

1.6. LE MISURE DI TENDENZA CENTRALE

Le rappresentazioni grafiche forniscono una sintesi visiva delle caratteristiche fondamentali delle distribuzioni di frequenza. Rispetto alle cifre, **le figure forniscono impressioni** che sono percepite con maggiore facilità; ma nel contempo hanno il limite di essere **meno precise** e **meno ricche di particolari**.

Per i **caratteri qualitativi**, la tabella e le rappresentazioni grafiche esauriscono quasi completamente gli aspetti descrittivi, quando sia possibile leggere con esattezza le frequenze delle varie classi.

Per i **caratteri quantitativi**, si pone il problema di **sintesi oggettive** che possano essere elaborate matematicamente e quindi che siano **numeriche**, al fine di **un'analisi obiettiva che deve condurre tutti i ricercatori, con gli stessi dati, alle medesime conclusioni**.

Una serie di dati numerici è compiutamente descritta da **3 proprietà principali**:
1) la **tendenza centrale o posizione**;
2) la **dispersione o variabilità**;
3) la **forma**.

Queste misure descrittive sintetiche, riassuntive dei dati tabellari, sono chiamate
- **statistiche, quando sono calcolate su un campione di dati**,
- **parametri, quando descrivono la popolazione od universo dei dati**.

I ricercatori in ecologia e nelle scienze ambientali molto raramente conoscono tutta la popolazione; di conseguenza, i metodi statistici di norma utilizzati sono riferiti quasi esclusivamente alla descrizione, all'analisi e al confronto di **campioni**.

1.6.1 Le **misure di tendenza centrale o posizione** servono per individuare il valore intorno al quale i dati sono raggruppati; la tendenza centrale è la misura più appropriata per sintetizzare l'insieme delle osservazioni, se una distribuzione di dati dovesse essere descritta con un solo valore; è la prima indicazione della dimensione del fenomeno.
Le misure proposte sono essenzialmente 3: **la media, la moda e la mediana**. Più raramente ed in discipline specifiche si utilizzano altre misure, come l'intervallo medio.
La scelta della misura di tendenza centrale di una serie di dati dipende dalle caratteristiche della distribuzione e dal tipo di scala.

La **media aritmetica semplice** è la misura di tendenza centrale più comunemente utilizzata. Quando si parla solo di **media**, si intende la media aritmetica semplice. E' definita come la somma del valore di tutte le osservazioni, diviso il numero di unità.
Con simboli, è

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

e, con una notazione più generale, diventa

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

dove:

- $\overline{x}$ = media del campione
- $\mathbf{x_i}$ = **i**-esima osservazione della variabile X
- **n** = numero di osservazioni del campione
- $\sum_{i=1}^{n}$ = sommatoria di tutti gli $\mathbf{x_i}$ del campione.

La media può essere vista come il baricentro della distribuzione campionaria, quando ogni singola osservazione è rappresentata da un peso convenzionale, identico per tutte, lungo l'asse che riporta i valori su una scala di intervalli o di rapporti.

Per dimostrare graficamente che **la media aritmetica corrisponde al punto di bilanciamento o di equilibrio dei dati,** si supponga di avere 5 misure: 10,9 11,5 12,3 12,8 15,4.

La loro media

$$\overline{X} = \frac{10,9+11,5+12,3+12,8+15,4}{5} = 12,58$$

è uguale a 12,58.

La rappresentazione grafica dei dati e della media, riportata nella figura seguente, mostra otticamente come la somma della distanza dalla media dei valori collocati prima sia uguale alla somma della distanza dei valori collocati dopo.

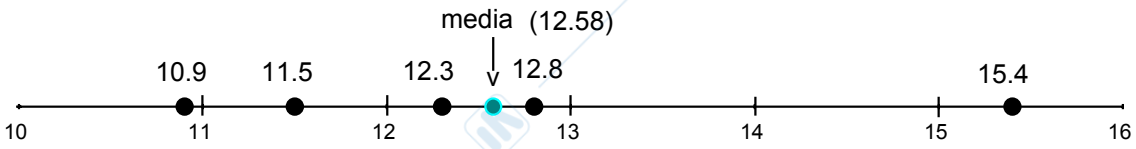


Figura 15. Rappresentazione grafica di 5 dati e della loro media aritmetica.

In una distribuzione di frequenza raggruppata in classi, come valore rappresentativo di ogni classe è preso il dato centrale, nell'assunzione che, entro ogni classe, i dati siano distribuiti in modo uniforme.

La **media aritmetica di distribuzioni di frequenza** raggruppate in classi, detta **media aritmetica ponderata**, è calcolata più rapidamente

con

$$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

dove:

- $\overline{x}$ = media della distribuzione in classi,
- $\mathbf{x_i}$ = valore medio della **i**-esima classe di intervallo,
- $\mathbf{f_i}$ = numero di osservazioni della classe **i**-esima classe,
- **n** = numero di classi,
- $\sum$ = sommatoria per tutte le **n** classi.

ESEMPIO. Da un gruppo di 25 dati, raggruppati nella seguente distribuzione in classi

Classe	$\mathbf{X_i}$	150-159	160-169	170-179	180-189	190-199
Frequenza	$\mathbf{f_i}$	3	5	8	6	3

calcolare la media.

Risposta. Con la formula della **media ponderata**

$$(\text{media})\overline{x} = \frac{(155 \cdot 3) + (165 \cdot 5) + (175 \cdot 8) + (185 \cdot 6) + (195 \cdot 3)}{3 + 5 + 8 + 6 + 3} = \frac{4385}{25} = 175,4$$

la media di tutto il campione risulta uguale a 175,4.

Le applicazioni della media aritmetica semplice e di quella ponderata sono numerose e derivano da alcune loro proprietà:

- le **grandezze additive sono le più frequenti in natura**;
- **la media aritmetica effettua la correzione degli errori accidentali** d'osservazione, per cui essa è la stima più precisa di misure ripetute;
- la **media aritmetica è la più semplice** delle medie algebriche.

Quando le quantità od i fattori causali non sono additivi oppure i dati sono ottenuti da rapporti, si ricorre ad altri tipi di medie; in questi casi, trovano un uso relativamente frequente nelle scienze ambientali la **media geometrica**, la **media armonica** e la **media quadratica**.

La **media geometrica semplice** è utilizzata quando le variabili non sono rappresentate da valori lineari, ma ottenuti da prodotti o da rapporti di valori lineari. Serve per il confronto di superfici o volumi, di tassi di accrescimento o di sopravvivenza, per quei valori appunto che sono espressi da rapporti.

Per il calcolo della media geometrica, è condizione necessaria che le quantità siano tutte positive. Se alcune fossero negative, si deve ricorrere al valore assoluto.

La media geometrica di n dati è uguale alla radice di ordine n (solo positiva) del prodotto degli n dati:

con simbologia matematica è

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

e può essere scritta anche come

$$\overline{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Una proprietà importante è che

- il logaritmo della media geometrica (log $\overline{x}$)

è uguale alla media aritmetica dei logaritmi dei dati $\left(\frac{\sum_{i=1}^{n} \log x_i}{n}\right)$**:**

$$\log \overline{x} = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

E' una proprietà che risulta utile quando si deve ricorrere alla trasformazione dei dati nei loro logaritmi, allo scopo di normalizzare la distribuzione ed applicare in modo corretto i test di inferenza. (Le trasformazioni dei dati sono discusse ampiamente nel capitolo finale dell'analisi della varianza).

La **media armonica** è la stima più corretta della tendenza centrale, per distribuzioni di dati in cui devono essere usati gli inversi. E' utilizzata quando i valori di X sono espressi come rapporti di un totale costante od in misure di tempi di reazione.
La media armonica è data da

$$m_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

La **media quadratica** è la radice quadrata della media aritmetica dei quadrati:

$$m_q = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

Sotto l'aspetto matematico può essere calcolata per valori positivi, nulli o negativi; ma essa ha senso come misura di tendenza centrale solamente se i valori sono positivi o nulli. E' un indice che trova applicazioni quando si analizzano superfici.

1.6.2 **La mediana** è il valore che ***occupa la posizione centrale in un insieme ordinato di dati***. E' una **misura robusta**, in quanto **poco influenzata dalla presenza di dati anomali**. La sua utilizzazione è **indispensabile nel caso di scale ordinali o di ranghi**.
La sue caratteristiche più importante sono due:
- è calcolata sul numero di osservazioni; si ricorre al suo uso quando si vuole attenuare l'effetto di valori estremi o comunque prendere in considerazione solo l'informazione fornita dai ranghi;
- in una distribuzione o serie di dati, ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.

Come la media è la misura di tendenza centrale nella statistica parametrica, la mediana è la misura di posizione o tendenza centrale utilizzata in quasi tutti i test non parametrici.

Per **calcolare la mediana di un gruppo di dati**, occorre
1 - disporre i valori in una fila ordinata in modo crescente oppure decrescente e contare il numero totale **n** di dati;
2 - se il numero (**n**) di dati è dispari, la mediana corrisponde al valore numerico del dato centrale, quello che occupa la posizione **(n+1)/2**;

3 – se il numero (**n)** di dati è pari, la mediana è stimata utilizzando i due valori centrali che occupano le posizioni **n/2** e **n/2+1**; con poche osservazioni, come mediana viene assunta la media aritmetica di queste due osservazioni intermedie; con molte osservazioni raggruppate in classi, si ricorre talvolta alle proporzioni.

ESEMPIO. Calcolare la mediana nella serie di 6 dati: 10,1 10,8 13,1 13,9 14,2 14,5 .
Risposta: Il numero di osservazioni è pari e i due valori centrali sono 13,1 e 13,9; la mediana è individuata dalla loro media aritmetica e quindi è uguale a 13,5.

Per meglio comprendere le differenze tra media aritmetica e mediana, con la stessa serie di 6 dati (10,1 10,8 13,1 13,9 14,2 14,5) in cui
- la media è 12,85 e
- la mediana 13,5

la rappresentazione grafica evidenzia come la media sia il baricentro della distribuzione e la mediana sia collocata tra i valori più addensati.

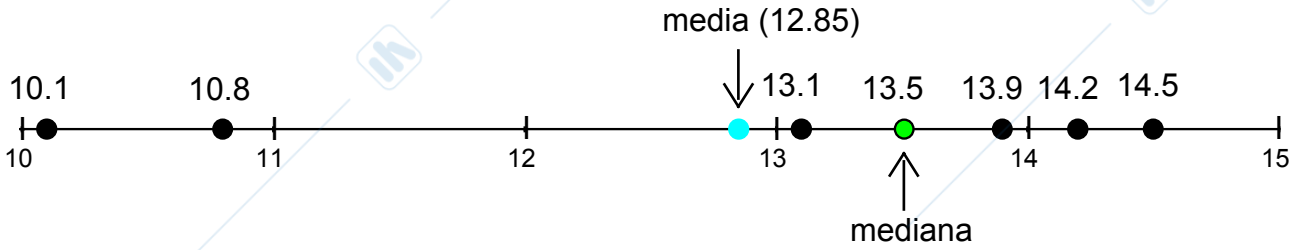


Figura 16. Rappresentazione grafica della media e della mediana di 6 dati.

Nella precedente figura 16, il grafico mostra come, nel caso di dati distribuiti in modo non simmetrico, la mediana rappresenti in modo più adeguato della media l'addensamento dei dati, il valore "normale o tipico" della serie. La media infatti è maggiormente influenzata dalla presenza dei due valori più distanti, che la allontanano dal gruppo dei valori più frequenti e la rendono diversa da essi. Se i due valori anomali fossero più vicini (o più lontani) rispetto agli altri 4, la media cambierebbe mentre la mediana rimarrebbe invariata.

1.6.3 La **moda** (detta più raramente anche **dato prevalente**) è **il valore più frequente di una distribuzione**. Essa non è influenzata dalla presenza di nessun valore estremo; tuttavia viene utilizzata solamente a scopi descrittivi, perché **è meno stabile e meno oggettiva delle altre misure di tendenza centrale**. Può infatti differire nella stessa serie di dati, quando si formano classi di distribuzione con ampiezza differente. Per individuare la moda entro una classe di frequenza, non conoscendo come i dati sono distribuiti, si ricorre all'ipotesi della **uniforme ripartizione**.

Oltre alle distribuzioni di frequenza che hanno una sola moda e che si chiamano **distribuzioni unimodali**, si trovano distribuzioni di frequenza che presentano due o più mode; sono denominate **distribuzioni bimodali o plurimodali**.

Le distribuzioni plurimodali possono essere il risultato della scarsità di osservazioni o dell'arrotondamento dei dati; di norma, sono dovute alla sovrapposizione di più distribuzioni con tendenza centrale differente. Per esempio, misurando le altezze di un gruppo di giovani in cui la parte maggiore sia formata da femmine e la minore da maschi si ottiene una distribuzione bimodale, con una moda principale ed una secondaria, come la seguente.

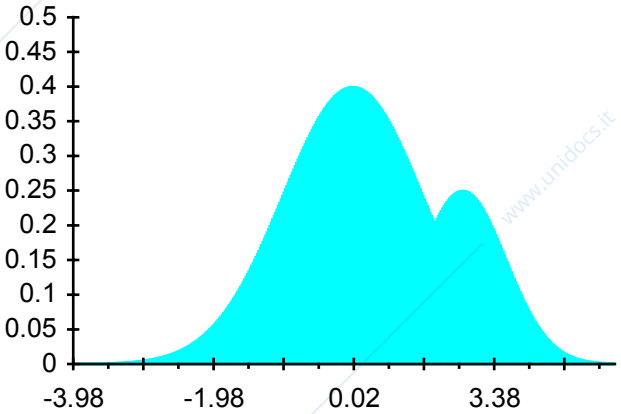


Figura 17. Distribuzione bimodale

Quando la distribuzione dei dati evidenzia due o più mode, il ricercatore deve quindi sospettare che i dati non siano omogenei, ma formati da altrettanti gruppi con differenti tendenze centrali. E' pertanto **errato fondare le analisi sulla media generale della distribuzione, poiché non è vera l'assunzione fondamentale che siano dati tratti dallo stesso universo o popolazione con una sola tendenza centrale**.

La media di una distribuzione bimodale, formata in quota pari da maschi e da femmine, sarebbe un valore "assurdo" che non descrive né i maschi né le femmine, ma un individuo inesistente, non essendo né maschio né femmina.

L'**intervallo medio** è semplicemente la media aritmetica tra il valore minimo e quello massimo. Ha il grande vantaggio di essere calcolato molto rapidamente, anche con un numero molto elevato di dati. Deve essere utilizzato con estrema cautela e solamente quando non esistono valori erratici o anomali: la presenza di un solo dato che si differenzia sensibilmente da tutti gli altri determina un valore dell'intervallo medio molto distorto, come misura della tendenza centrale.
In questi casi, può essere usata con maggiore correttezza la **media interquartile**, definita come la media fra il 1° e il 3° quartile, che risente in misura molto più ridotta della presenza di valori estremi.
Nelle scienze che studiano l'ambiente, l'intervallo medio era utilizzato in alcune discipline come la meteorologia. Poteva essere utile nel caso di una serie di dati sulla temperatura, ove non esistono mai valori anomali; infatti supponendo che in una giornata la temperatura minima sia stata di 10 gradi e quella massima di 20 gradi, il calcolo della media è rapidissimo (15) ed il valore si avvicina notevolmente alla media aritmetica, che richiederebbe un numero elevato di osservazioni e un disegno sperimentale accurato.
Per analogia, in meteorologia sovente questo metodo è stato utilizzato anche per il calcolo della precipitazione media mensile. E' un procedimento criticabile, addirittura errato: in questo caso si tratta di un fenomeno con elevatissima variabilità, con la presenza di valori che possono essere anomali e che influenzano fortemente sia l'intervallo medio che la media interquartile.

Oltre alla media, alla mediana e alla moda, insieme all'intervallo medio e alla media interquartile tra le misure di tendenza centrale può essere ricordata anche **la trimedia** proposta da **Tuckey** e calcolata come

$$T = (Q_1 + 2Q_2 + Q_3)/4$$

dove Q_2 è la mediana, Q_1 e Q_3 sono rispettivamente le mediane della prima metà e della seconda metà dei dati ordinati, detti anche primo e terzo interquartile.
E' un metodo che potrebbe essere utile quando si dispone di materiale molto variabile o con una distribuzione molto asimmetrica. Per esempio, le misure dell'inquinamento atmosferico presentano vari picchi anomali; la tendenza centrale potrebbe essere espressa dalla trimedia di Tuckey. Ma anche questa misura rientra tra le proposte che hanno avuto scarso seguito.
Le misure classiche, presenti in quasi tutte le discipline ed utilizzate senza sollevare obiezioni, sono media (aritmetica), mediana e moda.

1.7. MISURE DI DISPERSIONE O VARIABILITA'

La **dispersione o variabilità è la seconda importante caratteristica** di una distribuzione di dati. Essa definisce la forma più o meno raccolta della distribuzione intorno al valore centrale e fornisce indicazioni sul tipo di test da applicare; nei capitoli successivi verrà dimostrato come per confrontare le medie di due o più campioni sia richiesta l'omogeneità della varianza.

1.7.1 La prima misura ad essere stata storicamente utilizzata per descrivere la dispersione o variabilità dei dati è **il campo o intervallo di variazione**, definito come **la differenza tra il valore massimo e quello minimo**.

Intervallo di variazione = Valore massimo - valore minimo

Ha il grande vantaggio di essere un metodo intuitivo e molto semplice, in particolare quando i dati sono ordinati.

Tra gli inconvenienti di questa misura sono da prendere in considerazione:

- l'**incapacità di sapere come i dati sono distribuiti** entro l'intervallo, in particolare di dedurre la presenza di valori anomali;
- la sua **dipendenza dal numero di osservazioni**. All'aumentare del numero dei dati, cresce anche la probabilità di trovare un valore minore del minimo precedente ed uno maggiore di quello massimo precedente.

L'intervallo di variazione è quindi una misura poco efficiente della dispersione dei dati: per un confronto omogeneo tra distribuzioni, sarebbe necessario avere campioni delle stesse dimensioni, una condizione operativa eccessivamente limitante per la ricerca e l'analisi dei dati.

1.7.2 **La differenza interquartile** (figura 18), la differenza tra il 3° (Q_3) ed il 1° (Q_1) quartile ha il vantaggio di eliminare i valori estremi, ovviamente collocati nelle code della distribuzione. Tuttavia le proprietà di questa semi-differenza, chiamata anche **scarto interquartile**, non sono sostanzialmente differenti da quelle del campo di variazione.

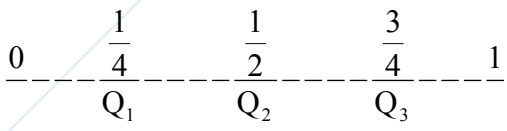


Figura 18. Differenza interquartile = 3°quartile (Q_3) - 1°quartile (Q_1)

Come **misure di posizione non-centrale**, ma con finalità esclusivamente descrittive, sono spesso usati i **quantili**, chiamati anche **frattili**, in quanto ogni sottogruppo contiene la stessa frazione di osservazioni. Quelli più comunemente usati sono i **decili**, che classificano i dati ordinati in decine, ed i **percentili**, che li suddividono in centesimi. Con i quantili, si possono individuare quali sono i valori che delimitano, nel margine inferiore o superiore della distribuzione, una percentuale o frazione stabilita di valori estremi. Per esempio, nello studio dell'inquinamento, come di qualunque altro fenomeno, può essere utile vedere quali sono le zone o i periodi che rientrano nell'1, nel 5 o nel 10 per cento dei valori massimi o minimi. **A valori così rari, facilmente corrispondono cause anomale, che di norma è interessante analizzare in modo più dettagliato. Nello studio di qualunque fenomeno biologico od ecologico, le misure particolarmente piccole o eccezionalmente grandi rispetto ai valori normali quasi sempre evidenziano cause specifiche, meritevoli di attenzione.**
Quando la forma della distribuzione è ignota o risulta fortemente asimmetrica, l'uso dei quantili fornisce indicazioni operative semplici e robuste per individuare i valori **più frequenti,** da ritenersi **"normali"** e quelli **meno frequenti od "anomali"**.

Gli **scarti dalla media** sono la misura più appropriata della variabilità di un insieme di dati. Ma poiché la loro somma è sempre nulla per definizione, in quanto la media è il baricentro della distribuzione, è necessaria una trasformazione che potrebbe essere attuata in due modi:
a) **gli scarti assoluti dalla media**;
b) **i quadrati degli scarti dalla media**.

1.7.3 Lo **scarto medio assoluto** (S_m) **dalla media** $(\bar{x})$ per dati semplici

è dato da

$$S_m = \frac{\sum |x_i - \bar{x}|}{n}$$

e per raggruppamenti in classi è ottenuto con

$$S_m = \frac{\sum |x_i - \bar{x}| \cdot n_i}{n}$$

dove

- x_i = valore dell'**i**-esimo dato in una distribuzione semplice,
- $\bar{x}$ = valore centrale della classe in una distribuzione di frequenza,
- **n** = numero totale di dati,
- n_i = numero di dati della classe **i** in una distribuzione di frequenza.

Un indice analogo, usato nelle discipline sociali ed economiche per valutare la diversità tra due distribuzioni di frequenze relative, è

l'indice semplice di dissomiglianza (D)

$$D = \frac{\sum_{i=1}^{k} |f_{1i} - f_{2i}|}{2}$$

dove **1** e **2** sono i due gruppi e **k** sono le classi.

D è uguale a 0 quando le due distribuzioni di frequenza relativa sono identiche e uguale a 1 quando la prima distribuzione è tutta concentrata in una classe e l'altra distribuzione in una classe diversa.

1.7.4 In alcuni **test di statistica non parametrica**, come misura di dispersione è utilizzato **lo scarto medio assoluto dalla mediana**, che è la media degli scarti assoluti dei singoli dati dalla loro mediana; le formule sono uguali alle due precedenti, sostituendo la mediana alla media.

E' proprietà specifica della mediana rendere minima la somma degli scarti assoluti. Di conseguenza, lo scarto medio assoluto dalla mediana è sempre inferiore allo scarto medio assoluto dalla media; i due valori sono uguali solamente quando la distribuzione è simmetrica e quindi media e mediana coincidono.

1.7.5 La **Devianza** o **Somma dei Quadrati (SQ)** degli scarti dalla media (**SS = Sum of Squares**, in inglese) **è la base delle misure di dispersione dei dati, utilizzate in tutta la statistica parametrica**. Tutta la statistica parametrica è fondata sulla devianza e sulle misure da essa derivate.

$$(1) \qquad \text{devianza } (SQ) = \sum \left(x_i - \bar{x}\right)^2$$

L'equazione precedente è la **formula di definizione od euristica**. Spesso è poco pratica, in particolare quando la media è un valore frazionale, con vari decimali. Diviene allora conveniente ricorrere a un'altra formula, algebricamente equivalente, che permette di effettuare i calcoli manuali in tempi più brevi e con una sola approssimazione finale, chiamata **formula empirica od abbreviata**:

$$(2) \qquad \text{devianza } (SQ) = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

dove:

- $\sum x^2$ = sommatoria dei valori dopo che ogni osservazione è stata elevata al quadrato,
- $\left(\sum x\right)^2$ = sommatoria di tutti i dati, elevata al quadrato,
- **n** = numero di osservazioni sulle quali è stata calcolata la somma.

ESEMPIO. Calcolare con la formula euristica (1) e con quella abbreviata (2) la devianza (SQ) dei 6 numeri seguenti: 5, 6, 7, 7, 8, 10.

Risposta.

1. Con la formula euristica, si deve calcolare dapprima la media:

$$\bar{x} = \frac{5+6+7+7+8+10}{6} = \frac{43}{6} = 7,1\overline{6}$$

ed in seguito la devianza (SQ),

intesa come Somma dei Quadrati degli scarti di ogni valore dalla media:

$$devianza\ \ (SQ) = \sum\left(x_i - \bar{x}\right)^2 =$$
$$= (5-7,1\overline{6})^2 + (6-7,1\overline{6})^2 + (7-7,1\overline{6})^2 + (7-7,1\overline{6})^2 + (8-7,1\overline{6})^2 + (10-7,1\overline{6})^2 =$$
$$= 4,665 + 1,3456 + 0,0256 + 0,0256 + 0,7056 + 8,0656 = 14,8356$$

2. Con la formula abbreviata, calcolare direttamente il valore della devianza (SQ), dopo aver fatto sia la somma dei dati precedentemente elevati al quadrato, sia il quadrato della somma dei dati, secondo l'annotazione algebrica seguente

$$\text{devianza}\ \ (SQ) = \sum x^2 - \frac{\left(\sum x\right)^2}{n} =$$
$$= (25+36+49+49+64+100) - \frac{43^2}{6} = 323 - \frac{1849}{6} = 323 - 308,1\overline{6} = 14,84$$

I due valori della devianza spesso non risultano identici, in particolare quando stimati con più cifre decimali, a causa dell'approssimazione con la quale è calcolata la media, se non risulta un valore esatto. In questi casi, è da ritenersi corretta la stima fornita dalla formula abbreviata, che non richiede approssimazioni nei calcoli intermedi.

E' utile ricordare che, **per distribuzioni di dati raggruppati in classi, la formula euristica** diventa

$$devianza\ \ (SQ) = \sum\left(\bar{x}_i - \bar{\bar{x}}\right)^2 n_i$$

dove

- $\bar{x}_i$ è il valore centrale di ogni classe e
- $\bar{\bar{x}}$ è la media generale della distribuzione.

Il valore della devianza dipende da 2 caratteristiche della distribuzione: gli scarti di ogni valore dalla media ed il numero di dati. La prima è una misura della dispersione o variabilità dei dati ed è l'effetto che si intende stimare; la seconda è un fattore limitante per l'uso della devianza, in quanto un confronto tra 2 o più devianze richiederebbe campioni con lo stesso numero di dati. Pertanto, per una misura di dispersione dei dati che sia indipendente dal numero di osservazioni, si ricorre alla varianza.

1.7.6 **La varianza o Quadrato Medio** (QM, in italiano; MS da Mean Square, in inglese) **è una devianza media** o devianza rapportata al numero di osservazioni.

La **varianza di una popolazione (1)**, il cui simbolo è σ^2, è ottenuta dividendo la devianza per **n**, **il numero di osservazioni**.

$$\textbf{(1)} \quad \sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

La **varianza di un campione (2)**, il cui simbolo è s^2, è ottenuta dividendo la devianza per **n-1, il numero di gradi di libertà**.

$$\textbf{(2)} \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Ovviamente, quando **n** è grande le differenze tra varianza della popolazione e varianza del campione sono minime; quando **n** è piccolo, le differenze sono sensibili.

E' importante ricordare che quando si parla di **inferenza**, cioè quando si utilizzano i **dati di un campione per conoscere le caratteristiche della popolazione**, si usa sempre la **varianza campionaria**.

Le giustificazioni logiche dell'uso di **dividere la devianza per n-1, detta anche correzione di Student**, sono lunghe e complesse: la più semplice si basa sul fatto che **n-1 è il numero di osservazioni indipendenti, chiamato gradi di libertà**, abbreviato abitualmente in **gdl** o **df** (da ***degree freedom***). Poiché la somma degli scarti dalla media è uguale a 0, **l'ultimo valore di una serie è conosciuto a priori,** non è libero di assumere qualsiasi valore, quando siano già noti i precedenti n-1 valori.

Come concetto generale introduttivo, si può dire che **il numero di gradi di libertà è uguale al numero di dati meno il numero di costanti che sono già state calcolate o di informazioni che siano già state estratte dai dati**. Nel caso specifico della varianza, la costante utilizzata per calcolare gli scarti è la media: quindi i gradi di libertà sono **n-1**.

Mentre la media è un valore lineare, la varianza è un valore al quadrato; per stime associate alla media o per confronti con essa, è necessario ricondurla a un valore lineare.

1.7.7 Lo **scarto quadratico medio** o **deviazione standard**, il cui simbolo è σ nel caso della popolazione ed **s** nel caso di un campione, **è la radice quadrata della varianza**. Il termine ***standard deviation*** e il suo simbolo σ (la lettera greca sigma minuscola) sono attribuiti al grande statistico inglese Karl **Pearson** (1867 – 1936) che l'avrebbe coniato nel 1893; in precedenza era chiamato ***mean error***. In alcuni testi di statistica è abbreviato anche con **SD** ed è chiamato ***root mean square deviation*** oppure ***root mean square***,

E' una misura di distanza dalla media e quindi ha sempre un valore positivo. E' una misura della dispersione della variabile casuale intorno alla media.

Nel caso di **un campione**, a partire da una serie di dati la **deviazione standard, il cui simbolo è s,** può essere calcolata come:

$$\text{deviazione standard } (\mathbf{s}) = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 n_i}{n-1}}$$

dove

- $\mathbf{x_i}$ = valore del dato in una distribuzione semplice,
- $\bar{x}$ = valore centrale della classe in una distribuzione di frequenza,
- $\mathbf{n_i}$ = numero di dati della classe i in una distribuzione di frequenza,
- **n** = numero totale di dati.

1.7.8 L'**errore standard** (***standard error***, in inglese) è indicato con **es** e misura la dispersione delle medie calcolate su **n** dati, come la deviazione standard serve per la dispersione dei dati.

L'errore standard **es** è

$$es = \frac{s}{\sqrt{n}}$$

ESERCIZIO. Calcolare media, devianza, varianza e deviazione st. e errore st. di : 9 6 7 9 8 8.

Risposta:
media = 7,833; devianza = 6,8333; varianza = 1,367; deviazione st. = 1,169; errore standard = 0,477

Per l'uso della varianza, che sarà fatto nei capitoli dedicati all'inferenza, è importante comprendere che **la varianza tra una serie di dati rappresenta una misura di mutua variabilità tra di essi**.

Essa può essere calcolata in tre modi:
1 - come **la metà della media aritmetica del quadrato di tutti gli n(n-1)/2 scarti possibili tra coppie di osservazioni**,
2 - **mediante gli scarti tra i dati e la loro media,**
3 - **mediante la formula abbreviata**.

1 - Il primo metodo utilizza gli scarti tra tutte le possibile coppie di dati; è una procedura molto lunga, che serve per comprendere il **reale significato della varianza tra dati o tra medie:**

$$s^2 = \frac{1}{2} \cdot \frac{\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} (x_i - x_j)^2 \cdot f_i \cdot f_j}{\frac{n(n-1)}{2}}$$

2 - Il secondo metodo rappresenta la **formula euristica**, quella che definisce la varianza, come confronto con il valore medio:

$$s^2 = \frac{\sum_{j=1}^{J} (x_j - \bar{x})^2 \cdot f_i}{n-1}$$

3 - Il terzo metodo è una delle varie versioni della **formula abbreviata**, quella che serve per semplificare i calcoli manuali e ridurre i tempi per il calcolo

$$s^2 = \frac{\sum_{j=1}^{n} x_j^2 f_j - \frac{\left(\sum_{j=1}^{n} x_j f_j\right)^2}{n}}{n-1}$$

ESEMPIO. Calcolare la varianza di 6 dati (5, 6, 7, 7, 8, 10) mediante le 3 formule proposte, per dimostrare empiricamente la loro equivalenza (ricordando che, in questo esempio, $f_i = 1$).

Risposta:

1. Utilizzando **gli scarti assoluti (j - i) tra tutte le possibili coppie di dati**, riportati nella matrice triangolare sottostante:

j\i	***5***	***6***	***7***	***7***	***8***	***10***
5	0					
6	1	0				
7	2	1	0			
7	2	1	0	0		
8	3	2	1	1	0	
10	5	4	3	3	2	0

si ottiene

$$s^2 = \frac{1}{2} \cdot \frac{1^2 + 2^2 + 2^2 + 3^2 + 5^2 + 1^2 + 1^2 + 2^2 + 4^2 + 0^2 + 1^2 + 3^2 + 1^2 + 3^2 + 2^2}{15} = \frac{90}{30} = 3$$

2. Mediante **gli scarti dalla media** ($\overline{X} = 7{,}1\overline{6}$)

si ottiene

$$s^2 = \frac{(5 - 7{,}1\overline{6})^2 + (6 - 7{,}1\overline{6})^2 + \cdots + (10 - 7{,}1\overline{6})^2}{6 - 1} = \frac{15}{5} = 3$$

3. Ricorrendo alla **formula ridotta**

si ottiene

$$s^2 = \frac{(5^2 + 6^2 + \cdots + 10^2) - \frac{43^2}{6}}{6-1} = \frac{323 - 308{,}1\overline{6}}{5} = \frac{14{,}83}{5} = 2{,}9\overline{6}$$

1.7.9 Il <u>coefficiente di variazione</u> (*coefficient of variation* oppure *coefficient of variability*) è una **misura relativa di dispersione,** mentre le precedenti erano tutte misure assolute. E' quindi particolarmente utile ricorrere ad esso, quando si intende confrontare la variabilità di due o più gruppi con medie molto diverse oppure con dati espressi in scale diverse.

Consideriamo come esempio il confronto tra la variabilità di due specie animali con dimensioni medie sensibilmente diverse, come tra i cani e i cavalli. La varianza tra cavalli di razze diverse è superiore a quella esistente tra i cani, perché gli scarti assoluti dalla media della specie sono maggiori. Ma spesso il problema consiste nel fare un confronto relativo tra variabilità e dimensioni medie delle due specie; allora il rapporto tra il cane di dimensioni maggiori e quello di dimensioni minori risulta superiore a quello esistente nei cavalli.

<u>Il Coefficiente di Variazione</u> (CV oppure semplicemente con **V** in molti testi recenti**)** misura la **dispersione percentuale in rapporto alla media.**

$$\textbf{<u>Per una popolazione</u>}: \ \mathrm{CV} = \left(\frac{\sigma}{\mu}\right) \cdot 100$$

dove

- σ = deviazione standard della popolazione
- μ = media della popolazione

$$\textbf{<u>Per un campione</u>}: \ \mathrm{CV} = \left(\frac{s}{\overline{X}}\right) \cdot 100$$

dove

- **s** = deviazione standard del campione
- $\overline{X}$ = media del campione

Quando è calcolato su dati campionari, in particolare se il numero di osservazioni è limitato, il coefficiente di variazione CV deve essere corretto di una quantità 1/4N, dove N è il numero di osservazioni del campione. Di conseguenza,

<u>il coefficiente di variazione corretto V'</u> diventa

$$CV' = CV\left(1 + \frac{1}{4N}\right)$$

La figura successiva (tratta da pag. 16 di George W. **Snedecor**, William G. **Cochran** del 1974, ***Statistical Methods***, Iowa University Press Ames, Iowa, U.S.A. sixth edition , seventh printing, pp. XIV + 593) è utile per spiegare i concetti già presentati:

- con la linea tratteggiata descrive **l'altezza media** di gruppi di ragazze da 1 a 18 anni d'età, che varia da circa 70 cm. ad oltre 170 cm.; la scala di riferimento è riportata sulla sinistra e varia da 60 a 200 centimetri;
- con la linea formata da tratti e punti alternati descrive la **deviazione standard dell'altezza** di ogni gruppo d'età; la scala è riportata sulla destra (parte superiore) e i valori variano da 0 a 7;
- con la linea continua descrive il **coefficiente di variazione**; la scala è riportata a destra nella parte inferiore più esterna ed i valori variano da 2 a 5.

E' importante osservare come la media e la sua deviazione standard aumentino in modo quasi correlato, mentre il CV resta costante intorno al 4%:

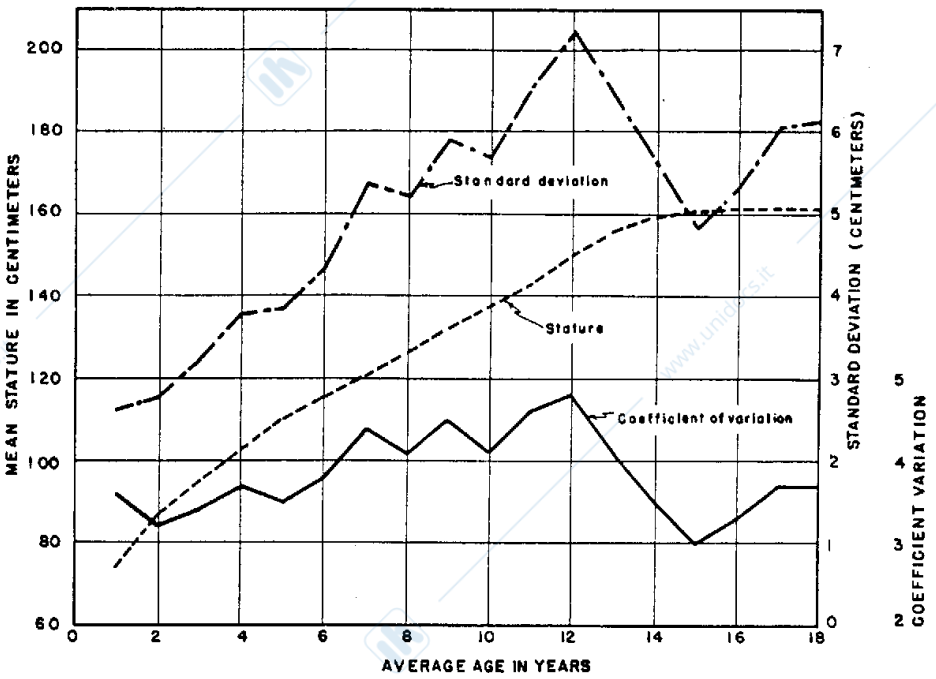


FIG. 2.17.1—Graph of 3 time series; stature, standard deviation, and coefficient of variation of girls from 1 to 18 years of age. See reference (1).

In natura, il coefficiente di variazione tende ad essere costante per ogni fenomeno, con valori che abitualmente oscillano tra il 5% e il 15%. Valori esterni a questo intervallo possono fare sorgere il sospetto di essere in presenza di un errore di rilevazione o di calcolo; si tratta comunque di situazioni non usuali che occorrerebbe spiegare, individuandone la causa. Nell'esempio precedente, si tratta di individui della stessa età.

Se il materiale biologico in esame ha un CV troppo basso (2-3 %), si può sospettare l'esistenza di un fattore limitante che abbassa notevolmente od elimina la variabilità, come la presenza di omogeneità genetica congiunta ad una situazione ambientale uniforme; viceversa, un CV molto alto (50%) è indice della presenza di condizioni anomale o molto differenti contemporaneamente per più fattori.

Per l'uomo, il coefficiente di variazione dell'altezza è stato calcolato tra il 40% e il 45%, testimoniando l'esistenza nella specie di grandi differenze, dovute sia a cause genetiche che ambientali (alimentazione, condizioni sanitarie, ecc.).

Quando per misurare lo stesso fenomeno si utilizzano scale differenti, ad esempio l'altezza misurata in centimetri o in pollici, **la media e la deviazione standard cambiano, ma il CV resta uguale**.

Esso può essere calcolato anche per campioni; ma quando il numero di dati è limitato, la sua stima può indurre in errore.

In laboratorio per valutare la qualità dei reagenti, spesso si ricorre al **C.V**.: i reagenti che determinano il CV minore sono quelli di qualità superiore, poiché forniscono risposte meno variabili in rapporto ai valori medi.

1.7.10 La varianza in dati raggruppati: correzione di Sheppard

La varianza calcolata in una distribuzione di frequenza di misure continue è approssimata; la sua stima è fondata sull'ipotesi di distribuzione uniforme entro ogni classe e quindi si presume che il valore centrale di ogni classe corrisponda alla sua media. In realtà, la varianza calcolata sui dati reali e quella stimata a partire dal raggruppamento in classi non sono uguali.

Quando la distribuzione dei dati è normale, entro ogni classe i valori più vicini alla media generale sono sempre più numerosi di quelli più distanti, collocati verso gli estremi.

Come già evidenziato, per il calcolo della media, le approssimazioni nella parte sinistra del valore centrale compensano le approssimazioni fatte nella parte destra: la media calcolata direttamente dai dati e quella calcolata con il raggruppamento in classi hanno solo differenze casuali, di solito di entità ridotta.

Il coefficiente di variazione è un numero puro, svincolato da ogni scala di misura e dalla tendenza centrale del fenomeno studiato. Secondo molti, appunto perché un rapporto, avrebbe significato solamente se calcolato per variabili misurate con una scala di rapporti.

Per il calcolo della varianza, le approssimazioni di segno opposto nelle due parti della media sono elevate al quadrato: di conseguenza, non si compensano, ma si sommano. In una popolazione con un numero molto alto di dati, **la varianza calcolata dal raggruppamento in classi è sistematicamente maggiore di quella reale, quella calcolata direttamente dai dati originari. Le differenze crescono all'aumentare della misura dell'intervallo di ogni classe, poiché aumenta l'imprecisione.**

Pertanto si deve apportare una correzione, detta
correzione di Sheppard, proposta appunto da W. F. **Sheppard** nel 1898 sulla rivista ***Proceeding London Mathematical Society*** e riportata in vari testi, tra cui ***Statistical Methods*** di George W. **Snedecor** e William G. **Cochran** (1967, Iowa State University Press).
Consiste nel **sottrarre alla varianza calcolata un valore pari a**

$$\frac{h^2}{12}$$

per cui

$$\sigma^2_{\text{reale}} = \sigma^2_{\text{calcolata}} - \frac{h^2}{12}$$

dove

- **h** è l'ampiezza delle classi e
- **12** è una costante.

ESEMPIO. In una distribuzione di frequenza, in cui le classi hanno ampiezza costante con intervallo h = 10, è stata calcolata una varianza $s^2 = 50$; la varianza corretta, quella che si sarebbe ottenuta utilizzando i singoli valori, secondo **Sheppard** dovrebbe essere

$$\sigma^2_{\text{reale}} = 50 - \frac{10^2}{12} = 50 - 8,3\overline{3} = 41,6\overline{6}$$

uguale a 41,66 come risulta dal calcolo mostrato.

Questa relazione è ritenuta valida per le popolazioni.

Con campioni formati da pochi dati, non è facile, spesso non è possibile, verificare se la distribuzione sperimentale utilizzata rispetti **le tre condizioni** fissate da **Sheppard** per applicare la correzione:

- **essere continua;**
- **avere un intervallo di ampiezza finito;**
- **tendere a zero in modo graduale nelle due code della distribuzione.**

Quando si dispone solo di piccoli campioni, la correzione potrebbe essere non adeguata alla forma della distribuzione e determinare un errore maggiore.

Di conseguenza, per piccoli campioni come quelli usati nella ricerca ambientale, **la quasi totalità dei ricercatori preferisce non applicare questa correzione, ma usare direttamente la varianza calcolata dalla distribuzione di frequenza,** in qualunque modo sia stato fatto il raggruppamento in classi.

1.8. INDICI DI FORMA: SIMMETRIA E CURTOSI

Gli **indici di forma di una distribuzione riguardano 2 caratteristiche: la simmetria e la curtosi**.

A differenza di quanto avvenuto nello studio della variabilità, nell'analisi della forma di una distribuzione statistica le misure elaborate sono rimaste rudimentali e le stesse definizioni sono sovente equivoche. Inoltre **l'uso degli indici di forma** non rientra nei test d'inferenza, ma **è limitato alla semplice descrizione della forma della distribuzione**.

Nelle distribuzioni unimodali si ha simmetria quando media, moda e mediana coincidono; se la distribuzione è bimodale, possono essere coincidenti solamente la media aritmetica e la mediana.

Di norma, le distribuzioni dei dati sono unimodali; pertanto, l'analisi della simmetria è accentrata su di esse.

In una distribuzione,

- l'**asimmetria** è detta **destra** (più correttamente, **a destra**) quando i valori che si allontanano maggiormente dalla media sono quelli più elevati, collocate a destra dei valori centrali (figura 19); nell'asimmetria destra, la **successione** delle 3 misure di tendenza centrale da sinistra a destra è: **moda, mediana, media**;
- l'**asimmetria** è detta **sinistra** (o **a sinistra**) quando i valori estremi, quelli più distanti dalla media, sono quelli minori (figura 20). Nell'asimmetria sinistra, la **successione** delle 3 misure di tendenza centrale da sinistra a destra è invertita rispetto all'ordine precedente: **media, mediana, moda.**

Quando media, mediana e moda non coincidono, la distribuzione è asimmetrica; ma quando queste tre misure coincidono non sempre la distribuzione è simmetrica. **Per avere una distribuzione simmetrica, la perfetta coincidenza delle tre misure di tendenza centrale è condizione solo necessaria, non sufficiente**.

Infatti, supponendo di analizzare una distribuzione come

- 16 20 20 20 30 30

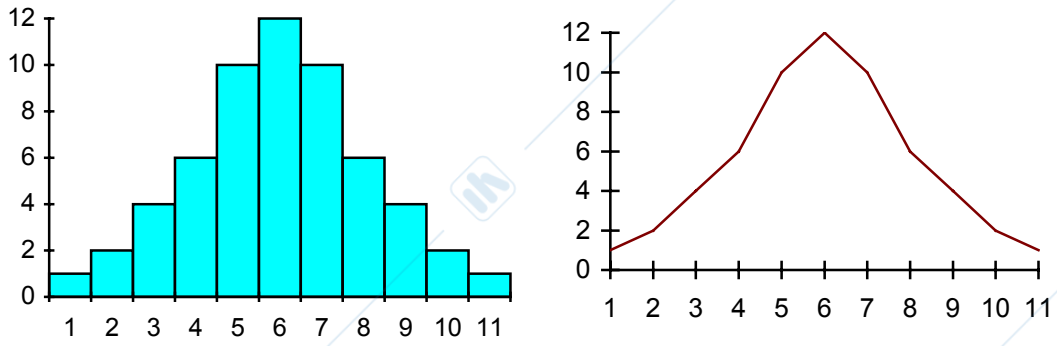
troviamo che

- la media (140/7 = 20),
- la mediana (su 7 valori è il 4° = 20) e
- la moda (il valore più frequente è 20)

sono coincidenti (20); ma, come si evidenzia dalla semplice lettura dei dati, la sua forma non è simmetrica poiché i dati non declinano in modo regolare ed identico dalla tendenza centrale verso i due estremi.

Un **altro metodo proposto per valutare la simmetria** utilizza la distanza delle classi di frequenza dalla mediana: **una distribuzione è simmetrica, se i valori che sono equidistanti dalla mediana hanno la stessa frequenza**. Ma è possibile dimostrare che si tratta di una condizione che si realizza sia in distribuzioni unimodali che plurimodali; è quindi una definizione che non caratterizza la distribuzione simmetrica in modo biunivoco, non è vera esclusivamente in una distribuzione normale.

I grafici di seguito riportati evidenziano la forma di una distribuzione simmetrica (Fig. 18),



.

Figura 18. Distribuzioni simmetriche, con istogrammi e con poligoni

quella di una distribuzione destra o positiva (Fig. 19)

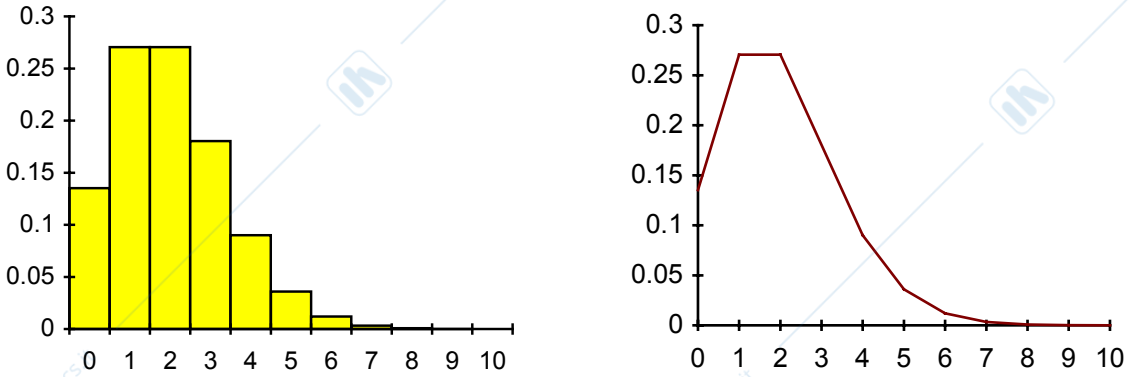


Figura 19. Distribuzioni con asimmetria a destra, con istogrammi e con poligoni

e quella di una distribuzione sinistra o negativa (Fig. 20).

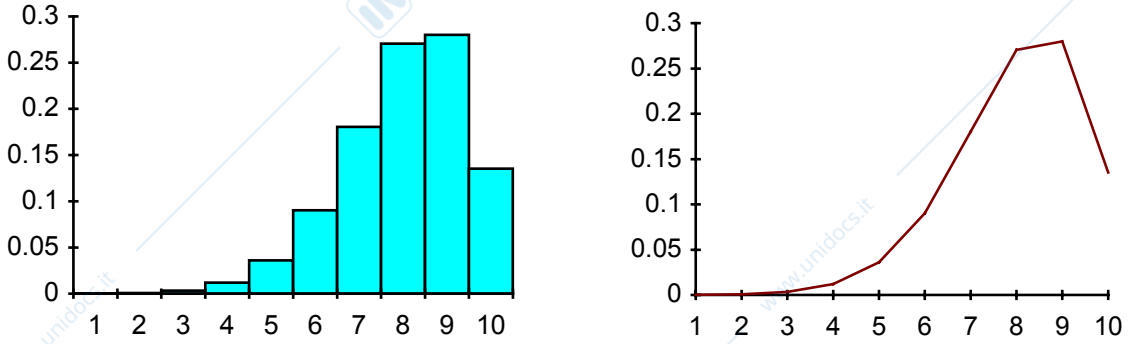


Figura 20. Distribuzioni con asimmetria a sinistra, con istogrammi e con poligoni.

Analizzando la distribuzione in classi dei dati di un campione, è possibile osservare un'**asimmetria causata da un numero ridotto di osservazioni** oppure da **modalità non adeguate nel raggruppamento in classi**, di solito eccessivo; in questi casi, si parla di **asimmetria falsa**, da distinguere dalla **asimmetria vera, che può esistere solo nella distribuzione reale di una popolazione**.

Una proprietà essenziale degli indici di asimmetria di una distribuzione è che essi dovrebbero essere uguali a zero quando, e solamente quando, la distribuzione è simmetrica.

E' una proprietà che si realizza per gli indici abituali di variabilità o dispersione, come la devianza, la varianza e le misure derivate; esse sono nulle quando, e solamente quando, tutti i valori sono uguali, e quindi non esiste variabilità; quando non sono nulle, esiste una variabilità, che cresce all'aumentare del valore dell'indice.

Gli indici di simmetria non godono della stessa proprietà: quando la distribuzione è simmetrica sono nulli; ma possono essere nulli anche per distribuzioni non simmetriche.

Per valutare l'asimmetria di una distribuzione, si possono usare

- **misure dell'asimmetria assoluta**

misure di asimmetria relativa.

Gli indici di **asimmetria assoluta si esprimono con le distanze tra la media e la moda o la mediana**.

Una misura assoluta, usata frequentemente, è **la differenza (d) tra la media e la moda**:

d = media - moda

La differenza è:

d = 0, se la curva è simmetrica;

d > 0, se la curva ha asimmetria positiva (o destra : media > mediana > moda);

d < 0, se la curva ha asimmetria negativa (o sinistra : media < mediana < moda).

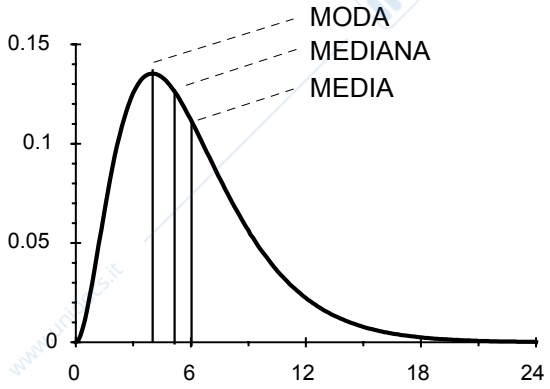


Figura 21. Asimmetria destra o positiva (d>0)

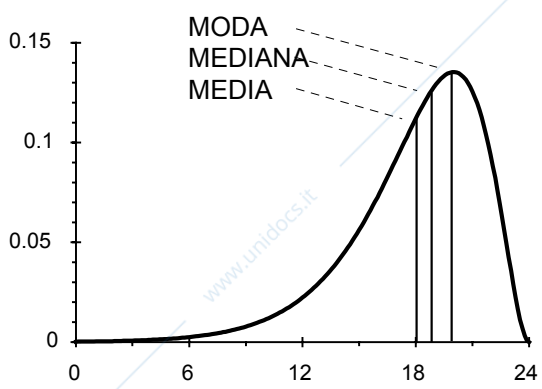


Figura 22. Asimmetria sinistra o negativa (d<0)

E' possibile valutare in modo molto semplice ed empirico **il grado d'asimmetria** di una distribuzione; essa è ritenuta **moderata** se

Moda = Media - 3(Media - Mediana)

ed è ritenuta **forte** se è sensibilmente maggiore di tale valore.

Per ottenere una misura del grado di asimmetria, che possa essere confrontato con quello di qualsiasi altra distribuzione in quanto indipendente dalle dimensioni delle misure, occorre utilizzare **indici relativi**, quali

skewness di Pearson;

γ_1 di Fisher;

β_1 di Pearson.

L'indice **skewness di Pearson (sk)** è un rapporto: la differenza (**d**) tra la media e la moda è divisa per la deviazione standard (σ) o scarto quadratico medio. Nel caso di una distribuzione campionaria, dove la deviazione standard è indicata con **s**, è

$$sk = \frac{d}{s}$$

Come per il valore **d** precedente, **sk** può essere nullo, positivo o negativo secondo la forma della distribuzione.

Essendo **un rapporto tra misure statistiche della stessa distribuzione**, è divenuto una **misura adimensionale, indipendente dal valore assoluto degli scarti dalla media**; quindi può essere utilizzato per il confronto tra due o più distribuzioni.

Un altro indice di simmetria, proposto da A. L. **Bowley** nel 1920 (vedi il testo ***Elements of Statistics***, Charles Scribner's Sons, New York), chiamato appunto **Bowley coefficient** e riproposto in alcuni programmi informatici, utilizza i quartili (

$$\text{Skewness} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

dove

- Q_2 = valore della mediana o del secondo Quartile
- Q_1 = valore del primo quartile,
- Q_3 = valore del terzo quartile.

Il valore ottenuto è uguale a zero se la distribuzione è perfettamente simmetrica, mentre è negativo o positivo in rapporto al tipo di asimmetria, in modo analogo alle formule precedenti.

Anche la **curtosi**, la concavità della distribuzione di dati (più ampiamente spiegata nel prosieguo del paragrafo), può essere misurata con **i quantili** o meglio gli **ottili**, come proposto da J. J. A. **Moors** nel 1988 (nell'articolo ***A quantile alternative for kurtosis***, su **Statistician** Vol. 37, pp. 25-32).

Il concetto di base è che la curtosi può essere interpretata come una misura di **dispersione intorno ai due valori limite dell'intervallo** $\mu \pm \sigma$, con una misura alternativa a quella parametrica, che è fondata sul momento di quarto ordine. Quando si confrontano testi differenti, l'interpretazione della curtosi è controversa, poiché molti la descrivono come un eccesso o un difetto di valori solo nella classe centrale, mentre la differenza dalla normale riguarda più punti.

Utilizzando i **quantili** o meglio gli **ottili** (indicati ovviamente con ***E*** nel testo in inglese), si ottiene

$$\mathbf{T} = \frac{(E_3 - E_1) + (E_7 - E_5)}{E_6 - E_2}$$

dove

- E_1 è la metà di un quartile, cioè il valore che occupa il 12,5 percentile,
- E_2, E_3, … sono multipli di esso, fino a E_7 corrispondente al 87,5 percentile.

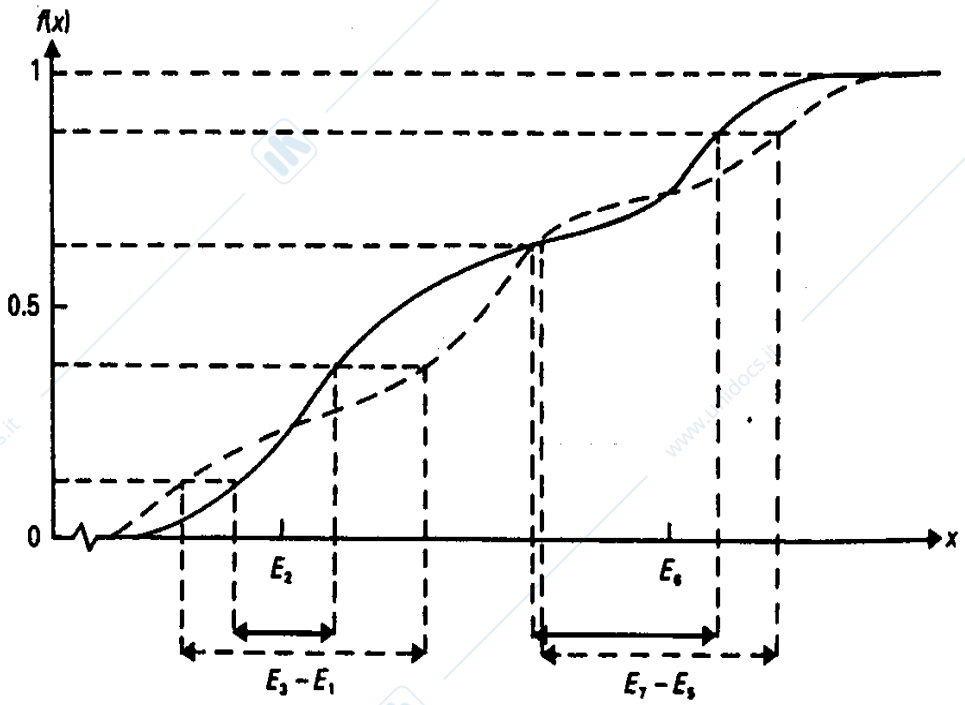


Figura 23. Grafico degli ottili

Dal grafico risulta con evidenza come **i due termini del numeratore, $(E_3 - E_1)$ e $(E_7 - E_5)$, valutano la concentrazione di dati intorno a E_6 e E_2, corrispondenti ai punti** $\mu \pm \sigma$. L'indice T è analogo alle misure di dispersione e simmetria più familiari fondate sui quantili e offre gli

vantaggi simili; inoltre esclude i valori estremi e quindi è più robusto dell'indice parametrico fondato sui momenti..

Il **denominatore (E_6 e E_2) è una costante di normalizzazione**, che garantisce **l'invarianza dell'indice T**, nel caso di **trasformazioni lineari** (Vedere capitolo sulle trasformazioni).

Per distribuzioni che sono simmetriche intorno a 0, la formula precedente

può essere semplificata in

$$\mathbf{T} = \frac{R_7 - E_5}{E_6}$$

Gli indici relativi della forma di una distribuzione attualmente più diffusi sono derivati dai **momenti**.

I momenti (m) di ordine k rispetto ad un punto c sono calcolati con

$$m_k = \frac{\sum (x_i - c)^k}{n}$$

per una serie di dati

e con

$$m_k = \frac{\sum (x_i - c)^k \cdot f_i}{n}$$

per una distribuzione di frequenza suddivisa in classi.

Abitualmente, con **c** si indica l'origine (**c = 0**) oppure la media (**c = media**). Nel primo caso, si parla di **momento rispetto all'origine**; nel secondo, di **momento centrale**.

Il momento di ordine 1 (k = 1) rispetto all'origine dei valori **(c = 0) è la media**;

il momento centrale (c = m) di ordine 1 (k = 1) é uguale a 0 (è la somma degli scarti dalla media).

Il momento centrale (c = m) di ordine 2 (k = 2) è la varianza.

Nello stesso modo del momento centrale di secondo ordine (**m_2**), si possono calcolare i momenti centrali di ordine terzo (**m_3**), quarto (**m_4**), quinto (**m_5**),...ennesimo (**m_n**).

I momenti centrali di ordine dispari (m_3, m_5,...) sono utilizzati per indici di simmetria.

Essi sono **nulli per distribuzioni simmetriche e differiscono da zero quando le distribuzioni non sono simmetriche**; quanto maggiore è l'asimmetria, tanto più il valore del momento centrale di ordine dispari è grande. Inoltre, in distribuzioni con **asimmetria destra ha un valore positivo** ed in quelle con **asimmetria sinistra ha un valore negativo**.

Per queste sue caratteristiche, il momento centrale di terzo ordine ($\mathbf{m_3}$) è adeguato per valutare la simmetria o asimmetria di una distribuzione; ma esiste il limite che il suo valore dipende dalla scala utilizzata.

Per ottenere una **misura relativa, adimensionale**, che permetta i confronti tra più distribuzioni, bisogna dividere $\mathbf{m_3}$ per il cubo dello scarto quadratico medio.

E' **l'indice γ_1** ***(coefficient of skewness)*** **di Fisher**

$$\gamma_1 = \frac{m_3}{\sigma^3}$$

detto anche il **momento standardizzato di terzo ordine** e che mantiene le proprietà precedentemente descritte.

I momenti centrali di ordine dispari sono nulli, quando la distribuzione è simmetrica; sono positivi o negativi rispettivamente quando vi è asimmetria destra o sinistra.

Nel caso di una distribuzione di dati sperimentali, l'indice di asimmetria (*coefficient of skewness*) è indicato con g_1.

Per **valutare il grado di asimmetria**, è convenzione che si abbia una distribuzione ad asimmetria

forte, quando $|\gamma_1| > 1$;

moderata, quando $½ < |\gamma_1| < 1$;

trascurabile, quando $0 < |\gamma_1| < 1/2$.

L'indice di asimmetria **$ß_1$ di Pearson** , storicamente antecedente al γ_1 di Fisher, è stato definito come

$$\beta_1 = \left(\frac{m_3}{\sigma^3}\right)^2$$

Quando calcolato su una distribuzione sperimentale è indicato con $\mathbf{b_1}$.

Fisher ha solo semplificato l'indice di Pearson,

mediante la relazione semplice

$$\gamma_1 = \sqrt{\beta_1}$$

ma nella pratica della statistica si è affermato il suo metodo.

E' utile ricordare quanto già ripetuto alcune volte:

- nel caso di distribuzioni simmetriche i 3 indici sk, γ_1, $ß_1$ danno un risultato uguale a 0;

- ma non sempre vale l'inverso, non sempre un indice di asimmetria uguale a 0 caratterizza la simmetria perfetta di una distribuzione di dati.

Quando si descrive la forma delle curve unimodali simmetriche, con il termine **curtosi** (dal greco kurtos, che significa curvo o convesso) si intende il **grado di appiattimento, rispetto alla curva normale o gaussiana** (le cui caratteristiche saranno discusse in modo più approfondito nel capitolo dedicato alle distribuzioni teoriche).

Nella valutazione della curtosi, una **distribuzione unimodale simmetrica** è detta:

- **mesocurtica** , quando ha forma uguale alla distribuzione normale;
- **leptocurtica** (figura 23), quando ha un eccesso di frequenza delle classi centrali, una frequenza minore delle classi intermedie ed una presenza maggiore delle classi estreme; è quindi una distribuzione più alta al centro e agli estremi e più bassa ai fianchi; la caratteristica più evidente è l'eccesso di frequenza dei valori centrali;
- **platicurtica** (figura 24), quando rispetto alla normale presenta una frequenza minore delle classi centrali e di quelle estreme, con una frequenza maggiore di quelle intermedie; è quindi una distribuzione più bassa al centro e agli estremi mentre è più alta ai fianchi; la caratteristica più evidente è il numero più ridotto di valori centrali.

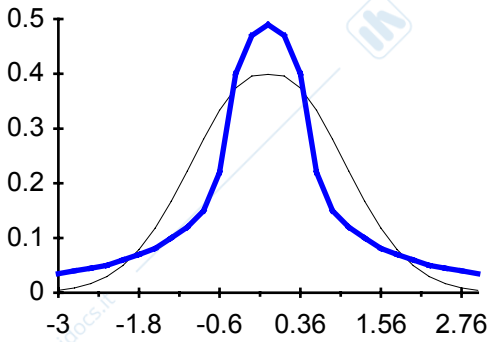


Figura 24. Distribuzione leptocurtica rispetto alla mesocurtica

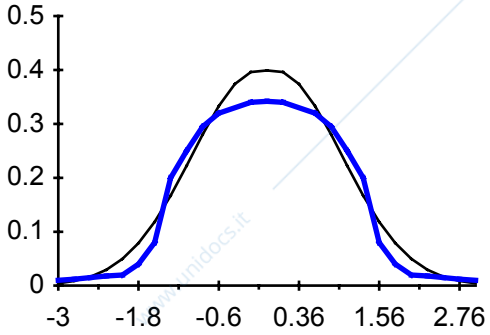


Figura 25. Distribuzione platicurtica rispetto alla mesocurtica

L'**indice di curtosi** è il risultato di un confronto, è un rapporto; quindi, **è una misura adimensionale**.

I **due indici di curtosi** più utilizzati sono analoghi a quelli di asimmetria:

- **l'indice γ_2 di Fisher** (g_2 in una distruzione di dati osservati)
- **l'indice β_2 di Pearson** (b_2.in una distribuzione di dati osservati)

L'**indice γ_2 di Fisher** è fondato sul rapporto

$$\gamma_2 = \frac{m_4}{\sigma^4}$$

Se la distribuzione è perfettamente normale, il risultato del calcolo è uguale a 3; è maggiore di 3 se la distribuzione è leptocurtica, mentre è minore di 3 se la distribuzione è platicurtica.

Per spostare la variazione attorno allo 0, l'indice di curtosi di Fisher è scritto come

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3$$

Ovviamente, **il risultato diviene**

0, se la distribuzione è normale o mesocurtica,

positivo, se la distribuzione è leptocurtica o ipernormale,

negativo, se la distribuzione è platicurtica o iponormale

Mentre l'indice γ_1 può variare tra $\pm \infty$, l'indice γ_2 può variare tra -2 e $+\infty$; non è quindi possibile associare ad esso una gradazione in valore assoluto che valuti l'intensità della curtosi. Come già precedentemente discusso, le condizioni che γ_1 e $\gamma_2 = 0$ sono necessarie ma non sufficienti, affinché la curva sia simmetrica e mesocurtica.

L'**indice β_2 di Pearson** è il rapporto fra il momento centrale di quarto ordine e la deviazione standard, elevato alla quarta potenza:

$$\beta_2 = \frac{m_4}{\sigma^4}$$

Il suo legame con γ_2 di Fisher è semplice, di tipo lineare:

$$\beta_2 = \gamma_2 + 3$$

Come l'indice γ_2 varia attorno a 0, β_2 varia attorno a 3.

Tutti gli indici presentati, dalle misure di tendenza centrale a quelle di dispersione e di forma, sono validi sia per variabili discrete che continue, con l'ovvia approssimazione data dal raggruppamento in classi.

Quando simmetria e curtosi sono stimate non sulla popolazione (γ_1 e γ_2) ma su un campione (quindi indicate con i corrispondenti simboli latini g_1 e g_2),

g_1 in valore assoluto tende a sottostimare γ_1 ($|g_1| < |\gamma_1|$); infatti è possibile dimostrare che, in un campione di dimensioni **n**, non supera il valore della radice di **n**

$$|g_1| \leq \sqrt{n}$$

Problemi simili esistono per la stima di γ_2 in piccoli campioni con forte curtosi.

I limiti di g_2 sono

$$\frac{-2(n-1)}{n-3} \leq g_2 \leq N$$

Con dati campionari, simmetria e curtosi sono ovviamente calcolati da distribuzioni di frequenza raggruppate in classi.

Definendo

- **k** = numero di classi di frequenza
- f_i = frequenza della classe i,
- $\bar{x}_i$ = valore centrale della classe i
- $\bar{\bar{x}}$ = media generale del campione
- **s** = deviazione standard del campione

e da essi avendo ricavato

$$\sum x^3 = \sum_{i=1}^{k} f_i \left(\bar{x}_i - \bar{\bar{x}}\right)^3$$

e

$$\sum x^4 = \sum_{i=1}^{k} f_i \left(\bar{x}_i - \bar{\bar{x}}\right)^4$$

si calcola g_1 con

$$g_1 = \frac{n \cdot \sum x^3}{(n-1)\cdot(n-2)\cdot s^3}$$

e g_2 con

$$g_2 = \frac{(n+1)\cdot n \cdot \sum x^4}{(n-1)\cdot(n-2)\cdot(n-3)\cdot s^4} - \frac{3\cdot(n-1)^2}{(n-2)\cdot(n-3)}$$

I valori di g_1 e g_2 sono adimensionali: in altri termini, il risultato è identico, qualunque sia la misura utilizzata o la trasformazione applicata alla variabile X.

Alla fine del capitolo 10 sono riportati i test proposti da **Snedecor** e **Cochran** sul loro testo ***Statistical Methods***, per valutare la significatività di $\mathbf{g_1}$ e $\mathbf{g_2}$ in campioni di grandi dimensioni (oltre 100 dati).

1.9. ACCURATEZZA, PRECISIONE E SCELTA DEL NUMERO DI CIFRE SIGNIFICATIVE

Un conteggio di poche unità fornisce una misura precisa: con alta probabilità, la sua ripetizione determina lo stesso valore. Un conteggio ripetuto di un campione grande, formato da varie centinaia o migliaia di unità, difficilmente conduce allo stesso risultato, per la frequenza con la quale si commettono errori. D'altronde tale conteggio non è sostanzialmente modificato, se gli individui risultassero 15.612 oppure 15.623.

Quando si utilizza una scala continua, cioè da uno strumento si ricava la misura di una lunghezza o di un peso, quasi sempre la ripetizione conduce a valutazioni differenti. Inoltre, a causa della variabilità biologica ed ambientale, se effettuate su più individui e non determinate con troppa approssimazione, queste misure non conducono mai a risultati identici.

Quando si dispone di misure ripetute, la distribuzione dei valori può essere rappresentata e quantificata con gli indici di statistica descrittiva già presentati. Essi servono per rispondere a due domande:

- Quale è il valore reale del fenomeno?
- Come descrivere la variabilità del fenomeno o l'errore commesso nella sua misura?

Al momento della raccolta dei dati, occorre quindi tenere presente che i valori devono essere misurati con la **precisione utile per fornire una risposta accurata** alle due domande precedenti.

E' ovvio che quando la misura è approssimata, come il peso di una persona che sia arrotondato al chilogrammo, è impossibile valutare l'errore di una bilancia, se esso è limitato a uno o al massimo due ettogrammi. Nello stesso modo, ma simmetricamente, è assurdo pretendere misure precise al grammo con una bilancia pesa – persone.

Tale concetto è insito nella convenzione con la quale i dati sono espressi. Se si afferma che un individuo pesa 68 Kg., non si intende dire che è esattamente Kg. 68,000 (cioè 68.000 gr), ma un valore compreso tra Kg. 67,5 e 68,5 o, se si vuole una misura più precisa, tra Kg. 67,50 e 68,49. Sulla base dello stesso principio, se si scrive che quell'individuo è alto 1,71 metri, si intende affermare un valore con approssimazione al centimetro; quindi che egli varia tra mm. 1705 e 1715 (o meglio 1,7149).

In un conteggio di alcune migliaia di individui, 15.000 non indica che essi sono esattamente quella cifra, ma un dato compreso tra 14.500 e 15.500 (15.499). Un conteggio di 150 individui indica che essi variano tra 145 e 155 unità.

Se si vuole fornire una misura estremamente precisa, come può essere un peso fornito in grammi o un'altezza in millimetri, si dovrebbe acquistare uno strumento più sofisticato di quello abitualmente in commercio. E' un'operazione che ha un costo, che richiede maggiore attenzione, e potrebbe rappresentare un dispendio inutile di tempo e risorse. Anche le operazioni statistiche avrebbero poi un appesantimento inutile, quando non dannoso, per calcolare gli indici di tendenza centrale, dispersione, simmetria e curtosi.

Da queste osservazioni fondate sulla pratica quotidiana deriva che, al momento della raccolta dei dati, si pone un **problema pratico non banale**. E' necessario definire quale è il **numero di cifre significative che è utile raccogliere**; ricordando che

- una approssimazione troppo grande conduce a valori identici e fa perdere una parte importante dell'informazione,
- una rilevazione troppo fine aumenta inutilmente i costi e i tempi della ricerca, senza accrescere realmente l'informazione sul fenomeno studiato.

Le soluzione di questo dilemma dipende dall'**errore che si accetta di commettere.** Esso è legato ai concetti di **precisione** e di **accuratezza di una misura**, che nel linguaggio comune sono sinonimi. In realtà, nel linguaggio statistico hanno significati differenti.

L'**accuratezza è la vicinanza di un valore misurato al suo valore reale** (***accuracy is the closeness of a measured value to its true value***) e in buona parte dipende dallo strumento. Per esempio, nelle misure ecologiche, il metodo delle trappole per la stima della quantità di individui in una popolazione e quello del C-14 per la produzione di plancton nell'oceano tropicale hanno una accuratezza molto bassa; cioè possono essere molto distanti dal valore reale. Ogni stima, ripetuta nelle stesse condizioni, può dare la metà del valore reale. Uno strumento o un reagente che forniscono una risposta sbagliata spesso sono tarati in modo non corretto e sono definiti inaccurati; i valori ottenuti sono ***biased***.

Nel processo di misurazione con uno strumento inaccurato si commette un errore sistematico, chiamato appunto ***bias***. Esso rappresenta un problema importante e ricorrente, in molte tecniche di stima di una quantità. In varie discipline, il progresso è spesso collegato alla ricerca di metodi di misurazione più accurati.

La **precisione è la vicinanza di misure ripetute al medesimo valore** (***precision is the closeness of repeated measuments to the same item***). Spesso dipende dalla capacità del tecnico di ripetere la misurazione con le stesse modalità e ha origine dalla sua esperienza o abilità.

I concetti di ***accuracy***, ***precision*** e ***bias*** applicati a una misura di peso sono rappresentati graficamente nella figura successiva tratta dall'ottimo testo, per la stima di popolazioni animali e vegetali, di Charles J. **Krebs** del 1999 (***Ecological methodology***, 2nd ed. Menlo Park, CA, Cummings, pp. XII, 620),

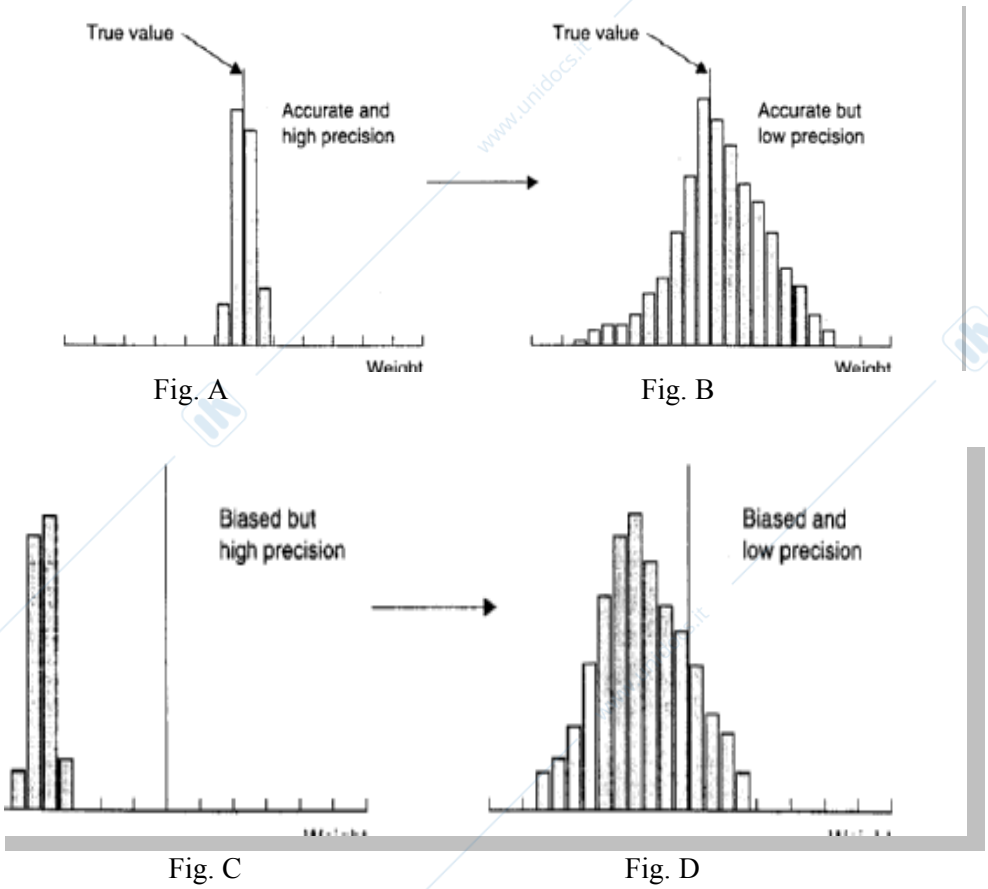


Fig. A Fig. B

Fig. C Fig. D

Figura 26. Illustrazione grafica dei concetti di **accuratezza** (***accuracy***) e **precisione** (***precision***) in misure di peso.

Nella figura A le misure sono **accurate**, vicine al valore vero (***true value***), e **molto precise** (***high precision***).

Nella figura B le misure sono **accurate ma poco precise** ***(low precision***), cioè differenti tra loro.

Nella figura C le misure **sono non accurate** (***biased***) ma **molto precise** (***high precision***).

Nella figura D le misure s**ono non accurate** (***biased***) e con **poco precise** (***low precision***).

Un valore può essere **preciso ma inaccurato**.
Un esempio didattico riportato in alcuni testi, applicato a una misura bidimensionale, è il tiro ad un bersaglio con un'arma, dove la media delle varie prove permette di misurarne l'accuratezza e la loro variabilità la precisione.
Se tutti i colpi centrano esattamente il bersaglio o sono molto vicini a esso, con media esattamente sul centro, si ha accuratezza (il fucile è tarato esattamente per le caratteristiche visive di chi spara) e precisione (il tiratore è abile).
Se i colpi sono tutti nello stesso posto, ma lontani dal centro del bersaglio, si ha inaccuratezza o misure biased, ma precisione. Il fucile è tarato male, ma il tiratore sa sparare.
Se i colpi sono molto dispersi intorno al bersaglio e la loro media coincide con quella del centro, si ha accuratezza ma bassa precisione: il fucile è tarato esattamente, ma il tiratore non sa sparare con precisione (ad esempio, gli trema la mano).
Se i colpi formano una rosa molto ampia e la loro media è distante dal centro, si ha inaccuratezza e bassa precisione: lo strumento è *biased* e l'individuo non sa usarlo correttamente.

Per valutare in modo appropriato questi due fenomeni, soprattutto la dispersione o variabilità dei dati, è importante che **le misure raccolte abbiano un numero di cifre significative che sia adeguato**. E' un concetto sfumato, per la soluzione del quale sono state proposte varie metodologie, che si fondano sulla variabilità delle misure campionarie. Il testo di **Krebs** già citato, sia nella prima edizione del 1989 (pubblicato da Harper Collins) sia nella seconda del 1999 (pubblicata da Benjamin/Cummings), riporta due metodi
A - il primo attribuito a **Sokal e Rohlf**,
B - il secondo a **Barford**.
che non forniscono risultati identici. Il secondo è giudicato più conservativo, cioè induce a raccogliere misure con un numero minore di cifre significative.

A - Il **metodo di Sokal e Rohlf**, proposto sia nel loro testo del 1981 (***Biometry***, edito da **W. H. Freeman**, San Francisco) che nella edizione del 1995 (***Biometry***, editi da **W. H. Freeman and Co.**, New York), non ha alcuna base teorica ma è fondato solo sull'esperienza. Secondo alcuni autori si richiama al **buon senso statistico,** che ovviamente possono avere solo persone con esperienza. Secondo altri, più critici, è una **indicazione a "lume di naso**"; ma questi autori non hanno formulato proposte migliori.

R. R. **Sokal** e F. J. **Rohlf** affermano che, **tra la misura minore e quella maggiore, il numero di valori possibili dovrebbe variare tra 30 e 300**. Per ottenere questo risultato, è sufficiente che **il campo di variazione (*range*) delle misure da effettuare sia diviso prima per 30 e poi per 300**.
Ad esempio, per misurare l'altezza di un gruppo di giovani, dove è noto che il fenomeno varia approssimativamente da 150 a 200 cm,
- dopo aver determinato il **range o campo di variazione**

range = valore massimo – valore minimo

50 = 200 - 150

- si ricavano il **livello minimo di misurazione** con il rapporto

livello minimo di misurazione = range / 30

livello minimo di misurazione = cm. 50 / 30 = cm. 1,67

dove essa risulta uguale a cm. 1,67

- e il **livello massimo di misurazione** con il rapporto

livello massimo di misurazione = range / 300

livello massimo di misurazione = cm. 50 / 300 = cm. 0,167

dove essa risulta uguale a cm. 0,167.

In termini semplici, quasi banali, è possibile affermare che in questa ricerca il livello di misurazione può variare tra un minimo di circa 1,5 cm. e un massimo di circa 2 mm. Poiché è conveniente ottenere circa 50 valori differenti, un numero compreso nell'intervallo tra 30 e 300, il livello di misurazione può essere il cm. Ma sarebbe ugualmente accettabile volere 100 misure differenti, cioè l'approssimazione a cm 0,5.
Invece, se la misurazione avesse come unità 2 cm o peggio ancora 5 cm, si otterrebbero rispettivamente solo 25 e 10 possibili valori differenti, un numero troppo basso. All'opposto, misure che rilevino il mm determinerebbero 500 possibili valori, un numero eccessivamente alto.

ESEMPIO. Qual è il numero di cifre significative per effettuare misurazioni di un fenomeno che varia approssimativamente da gr. 3 a 5?

Risposta. Il campo di variazione è

range = gr. 5 – gr. 3 = gr. 2

uguale a gr. 2.
Il livello minimo di misurazione è

livello minimo di misurazione = gr. 2 / 30 = gr 0,0667

uguale a gr.0,067

mentre il livello massimo è

livello massimo di misurazione = cm. 2 / 300 = cm. 0,0067

uguale a gr. 0,0067.

In altri termini,

- con un solo decimale potrei avere solo 20 valori differenti: è un numero troppo basso;
- con due decimali 200 valori, cioè un numero compreso tra 30 e 300;
- con tre cifre 2000 valori: un numero troppo alto.

In conclusione, la misura corretta dovrebbe valutare il centesimo di grammo; se il numero appare eccessivo per la precisione dello strumento è possibile accettare una misura approssimata a due centesimi (corrispondente a 100 possibili valori differenti) o al massimo a 5 centesimi di grammo (corrispondenti a 40 valori).

B - Il metodo proposto da N. C. **Barford** nel suo testo del 1985 (***Experimental Measurements***: ***Precision, Error and Truth.*** John Wiley & Sons, New York) è fondato sulla **stima dell'errore standard** (che misura la dispersione delle medie di n dati).

Poiché l'**accuratezza relativa dell'errore standard** (***relative accuracy of standard error***) può essere determinata, in modo approssimato, sulla sola base del numero di dati (**n**) che si vogliono raccogliere

$$\text{Accuratezza relativa di es} \cong \frac{1}{\sqrt{n-2}}$$

una volta che sia stata effettuata una misurazione pilota dell'errore standard è possibile stimare **l'intervallo di variazione dell'errore probabile dell'errore standard** (***range of probable error of the standard error***) attraverso la relazione

Errore probabile dell'errore standard = (es) • (accuratezza relativa di es)

Da esso si ricava **il campo di variazione probabile dell'errore standard** e si deduce logicamente il **numero di cifre significative**.

Riprendendo l'esempio di Krebs, se

- si intendono determinare 100 misure (n = 100),
- e, con uno studio pilota, sono state stimate sia la media del campione ($\bar{x}$ = 173,86 mm),
- sia la sua deviazione standard (s = 12,26)

dopo aver ricavato l'**errore standard**

con

$$es = \frac{s}{\sqrt{n}} = \frac{12{,}26}{\sqrt{100}} = 1{,}226$$

e **la sua accuratezza relativa**

con

$$\text{accuratezza relativa di es} = \frac{1}{\sqrt{100-2}} = 0{,}1010$$

si perviene dapprima alla **stima dell'errore probabile di es**

mediante la relazione

$$\text{errore probabile di es} = (\text{es}) \bullet (\text{accuratezza relativa di es})$$

$$\text{errore probabile di es} = \pm(1{,}226)\cdot(0{,}1010) = \pm 0{,}1238 \text{ mm}$$

e infine si ottiene **il campo probabile di variazione dell'errore standard**

mediante

$$\text{campo probabile di variazione dell'errore standard} = (\text{es}) \pm (\text{errore probabile di es})$$

$$\text{campo probabile di variazione dell'errore standard} = 1{,}226 \pm 0{,}1238$$

dalla quale si ricava che esso può variare tra 1,3498 (1,226 + 0,1238) e 1.1022 (1,226 – 0,1238).
In altri termini esso ha un range

$$\text{range} = 1{,}3498 - 1{,}1022 = 0{,}2476$$

di circa 0,25 mm.
Per cogliere una differenza tra i due limiti (circa 1,3 mm e circa 1,1 mm) le misure devono avere la precisione di almeno una cifra decimale, cioè 0,1 mm.

Con questa impostazione, la scelta del numero di cifre significative dipende

- dal numero di dati che si intendono raccogliere,
- oltre che dalla naturale variabilità del fenomeno (s).

L'errore standard e la sua accuratezza diminuiscono in valore all'aumentare del numero di misure (n) che si vogliono effettuare.
Di conseguenza,

- il numero di cifre significative diminuisce se n diminuisce: ci si accontenta di una misura meno precisa e quindi anche il numero di cifre significative può essere minore,
- il numero di cifre significative cresce all'aumentare di n.

Per esempio, con la stessa deviazione standard (s) e solo 10 misure (n = 10), l'errore standard probabile è ± 0,433 e quindi il campo di variazione dell'errore standard è tra 0,8 mm e 1,7 mm. Misure rilevate con una cifra decimale (mm 0,1) sono ampiamente sufficienti per cogliere queste differenze.

1.10. METODI PER CALCOLARE UN GENERICO QUANTILE DA UNA SERIE DI DATI

Nella statistica ambientale, è diffuso l'uso dei quantili, per due scopi:

- individuare la collocazione di un dato entro una serie di rilevazioni,
- stimare il valore di uno specifico percentile, come avviene frequentemente per la mediana o il primo e il terzo quartile.

Disponendo di uno o più dati, è utile conoscere la loro collocazione entro una distribuzione sperimentale precedente, come indicazione approssimata e descrittiva della sua posizione. In questo caso, la soluzione è semplice: il percentile è la posizione o rango occupata da quel valore, rapportata appunto a 100.

Più complesso è stimare il valore esatto di un determinato percentile, con differenze metodologiche tra una serie limitata di dati o una distribuzione di frequenza.

Il concetto di stima di un quartile appare semplice; ma i metodi di calcolo non sono così banali ed unanimi come possono apparire.

Non esiste un metodo unico, con risultati universalmente accettati, in quanto

- **le varie proposte rintracciabili nei testi mostrano tutte inconvenienti od illogicità di tipo differente**.

Quando il numero di osservazioni è alto, tutti i metodi forniscono risposte simili, spesso coincidenti; ma quando il numero di dati è limitato e sono presenti valori anomali, i risultati differiscono anche in modo sensibile.

A dimostrazione dei concetti appena espressi, si supponga di disporre di una serie di dati fortemente asimmetrica e con valori anomali, quale

1, 7, 4, 2, 50, 51,

in cui **n = 6,**

per calcolare un generico quantile **Px**.

Dopo aver ordinato gli **n** dati in modo crescente, ottenendo

1, 2, 4, 7, 50, 51,

un primo metodo richiede di

1 - Calcolare **R,** che è dato da

$$\mathbf{R = ((n - 1) \bullet Px) + 1}$$

Con **n = 6** dati e **Px** supposto uguale al 3° quartile, (3/4 oppure 75/100, espresso nell'intervallo 0-1) e quindi **Px = 0,75**

$$R = ((6 - 1) \bullet 0{,}75) + 1 = 3{,}75 + 1 = 4{,}75$$

si ottiene **R** = 4,75.

Il valore di **R** (che nell'esempio è uguale a 4,75) indica che il quantile da stimare si trova tra il 4° e il 5° valore nella serie ordinata dei dati ed esattamente nella posizione 0,75 della distanza tra i valori di rango 4 e rango 5. Per l'individuazione di tale valore, il metodo qui presentato (valido anche per la mediana con **Px** = 0,5) chiede ulteriori passaggi, quali

2 – Prendere **I**, la parte intera di **R**,

$$\mathbf{I = Int\ (\ R\ )}$$

per cui, nell'esempio,

$$I = Int\ (4{,}75) = 4$$

I risulta uguale a 4.

3 – Calcolare **D** per differenza tra **R** e **I**

$$\mathbf{D = R - I}$$

che, sempre con i 6 dati dell'esempio

$$D = 4{,}75 - 4 = 0{,}75$$

risulta uguale a 0,75.

4 – Individuare nella serie ordinata dei dati $\mathbf{X_{(I)}}$ e $\mathbf{X_{(I+1)}}$

cioè (con **I = 4**) i valori che occupano il rango 4° e 5°, per cui, con i dati dell'esempio,

$$X_{(4)} = 7 \quad \text{e} \quad X_{(5)} = 50$$

5 - La stima del quantile (**Q**) è determinata dalla relazione

$$\mathbf{= (1 - D) \bullet X_{(I)} + D \bullet X_{(I+1)}}$$

Con i dati dell'esempio, il 3° quartile ($\mathbf{Q_{0,75}}$) è

$$\mathbf{Q_{0,75}} = (1 - 0{,}75) \times 7 + 0{,}75 \times 50 = 1{,}75 + 37{,}5 = 39{,}25$$

uguale a 39,25.

Dopo aver calcolato che il quantile ($Q_{0,75}$) desiderato si trova in posizione 4,75 su 6 dati, **una variante del primo metodo** appena descritto è fondata sull'interpolazione lineare a 0,75 tra il valore che occupa il 4° rango ($X_{(4)} = 7$) e quello che occupa il 5° rango ($X_{(5)} = 50$). Dopo averne stimato la differenza **d**

$$\mathbf{d} = X_{(I+1)} - X_{(I)} = 50 - 7 = 43$$

si calcola la quota dovuta alla proporzione **P** (0,75) che eccede il rango **I** mediante la proporzione

$$P = 43 \times 0,75 = 32,25$$

e viene sommata al valore del rango **I**

$$Q_{0,75} = 7 + 32,25 = 39,25$$

per ottenere un valore (39,25) uguale al precedente.

Un secondo metodo calcola il quantile **Px** mediante la relazione

$$\mathbf{Rx = n \bullet Px + 0,5}$$

per cui il 75° percentile o terzo quartile con **n = 6** dati è

$$R_{0,75} = 6 \times 0,75 + 0,5 = 5,0$$

esattamente il 5° valore.

Con i 6 dati dell'esempio precedente $Q_{0,75}$ risulta uguale a 50.

Per la quota eccedente l'intero **I**, quando esiste, si può usare **l'interpolazione come calcolata prima**, fra il valore $X_{(I)}$ e $X_{(I+1)}$.

Altri autori, con **un terzo metodo**, definiscono il valore Qx del quantile Px nei termini della relazione

$$Rx = Px \bullet (\mathbf{n} + 1)$$

per cui il 75° percentile o terzo quartile con **n = 6** dati è

$$Rx = 0,75 \times (6 + 1) = 5,25$$

il valore che occupa la posizione 5,25.

Di conseguenza $Q_{0,75}$ può essere stimato per interpolazione, tra il 5° e il 6° valore, risultando

$$Q_{0,75} = 50 + 0,25\ (51 - 50\ ) = 50,25$$

uguale a 50,25.

Anche **<u>questo metodo presenta varianti</u>**, fondate sulla logica di non voler stimare un valore che pretende di essere molto più preciso di quanto siano oggettivamente i dati:

- una prima è l'arrotondamento all'intero più vicino, per cui è il 5° valore e $Q_{0,75}$ risulta uguale a 50,
  una seconda è l'interpolazione come media tra i due valori, calcolando quindi $Q_{0,75}$ uguale a 50,5.

Anche il primo metodo, al quale ricorrono vari programmi informatici a grande diffusione, presenta inconvenienti logici, come evidenzia l'esempio seguente.

Le misure dell'inquinamento idrico spesso sono fornite come medie mensili; in Italia spesso manca il dato di agosto, coincidente con il mese di ferie. Calcolare il 9° decile della serie di 11 valori

12, 10, 8, 7, 14, 27, 29, 21, 14, 11, 9

Dopo aver ordinato per rango i valori

7, 8, 9, 10, 11, 12, 14, 14, 21, 27, 29

il 90° percentile

$$R_{0,9} = (11 - 1) \times 0,9 + 1 = 9 + 1 = 10$$

risulta il 10° valore, per cui $Q_{0.9}$ è uguale a 27.

Se è corretto che la mediana o $R_{0,5}$ sia uguale esattamente al sesto valore, è indubbiamente una stima approssimata che tutti i decili da 1 a 9, come indicano i calcoli, risultino esattamente i valori che occupano le posizioni dalla seconda alla decima.

E' utile ricordare quanto affermato da Peter **Armitage** e Geoffry **Berry** (in ***Statistica Medica, metodi statistici per la ricerca in medicina***, 3a edizione, in italiano, McGraww-Hill Libri Italia srl, Milano 1996, a pag. 33):

- " ***Si noti che non esiste un'unica procedura standard nel calcolo dei quartili*** *(e dei quantili).* ***Le diverse convenzioni conducono, comunque, a*** *piccole e insignificanti* ***differenze tra i risultati finali***".

Con eccezione della sola mediana, non appare possibile definire quale sia il metodo migliore. Le differenze tra i diversi risultati, come nel caso di dati fortemente anomali, possono anche essere di quantità rilevanti, contrariamente a quanto affermato da Armitage; ma è un'incertezza insita nella variabilità delle osservazioni campionarie e nel numero limitato di osservazioni.

Di conseguenza,

- **è evidente la difficoltà di pervenire a conclusioni generali e condivise, attraverso analisi fondate sui quantili**.

1.11. RAPPRESENTAZIONI GRAFICHE E SEMI-GRAFICHE DELLE DISTRIBUZIONI: DATA PLOT, BOX-AND-WHISKER, LINE PLOT, STEM-AND-LEAF

Per evidenziare le caratteristiche di una tabella o di un semplice elenco di dati, sono utili anche altre rappresentazioni grafiche o semigrafiche. Eccetto il box-and-whisker, possono essere costruiti con facilità, anche senza l'aiuto di programmi informatici. Nella terminologia internazionale, quelli di uso più comune sono

A – ***data plot,***

B – ***box-and-whisker***,

C – ***line plot***,

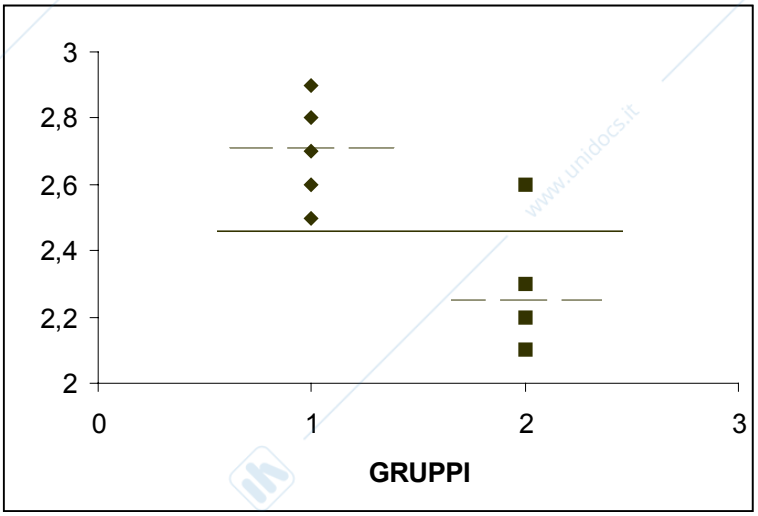
D – ***stem-and-leaf***.

A – Il metodo più semplice e intuitivo è la rappresentazione del valore di ogni singolo dato con un punto: il ***data plot***, chiamato anche ***dotplot***.

E' un diagramma cartesiano, in cui

- ogni **singolo dato (*data*)** è **rappresentata da un punto (*plot*)** o altro segno convenzionale che
- sull'**asse verticale** indica il **valore osservato**,
- e sull'**asse orizzontale riporta** il **nome**, il **numero** o un **simbolo di identificazione del gruppo**.

Ad esempio, la figura



è il **data plot** della tabella

Gruppo 1	2,7	2,8	2,9	2,5	2,6	2,7	2,8	---
Gruppo 2	2,2	2,1	2,2	2,3	2,1	2,2	2,3	2,6

Nel grafico, il numero di punti risulta minore di quello dei dati riportati in tabella, perché alcuni hanno valori identici e quindi sono graficamente sovrapposti.

B - BOX-AND-WHISKER

I **diagrammi *Box-and-Whisker*** (**scatola-e-baffi**), chiamati anche più rapidamente ***boxplot***, sono stati presentati in modo organico per la prima volta da John W. **Tukey** nel suo testo del 1977 (***Exploratory Data Analysis***, pubblicato da Addison-Wesley, Reading, Mass.).

Sono un metodo grafico diffuso recentemente e reso di uso corrente dai programmi informatici, che possono costruirlo con rapidità. La quantità di informazioni che forniscono è elevata. Molto più raramente, sono chiamati anche ***five number summary*** poiché, nella loro forme più semplice, riassumono in cinque numeri le informazioni contenute nella distribuzione:

- la **mediana**,
- il **primo** e il **terzo quartile**;
- il **valore minimo** e quello **massimo**.

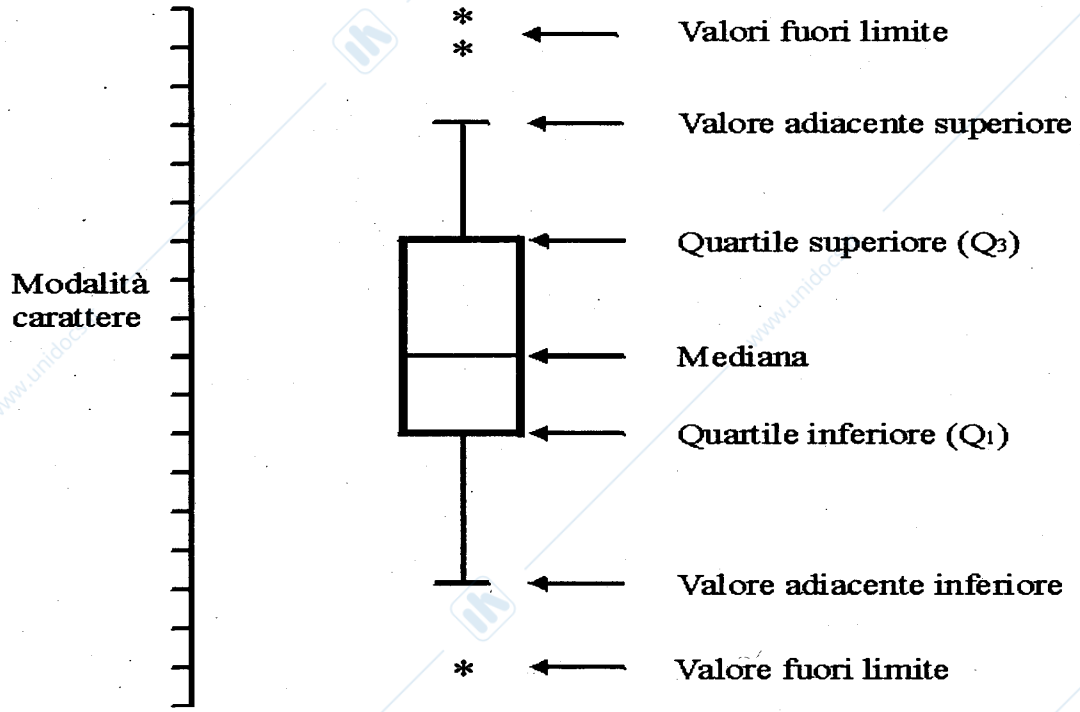


Figura 27. Box-and-Whisker come proposto da Tukey nel 1977.

Servono per rappresentare visivamente **quattro caratteristiche fondamentali** di una **distribuzione statistica** di dati campionari:

1 - la misura di **tendenza centrale**, attraverso la mediana e/o la media;

2 - il grado di **dispersione o variabilità dei dati**, rispetto alla mediana e/o alla media;

3 – **la forma della distribuzione** dei dati, in particolare la **simmetria**;

4 – sia la semplice **presenza** che **l'individuazione specifica** di ogni **valore anomalo** o ***outlier***.

Secondo il metodo proposto da **Tukey** nel 1977, riportato nella figura precedente con i termini in italiano, un diagramma ***Box-and-Whisker*** o ***boxplot*** è costruito a fianco di una scala, che riporta le modalità o i valori del carattere.

La sua realizzazione richiede una serie di passaggi logici, che può essere riassunta in uno schema composto da 8 punti, dalla quali derivano gli elementi metodologici:

1 - Ha **origine da una linea orizzontale**, interna alla scatola, che rappresenta la **mediana** (***median***).

2 - La **scatola** (***box***) è **delimitata da due linee orizzontali**:

- la **linea inferiore**, indicata con $\mathbf{Q_1}$, che rappresenta il **primo quartile** o **quartile inferiore** (***lower quartile*** oppure più raramente ***lower fourth***);
- la **linea superiore**, indicata con $\mathbf{Q_3}$, che rappresenta il **terzo quartile** o **quartile superiore** (***upper quartile*** o più raramente ***upper fourth***).

Quartiles e ***fourths*** nel linguaggio di Tukey non sono esattamente sinonimi: i ***fourths*** **sono quartili approssimati**, che segnano i limiti del box. Ma sono distinzioni spesso ignorate, nel linguaggio scientifico più diffuso nella statistica applicata.

3 - La **distanza tra il terzo (Q_3) e il primo quartile (Q_1)**, detta **distanza interquartilica** (***interquartile range*** o **IQR**), è una misura della dispersione della distribuzione.

E' utile soprattutto quando sono presenti valori anomali, poiché

- **tra il primo e il terzo quartile** ($Q_3 - Q_1$) per costruzione sono **compresi il 50% delle osservazioni collocate intorno al valore centrale**.

Un intervallo interquartilico piccolo indica che la metà delle osservazioni ha valori molto vicini alla mediana. L'intervallo aumenta al crescere della dispersione (varianza) dei dati.

Inoltre, esso fornisce informazioni anche sulla forma della distribuzione (soprattutto sulla simmetria):

- se la linea inferiore e la linea superiore della scatola (cioè Q_1 e Q_3) hanno distanze differenti dalla mediana, la distribuzione dei valori è asimmetrica.

4 - Le **linee che si allungano dai bordi della scatola** e che **si concludono con altre due linee orizzontali**, i **baffi** (***whiskers***), delimitano gli intervalli nei quali sono collocati

- i **valori minori di Q_1** (nella parte inferiore)
- e **quelli maggiori di Q_3** (nella parte superiore).

Questi punti estremi, evidenziati dai baffi, in italiano spesso sono chiamati **valori adiacenti**.

5- Indicando con **r** la **differenza interquartilica**

$$r = Q_3 - Q_1$$

si definiscono le quantità che individuano

- il **Valore Adiacente Inferiore** (**VAI**), definito come il **valore osservato più piccolo che sia maggiore o uguale a Q_1 - 1,5r**:

$$VAI \geq Q_1 - 1{,}5r$$

- il **Valore Adiacente Superiore** (**VAS**) definito come il **valore osservato più grande che risulta minore o uguale a Q_3 + 1,5r**:

$$VAS \leq Q_3 + 1{,}5r$$

Una attenzione particolare deve essere posta alla **quantità 1,5 delle due formule**, per calcolare il VAI e il VAS. E' stato proposto da **Tukey**, ma non ha particolari proprietà.

Alla domanda del perché avesse indicato 1,5 e non un altro valore, la risposta di **Tukey** è stata che la sua scelta era fondata sulla **sua esperienza**, sul **suo "buon senso" statistico**.

6 - Se i due valori estremi sono contenuti entro l'intervallo tra VAI e VAS, i due baffi rappresentano i valori estremi e **nei dati raccolti non sono presenti valori anomali**.

7 - I valori esterni a questi limiti sono definiti **valori anomali** (***outliers***).

Nella rappresentazione grafica del ***box-plot***, gli ***outliers*** **sono segnalati individualmente**, poiché - costituiscono una **anomalia importante** rispetto agli altri dati della distribuzione

- e **nella statistica parametrica il loro peso sulla determinazione quantitativa dei parametri è molto grande.**

I valori che

- si discostano dalla mediana **tra 1,5 e 3 volte la distanza interquartile** possono essere considerati nella **norma**,

- mentre quelli che si discostano oltre **3 volte la distanza interquartile** dovrebbero essere molto rari e meritano una verifica ulteriore, per **distinguere con sicurezza gli outliers da quelli che possono essere stati determinati da banali errori di misura o di trascrizione**.

Gli **ipotetici outlier** dovrebbero essere sempre verificati, per capire le cause che li hanno determinati e quindi apportare le eventuali correzioni, se si trattasse di errori effettivi.

8 - Anche i due valori adiacenti (VAI e VAS), con la loro **distanza dai quartili**

$$\mathbf{Q_1 - VAI} \quad \text{e} \quad \mathbf{VAS - Q_3}$$

forniscono informazioni

- sia sulla **variabilità dei dati** sia sulla la **forma della distribuzione**.

Se **la distribuzione è normale**,

- nel box-plot le distanze tra ciascun quartile e la mediana saranno uguali
- e avranno lunghezza uguale le due linee che partono dai bordi della scatola e terminano con i baffi.

In una **distribuzione normale**, i **due baffi (*whiskers*) distano dalla mediana (me) una quantità pari a 2,69796 volte la deviazione standard (σ)**.

Questo intervallo comprende il **99,3% delle osservazioni** e

per valori estremi ha

$$me \pm 2{,}69796 \cdot \sigma$$

I diagrammi Box-and-Whiskers hanno avuto una serie di adattamenti e evoluzioni.

Tra le versioni più diffuse nei programmi informatici internazionali, sono da ricordare due tipi:

- quelli che impiegano la mediana come valore di tendenza centrale ed utilizzano la distribuzione dei quartili o dei percentili e si rifanno al modello descritto;
- quelli che riportano la media, insieme con l'errore standard e la deviazione standard.

I primi forniscono una descrizione non parametrica della forma della distribuzione, evidenziando dispersione e simmetria. I secondi rappresentano indici parametrici, presupponendo una distribuzione normale. Essi evidenziano sia la dispersione dei dati sia quella della media campionaria (questi argomenti saranno trattati in modo dettagliato quando si discuterà l'intervallo fiduciale o di confidenza).

Nei due Box-and-Whisker della figura 28, il valore di riferimento centrale è la mediana, la scatola delimita il primo ed il terzo quartile, mentre i baffi individuano il valore minimo e quello massimo.

Le due distribuzioni non sono perfettamente simmetriche: la loro mediana non è equidistante dal 1° e dal 3° quartile, individuato dall'altezza della scatola, né dal valore minimo e massimo, rappresentato dai baffi.

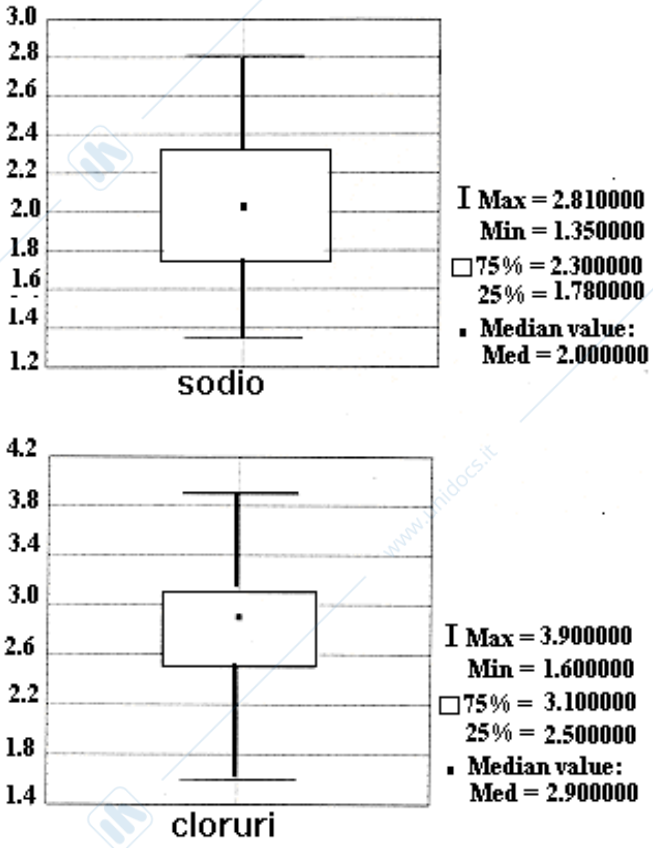


Figura 28. Box-and-Whisker con misure non parametriche, utilizzando una distribuzione di dati.

La distribuzione dei dati del sodio (i cui valori dettagliati sono riportati nell'ultimo paragrafo di questo capitolo) ha una asimmetria positiva o destra, mentre la distribuzione dei valori dei cloruri (riportati stessa tabella) ha una asimmetria sinistra o negativa. La rappresentazione in istogrammi e la misura del grado di asimmetria descrivono una lieve alterazione rispetto ad una distribuzione perfettamente normale.

Nei due boxplot della figura 29, **il valore di riferimento è la media**, la scatola riporta la distanza dell'**errore standard** e i **baffi indicano una distanza di una deviazione standard**.
Sono misure parametriche di dispersione rispettivamente della media e delle singole osservazioni, che saranno discusse dopo la presentazione della distribuzione normale e del test t di Student.
I **baffi** (***whisker***) riportano gli estremi che comprendono circa i 2/3 della distribuzione dei dati, mentre la **scatola** (***box***) fornisce gli estremi che comprendono i 2/3 delle medie che hanno identica variabilità.

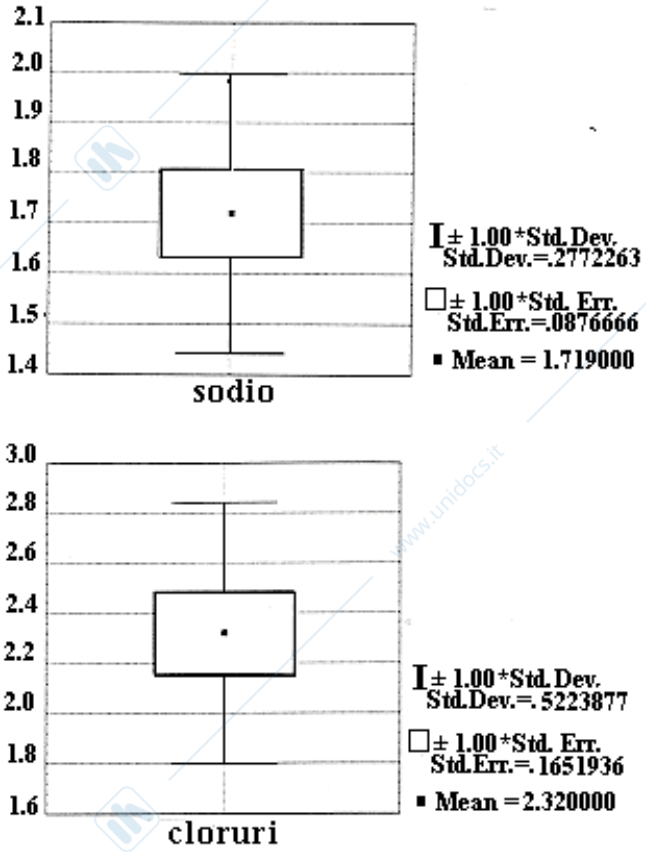


Figura 29. Box-and-Whisker con misure parametriche.

Una **forma più complessa**, specifica per **evidenziare gli outlier** in una distribuzione di dati, è riportata nel capitolo dedicato in parte alle metodologie specifiche per individuare statisticamente gli ***outliers***.

Nella pagina seguente, come applicazione di un esempio a partire dai dati rilevati, è riportato uno dei metodi più semplici di costruzione manuale di un generico Box-and-Wiskers.

Il diagramma Box-and-Wiskers può essere costruito indifferentemente in modo verticale, come i precedenti, oppure in modo orizzontale, come il successivo. Il primo metodo è più frequente, in quanto facilita il confronto tra due o più grafici

ESEMPIO 1. Con la seguente serie di 15 dati,

93	87	100	82	91	18	34	85	27	52	59	78	54	68	61

il ***Box-and Whiskers Plot*** risulta

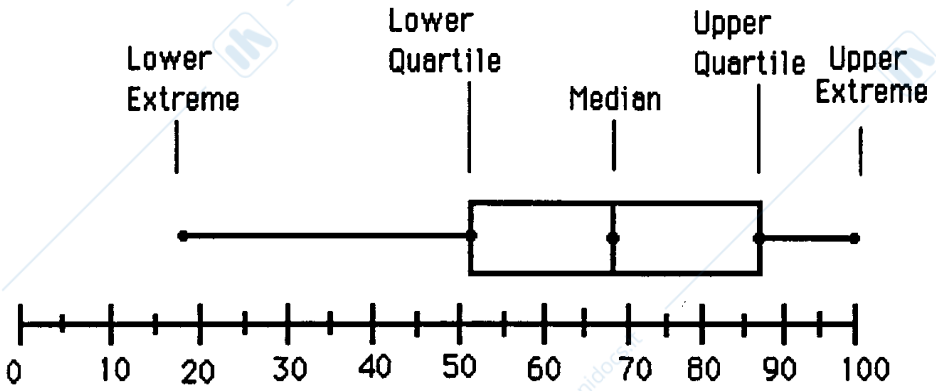


Figura 30. Box-and-Whishers Plot non parametrico

Per la sua costruzione, la serie di passaggi logici è:

1 - Dopo aver **ordinato i valori in modo crescente**, allo scopo di semplificare le operazioni richieste dal metodo

1	2	3	4	5	6	7	**8**	9	10	11	12	13	14	15
18	27	34	52	54	59	61	**68**	78	82	85	87	91	93	100

2 - si identifica **la mediana** che, su 15 dati, è esattamente l'ottavo valore:

Mediana = 68

3 – Considerando solo la **prima metà**, i sette valori minori della mediana, si individua la loro nuova mediana, che rappresenta il **primo quartile** (Q_1) o **quartile inferiore** (***lower quartile, lower fourth***); con 7 dati,

18	27	34	**52**	54	59	61

è esattamente il quarto valore:

Primo quartile = 52

4 – Considerando solo la **seconda metà**, i sette valori maggiori della mediana, si individua la loro nuova mediana, che rappresenta il **terzo quartile** (Q_3) o **quartile superiore** (***upper quartile***, ***upper fourth***): anche negli altri 7 dati

78	82	85	87	91	93	100

è esattamente il quarto valore

Terzo quartile = 87

La scelta della mediana e dei due quartili è stata semplice poiché i dati utilizzati sono sempre risultati dispari: **se fossero stati pari**, come nel caso seguente

78	82	85	87	91	93

la mediana relativa sarebbe caduta tra il terzo valore (85) e il quarto (87);
quindi identificata dalla loro media: 86.

5 – Dalla differenza tra il terzo e il primo quartile si ricava la **distanza interquartile** (***interquartile range*** o **IQR**)

distanza interquartile (IQR) = 87 – 52 = 35

6 – Infine di individuano i **due estremi**:

- il **valore minimo** o **estremo inferiore** (***lower estreme***) è 30;
- il **valore massimo** o **estremo superiore** (***upper extreme***) è 100.

C – LINE PLOT

Un secondo tipo di **rappresentazione semigrafica** è il **diagramma a linee o *line plot***, in italiano più frequentemente chiamato **diagramma a barre**.

Esso rappresenta il modo più facile e immediato per organizzare i dati.

La sua costruzione è molto semplice:

- la **linea orizzontale** rappresenta **i valori rilevati**, riportati in modo completo e ordinati dal minore al maggiore;
- ogni valore rilevato è individuato da una X, riportato in corrispondenza del valore rappresentato sull'asse orizzontale; il numero di X corrispondente a ogni punteggio indica quante volte un valore compare tra quelli rilevati.

ESEMPIO 2. Dai seguenti 30 valori

58	30	37	34	36	40	35	49	54	39	47	47	50	54	48
47	35	40	38	47	48	34	40	46	49	47	35	48	47	46

per costruire il ***line plot***

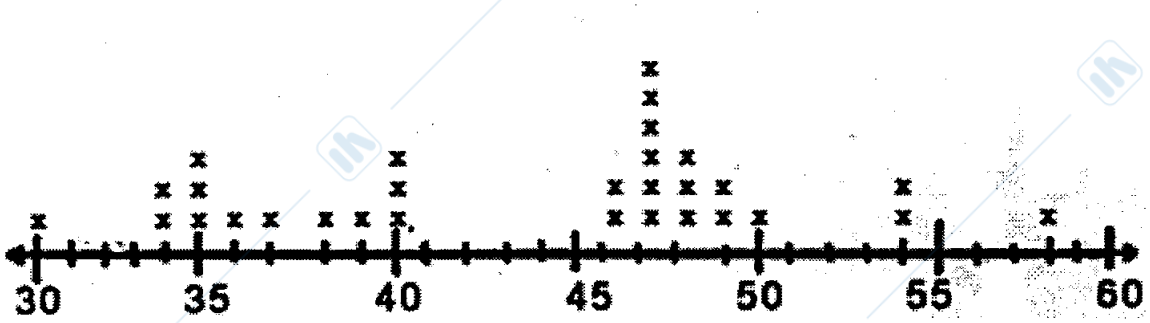


Figura 31. Line plot della tabella precedente

si richiede la serie di passaggi logici seguente:

1 – dopo aver ordinato i dati per rango

30	34	34	35	35	35	36	37	38	39	40	40	40	46	46
47	47	47	47	47	47	48	48	48	49	49	50	54	54	58

2 – si conta **quante volte compare ogni valore**, compreso tra il minimo e il massimo.
Minimo e massimo del grafico, allo scopo di descrive un intervallo completo, possono iniziare prima del valore più basso e terminare dopo il valore più alto tra quelli che sono stati effettivamente osservati.

D – STEM-AND-LEAF

Il **diagramma a ramo e foglia** (***stem-and-leaf plot***, ***stem-and-leaf display***) è un terzo tipo di tecnica semi-grafica, che può essere descritta come un incrocio tra un istogramma e una tabella di frequenza. E' chiamata, più brevemente, pure ***stem-plot*** o ***stemplot***.

Anche questa metodologia è stata presentata in modo organico da John W. **Tukey** nel suo testo del 1977 ***Exploratory Data Analysis*** (pubblicato da Addison-Wesley, Reading, Mass.). E' parimenti riportata in molti programmi informatici dell'ultima generazione. E' una diffusione tra gli utenti della statistica che ha alimentato anche la richiesta di conoscenza dei concetti sui quali questi metodi sono fondati.

Il metodo è utile per una prima descrizione di una distribuzione di dati. Inoltre, può essere di aiuto anche per valutare il livello di precisione con il quale i dati sono stati raccolti.

La figura seguente

```
14 : 2
14 : 555
14 : 67777
14 : 889
15 : 000000111111
15 : 22222222222333333333333333
15 : 44444444445555555555555555555
15 : 66666666666666666667777777777777777777
15 : 88888888888888888888888888888999999999999999999
16 : 00000000000000000001111111111111111111
16 : 22222222222222222233333333333333333333333333333
16 : 444444444444444455555555555555555
16 : 666666666667777777
16 : 888888999999999
17 : 00000000000111
17 : 333
17 : 4
17 : 67
17 : 88
```

Figura 32 - *A stem-and-leaf plot for the height of 351 elderly woman*

è un esempio tratto dal volume (pag. 363) di B. S. **Everitt** del 2002 ***The Cambridge Dictionary of Statistics*** (2nd edn. Cambridge University Press, UK, IX + 410 p.).

Il grafico ***stem-and-leaf*** può essere utilizzato sia per **variabili discrete** sia per **variabili continue**.
E' simile a un grafico a barre.

I principi di costruzione sono semplici:
- **ogni numero** è diviso in **due parti**: il **ramo** (***stem***) e **la foglia** (***leaf***):
- il **ramo è il numero**, collocato **a sinistra**, che include **tutte le cifre eccetto l'ultima**;
- la **foglia**, collocata **a destra**, è sempre **un numero con una cifra sola** (***single digit***), che può essere **esclusivamente l'ultima** di tutto il numero.

Anche questo grafico ha lo scopo di mostrare le **caratteristiche fondamentali di una distribuzione** di dati:
- **valore minimo** e **massimo** e quindi l'**intervallo di variazione**,
- **i valori più frequenti** o più comuni,
- la presenza di **uno o più picchi**,
- la **forma** della distribuzione, in relazione soprattutto alla **simmetria**,
- la **presenza di *outlier*** o valori anomali, quelli troppo distanti dal gruppo principale di valori.

Le modalità di organizzazione dei dati in grafici ***stem-and-leaf*** variano in funzione dei diversi modi con i quali i dati possono essere stati rilevati:
- con **una sola cifra** oppure con **più cifre**,
- numeri solo **interi** oppure con **uno o più decimali**.

Hanno il vantaggio pratico di potere essere costruiti, nel passato, con una macchina da scrivere. Attualmente, con un programma di video-scrittura, senza ricorrere all'uso di programmi statistici di elaborazione dei dati.
Gli esempi successivi illustrano casi differenti, per le caratteristiche sia dei valori rilevati sia della loro distribuzione di frequenza. L'applicazione dei principi prima enunciati richiede soluzioni tecniche diverse, in funzione delle differenti situazioni sperimentali.

ESEMPIO 3. Si assuma di aver effettuato 30 rilevazioni della concentrazione di CO (mg/mc) lungo una strada con traffico, riportati in modo ordinato e crescente nella tabella seguente:

6,7	8,5	8,7	8,8	9,1	9,1	9,3	10,2	10,2	10,3
10,5	10,5	10,9	11,2	11,4	11,5	11,7	11,7	11,7	12,8
13,2	13,3	13,5	14,0	14,1	14,2	20,0	20,5	21,5	22,0

Per costruire un **diagramma *stem-and leaf*** è utile seguire alcuni passaggi logici e metodologici.

1 - Dapprima nei valori rilevati si devono individuare le cifre che formano gli ***stem*** e i valori che formano le ***leaf***:

- i primi sono quelli che danno una misura approssimata del fenomeno, in questo caso, la parte intera del valore rilevato;
- i secondi sono quelli che rendono la stima più precisa, in questo caso i valori decimali, poiché ne è stato rilevato solamente uno.

2 – Successivamente, **i valori *stem* sono ordinati modo crescente lungo un'asse verticale**, riportando **anche le classi vuote**.

3 – Le cifre che formano le ***leaf*** sono riportate **in ordine crescente lungo l'asse orizzontale**, costruito **lateralmente ai valori stem**.

La disposizione dei numeri assume la forma della figura successiva, che ha l'aspetto grafico di una tabella:

Stem	Leaf
6	7
7	
8	5 7 8
9	1 1 3
10	2 2 3 5 5 7
11	2 4 5 7 7 7
12	8
13	2 3 5
14	0 1 2
15	
16	
17	
16	
18	
20	0 5
21	5
22	0

E' una specie di istogramma il cui l'asse delle ascisse è verticale e quello delle ordinate, nel quale sono riportate le frequenze, è orizzontale.

Rispetto ad esso, **spesso è caratterizzato da un numero di classi differente da quello richiesto per un istogramma corretto**. Nella costruzione di un ***stem-and-leaf*** abitualmente non si pone particolare attenzione a questo aspetto, che invece è di importanza rilevante nell'istogramma, che dovrebbe assumere forma normale.

In questa rappresentazione grafica,

- l'altezza di ogni classe è fornito dal numero di decimali riportati di fianco alla parte intera, che corrisponde al totale delle ***leaves*** rilevati per lo stesso ***stem***.

La lettura dettagliata della rappresentazione semigrafica riportata nella pagina precedente permette di ricavare varie informazioni, che è utile elencare in modo dettagliato:

1 - **l'intervallo di variazione** del fenomeno: da 6 a 22;

2 - **gli stem modali**: i valori 10 e 11;

3 - **la mediana**: tra 10,5 e 10,7 trattandosi di 30 dati (quindi un numero pari);

4 - **i quantili più utili** ad una descrizione dettagliata: l'80% dei valori è compreso tra 8,7 (il 10° percentile) e 20,5 (il 90° percentile);

5 - **la forma della distribuzione**: fortemente asimmetrica a destra (nei valori alti) e forse bimodale;

6 - **la presenza di outliers** (valori anomali rispetto alla distribuzione; se essi distano molto dagli altri ***stem***, non è necessario riportare tutti i valori ***stem*** intermedi);

7 - **la precisione con la quale i dati sono stati rilevati**: gli ultimi 4 valori sembrano arrotondati alla mezza unità (terminano infatti con 0 e 5), mentre i primi sembrano stimati con una precisione al decimale (sono infatti presenti tutti i valori da 1 a 9).

La costruzione di un diagramma ***stem-and-leaf*** deve essere adattata alle dimensioni del campione e alle caratteristiche dei dati raccolti.

Ad esempio, sempre nella misura della qualità dell'aria, i valori guida o livelli di attenzione sono da 100-150 mcg/mc come valore medio di 24 ore per SO_2. Le misure possono quindi essere approssimate all'unità; di conseguenza, gli ***stem*** possono essere indicati dalle decine e le ***leaf*** dalle unità.

Se i dati sono stati raccolti con troppa approssimazione, ad esempio i dati di CO rilevati con arrotondamento dell'unità, non è più possibile costruire un diagramma come quello presentato. Non sempre a posteriori è possibile costruire questo diagramma. Simmetricamente, se i dati sono raccolti con precisione eccessiva rispetto alla loro variabilità, ad esempio i valori di CO alla seconda cifra decimale, risulta necessario arrotondarli.

Gli esempi seguenti, sono tratti da indicazioni dei manuali di programmi informatici accessibili gratuitamente in internet.

ESEMPIO 4. Con la seguente serie di dati

23	12	6	19	7	10	15	12	25	21

costruire lo ***stem-and-leaf***:

Stem	Leaf
0	6 7
1	0 2 2 5 9
2	1 3 5

ESEMPIO 5. Con la seguente serie di dati

8	9	3	7	2	0	4	0	1	3	5	9	5	7	9

costruire lo ***stem-and-leaf***:

STEM	LEAF
0(0)	0 0 1 2 3 3 4
0(5)	5 5 7 7 8 9 9 9

Quando gli ***stem*** sono pochi, per ottenere una distribuzione di frequenza che riesca a descrivere almeno le caratteristiche più importanti della distribuzione di frequenza, è possibile e vantaggioso raddoppiarli, come è stato fatto in questo grafico.
Poiché **gli intervalli degli *stem* sono uguali**, le ***leaf*** hanno uguali possibilità di cadere nei due ***stem***.

ESEMPIO 6. Con la seguente serie di dati

27	25	31	29	18	32	28	28	26	27
24	10	26	20	22	39	27	28	31	22
21	24	19	28	27	24	25	29	26	28

si può costruire

A - sia lo ***stem-and-leaf***:

Stem	Leaf
1	0 8 9
2	0 1 2 2 4 4 4 5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
3	1 1 2 9

B - sia il successivo ***stem-and-leaf*** che, per le caratteristiche dei dati, appare preferibile:

Stem	Leaf
1(0)	0
1(5)	8 9
2(0)	0 1 2 2 4 4 4
2(5)	5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
3(0)	1 1 2
3(5)	9

Anche in questo caso gli ***stem*** sono pochi, per ottenere una distribuzione di frequenza sufficiente mete dettagliata. E' conveniente raddoppiarli come in questo grafico, in cui tutte le ***leaf*** hanno uguali possibilità di essere attribuite ai vari ***stem***.

ESEMPIO 7. Con la seguente serie di dati, che contengono decimali,

61,5	62,3	59,2	60,9	61,4	59,8	60,5	60,7	61,1	59,0
61,4	58,4	60,8	60,2	62,7	60,0	59,3	61,9	56,3	61,6
62,1	61,2	59,0	58,9	60,4	65,7	62,2	61,9	58,4	61,7

si ottiene lo ***stem-and-leaf***:

Stem	Leaf
56	3
57	
58	4 4 9
59	0 0 2 3 8
60	0 2 4 5 7 8 9
61	1 2 4 4 5 6 7 9 9
62	12 3 7
63	
64	
65	7

Anche in analisi più complesse, i tre metodi presentati permettono il confronto tra distribuzioni differenti.

Inoltre, in particolare quando i dati sono numerosi, essi sostituiscono la distribuzione dei valori campionari.

Per un **riassunto dei metodi illustrati** e un per **confronto tra le potenzialità** descrittive delle diverse caratteristiche delle distribuzioni dei dati, con il metodo

- dei ***data plots***, *dei* ***box-and-whisker plots*** e degli ***stem-and-leaf dispalys***

sono riportate le tre figure successive, tratte dal testo del 1991 di David Caster **Hoaglin** (nato nel 1944, della Harvard University), Frederick **Mosteller** (nato nel 1916, della Harvard Univerity) e John W. **Tukey** (nato nel 1915, della Princeton University) intitolato ***Fundamentals of Exploratory Analysis of Variance*** (A Wiley-Interscience Publication, John Wiley $ Sons, Inc. New York, XVII + 430 p.).

I dati sono stati pubblicati dalla rivista ***Consumer Reports***, April 1990, (pp.: 234-257).

I tre metodi sono anche utili indicazioni di come è possibile pubblicare i dati originali, in alternativa alla tabelle.

Nell'esempio riportato, si tratta dei galloni di carburante necessari per percorrere 15.000 miglia. Il consumo è stato stimato per 1990 auto, classificate in 7 categorie, sulla base di una estrapolazione fondata sui dati raccolti in test standard di guida.

Seppure con forme differenti, che evidenziano diversamente le caratteristiche della stessa distribuzione dei dati, la semplice lettura di ognuno dei tre grafici mostra:

1 - il **consumo medio** nettamente minore delle smalls cars e quello chiaramente maggiore delle ultime tre categorie;

2 - la **diversa variabilità** delle sette categorie di auto: la seconda categoria (sporty cars) e la settima (sport/utility) hanno una variabilità nettamente maggiore, in particolare se contrapposta a quella della categoria quattro (medium cars) e sesta (small vans).

3 - poiché le categorie sono state ordinate su una scala di rango della cilindrata, si evidenzia **una tendenza sistematico all'aumento dei valori medi e mediani,** passando dalle categorie inferiori (a sinistra) a quelle superiori (a destra).

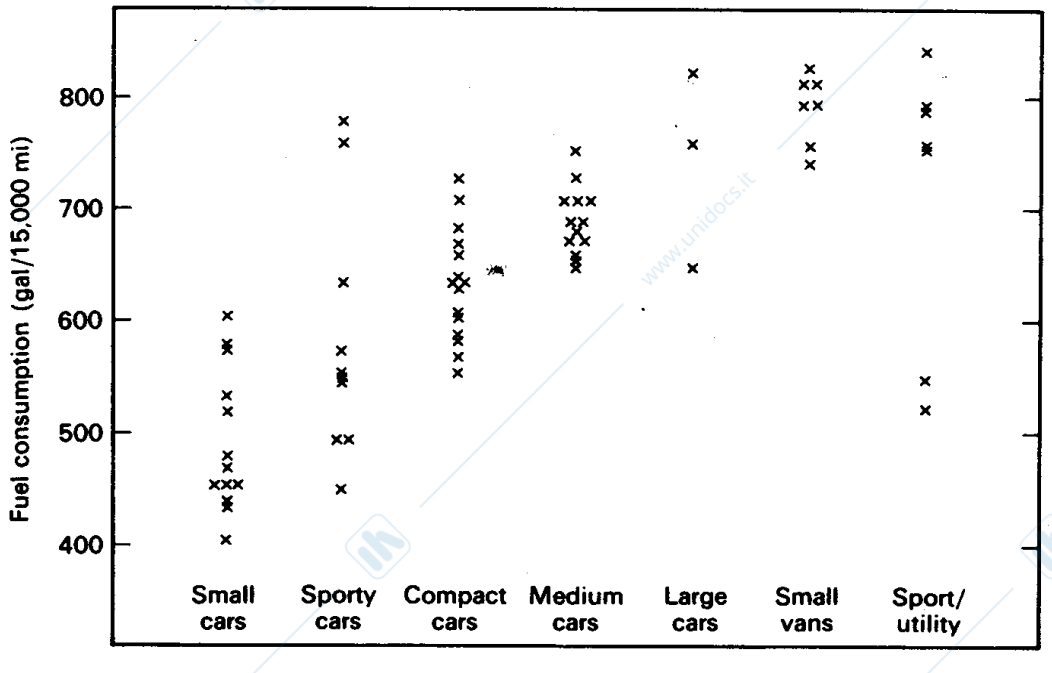


Figura 33. Data plot o dotplot

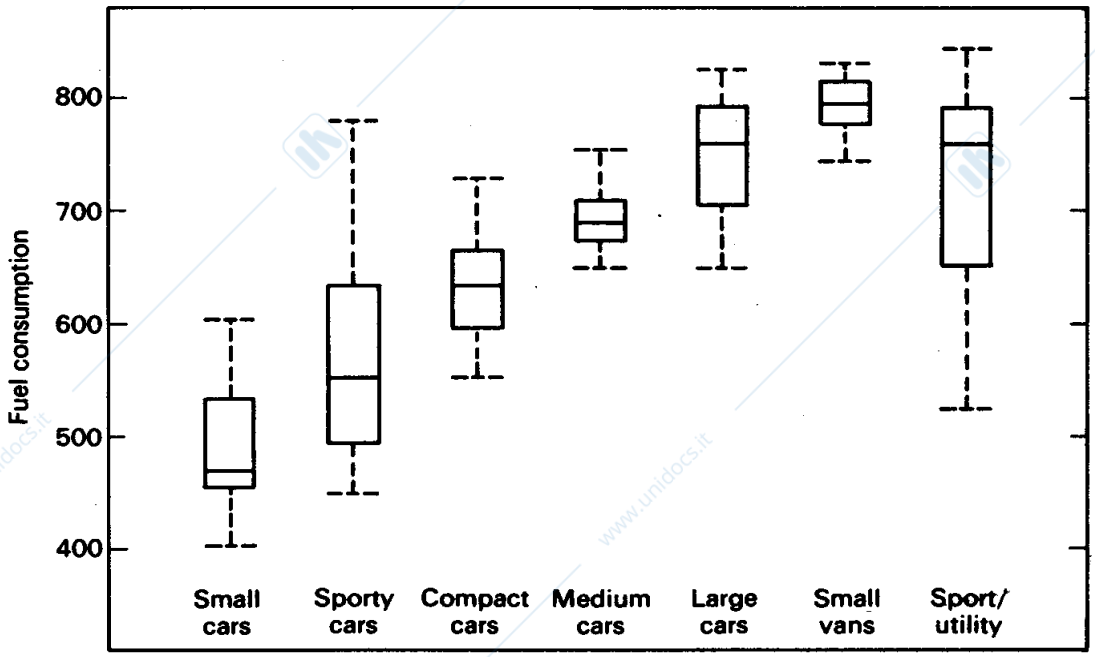


Figura 34. Box and whisker

Small Cars		Sporty Cars	Compact Cars	Medium Cars	Large Cars	Small Vans	Sport/ Utility
4*	0						
t	3						
f	4555	5					
s	7						
4·	8	99					
5*							
t	23						2
f		455	5				5
s	7	7	7				
5·	8		89				
6*	0		01				
t		3	333				
f			4	55	5		
s			67	677			
6·			8	899			
7*			1	111			
t			3	3			
f				5		4	5
s		6			6	6	6
7·		8				99	99
8*						11	
t					2	3	
f							4

Figura 35. Stem and leaf

La figura ***stem and leaf*** è una ulteriore variazione dei due modelli presentati: le dieci possibili foglie (***leaves***) sono classificate in cinque gruppi, indicati da simboli differenti:

- le foglie 0 e 1 con il simbolo * (asterisco),
- le foglie 2 e 3 con il simbolo t (lettera t),
- le foglie 4 e 5 con il simbolo f (lettera f),
- le foglie 6 e 7 con il simbolo s (lettera s),
- le foglie 8 e 9 con il simbolo • (punto).

A conclusione della presentazione di questi metodi di **rappresentazione grafica e semigrafica**, è utile riportare l'avvertenza dei tre autori citati sull'uso di programmi informatici, in particolare nella costruzione di ***box-plots***, ma estensibile a tutti i metodi:

- **spesso forniscono una impressione errata dei dati**;
- **per una impostazione corretta, è sempre necessario leggere attentamente la documentazione che accompagna il software**.

Segnalando l'articolo di M. **Frigge**, D. C. **Hoaglin** e B. **Iglewicz** del 1989 ***Some implementation of the boxplot*** (pubblicato su **The American Statistician** Vol. 43, 50-54), **Hoaglin**, **Mosteller** e **Tukey** nel loro articolo prima citato, dopo aver illustrato il **modello standard** di ***box and whiskers***, affermano (pag. 45): ***Some popular computer software produces boxplots that depart from the above standard in unexpected ways. To avoid getting a mistaken impression of the data, it may be necessary to check the documentation that accompanies the software***.

1.12. ESERCIZI SULLE MISURE DI TENDENZA CENTRALE, DISPERSIONE, SIMMETRIA E CURTOSI

ESEMPIO 1. In 36 laghi degli Appennini Settentrionali è stato prelevato un campione d'acqua e sono state misurate le concentrazioni di Sodio e di Cloruri, espresse in mg/l, (vedi tabella seguente):

Lago	Sodio	Cloruri
1	1,78	1,60
2	1,63	1,80
3	1,85	2,90
4	2,10	2,90
5	1,35	2,90
6	1,40	2,90
7	1,82	2,00
8	1,35	2,00
9	2,06	2,00
10	1,85	2,20
11	1,51	2,30
12	2,00	2,30
13	2,02	2,80
14	1,90	2,80
15	1,60	2,80
16	2,18	2,50
17	1,82	2,50
18	1,90	2,50
19	1,75	2,60
20	2,11	2,60
21	2,30	2,60
22	1,95	2,70
23	2,60	2,90
24	2,44	2,90
25	2,18	3,00
26	2,51	3,10
27	2,37	3,10
28	2,54	3,30
29	2,06	3,30
30	2,77	3,40
31	2,31	3,40
32	2,81	3,60
33	2,33	3,70
34	1,45	3,80
35	1,78	3,80
36	2,09	3,90

Calcolare le misure della tendenza centrale, della variabilità e degli indici di forma;
- rappresentare graficamente i dati in istogrammi.

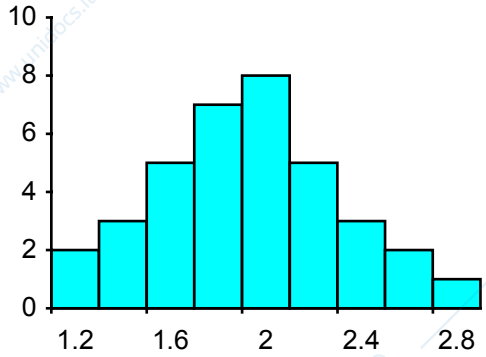
Risposta. Le statistiche calcolate dai programmi informatici comprendono varie misure di tendenza centrale, di dispersione, di simmetria e di curtosi. Quelle di seguito riportate presuppongono una distribuzione normale e sono fondate sulla media e sui momenti della distribuzione. Esistono programmi che utilizzano la mediana come misura della tendenza centrale e ricorrono ai quantili per descrivere la dispersione e la simmetria, come nel caso dei boxplot riportati in figure precedenti.
I programmi informatici forniscono una serie di valori, che descrivono compiutamente i dati campionari, come la tabella seguente (tra parentesi è riportato il termine inglese):

	Sodio	Cloruri
Numero di dati (Count, N. of data)	36	36
Somma (Sum)	72,87	101,4
Minimo (Minimum)	1,35	1,6
Massimo (Maximum)	2,81	3,9
Intervallo (Range)	1,46	2,3
Media (Mean)	2,024	2,817
Media geometrica (Geometric Mean)	1,987	2,756
Media armonica (Harmonic Mean)	1,949	2,692
Devianza (Sum of Squares)	5,29	11,76
Varianza (Variance, Mean Square)	0,151	0,336
Deviazione standard (Standard Deviation)	0,389	0,58
Errore standard (Standard Error)	0.065	0,097
Curtosi (Kurtosis)	-0,655	-0,53
Asimmetria (Skewness)	0,084	-0,015

Per valutare in modo più dettagliato e completo le caratteristiche delle 36 misure di sodio e cloruri presenti nei laghi campionati, è utile anche la loro rappresentazione in istogrammi. Quasi sempre sono forniti dai medesimi programmi informatici che calcolano anche gli indici già presentati.

Nei due istogrammi, i valori riportati sull'asse delle ascisse individuano la media della classe di riferimento. Nel primo grafico, sono riportati in modo alternato per evitare una eccessiva densità di numeri che renderebbe poco agevole la lettura. Sull'asse delle ordinate sono riportate le frequenze assolute.

Notare come i rapporti tra l'altezza e la lunghezza dell'istogramma rispondano ai criteri di eleganza grafica, già presentati.



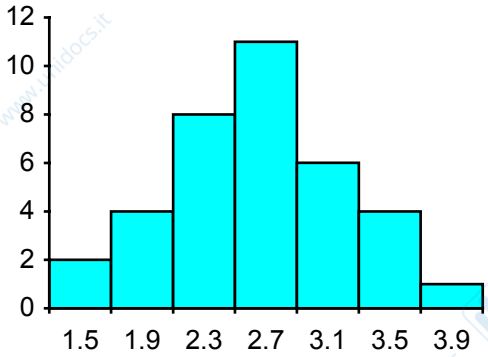


Figura 36

Istogramma delle concentrazioni del Sodio Istogramma delle concentrazioni dei Cloruri

Le due serie di valori hanno una distribuzione normale molto vicino alla normale, con curtosi negativa ed una leggerissima asimmetria, negativa per il sodio e positiva per i cloruri. Per analisi e confronti, possono essere applicati i test parametrici.

ESEMPIO 2. In 4 laghi degli Appennini settentrionali, sono state stimate le densità dei principali taxa fitoplanctonici riportati in tabella:

	Clorophyceae	Cryptophyceae	Crysophyceae	Diatomophyceae	Dinophyceae
Lago 1	179857	83497	30891	166861	25600
Lago 2	120893	29000	136791	27500	28000
Lago 3	198043	54454	82770	38712	54734
Lago 4	57496	42980	66440	34356	31270

Rappresentare i dati nelle forme grafiche di uso più comune.

Risposta. Sono distribuzioni di caratteri qualitativi. Le rappresentazioni grafiche adeguate sono i rettangoli distanziati e i diagramma a torta: può essere scelta una delle varie versioni, secondo la caratteristica da evidenziare.

La figura 30 riporta le 5 specie presenti nei 4 laghi con le due differenti versioni di rettangoli distanziati; potrebbero essere utilmente applicati anche gli ortogrammi in una delle due forme equivalenti.

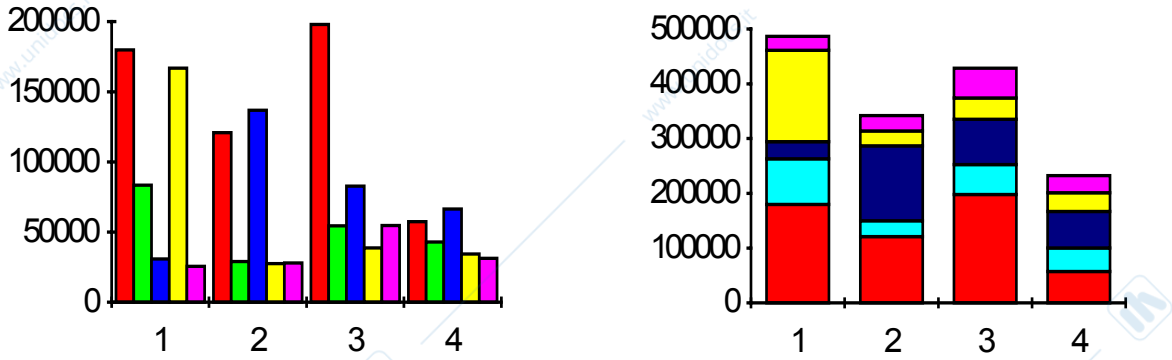


Figura 37. Istogrammi

Lago 1 Lago 2

Figura 38. Diagrammi a torta delle specie presenti nei laghi 1 e 2

I diagrammi a torta dimostrano di essere di più difficile lettura. Senza le percentuali, diventa difficile un confronto tra il numero di individui delle diverse specie presenti (ad esempio nel lago 1 è impossibile distinguere se è più alta la frequenza della specie 4 o della specie 1).\

Lago 3 Lago 4

Figura 39. Diagrammi a torta delle specie presenti nei laghi 3 e 4.

Il confronto visivo tra i 4 diagrammi a torta mostra quanto sia difficile cogliere le differenze tra le 4 distribuzioni, meglio evidenziate dai grafici precedenti, riportati nella figura 37.

ESEMPIO 3. Date due serie di dati relative ai campioni A e B

A:	5	7	2	4	3	
B:	15	11	9	8	10	12

calcolare per ognuna di esse
- le misure della tendenza centrale,
- le misure della dispersione,
- gli indici di forma.

Sono stati proposti 2 esercizi con pochi dati, per evitare che i calcoli manuali richiedano troppo tempo. Si chiede di
- calcolare la devianza secondo la formula euristica e quella abbreviata,
- calcolare la varianza sia con la formula euristica che con quella che considera tutti i possibili scarti

tra coppie di dati.

Risposta.

Sono riportati i risultati di un programma informatico:

Statistiche dei 2 campioni	A	B
Numero di dati (Count, N. of data)	5	6
Somma (Sum)	21	65
Minimo (Minimum)	2	8
Massimo (Maximum)	7	15
Intervallo (Range)	5	7
Media (Mean)	4,2	10,833
Media geometrica (Geometric Mean)	3,845	10,60
Media armonica (Harmonic Mean)	3,506	10,398
Devianza (Sum of Squares)	14,7	30,83
Varianza (Variance, Mean Square)	3,7	6,167
Deviazione standard (Standard Deviation)	1,924	2,483
Errore standard (Standard Error)	0,86	1,014
Curtosi (Kurtosis)	-1,005	-0,605
Asimmetria (Skewness)	0,084	0,636