Capitolo III

ELEMENTI DI STATISTICA DESCRITTIVA

1. Aspetti introduttivi

In statistica i dati sono ottenuti generalmente con operazioni ripetitive (osservazioni o misure) effettuate su certe variabili quantitative come il peso, la resa delle colture, il reddito, eccetera, o anche raccogliendo i risultati di osservazioni intorno ad un certo evento o carattere qualitativo (sesso, stagione, giorni della settimana, ecc.). In tutti questi casi, i dati ottenuti appaiono in una forma disorganizzata che può essere non adatta ad investigazioni scientifiche. Onde facilitare l'analisi e la interpretazione dei dati raccolti, è conveniente dividere l'intero gruppo di misure od osservazioni in *sottogruppi o classi* in modo tale che il numero di osservazioni che ricadono in ogni classe sia mostrato separatamente. Questi numeri sono chiamati "*frequenze*" relative alle classi corrispondenti e le misure od osservazioni o quantità variabili che formano i dati statistici sono chiamate "*modalità della variabile*", mentre con il termine "variabile" si indica il fenomeno allo studio. Tali raccolte di dati organizzate sono chiamate "*distribuzioni delle frequenze osservate*" e il loro scopo è quello di mostrare la natura della distribuzione di un fenomeno o carattere variabile lungo l'intervallo di valori o modalità con le quali il carattere si manifesta (ad esempio, il carattere sesso si manifesta con le due modalità maschio e femmina; il peso corporeo di un adulto si manifesta con le modalità 50 kg, 60 kg, ecc.).

Quando i dati sono sistemati in una forma organizzata, spesso è possibile costruire un modello matematico che può essere usato per studiare le proprietà di questo gruppo di osservazioni e trarre conclusioni concernenti il fenomeno investigato. Infatti, proprio utilizzando i modelli matematici adottati per descrivere il funzionamento del fenomeno studiato, attraverso le sue

interrelazioni con le variabili considerate influenti sul modo in cui esso si manifesta, lo statistico può fare delle predizioni sulle frequenze con cui ci si può attendere che si verifichino i diversi risultati possibili.

2. Tipi di dati e scale di misurazione

Esistono fondamentalmente due tipi di variabili statistiche, alle quali sono associati due tipi di dati: variabili *qualitative* e variabili *quantitative*:

- *i dati qualitativi o attributi sono generati da risposte categoriali* (es.: con prove sulla tossicità, le cavie muoiono o sopravvivono; con un farmaco, entro un tempo prefissato i pazienti guariscono o restano ammalati; con esperimenti sulle leggi dell'ereditarietà di Mendel, si hanno fiori rossi o fiori bianchi);

- *i dati quantitativi sono il risultato di risposte numeriche* (es.: per un'analisi del dimorfismo animale, si misurano le dimensioni di organi o il peso complessivo di alcuni maschi e di alcune femmine). *I dati quantitativi possono essere discreti o continui*: i primi derivano da un *conteggio* (es.: quante foglie sono attaccate ad un ramoscello) o da un *ordinamento* dei valori dal più piccolo al più grande o viceversa, a prescindere dalla intensità effettiva del fenomeno (es.: 1°, 2°, 3°,....,n°); i secondi da un processo di *misurazione* (es.: quanti centimetri è lungo il ramoscello).

Questa suddivisione, ormai storica nella presentazione ed elaborazione statistica dei dati, è stata resa più chiara e funzionale dalla classificazione delle scale di misurazione proposta da S.S. Stevens nel 1946 e divulgata soprattutto da S. Siegel, nel suo manuale di "*Statistica non parametrica*" del 1956.

Le misure possono essere raggruppate in 4 tipi di scale, che godono di proprietà formali differenti; di conseguenza, esse ammettono operazioni differenti. Come per le altre discipline, una scala di misurazione dei fenomeni ecologici ed ambientali può essere: 1) nominale o classificatoria; 2) ordinale o per ranghi; 3) ad intervalli; 4) di rapporti.

2.1 *La scala nominale o classificatoria*

E' il livello più basso di misurazione; viene utilizzata quando i risultati possono essere classificati o raggruppati in *categorie qualitative* o per *attributi*, eventualmente identificati con simboli.

In una popolazione si possono distinguere gli individui in maschi e femmine e contare quanti appartengono ai due gruppi; oppure possono essere suddivisi e contati secondo la loro specie, con una classificazione a più modalità.

Nella *scala nominale o qualitativa*, esiste una sola relazione, quella di *identità*: le unità statistiche attribuite a categorie diverse sono tra loro differenti, mentre tutte quelle della stessa categoria sono tra loro equivalenti, rispetto alla proprietà utilizzata nella classificazione. L'attribuzione di numeri per identificare le varie 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, calcolandone parametri caratteristici di sintesi. 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 e non si possiede una informazione maggiore.

L'operazione ammessa è il *conteggio* degli individui o dei dati presenti in ogni categoria. Tale conteggio può essere presentato come frequenza assoluta, cioè come numero assoluto di unità statistiche appartenenti ad una categoria, oppure in termini relativi, come frequenza relativa o percentuale, quando si vuole eliminare dal confronto tra due distribuzioni l'influenza determinata dal diverso numereo totale di unità statistiche da esse considerato. I quesiti statistici che possono essere posti correttamente riguardano, dunque, le frequenze, sia assolute che relative. Sono possibili confronti tra frequenze osservate oppure tra le frequenze osservate e le rispettive frequenze attese sulla base di leggi, ipotesi od altro.

Una classe è significativamente più numerosa dell'altra? Le varie classi hanno tutte lo stesso numero di individui, escludendo le variazioni casuali? I risultati ottenuti da un esperimento sulle leggi di Mendel sono in accordo con la sua distribuzione teorica?

2.2 *La scala ordinale o per ranghi*

Rappresenta una misurazione che contiene una quantità di informazione immediatamente superiore a quella nominale; alla proprietà precedente di *equivalenza* tra gli individui della stessa classe, *si aggiunge una gradazione tra le classi o tra individui* con misure diverse. Con la scala nominale, si ha la sola informazione che gli individui appartenenti a gruppi differenti sono tra loro diversi , ma non è possibile stabilire tra essi un ordine. Con la scala per ranghi, le differenti classi *possono essere ordinate sulla base dell'intensità del fenomeno* .

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; ma rimane ancora impossibile confrontare e misurare la quantità di differenza esistente tra loro.

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

Contengono una scala ordinale anche misure che sono rappresentate con simboli, come

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

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 (si veda il capitolo IV), i cui parametri essenziali sono la media e la varianza, poiché si fondano sulle differenze di ogni osservazione dalla media.

Tuttavia, questa indicazione di massima 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*. Tra gli studiosi, permane una ampia varietà di opinioni su quando il numero sia sufficientemente elevato od ancora troppo ridotto.

2.3 *La scala ad intervalli*

Alle due caratteristiche della scala ordinale aggiunge quella di *misurare le distanze tra tutte le coppie di valori*.

La scala ad 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, ed il tempo, misurato secondo calendari differenti. Le misure della temperatura, oltre a poter essere facilmente ordinate secondo l'intensità del fenomeno, godono della proprietà che le differenze tra loro sono direttamente confrontabili e quantificabili; le date in un calendario gregoriano, islamico, ebraico o cinese possono essere tra loro ordinate dalla più antica a quella più recente e le differenze temporali possono essere misurate con precisione oggettiva.

Ma la scala ad intervalli ha alcuni limiti: non gode di altre proprietà. Ad esempio, una temperatura di 80 gradi non è il doppio di una di 40 gradi, quando riferita alla temperatura corporea: se una persona ponesse la mano destra in una bacinella con acqua a 80 gradi e la mano sinistra in una con acqua a 10 gradi, non direbbe certamente che la prima scotta 8 volte più della seconda, ma solo che la prima è bollente e la seconda è fredda.

In una scala ad intervalli, *solo le differenze tra i valori sono quantità continue ed isomorfiche* 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 in misure alte e basse, sopra o sotto un valore prefissato). Pertanto, la scala d'intervalli gode anche delle proprietà definite dalle due scale precedenti.

2.4 *La scala di rapporti*

Ha il vantaggio di avere un'*origine reale*. Alle variabili misurate con una scala di rapporti, il tipo di misurazione più sofisticato e completo, può essere applicato qualsiasi test statistico. Come si vedrà in seguito, possono essere utilizzati anche la *media geometrica* ed il *coefficiente di variazione*, i quali richiedono che il punto 0 sia reale e non convenzionale.

Sono tipiche scale di rapporti l'altezza, la distanza, l'età, il peso, il reddito, 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.

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 rilevante e crescente, rispetto alle scale precedenti, della quantità d'informazione; 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.

Nella scala *nominale*, esistono solo relazioni di *equivalenza*; in quella *ordinale*, si aggiungono relazioni di *minore o maggiore di*; quella *ad intervalli*, oltre alle due precedenti possiede quella di *rapporto tra ogni coppia d'intervalli*; infine la scala *di rapporti* gode anche della relazione di *rapporto conosciuto tra ogni coppia di valori*.

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 in realtà richiedono elaborazioni diverse ed impongono il ricorso a verifiche 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 sono scale diverse. Nel primo caso, si ha una classica scala di rapporti; nel secondo, è possibile utilizzare le stesse procedure statistiche e gli stessi test 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 nella qualità e non hanno reale significato né i rapporti né le differenze tra loro.

3. Classificazione in tabelle

Un insieme di misure si chiama *serie statistica* o *serie dei dati*. La serie non ordinata delle rilevazioni od osservazioni risulta un insieme disordinato di numeri e non permette di evidenziare o cogliere rapidamente 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* permettono di individuare immediatamente il *campo od intervallo di variazione* del carattere in esame.

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.

Diversi metodi possono essere usati per classificare un certo insieme di dati in una tabella di frequenza, ma in generale si suppone che quando i dati sono raggruppati in classi di valori o modalità della variabile, tutte le misure in un dato intervallo che definisce la classe abbiano il valore rappresentato dal punto centrale dell'intervallo stesso. Questo valore è chiamato *modalità o indice della classe* ed è generalmente indicato con x_i. Il numero delle osservazioni che sono comprese nella i-esima classe o intervallo è indicato con f_i. Il numero totale delle osservazioni è spesso indicato con n oppure N.

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

Tabella III.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, semplice ed intuitivo in una distribuzione discreta, consiste nel definire le classi di valori del carattere. E' sufficiente identificare il valore minimo e quello massimo, contando quante volte compare ogni modalità di espressione.

Nella Tabella III.1 il valore minimo è 0 e quello massimo è 9. L'insieme delle informazioni ivi contenute può essere presentata in una tabella impostata come la seguente, contando per ogni classe quanti sono i rami con un numero di foglie uguali a quelle della classe.

Tabella III.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 questa tabella la frequenza assoluta di classe è il numero di volte con la quale compare ogni valore, essendo pari a 45 il totale delle frequenze assolute. La frequenza relativa di classe è la frequenza assoluta di classe divisa per il numero totale; il totale delle frequenze relative è pari all'unità, mentre se le frequenze relative sono cifre percentuali il loro totale è pari a 100. La frequenza cumulata di una classe è la somma di tutte le frequenze delle classi minori con quella della classe stessa. La distribuzione cumulata delle frequenze può essere stimata sia con quelle assolute sia con quelle relative

La trasformazione da frequenza assoluta a frequenza relativa risulta utile quando si vogliono confrontare 2 o più distribuzioni, con un differente numero complessivo di osservazioni, perché il confronto con le frequenze relative non dipende dal numero totale di unità statistiche considerato per ciascuna distribuzione.

La frequenza cumulata offre informazioni importanti quando si intende stimare il numero totale di osservazioni inferiori (o superiori) ad un valore prefissato.

Ad esempio: dall'ultima riga della tabella III.2 risulta che 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 o ad ogiva*, 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 III.2, 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 III.2 rispetto alla III.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.

Esempio III.1

Consideriamo un esempio di tavola di frequenza per classi di modalità. Supponiamo di avere un gruppo di pannocchie di mais che classifichiamo secondo la loro lunghezza, con il seguente risultato:

N° di ordine i	Limiti della classe (in centimetri)	Modalità o indice della classe : x_i	Frequenza o numero dei casi della classe: f_i	Frequenze cumulate $\sum_{k=1}^{k=i} f_k$
1	5 |--- 9	7	1	1
2	9 |--- 13	11	4	5
3	13 |--- 17	15	20	25
4	17 |--- 21	19	70	95
5	21 |--- 25	23	100	195
6	25 |--- 29	27	50	245
7	29 |--- 33	31	5	250
$\sum_{i=1}^{i=7} f_i$			250	

Nella tabella sopra riportata vediamo che l'intervallo di classe è considerato pari a 4 centimetri e l'intero gruppo delle 250 pannocchie è classificato in 7 classi. Per trovare il valore centrale di ogni classe consideriamo semplicemente la semisomma degli estremi di ciascuna classe. Ciascuna classe contiene tra i due estremi il segno |--- il quale sta ad indicare che l'attribuzione di ciascuna pannocchia alla rispettiva classe di appartenenza è avvenuta sulla base del confronto tra la sua lunghezza e gli estremi di ciascuna classe, facendo in modo di inserire la pannocchia in quella classe per la quale risultasse che la lunghezza misurata era maggiore o uguale all'estremo inferiore della classe e minore dell'estremo superiore della stessa classe.

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 dipendenza 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 manifesta la forma della distribuzione.

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

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

e 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

$$h = \frac{3,5.s}{\sqrt{N}}$$

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, che verranno 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 III.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 III.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 i valori osservati, ma li devono ovviamente comprendere. Sulla base del numero di dati, si decide il numero di classi.

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. Nel caso specifico, trattandosi di 40 dati, si potrebbero individuare 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. 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 *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 le classi iniziale e terminale non devono essere classi aperte (come "minore di 80" quella iniziale e "maggiore di 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; si perde 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 III.3 possono essere riportati in modo più schematico e più comprensibile, come nella seguente tabella III.4.

Tabella III.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. ass.	n_i	1	3	10	12	7	5	2
freq. rel. perc.	f_i	2,5	7,5	25,0	30,0	17,5	12,5	5,0
freq. perc. cumulate	$\sum_{k=1}^{k=i} f_k$	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 III.3, secondo le indicazioni di Sturges il numero di classi C avrebbe dovuto essere

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

dalla quale si deduce anche un'ampiezza di $\mathbf{\frac{140}{6,34} \cong 22}$.

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

$$\mathbf{h = \frac{3,5.s}{\sqrt{N}} = \frac{3,5.28,618}{6,3246} = 15,837}$$

dalla quale si deduce il numero di classi con $\mathbf{\frac{140}{15,837} = 8,84}$.

Ovviamente, il numero di classi calcolato (8,84) deve essere arrotondato all'unità successiva (9). Secondo i due metodi proposti, con i dati della tabella III.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, asimmetria, appiattimento*), 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, tale procedura comporta una approssimazione, poiché non è vera l'ipotesi operativa implicita che entro ogni classe i dati siano distribuiti in modo uniforme.

4. 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*, *asimmetria* e *appiattimento*). Insieme con i vantaggi di fornire una visione sintetica e di essere di facile lettura, hanno l'inconveniente fondamentale di mancare di precisione e soprattutto di essere soggettive, di permettere letture diverse degli stessi dati. Pertanto, ai fini di una elaborazione mediante i test statistici e di un confronto dettagliato dei parametri è preferibile la tabella, che riporta i dati esatti.

Le rappresentazioni grafiche proposte sono numerose e debbono essere scelte in rapporto al tipo di dati e 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 di frequenza*. Gli istogrammi sono grafici a barre verticali (per questo detti anche

diagrammi a rettangoli accostati). 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. I lati dei rettangoli sono costruiti in corrispondenza degli estremi di ciascuna classe.

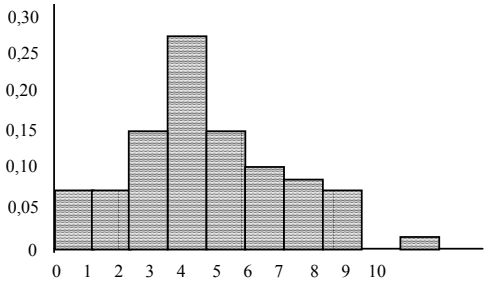


Figura III.1 . Istogramma dei dati di Tab. III.2 (frequenze relative)
Passo = 20; Classi = 7)

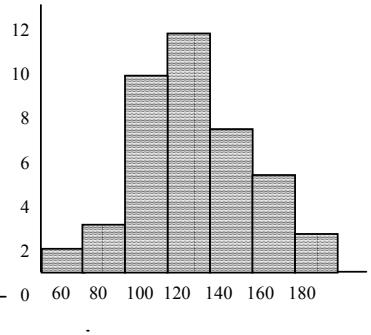


Figura III.2 . Istogramma dei dati di Tab. III.4 (Valore iniz.=60; Valore finale =199;

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. 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.

Le due figure seguenti riportano la somma di due classi di tabella III.4: tale somma è rappresentata nel primo caso con un grafico errato e nel secondo caso nella sua versione corretta, che richiede il valore medio delle classi raggruppate.

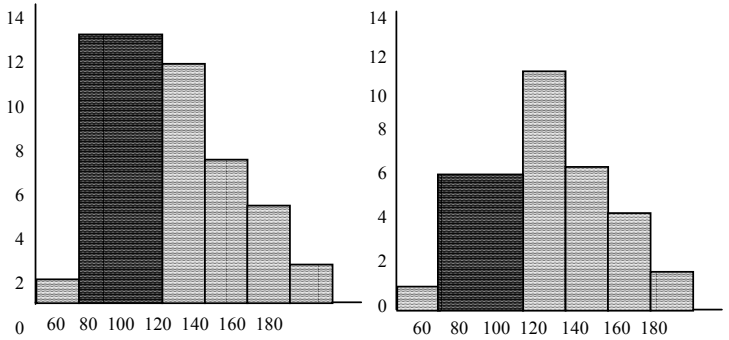


Figura III/3. Istogrammi dei dati di Tab. III.4
Somma della 2ª e 3ª classe della figura 2
Errata Corretta

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. Tuttavia, 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 prassi che tutto il disegno dell'istogramma possa 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.

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, ecologici 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. 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; una forte alterazione, che scompare con suddivisioni in classi meno frammentate, è una indicazione di un possibile errore tecnico di rappresentazione dei dati.

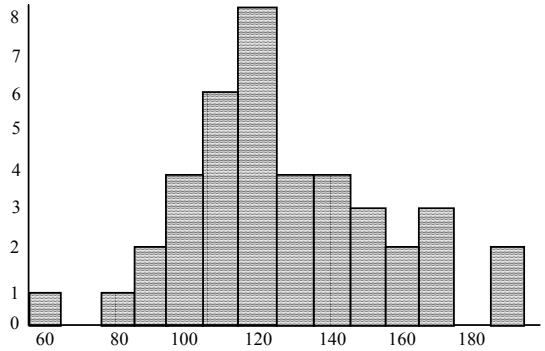


Figura III.4. Istogramma dei dati di Tab. III.4
(Valore iniz. = 60; Valore finale = 199; Passo = 10; Classi = 14)
(Rappresentazione grafica non appropriata, per eccessiva suddivisione in classi.)

I *poligoni di frequenza* sono figure simili agli istogrammi e sono utilizzati di norma per la rappresentazione di *valori relativi* o di *percentuali*. 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; l'area sottesa deve dare un totale

uguale a 1 con le frequenze relative ed uguale a 100 quando si riportano le percentuali. La linea spezzata deve essere unita all'asse orizzontale sia all'inizio che alla fine, per racchiudere l'area della distribuzione. E' un procedimnto che viene ottenuto con un artificio, utilizzando un istogramma come punto di partenza. Si unisce il valore centrale della prima classe con il valore centrale di una precedente classe fittizia di valore 0; tutti gli altri segmenti sono ottenuti unendo i punti centrali di ogni classe, fino all'ultima; 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 primo poligono di seguito riportato è stato costruito sulla base dell'istogramma della figura III.2, con i dati della tabella III.4, spostando le classi sull'asse delle ascisse per comprendere i nuovi estremi della distribuzione.

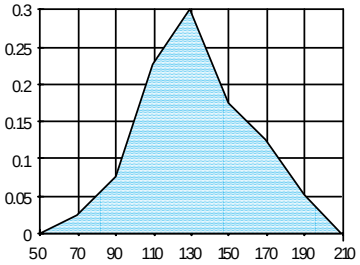


Figura III.5. Poligono dei dati di Tab. III.4

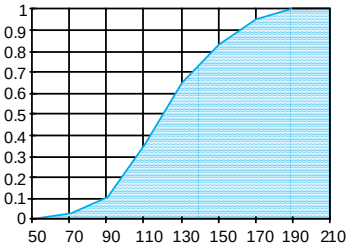


Figura III.6. Poligono cumulato di Tab. III.4

Le distribuzioni cumulate sono rappresentate sia con *istogrammi cumulati* che con *poligoni cumulati.* Non forniscono informazioni 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.

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. Come si vedrà in seguito, il valore dell'asse orizzontale che corrisponde al 50% dei valori identifica la *mediana* (riportato come linea tratteggiata nella figura precedente 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).

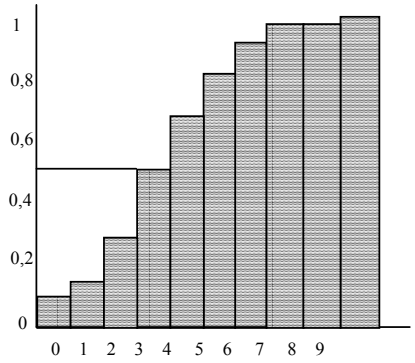


Figura III.7. Istogramma cumulato dei dati di Tab. III.2

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 sono tra loro *non contigui,* distaccati; 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.

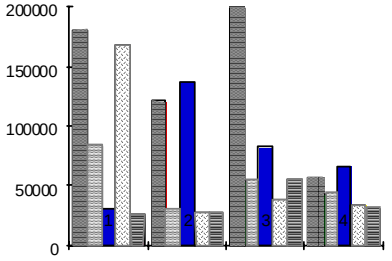


Figura III.8. Rettangoli distanziati

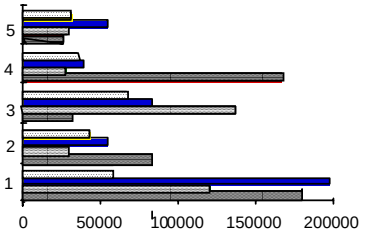


Figura III.9. Ortogramma

Nel caso di dati qualitativi o nominali, non esiste una logica specifica nell'ordine delle classi. Sovente, ma non obbligatoriamente, per convenzione, i rettangoli o le colonne 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. Avendo basi uguali, le aree sono proporzionali alle altezze; pertanto, anche i diagrammi a rettangoli distanziati sono rappresentazioni areali.

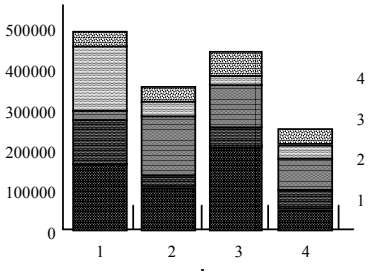


Figura III.10. Rettangoli distanziati

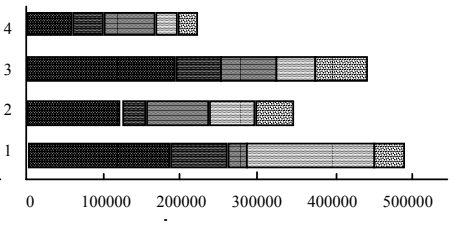


Figura III.11. Ortogramma

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. 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. Secondo alcuni esperti di percezione dei grafici, queste figure vengono lette con maggiore facilità rispetto ai rettangoli distanziati e meglio rappresentano le informazioni contenute in distribuzioni di frequenza di dati qualitativi.

Gli *areogrammi* sono grafici in cui le frequenze o le quantità riportate in una distribuzione 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; 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 siano rilevanti. Per differenze piccole, si dimostrano meno efficaci degli ortogrammi.

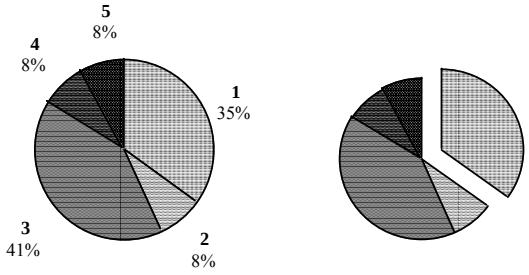


Figura III.12. Diagrammi circolari

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* devono essere proporzionali alle frequenze percentuali (x%) che vogliono rappresentare, in accordo con la relazione *a : 360 = x% : 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, o 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.

Questi diagrammi a figure hanno tuttavia il *grave inconveniente di prestarsi a trarre in inganno con facilità* il lettore inesperto di statistica.

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 alla seconda. L'occhio coglie complessivamente non l'altezza di ogni figura ma la sua superficie, che è il quadrato del valore: 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 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.

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, dove è necessario porre in evidenza la differenza, dove è più importante l'impressione che non la precisione ed una rappresentazione corretta della realtà. Per i ricercatori e gli specialisti, è invece fondamentale che la rappresentazione dei dati sia oggettiva.

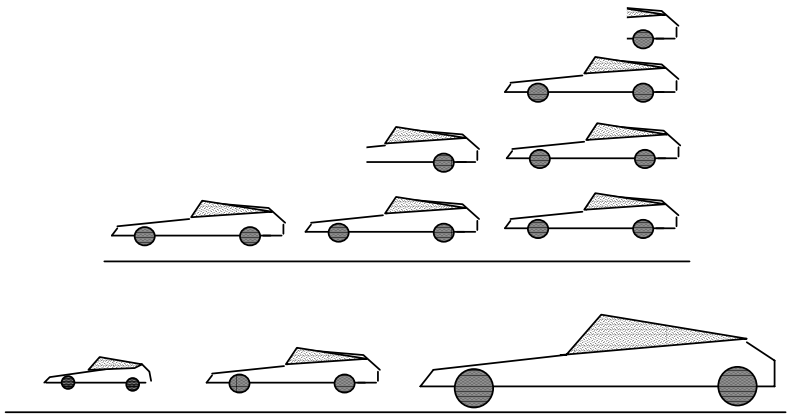


Figura III.13. Pittogramma della produzione mensile di auto di 3 case automobilistiche: la prima ha prodotto 100 mila auto, la seconda 180 mila e la terza 320 mila.

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 della figura che deve essere proporzionale, non la lunghezza.

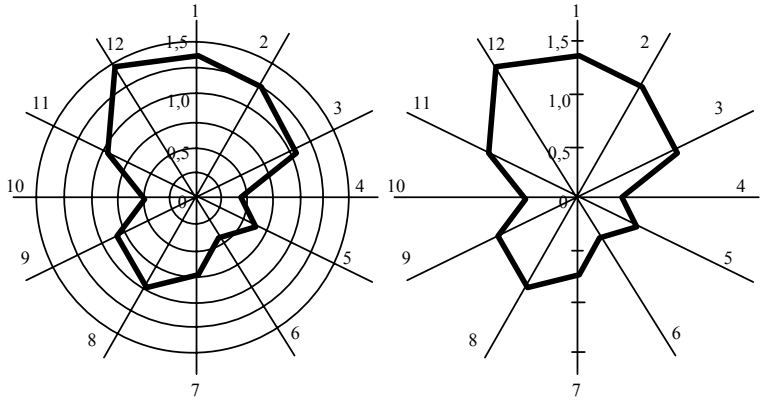
Molte discipline ricorrono a rappresentazioni grafiche specifiche, che possono essere utili per *statistiche territoriali*. Per rappresentare il numero di soggetti presenti in vari località, in geografia si ricorre al *cartogramma*: evidenzia distribuzioni territoriali mediante carte geografiche, in 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 sono sovente ridotti a schemi, illustrati nelle didascalie, per cui un solo cerchio bianco indica una quantità di base, due cerchi concentrici indicano una quantità maggiore e così via; altrimenti, sono essenzialmente areogrammi e possono trarre in inganno, poiché l'area aumenta con il quadrato del 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. 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 collocare i 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). Il diagramma polare è ottenuto congiungendo i vari punti e l'intensità del fenomeno è data dalla distanza dal centro. La figura III.14 riporta due differenti impostazioni grafiche per costruire un diagramma polare sui valori medi mensili in Italia della radioattività beta totale nell'anno 1993.

Figura III.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 la rappresentazione di dati numerici, è possibile ricorrere anche a *diagrammi cartesiani*. Essi saranno illustrati nei succesivi capitoli per la presentazione grafica di dati bivariati, utilizzati quando per ogni unità statistica sono rilevati contemporaneamente le modalità di 2 variabili (ad esempio, il peso e l'altezza di un certo numero di individui). Ma possono essere usati anche per una sola variabile, collocando punti sul 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 più efficace rappresentazione, in particolare a persone non esperte dell'argomento trattato.

5. Elementi caratteristici descrittivi di una distribuzione statistica

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 ricche di particolari.

Per caratteri qualitativi, la tabella e le rappresentazioni grafiche esauriscono quasi completamente gli aspetti descrittivi, quando sia possibile leggere con precisione le frequenze delle varie classi. *Per i caratteri quantitativi, si pone il problema di sintesi oggettive che siano numeriche*. I grafici forniscono una descrizione che può essere espressa mediante una interpretazione

soggettiva (due ricercatori possono spiegare e valutare in modo differente lo stesso grafico).

Invece, come esigenza primaria della ricerca, *serve un'analisi obiettiva dei dati osservati, capace di condurre tutti i ricercatori, con gli stessi dati, alle medesime conclusioni*. Tale sintesi oggettiva può essere ottenuta mediante la stima di alcuni *parametri* o *statistiche* della distribuzione. Si può asserire, infatti, che una serie di dati numerici è compiutamente descritta da tre proprietà principali che sono:

1) la tendenza centrale o posizione;
2) la dispersione o variabilità;
3) la forma della distribuzione.

Nei vari campi del sapere i ricercatori molto raramente conoscono tutta la popolazione o universo dei dati di studio; di conseguenza, i metodi statistici di norma utilizzati sono riferiti quasi esclusivamente alla descrizione, all'analisi e al confronto di un numero limitato di unità statistiche o campione di osservazioni piuttosto che all'intera popolazione di dati. Ma in entrambi i casi le proprietà sopra richiamate costituiscono i primi elementi che occorre conoscere per avere almeno una nozione semplice del fenomeno investigato. Per fare una distinzione nei due casi, queste misure descrittive sintetiche, riassuntive dei dati tabellari, sono chiamate *statistiche*, quando sono calcolate su un campione di dati; sono chiamate *parametri*, quando descrivono la intera popolazione od universo delle unità statistiche.

5.1 *Le misure di tendenza centrale e di posizione*

Le medie di una distribuzione sono valori attraverso i quali si cerca di *sintetizzare* la variabile descritta dall'intera distribuzione di frequenza ed attorno ai quali tendono a distribuirsi tutti gli altri valori della variabile compresi nell'intervallo o campo di definizione del fenomeno osservato. Tuttavia, malgrado nella maggior parte delle distribuzioni le variabili mostrino una certa concentrazione delle frequenze attorno ad un valore centrale, talvolta esse possono assumere modalità talmente diverse da rendere impossibile la selezione di un singolo valore centrale per rappresentare un punto di concentrazione delle frequenze.

Quindi, *affinché un valore medio possa rappresentare una intera distribuzione di frequenza è necessario che esista una tendenza dei dati a concentrarsi attorno ad un singolo valore*.

Tra tutte le medie che è possibile determinare alcune di esse sono suscettibili di essere espresse come funzioni di tutti i termini della serie a cui si riferiscono e formano il gruppo delle *medie ferme o analitiche*.

Per altre, invece, il valore assunto non dipende da tutti i termini della serie a cui si riferiscono, ma dalla posizione che occupano nella serie: queste sono le cosidette *medie lasche o di posizione* .

5.1.1 Medie ferme o analitiche

Tra le medie ferme o analitiche, le più importanti sono la *media aritmetica*, la *media geometrica*, la *media armonica*, il *valore divisorio*.

A) La media aritmetica (M_1 oppure M oppure M_a): La *media aritmetica* può essere definita come *quel valore di un carattere quantitativo che, sostituito ad ognuna delle modalità osservate del carattere e sommato per tutti i casi lascia invariato l'ammontare totale del carattere.;* cioè, indicando con M_1 la media aritmetica, deve essere:

$$\sum_{i=1}^{i=n} x_i = x_1 + x_2 + x_3 + \ldots + x_n = M_1 + M_1 + M_1 + \ldots + M_1 = n.M_1$$

Da ciò risulta anche che *la media aritmetica di una variabile statistica X è definita come la somma dei valori assunti dalla variabile in tutti i casi rilevati, divisa per il numero totale degli stessi casi*.

Se ogni valore della variabile viene rilevato in un solo caso, si avrà la *media aritmetica semplice* della distribuzione espressa dalla relazione:

$$M_1 = \frac{\sum_{i=1}^{i=N} x_i}{N}, \quad \text{con } i = 1, 2, ..., N.$$

Se, invece, ogni determinazione x_i della variabile viene rilevato più di una volta e precisamente con la frequenza f_i , allora si avrà la *media aritmetica ponderata* della distribuzione di frequenza, la quale è definita dall'espressione:

$$M_1 = \frac{\sum_{i=1}^{i=n} x_i f_i}{\sum_{i=1}^{i=n} f_i}, \text{con } \sum_{i=1}^{i=n} f_i = N.$$

La *media aritmetica* gode di alcune proprietà molto importanti:

1) La *somma algebrica* degli scostamenti dalla media aritmetica di una distribuzione è *uguale a zero*, cioè la somma degli scostamenti positivi è uguale, in valore assoluto, alla somma degli scostamenti negativi.

2) La *somma dei quadrati* degli scostamenti dalla media aritmetica di una distribuzione *è un minimo* rispetto alla somma dei quadrati degli scostamenti da qualsiasi altro valore diverso dalla media aritmetica.

3) La media aritmetica di una successione a_1, a_2, ..., a_n, con pesi f_1, f_2, ..., f_n è l'ascissa del baricentro di un sistema di forze di gravità di intensità uguale alle f_i, applicate nei punti di ascissa a_i di un asse orizzontale.

4) Poiché, inoltre, sospendendo per il baricentro un asse orizzontale su cui sono applicate delle forze di gravità questo rimane in posizione orizzontale, si può dire che la media aritmetica gode anche della proprietà di costituire il centro di sospensione di un asse orizzontale su cui sono applicate, in punti di ascissa a_i le forze di gravità di intensità f_i.

L'uso della media aritmetica è particolarmente utile per il fatto che effettua la correzione degli errori accidentali di osservazione, applicandosi di preferenza a grandezze additive, per cui *è la stima più precisa di misure ripetute*. Essa è, inoltre, la più semplice delle medie algebriche.

Il termine "*valore atteso*" o "*valore sperato*" o "*speranza matematica*" che talvolta viene usato al posto di "*media aritmetica*" è giustificato dal ragionamento che segue.

Siano p_1, p_2, p_3, ..., p_n, le probabilità che si verifichino le variabili mutualmente escludentisi x_1, x_2, x_3, ..., x_n, definite, come si è già visto, dalle relazioni:

$$p_1 = \lim_{N\to+\infty} \frac{f_1}{N}; \; p_2 = \lim_{N\to+\infty} \frac{f_2}{N}; .....; p_n = \lim_{N\to+\infty} \frac{f_n}{N}.$$

Per valori crescenti di n avremo allora che:

$$M_1 = \lim_{n\to+\infty} \left[\sum_{i=1}^{i=n} x_i . \frac{f_i}{N} \right] = \sum_{i=1}^{i=n} x_i . p_i .$$

Ma abbiamo già visto che la espressione a destra dell'ultimo segno di uguaglianza è la "*speranza matematica*" o "*valore sperato*" della variabile x. *La media aritmetica è la misura di tendenza centrale generalmente utilizzata nella statistica parametrica.*

Esempio III.2

Si abbia la distribuzione di una data popolazione di individui secondo la statura, quale quella descritta nella tavola che segue e si faccia l'ipotesi che le intensità di ogni intervallo delle stature siano concentrate nel valore centrale di ogni classe; per le classi estreme aperte, si ipotizza che il valore centrale sia rispettivamente di 145 e 185 cm. I calcoli da eseguire per determinare la media aritmetica ponderata dell'intera distribuzione sono mostrati nella tavola stessa.

Classi di statura (cm)	Valori centrali delle classi (x_i)	Frequenze (n_i)	Prodotti ($n_i x_i$)
meno di 150	145	1.500	217.500
150|--160	155	41.200	6.386.000
160|--170	165	188.300	31.069.500
170|--180	175	125.900	22.032.500
180 e oltre	185	14.500	2.682.500
		371.400	62.388.000

Pertanto la media aritmetica ponderata sarà:

$$\bar{x} = \frac{62.388.000}{371.400} = 167,98.$$

B) La media geometrica (M_g o M_0). La *media geometrica* può essere definita come *quel valore di una variabile che, elevato ad una potenza pari al numero delle osservazioni, fornisce un risultato pari al prodotto dei singoli termini osservati, ciascuno elevato a potenza pari alla frequenza con cui il termine figura tra le osservazioni*, cioè, indicando con M_g o M_0 la media geometrica, deve essere:

$$M_0^n = x_1 . x_2 . x_3 ..... x_i ..... x_n = \prod_{i=1}^{i=n} x_i$$

Da ciò risulta che *la media geometrica semplice di una variabile statistica descritta da N termini è definita come la radice di ordine N del prodotto degli N termini*. In formule si avrà:

$$M_0 = \sqrt[N]{\prod_{i=1}^{i=N} x_i}\ , \qquad \text{dove i } = 1, 2, ..., N\ .$$

Se ciascuno degli n valori x_i della variabile si presenta con frequenza f_i tale che la somma di tutte le frequenze sia pari ad N, *la media geometrica ponderata della variabile statistica è definita come la radice di ordine N del prodotto degli n valori x_i ciascuno elevato alla potenza f_i.* In formule si avrà:

$$M_0 = \sqrt[N]{\prod_{i=1}^{i=n} x_i^{f_i}}\ , \qquad \text{dove} \quad \sum_{i=1}^{i=n} f_i = N.$$

La media geometrica è utilizzata quando le variabili non sono rappresentate da valori lineari, ma da valori ottenuti come prodotto o rapporto tra valori lineari. Essa serve per il confronto di superfici o volumi, oppure di tassi di variazione percentuale, cioè valori che sono appunto espressi da rapporti. *Per il calcolo della media geometrica è condizione necessaria che le quantità siano tutte positive*. Se fossero negative si dovrebbe far ricorso al loro valore assoluto.

La *media geometrica* gode delle proprietà seguenti:

1) *Il reciproco della media geometrica è uguale alla media geometrica dei reciproci dei termini*.

2) *La potenza emmesima della media geometrica è uguale alla media geometrica delle potenze emmesime dei termini*.

3) *Il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi dei termini*.Cioè:

$$\log M_0 = \frac{1}{N}\left(f_1 \log x_1 + f_2 \log x_2 + ... + f_n \log x_n\right) = \frac{1}{N}\sum_{i=1}^{i=n} f_i \log x_i$$

Esempio III.3

Supponiamo di impiegare 1 lira ad interesse composto ai seguenti tassi : i_1=0,05 nel primo anno; i_2=0,06 nel secondo anno; i_3=0,055 nel terzo anno; i_4=0,07 nel quarto anno; i_5=0,065 nel quinto anno. Il montante alla fine del primo anno sarà dato dalla relazione C_1=1+0,05; alla fine del secondo anno sarà C_2=(1+0,05)(1+0,06); alla fine del terzo anno sarà C_3=(1+0,05)(1+0,06)(1+0,055) e così via.

Ci si chiede qual'è il *tasso medio i* a cui capitalizzare la nostra lira per ottenere alla fine del quinquennio il montante C_5 che rappresenta, evidentemente, l'invariante del problema.

Da quanto detto, si deduce che deve essere:

$$(1+i)^5 = 1{,}05 \text{ x } 1{,}06 \text{ x } 1{,}055 \text{ x } 1{,}07 \text{ x } 1{,}065$$

da cui si vede che (1+i) è la media geometrica dei prodotti indicati nel secondo membro e non dei singoli tassi annui, per cui, mediante i logaritmi si calcola:

$$\log(1+i) = \frac{1}{5}\left(\log 1{,}05 + \log 1{,}06 + \log 1{,}055 + \log 1{,}07 + \log 1{,}065\right) = 0{,}025296$$

Risalendo al numero, si ha che i = 0,0599, pari a 5,99%.

C) La media armonica (M_{-1}). La *media armonica* può essere definita come *quel valore di una variabile il cui reciproco, sostituito al reciproco di ciascuno dei termini osservati, flascia invariata la somma dei reciproci dei singoli termini osservati ciascuno moltiplicato per la frequenza con cui il termine figura tra le osservazioni*, cioè, indicando con M_{-1} la media armonica, deve essere:

$$n.\frac{1}{M_{-1}} = \frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + ... + \frac{1}{x_i} + ... + \frac{1}{x_n} = \sum_{i=1}^{i=n} \frac{1}{x_i}$$

Da ciò risulta che *il reciproco della media armonica è pari alla media aritmetica dei reciproci dei termini*. E quindi è anche vero che *la media armonica è pari al reciproco della media aritmetica dei reciproci dei termini*.. Se ogni termine si presenta con frequenza unitaria, si avrà la *media armonica semplice*, definita dalla relazione:

$$M_{-1} = \frac{N}{\sum_{i=1}^{i=N} \frac{1}{x_i}}$$

Se, invece, ogni termine della variabile si presenta con frequenza pari ad f_i, si avrà la *media armonica ponderata*, espressa dalla relazione:

$$M_{-1} = \frac{\sum_{i=1}^{i=n} f_i}{\sum_{i=1}^{i=n} \frac{f_i}{x_i}}.$$

La media armonica è la stima più corretta della tendenza centrale per distribuzioni di dati in cui devono essere usati gli inversi o reciproci, come nel caso di misure dei tempi di reazione.

Esempio III.4

Si voglia conoscere il consumo medio annuo di rasoi usa-e-getta in Italia, mediante una ricerca diretta sui consumatori.

Non sarà opportuno chiedere: "Quanti rasoi consuma in media all'anno?" perché la domanda così formulata richiede una stima relativa ad un ampio intervallo di tempo; si potrà invece chiedere: "Quanti giorni le dura in media un rasoio?". Immaginiamo di esaminare le risposte di cinque persone:

1ª persona	10 giorni in media
2ª persona	6 giorni in media
3ª persona	30 giorni in media
4ª persona	5 giorni in media
5ª persona	14 giorni in media
Totale	65

La media aritmetica delle durate è 65:5=13 giorni. Ma da questo dato non è corretto ricavare il consumo medio annuo pari a 365:13=28,1 rasoi in media per persona, equivalente per i 5 consumatori considerati a 28,1x5=140,5 rasoi di consumo annuo. Infatti con i dati di partenza possiamo ricavare direttamente il consumo globale

Persone	Consumo annuo di rasoi
1ª	365:10=36,5
2ª	365: 6=60,8
3ª	365:30=12,2
4ª	365: 5=73,0
5ª	365:14=26,1
In complesso	208,6 rasoi

mentre in precedenza si era ottenuto il risultato di 140,5 rasoi. Con l'ultimo risultato il consumo pro-capite è 208,6:5=41,7 rasoi e la durata media 365:41,7=8,8 giorni. Questo valore si ottiene immediatamente come media armonica dei dati iniziali:

$$M_{-1} = \frac{5}{\frac{1}{10} + \frac{1}{6} + \frac{1}{30} + \frac{1}{5} + \frac{1}{14}} = 8,8.$$

Per comprendere il motivo per il quale si deve adoperare la media armonica e non quella aritmetica delle durate, occorre osservare che il problema riguarda il *consumo*, per cui si deve tenere conto che la prima persona consuma in un giorno 1/10 di rasoio, la seconda consuma 1/6 di rasoio e così via, per cui, nel complesso, le cinque persone consumano in un giorno la somma delle quantità suddette.

Il valore unico $1/\overline{x}$ da sostituire a questi consumi diversi, lasciando invariato il consumo complessivo delle 5 persone è dato, perciò, dalla equazione:

$$5\frac{1}{\overline{x}} = \frac{1}{10} + \frac{1}{6} + \frac{1}{30} + \frac{1}{5} + \frac{1}{14}$$

da cui si ricava la durata media facendo l'inverso della media dei consumi, cioè proprio la media armonica. Avendo quindi rilevato la durata dei rasoi invece dei consumo, bisogna tener conto che tra queste due quantità esiste una relazione inversa.

D) Il valore divisorio. *Il valore divisorio di una successione non decrescente è quel valore medio che si può dividere in due parti tali che la somma della prima parte con i termini che lo precedono risulti uguale alla somma della seconda parte con i termini che lo seguono*. In altri termini, data la successione $a_1, a_2, ..., a_{k-1}, a_k, a_{k+1}, ...a_n$, il valore divisorio è a_k se si verifica che che è:

$$a_1 + a_2 + ... + a_{k-1} < a_k + a_{k+1} + ... + a_n$$
$$a_1 + a_2 + ... + a_{k-1} + a_k > a_{k+1} + ... + a_n$$

Se invece delle disuguaglianze sussiste l'uguaglianza, allora il valore divisorio sarà pari alla semisomma dell'ultimo valore del primo membro e del primo valore del secondo membro della uguaglianza.

Esempio III.5

Si consideri la seguente successione: 2, 4, 5, 6, 8, 10, 12. Il valore divisorio è 8 e, infatti, risulta sia che 2+4+5+6<8+10+12, sia che 2+4+5+6+8>10+12.

Per la seguente distribuzione 3, 4, 5, 7, 8, 9, 10 le due disuguaglianze sono 3+4+5+7<8+9+10 la prima e 3+4+5+7+8>9+10 la seconda, per cui il valore divisorio è ancora 8, ma in questo caso si ha in più che la somma dei termini che precedono il valore divisorio è uguale alla somma dei termini che lo seguono, ossia è 3+4+5+7=9+10.

5.1.2 Medie lasche o di posizione

Tra le medie lasche, quelle di uso più frequente sono la *mediana*, la *moda*, la *semisomma degli estremi,* il *valore poziore*.

E) La mediana (M_e). Nell'ipotesi di aver posto tutte le unità statistiche esaminate secondo l'ordine crescente dei valori assunti dalla variabile, *la mediana è quel valore della variabile al di sotto ed al di sopra del quale si situa la metà del numero totale dei casi, essa cioè divide l'insieme delle unità in due parti di uguale frequenza*. Quindi la mediana è maggiore delle modalità del carattere che sono possedute dalla metà dei casi in esame e minore delle modalità possedute dall'altra metà dei casi. Se il numero dei termini è dispari, la mediana è il valore posseduto dal termine che occupa il posto di mezzo; se il numero dei termini è pari essa è uguale alla media aritmetica dei valori relativi ai due termini che occupano i due posti centrali.

Nel caso di dati raggruppati in n classi, può essere immediatamente calcolata la classe i cui appartiene la mediana. Il valore della mediana si calcola poi con la formula:

$$M_e = Cl_i + \frac{\left[\left(\frac{N}{2} - \sum_{k=1}^{k=i-1} f_k\right).A_i\right]}{f_i}, \qquad \text{con} \sum_{i=1}^{i=n} f_i = N,$$

dove Cl_i è il confine inferiore della classe i , A_i è la sua ampiezza e f_i è la frequenza della classe.

La *mediana* o *valore mediano* gode delle due seguenti proprietà:

1) Il *numero* degli scostamenti *positivi* è *uguale* al numero degli scostamenti *negativi*. Ciò vuol dire che in una distribuzione o serie di dati ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.

2) *La somma dei valori assoluti degli scostamenti è un minimo* in confronto alla somma dei valori assoluti degli scostamenti cui darebbe luogo un altro valore medio qualsiasi diverso dal valore mediano.

3) E' una misura *robusta* in quanto *non è influenzata dalla presenza di dati anomali* e in particolare non è influenzata dai valori estremi, ma soltanto dal *numero* delle osservazioni, per cui *si ricorre al suo uso quando si vuole attenuare l'effetto di valori anomali*.

La mediana è la misura di posizione o di tendenza centrale utilizzata in quasi tutti i tests non parametrici.

Esempio III.6

Nella seguente successione di numeri 5, 9, 6, 14, 11, il valore mediano è il 9. Nella seguente 5, 9, 6, 14, 11, 18 formata da un numero pari di termini la mediana è (9+11)/2=10.

Nel caso di variabili statistiche divise in intervalli, il metodo migliore è quello di costruire la distribuzione cumulativa delle frequenze. Ad esempio, consideriamo la distribuzione:

Classi	n_i	$N_i=\Sigma n_k$
50|--100	110	110
100|--200	400	510
200|--300	90	600
N=	600	

La mediana corrisponde alla modalità del termine che occupa il posto 600/2=300, quindi si tratta di un valore interno alla classe 100|--200.

Per individuarlo, si fa l'ipotesi di uniforme distribuzione delle unità all'interno della classe e si considera la proporzione:

$$(M_{-1} - 100): (200-100) = (300-110) : 400$$

dalla quale si ricava la mediana

$$M_{-1} = 190x100/400 + 100 = 147,5$$

che risulta leggermente inferiore al valore centrale della classe 100|--200 alla quale apparteneva il valore che lasciava da una parte e dall'altra lo stesso numero di termini.

La ricerca della mediana si può applicare all'intera serie di dati a disposizione, oppure ad una sola parte dei dati stessi posti in ordine non decrescente. In tale ultimo caso si hanno medie lasche derivate note come *quartili, decili* e *centili.*

Il primo quartile (Q_1) di una successione di termini disposti in senso non decrescente è quella quantità al di sotto della quale sta 1/4 ed al di sopra della quale stanno i 3/4 dei valori dati. Il secondo quartile (Q_2) coincide con la mediana della distribuzione. Mentre il terzo quartile (Q_3) è quella quantità al di sotto della quale stanno i 3/4 ed al di sopra della quale sta 1/4 dei valori dati. Una espressione più generale è quella che definisce *il primo quartile come la mediana dei valori inferiori alla mediana* e *il terzo quartile come il valore mediano dei valori superiori alla mediana della distribuzione.*

Un ragionamento di tipo analogo vale per i *quantili* chiamati anche *frattili*, perché ogni gruppo parziale contiene la stessa frazione di osservazioni. Quelli più comunemente usati sono i *decili*, che dividono i dati ordinati in decine, ed i *percentili,* che suddividono i dati in centesimi.

F) La moda o *norma* o *valore modale* o *valore normale* o *valore di massima frequenza* (M_d): Tra tutti i valori assunti dalla variabile studiata *si chiama "Moda" o "Valore modale" quella modalità della variabile che si presenta con la frequenza più elevata* e la classe in cui essa risulta compresa si chiama classe modale. Nel caso di dati raggruppati per classi, dopo aver individuato anzitutto la classe alla quale appartiene la moda, quest'ultima può essere immediatamente calcolata in base alla formula

$$M_d = Cl_i + \frac{\left[\left(f_i - f_{i-1}\right) A_i\right]}{\left(f_i - f_{i-1}\right) + \left(f_i - f_{i+1}\right)},$$

ove i simboli conservano lo stesso significato già detto.

La *moda* ha la proprietà di rendere *massimo il numero degli scostamenti nulli*. Essa *non è influenzata dalla presenza di valori estremi*, tuttavia viene *utilizzata solamente per scopi descrittivi*, perché è *meno stabile ed*

oggettiva di altre misure di tendenza centrale. Essa differisce sia da campione a

campione, sia quando con gli stessi dati si formano classi di distribuzione con ampiezza differente.

La *moda* e la *mediana* hanno anche la proprietà seguente: *se si fanno variare tutti i termini di una serie in base ad una certa legge, i valori normale e mediano della serie data corrispondono ai valori normale e mediano della nuova serie*.

Esempio III.7

Si abbia la seguente distribuzione del numero di frantoi in funzione della capacità annua di produzione di olio:

Capacità produttiva (q.li)	Numero di frantoi
150|--200	60
200|--300	115
300|--500	140
500|--750	75
750|--1000	15
Totale	405

A prima vista si potrebbe ritenere che la moda sia compresa nella classe 300|--500, ma ciò è falso, in quanto le classi hanno ampiezza differente. Per trovare la classe modale, quindi occorre anzitutto ridurre le varie classi ad una ampiezza uguale: ad esempio:

Capacità produttiva (q.li)	Numero di frantoi
150|--200	60
200|--250	57,5
250|--300	57,5
300|--350	35
350|--400	35
400|--450	35
450|--500	35
500|--550	15
550|--600	15
600|--650	15
650|--700	15
700|--750	15
750|--800	3
800|--850	3
850|--900	3
900|--950	3
950|--1000	3
Totale	405

Così facendo si scopre che la classe modale è la prima

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.

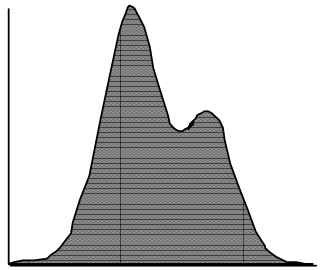


Figura III.17. Distribuzione bimodale

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.

G) La semisomma degli estremi è un altro valore medio, piuttosto grossolano come misura, ma che talvolta si considera per la estrema rapidità con cui si può calcolare. Essa rappresenta la sola possibilità di sintesi di una serie quando di quest'ultima si abbiano solo informazioni relative agli estremi

del campo di variazione del carattere. Questa media ha la proprietà di *rendere minimo lo scostamento massimo*.

Essa deve essere utilizzata 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 è utilizzato in alcune discipline come la meteorologia. Può essere utile nel caso di una serie di dati sulla temperatura, ove non esistono mai valori anomali. 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. Per molti statistici è 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, 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.

H) Il valore poziore. *Il valore poziore è quel valore che moltiplicato per la sua frequenza dà luogo ad un massimo*. E' questo un valore medio importante in certe circostanze, come quando si vuole mettere in luce quali siano le unità statistiche che forniscono il maggior contributo alla formazione dell'ammontare complessivo del carattere.

Esempio III.8

Si consideri il numero delle famiglie classificato secondo il numero dei figli avuti ad una certa data, come indicato dalla distribuzione ipotetica seguente:

Numero di figli	Numero di famiglie	Prodotto
1	1823	1823
2	1704	3408
3	1103	3309
4	651	2604
5	351	1755
6	188	1128
7	89	623
8	36	288
9 e oltre	26	260
Totale	5971	15198

Se si vuole mettere in luce quali sono le famiglie che più contribuiscono alla generazione successiva, si deve appunto ricercare quel valore che moltiplicato per la sua frequenza dà un massimo. Risulta quindi che il valore poziore di questa distribuzione è pari a 2, mentre il valore modale é 1, la mediana è 2 e la media aritmetica è 2,545 e il valore divisorio è 3.

I) Una espressione generale valida per qualunque media ottenibile a calcolo è la formula che definisce la media di potenze. *La media di potenze di indice r di una variabile statistica che si presenta con n modalità differenti, ciascuna avente frequenza f_i , è quel valore che si ottiene considerando la radice di ordine r della media aritmetica delle potenze r-esime delle singole determinazioni*. In simboli:

$$M_r = \sqrt[r]{\frac{\sum_{i=1}^{i=n} x_i^r f_i}{\sum_{i=1}^{i=n} f_i}} = \left(\frac{\sum_{i=1}^{i=n} x_i^r f_i}{\sum_{i=1}^{i=n} f_i} \right)^{\frac{1}{r}}$$

Nel caso particolare che sia r = -1, la media di potenze si riduce alla media armonica, se r = 0 riproduce la media geometrica, se r = 1 quella aritmetica e se r = 2 si ha la media quadratica (particolarmente usata in fisica). Anche in questo caso la media di potenze si definirà semplice o ponderata, a seconda che le frequenze siano tutte uguali all'unità oppure tra loro diverse. E' da osservare che M_r è una funzione continua e crescente con r. Come si è detto, essa in statistica definisce la *Media di potenza di ordine r* della variabile considerata, che è pari alla radice r-esima del *Momento di ordine r rispetto all'origine.*

5.1.3 Relazione tra media, moda e mediana

Come si è detto, la media aritmetica, la media geometrica e la media armonica sono parametri ottenuti a calcolo, cioè esse dipendono da tutti i valori o modalità assunti dalla variabile, mentre la moda, la mediana e le altre medie lasche dipendono dalla posizione relativa dei valori assunti dalla variabile ed il loro valore non cambia se si cambia il valore - ma non la frequenza - di qualcuno dei dati singoli che compongono l'insieme, in modo che l'ordine crescente o decrescente dei singoli dati resti immutato.

In una distribuzione simmetrica la media, la mediana e la moda coincidono con lo stesso valore. Quando la distribuzione è moderatamente asimmetrica la mediana cade tra la media e la moda.

Tra questi tre tipi di medie vale la relazione descritta dalla seguente formula:

$$M_d = M_1 - 3(M_1 - M_e)$$

La media geometrica è più piccola in valore della media aritmetica, ma è più grande della media armonica:

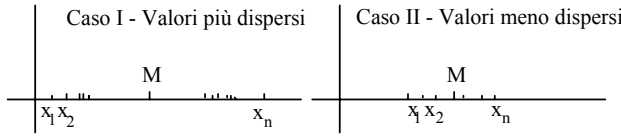
$$M_{-1} < M_0 < M_1.$$

5.2 *Le misure di dispersione o variabilità*

In alcune distribuzioni le variabili assumono valori più raggruppati intorno ad un valore medio che in altre distribuzioni. Quando si parla di variabilità di una variabile in una distribuzione, generalmente intendiamo riferirci alla variazione dei dati attorno ad una misura di tendenza centrale.

Generalmente si usa la parola *dispersione* per intendere la stessa cosa. Quindi, *si definisce dispersione o variabilità l'attitudine dei dati a disporsi intorno a un valore medio* .

Poiché in due diverse distribuzioni le variabili non saranno disperse nello stesso modo intorno ad un valore medio prestabilito, una misura accettabile della dispersione potrà aiutarci a caratterizzare meglio una particolare distribuzione di valori.



Le *misure della dispersione o della variabilità* più usate in statistica sono le seguenti.

A) Il campo di variazione: *è l'intervallo di valori compreso tra il più piccolo ed il più grande dei valori assunti dalla variabile*.

Intervallo di variazione = Valore massimo - valore minimo

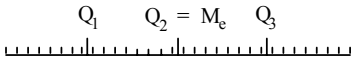
Ha il grande vantaggio di essere un metodo intuitivo e molto semplice, in particolare quando i dati sono ordinati. E' utile nelle discipline in cui i valori delle osservazioni hanno limiti noti e la semplice individuazione di quelli estremi serve per valutare la situazione; per esempio, un intervallo di variazione molto limitato nelle dimensioni di un gruppo di individui suggerisce che probabilmente appartengono alla stessa generazione, hanno un patrimonio genetico simile, sono cresciuti in condizioni ambientali omogenee.

Tra gli inconvenienti di questa misura sono da considerare sia l'incapacità di misurare come i dati sono distribuiti entro l'intervallo, sia la sua dipendenza dal numero di osservazioni, in particolare dalla presenza di valori anomali determinati da fattori eccezionali. 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 una misura poco efficiente della dispersione dei dati: per un confronto omogeneo tra distribuzioni, perché sarebbe necessario avere campioni delle stesse dimensioni,

cioè una condizione operativa eccessivamente limitante per la ricerca e l'analisi dei dati.

B) Lo scarto interquartilico o "deviazione interquartilica" o "semi-differenza interquartilica": *è uguale alla metà della distanza compresa tra il terzo quartile ed il primo quartile*. Il terzo quartile è detto anche quartile superiore ed è caratterizzato dal fatto che al di sopra di esso si situano il 25% di tutti i valori assunti dalla variabile, dopo averli ordinati per ordine crescente. Il primo quartile è detto anche quartile inferiore ed è caratterizzato dal fatto che al di sotto di esso si situano il 25% di tutti i valori assunti dalla variabile, dopo averli ordinati per ordine crescente. Evidentemente *il secondo quartile coincide con la mediana* della distribuzione, in quanto lascia al di sotto ed al di sopra di esso il 50% dei valori assunti dalla variabile.

I quartili dividono l'insieme dei valori assunti da una variabile in 4 gruppi eguali ed il 50% di essi sono compresi tra il primo ed il terzo quartile. Tanto più breve è l'intervallo tra dette due quantità, tanto minore sarà la variabilità della distribuzione.



Lo scarto interquartilico è quindi definito dalla relazione che esprime la semidifferenza tra il terzo ed il primo quartile:

$$Q_d = \frac{Q_3 - Q_1}{2}.$$

Per distribuzioni moderatamente asimmetriche vale la formula empirica in base alla quale la semi-differenza interquartilica è uguale ai due terzi dello scarto quadratico medio, cioè

$$Q_d = \frac{2}{3}\sigma.$$

Ha il vantaggio di eliminare i valori estremi, ovviamente collocati nelle code della distribuzione; è una misutra che dipende dal numero delle osservazioni.

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 nel 1, 5 o 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 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".

C) Lo scostamento medio assoluto dalla mediana: è pari alla *media dei valori assoluti degli scostamenti dalla mediana*. E' una misura di dispersione molto usata in alcuni test di statistica non parametrica. Poiché la mediana ha la proprietà di rendere minima la somma dei valori assoluti degli scarti da essa, *lo scostamento medio assoluto dalla mediana è sempre inferiore allo scarto medio assoluto dalla media aritmetica.* I due valori sono uguali solo per distribuzioni simmetriche, cioè quando media aritmetica e mediana coincidono. Alcuni autori utilizzano anche lo*scostamento probabile* o *mediano*, che è definito come la mediana degli scostamenti rispetto ad un dato valore che può essere la media aritmetica, la mediana o altro. Questo è quello scostamento che è maggiore della metà degli scostamenti più piccoli e minore della metà degli scostamenti più grandi.

D) La differenza semplice media: è uguale alla *media dei valori assoluti di tutte le possibili differenze che si possono istituire tra le quantità osservate*. Essa esprime il valore probabile della differenza che si otterrebbe tra due quantità scelte a caso dall'insieme delle quantità osservate. Se tra le differenze possibili si considerano oppure no anche quelle di ciascun termine da se stesso si parla di differenza semplice media *con o senza ripetizione* e tra i due indici la sola differenza consiste nel numero dei casi considerati, in quanto il numeratore sarà lo stesso, mentre al denominatore si avrà n^2 oppure $n(n-1)$.

E) La differenza quadratica media: è uguale alla *media dei quadrati dei valori di tutte le possibili differenze che si possono istituire tra le quantità osservate*. Le differenze medie si distinguono dagli altri indici di

variabilità, in quanto in esse non figura alcun altro valore centrale e ciò fa sì che queste e gli altri indici considerati trovino impiego appropriato in campi diversi.

F) Lo "scostamento semplice medio" o deviazione semplice media: *è uguale alla somma dei valori assoluti delle differenze tra ciascun valore della variabile statistica considerata ed un valore K prefissato* (generalmente si considera la media aritmetica, ma se ne possono considerare anche altri) *divisa per il numero delle differenze calcolate*. Quindi:

$$\text{Scostamento semplice medio rispetto a K} = = \frac{1}{N}\sum_{i=1}^{i=n}|x_i - K|.f_i$$

ove $N = \sum_{i=1}^{i=n} f_i$.

In questa espressione K può essere la media aritmetica, la mediana, la moda, la media quadratica o qualsiasi altro valore medio prescelto.

La deviazione media è una misura di dispersione che dipende da tutti i valori della variabile, ma non è una misura molto utilizzata, specialmente negli sviluppi teorici formali, a causa della presenza del valore assoluto degli scarti che offre non indifferenti difficoltà nel trattamento analitico della funzione (non derivabilità). Per questa ragione ad essa viene generalmente preferita un'altra misura della variabilità basata sul quadrato degli scarti dalla media.

G) Lo scostamento quadratico o "scarto quadratico" o "deviazione quadratica": è la misura di dispersione più utilizzata. Esso *è definito come la radice quadrata della media dei quadrati degli scostamenti di ciascun valore della variabile rispetto ad un valore K prefissato*. La formula che ne stabilisce la struttura è la seguente:

$$\text{Scostamento quadratico da K} = \sqrt{\frac{\sum_{i=1}^{i=n}(x_i - K)^2.f_i}{N}}$$

in cui, come precedentemente detto, K può essere la media aritmetica, la mediana o qualsiasi altro valore medio preferito. Nel caso in cui si sia scelta la media aritmetica come termine di riferimento degli scarti, lo scostamento quadratico medio si chiama anche deviazione standard e, in genere, viene sempre indicato

con la lettera greca σ (sigma) quando ci si riferisce ad una intera popolazione di valori oppure con la lettera latina s (esse), quando ci si riferisce ad un campione di valori tratto dalla popolazione di studio.

In certi casi, diviene più conveniente utilizzare come misura di variazione il quadrato della deviazione standard, che è denominato *varianza* ed ha quindi l'espressione seguente:

$$\sigma^2 = \frac{\sum_{i=1}^{i=n}(x_i - M)^2.f_i}{N}$$

Il numeratore della varianza è chiamato *devianza* ed ha una grande importanza in statistica, perché può essere scomposto in porzioni che sono di grande utilità per la teoria dell'analisi della varianza, come si vedrà in seguito.

Se K è la media aritmetica, per distribuzioni moderatamente asimmetriche vale la relazione:

$$\text{Scostamento semplice medio dalla media aritmetica} = \frac{4}{5}\sigma$$

Se con ${}^2\Delta_R$ si indica la differenza quadratica media con ripetizione (cioè considerando anche le differenze di ciascun valore con se stesso, il ché non altera il numeratore, ma aumenta il denominatore di tante unità quanti sono i valori considerati) dell'insieme di osservazioni considerato, si dimostra che valgono le relazioni:

$$\sigma = \frac{{}^2\Delta_R}{\sqrt{2}} \qquad \text{ossia} \qquad 2.\sigma^2 = {}^2\Delta_R^2 .$$

H) Il coefficiente di variazione. I valori della deviazione standard ottenuti per due o più distribuzioni le cui unità di misura sono molto diverse tra loro non sono comparabili. Per superare questa difficoltà, si ha la necessità di utilizzare una misura relativa della variabilità. Ciò si ottiene esprimendo una data misura della dispersione come percentuale della media da

cui è stato misurato lo scostamento. In questo modo si ottiene una misura relativa della variabilità chiamata *coefficiente di variazione*, che è ottenuto esprimendo lo scostamento quadratico medio in percentuale della media della distribuzione su cui esso è stato calcolato. In simboli:

$$V = \frac{\sigma}{M}.100$$

Altre misure di variabilità si possono avere anche con riferimento agli scarti delle singole modalità da altri tipi di medie, come si è detto, ma generalmente tali misure sono scarsamente utilizzate.

5.3 *Momenti di una distribuzione*

I momenti di una distribuzione possono considerarsi ad un tempo misure della variabilità ed anche misure che caratterizzano la forma delle distribuzioni statistiche. Per questa ragione essi sono qui trattati autonomamente rispetto agli altri argomenti di questo capitolo.

5.3.1. Definizione

Il k-esimo momento rispetto ad una origine arbitraria (A) è definito dalla espressione:

$$\nu_k = \frac{1}{N}\sum_{i=1}^{i=n}(x_i - A)^k.f_i$$

Ponendo A = M, il k-esimo momento rispetto alla media aritmetica (M) di conseguenza è definito come:

$$\mu_k = \frac{1}{N}\sum_{i=1}^{i=n}(x_i - M)^k.f_i$$

Inoltre, ponendo A = 0, si definisce il k-esimo momento rispetto all'origine degli assi (zero), il quale risulta essere pari a:

$$M_k = \frac{1}{N}\sum_{i=1}^{i=n}x_i^k.f_i$$

Nel caso di distribuzioni continue il posto del segno di sommatorio è preso dal segno di integrale. Quindi le tre espressioni finora scritte diventano:

$$\nu_k = \int_a^b (x - A)^k.f(x).dx,$$

$$\mu_k = \int_a^b (x - M)^k.f(x).dx,$$

$$M_k = \int_a^b x^k.f(x).dx,$$

in cui, al solito l'intervallo (a, b) indica il campo di definizione della funzione o distribuzione.

Da queste definizioni segue immediatamente che *il primo momento rispetto all'origine non è altro che la media aritmetica* e che *il secondo momento rispetto alla media è la varianza*. Il terzo ed il quarto momento rispetto alla media, divisi rispettivamente per il cubo e per la quarta potenza dello scostamento quadratico medio, sono utilizzati in statistica anche per misurare le caratteristiche di forma (asimmetria e appiattimento) delle distribuzioni statistiche di tipo campanulare. Infatti, ricordando che per la media vale la definizione:

$$M = M_1 = \frac{1}{N}\sum_{i=1}^{i=n}x_i.f_i, \text{ con } N = \sum_{i=1}^{i=n}f_i$$

mentre per la deviazione standard vale la definizione:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{i=n}(x_i - M)^2.f_i} = \sqrt{\mu_2},$$

si assume come misura dell'asimmetria (si veda il paragrafo 7 di questo capitolo) l'espressione:

$$\alpha_3 = \frac{1}{\sigma^3}\left[\frac{1}{N}\sum_{i=1}^{i=n}(x_i - M)^3 . f_i\right] = \frac{\mu_3}{\sigma^3},$$

mentre la misura dell'appiattimento (si veda il paragrafo 7 di questo capitolo) è ottenuta con la relazione:

$$\alpha_4 = \frac{1}{\sigma^4}\left[\frac{1}{N}\sum_{i=1}^{i=n}(x_i - M)^4 . f_i\right] = \frac{\mu_4}{\sigma^4}.$$

Dalle definizioni che precedono segue che per qualunque distribuzione sarà $\mu_1 = 0$, cioè il momento primo rispetto alla media è nullo ossia, in altre parole, *la media aritmetica gode della proprietà di rendere nulla la somma algebrica degli scarti*.

Ciò si vede con facilità considerando lo sviluppo seguente:

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{i=n}(x_i - M).f_i = \frac{1}{N}\sum_{i=1}^{i=n}x_i.f_i - \frac{1}{N}M\sum_{i=1}^{i=n}f_i = \frac{1}{N}\sum_{i=1}^{i=n}x_i.f_i - M = 0.$$

5.3.2. Calcolo dei momenti con il metodo abbreviato

Il calcolo dei momenti usando in ciascuna classe i valori originari delle variabili spesso implica la manipolazione di numeri molto grandi, per cui, se si vogliono prevenire tediose elaborazioni, è spesso conveniente utilizzare un metodo di calcolo più semplice, chiamato "*Metodo abbreviato*" o "*Metodo della media fittizia*", il quale ha appunto il vantaggio di far effettuare le operazioni necessarie con numeri meno grandi, riducendo così notevolmente la gravosità dei calcoli quando questi non possano essere effettuati utilizzando una apparecchiatura di calcolo automatico, per la quale evidentemente il problema della dimensione dei numeri non si pone affatto.

Il metodo abbreviato o della media fittizia consiste essenzialmente nell'assumere per vera, all'inizio di tutte le operazioni, una media

fittizia o di comodo per i calcoli e quindi nell'applicare ai risultati così ottenuti una correzione per ottenere sia la media vera sia i momenti di ordine più elevato.

La scelta migliore consiste nell'assumere come valore medio fittizio il valore della classe più vicino al centro della distribuzione, oppure, se possibile, il valore che corrisponde alla frequenza maggiore. In altre parole, invece di fare i calcoli usando i valori originari della variabile, si sostituiscono ad essi altri valori ottenuti assoggettando la variabile ad una trasformazione del tipo

$$x = \frac{v - A}{h}$$

in cui v rappresenta il valore centrale originario della classe, h rappresenta l'intervallo o ampiezza della classe ed A la media fittizia. La nuova variabile x assumerà quindi dei valori di dimensione ridotta rispetto a quelli originari, dimodoché i calcoli risulteranno semplificati.

A questo accorgimento si usa far ricorso spesso anche per *ridurre il numero dei decimali su cui si deve operare* quando il valore centrale sia appunto una cifra contenente numerosi decimali.

Ora cerchiamo la correzione che si deve apportare ai momenti calcolati rispetto alla variabile x per ottenere i momenti calcolati rispetto alla variabile originaria v. Per far questo, dalla trasformazione precedente, si ricava agevolmente:

$$v = h.x + A.$$

Ora, ricordando la definizione della media aritmetica, la relazione tra la media della variabile trasformata e quella della variabile originaria può essere determinata compiendo le dovute sostituzioni già dette e con dei passaggi molto semplici, come si può rilevare dalla espressione seguente:

$$M_v = \frac{1}{N}\sum_{i=1}^{i=n}v_i.f_i = \frac{1}{N}\sum_{i=1}^{i=n}(h.x_i + A).f_i = \frac{1}{N}h\sum_{i=1}^{i=n}x_i.f_i + \frac{1}{N}A\sum_{i=1}^{i=n}f_i = h.M_x + A$$

Per cui risulta la relazione:

$$M_v = h.M_X + A$$

secondo la quale *la media è assoggettata ad una trasformazione dello stesso tipo di quella subita dalla variabile*. Similmente, in base alla definizione del momento di ordine k rispetto alla media, sarà:

$$\mu_{k,v} = \frac{1}{N}\sum_{i=1}^{i=n}(v_i - M_v)^k.f_i$$

ed anche in questo caso, usando le relazioni stabilite per esprimere il legame tra le due variabili e quindi tra le loro medie aritmetiche, possiamo scrivere per il momento di ordine <u>k</u> la relazione:

$$\mu_{k,v} = \frac{1}{N}\sum_{i=1}^{i=n}\left[(h.x_i + A) - (h.M_x + A)\right]^k.f_i = \frac{1}{N}\sum_{i=1}^{i=n}(h.x_i - h.M_x)^k.f_i =$$

$$= \frac{1}{N}.h^k.\sum_{i=1}^{i=n}(x_i - M_x)^k.f_i = h^k.\mu_{k,x}$$

In definitiva tra i momenti della variabile trasformata v e quelli della variabile originaria x vale la relazione veramente semplice che segue:

$$\mu_{k,v} = h^k.\mu_{k,x}$$

dalla quale otteniamo le seguenti espressioni:

- per la varianza $\mu_{2,v} = h^2.\mu_{2,x}$
- per il terzo momento $\mu_{3,v} = h^3.\mu_{3,x}$
- per il quarto momento $\mu_{4,v} = h^4.\mu_{4,x}$
- per la deviazione standard $\sigma_v = h.\sigma_x$

- per l'indice di asimmetria $\alpha_{3,v} = \alpha_{3,x}$
- per l'indice di appiattimento $\alpha_{4,v} = \alpha_{4,x}$

5.3.3. <u>Relazione tra M_k e μ_k</u>

Per calcolare i momenti rispetto alla media di una data distribuzione dobbiamo anzitutto trovare le potenze seconda, terza e quarta o più elevate ancora delle deviazioni dalla media. Di solito queste deviazioni sono numeri frazionari e le loro potenze non sono comode da calcolare senza opportune apparecchiature elettroniche.

Per superare questo inconveniente possiamo esprimere i momenti rispetto alla media in termini dei momenti rispetto all'origine degli assi, questi ultimi essendo calcolabili con maggiore facilità specialmente quando le variabili originarie sono sostituite dalla variabile trasformata x in modo che essa assuma solo, o la maggior parte delle volte, valori espressi da numeri interi.

Le formule che permettono il calcolo *dei momenti rispetto alla media* utilizzando i *momenti rispetto all'origine* sono ottenute molto semplicemente con i seguenti sviluppi formali indicati nelle espressioni che seguono:

$$\mu_2 = \frac{1}{N}.\sum_{i=1}^{i=n}(x_i - M_1)^2.f_i = \frac{1}{N}.\sum_{i=1}^{i=n}\left(x_i^2.f_i - 2.x_i.M_1.f_i + M_1^2.f_i\right) =$$

$$= \frac{1}{N}.\sum_{i=1}^{i=n}x_i^2.f_i - 2.M_1.\frac{1}{N}.\sum_{i=1}^{i=n}x_i.f_i + M_1^2.\frac{1}{N}.\sum_{i=1}^{i=n}f_i =$$

$$= M_2 - 2.M_1^2 + M_1^2 = M_2 - M_1^2$$

cioè, si ha che il momento secondo rispetto alla media è pari al momento secondo rispetto all'origine diminuito del quadrato del momento primo rispetto all'origine; quindi sarà anche:

$$\sigma = \sqrt{M_2 - M_1^2} = \sqrt{\frac{\sum_{i=1}^{i=n} x_i^2 . f_i}{N} - \left(\frac{\sum_{i=1}^{i=n} x_i . f_i}{N} \right)^2} .$$

Le formule che esprimono μ_3, μ_4, ..., μ_k, possono ottenersi in maniera simile. In generale, poiché vale la uguaglianza $M^0 = M_0 = 1$, il momento μ_k potrà essere calcolato in base alla seguente formula:

$$\mu_k = \sum_{i=0}^{i=k} C(k,i).(-M)^i . M_{k-i}$$

di modo che per k = 3 avremo:

$$\mu_3 = M_3 - 3.M_2.M_1 + 2.M_1^3$$

mentre per k = 4 sarà:

$$\mu_4 = M_4 - 4.M_3.M_1 + 6.M_2.M_1^2 - 3.M_1^4$$

Quindi, per calcolare il momento di ordine k rispetto alla media occorre conoscere tutti i primi k momenti rispetto all'origine.

5.3.4. Le correzioni di Sheppard

Come si è visto i momenti di una data distribuzione sono calcolati usando come modalità i punti centrali degli intervalli delle varie classi di valori in cui è distribuita la variabile. Ma il raggruppamento di tutte le frequenze di una classe in corrispondenza del punto centrale di questa conduce a certi errori che possono essere corretti applicando ai momenti le formule denominate "*correzioni di Sheppard*".

Se con μ_k indichiamo il momento k-esimo calcolato rispetto alla media e con μ'_k rappresentiamo invece la forma corretta dello stesso momento, le

espressioni corrette dei momenti sono definite dalle seguenti relazioni, nelle quali h indica l'intervallo di ampiezza della classe:

- per il secondo momento $\mu'_2 = \mu_2 - \frac{h^2}{12}$
- per il terzo momento $\mu'_3 = \mu_3$
- per il quarto momento $\mu'_4 = \mu_4 - \frac{h^2}{2}.\mu_2 + \frac{7}{240}.h^4$

Queste espressioni sono note in statistica come *le correzioni dei momenti* suggerite da *Sheppard per ridurre l'influenza dell'errore dovuto al raggruppamento in classi dei dati originari* di una distribuzione. Tali correzioni, quindi non sono necessarie solo se si dispone delle misure relative alle singole unità statistiche.

Esempio III.9

Supponiamo di voler calcolare i primi quattro momenti dell'insieme di dati statistici seguente. In una prova sperimentale l'esame del peso in milligrammi di 100 semi di pino di montagna ha mostrato la seguente distribuzione:

Classi di peso mg	Frequenza o numero di semi
15 ¬ 35	3
35 ¬ 55	10
55 ¬ 75	17
75 ¬ 95	50
95 ¬ 115	12
115 ¬ 135	6
135 ¬ 155	2
	100

La tavola che segue mostra i calcoli dei primi quattro momenti della distribuzione suddetta usando il metodo abbreviato:

Classi di peso mg	Valore centrale v	Frequenza f	x	xf	x^2f	x^3f	x^4f
15 ¬ 35	25	3	- 3	- 9	27	- 81	243
35 ¬ 55	45	10	- 2	- 20	40	- 80	160
55 ¬ 75	65	17	- 1	- 17	17	- 17	17
75 ¬ 95	85	50	0	0	0	0	0
95 ¬ 115	105	12	1	12	12	12	12
115 ¬ 135	125	6	2	12	24	48	96
135 ¬ 155	145	2	3	6	17	54	16
		100	0	- 16	138	- 64	690

N = 100 , h = 20 A = 85, x = (v - 85)/20.

Media:

$M_x = (- 16)/100 = - 0,16$ $M_v = h.M_x + A$

$M_v = 20.(-0,16) + 85 = - 3,2 + 85 = 81,8$

Deviazione stardard o scostamento quadratico medio:

$M_{2,x} = 138/100 = 1,38$ $M^2_x = (- 0,16)2 = 0,0256$

$\_\mu_{2,x} = M_{2,x} - M_{2x} = 1,38 - 0,0256 = 1,35$

$\_\sigma_x = \sqrt{\mu_{2,x}} = \sqrt{1,35} = 1,16$

$\sigma_v = h.\sigma_x = 20 . 1,16 = 23,2$

Correzioni di Sheppard sul secondo momento (si ricorda che per la variabile x l'ampiezza h = 1, mentre per la variabile v l'ampiezza h = 20):

$\_\mu'_{2,x} = 1,35 - 1/12 = 1,35 - 0,08 = 1,27.$

$\sigma'_x = \sqrt{1,27} = 1,127$

$\sigma'_v = 20 . 1,127 = 22,54.$

Momento terzo:

$M_{3,x} = (- 64)/100 = - 0,64$ $M^3_x = (- 0,16)3 = 0,004$

$\_\mu_{3,x} = M_{3,x} - 3.M_{2,x} M_x + 2.M_{3x} =$

$= - 0,640 – 3.1,38.(- 0,16) + 2. (-0,004) =$

$= - 0,640 + 0,662 - 0,008 = 0,014.$

$\alpha_3 = 0,014/(1,16)3 = 0,014 / 1,566 = 0,009.$

Ma anche con le correzioni di Sheppard su _x avremo:

$\_\alpha'_3 = 0,014/(1,127)3 = 0,014 / 1,431 = 0,009.$

Momento quarto:

$M_{4,x} = 690/100 = 6,90$ $M^4_x = (- 0,16)4 = 0,0007$

$\mu_{4,x} = M_{4,x} - 4.M_{3,x}.M_x + 6.M_{2,x}.M^2_x - 3.M^4_x =$

$= 6,90 - 4.(-0,64).(- 0,16) + 6.1,38.0,0256 - 3.0,0007 =$

$= 6,90 - 0,4096 + 0,212 - 0,0021 = 6,70$

$\alpha_4 = 6,70/(1,16)^4 = 6,70/ 1,82 = 3,68.$

mentre con le correzioni di Sheppard sul quarto momento e per h = 1 sarà:

$\mu'_{4,x} = 6,70 - 1,35/2 + 7/240 =$

$= 6,70 - 1,35/2 + 7/240 = 6,70 - 0,675 + 0,029 = 6,054.$

$\alpha'_4 = \mu'_{4,x}/\sigma'^2_x = 6,054/(1,127)4 = 06,054 / 1,613 = 3,75.$

5.4. *Misure della forma delle distribuzioni*

La forma di una distribuzione può essere descritta da due aspetti: il primo si riferisce alla più o meno accentuata simmetria della distribuzione dei valori osservati attorno ad un valore centrale; la seconda si riferisce ad una concentrazione più o meno accentuata dei dati osservati, cioè delle frequenze, su certi valori centrali o periferici tali da determinare una forma più o meno appiattita della curva.

A differenza di quanto avvenuto nello studio della variabilità, nell'analisi della forma le misure elaborate sono rimaste rudimentali e le stesse definizioni sono sovente equivoche. Il loro uso non rientra nei test di inferenza, ma è limitato semplicemente alla descrizione della forma della distribuzione, senza possibilità alcuna di stabilire una relazione tra lo stesso indice nel campione e nell'intera popolazione da cui il campione è stato tratto.

5.4.1 Asimmetria

Una distribuzione si dice simmetrica quando, preso come riferimento un determinato valore centrale, *le frequenze dei valori inferiori a quest'ultimo si distribuiscono nello stesso modo in cui si distribuiscono i valori superiori ad esso*, ma in maniera speculare rispetto al valore centrale.

Per indicare il grado di simmetria di una data distribuzione rispetto ad un determinato valore centrale molti Autori hanno suggerito delle misure di asimmetria (in inglese *Skewness*) diverse tra loro, tra le quali si ricordano le seguenti.

1) La prima di esse rapporta alla distanza interquartilica la differenza tra lo scarto del terzo quartile dalla mediana e quello della mediana dal primo quartile, per cui si perviene alla seguente espressione formale:

$$S_k = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)} = \frac{Q_3 + Q_1 - 2.M_e}{Q_3 - Q_1}$$

Questo indice di asimmetria, noto come *Coefficiente di Yule,* varia evidentemente tra -1 e +1. Esso assume valori negativi quando la curva che descrive la distribuzione ha il ramo più allungato verso la zona ove sono situati i valori inferiori alla mediana; positivi nel caso contrario. Se l'indice assume il valore zero, allora la curva è simmetrica. Il valore -1 viene raggiunto quando $M_e = Q_3$, il valore +1 quando $M_e = Q_1$, il valore 0 quando M_e è equidistante da Q_1 e Q_3.

2) La seconda misura si basa sulla differenza tra media aritmetica e moda rapportata allo scostamento quadratico medio, cioè:

$$S_k = \frac{M - M_d}{\sigma}$$

Tale indice di asimmetria assume il valore zero per le curve simmetriche, mentre assume valori negativi quando il ramo più allungato della curva si trova sulla sinistra della moda (e quindi $M<M_d$) ed assume valori positivi nel caso contrario (e quindi $M>M_d$).

3) La terza misura è definita come rapporto tra il triplo della differenza tra media e mediana e lo scostamento quadratico medio della distribuzione, cioè:

$$S_k = \frac{3.(M - M_e)}{\sigma}$$

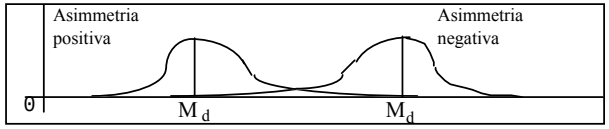
Anche questa volta l'indice assumerà il valore zero per le curve simmetriche, valore negativo per le distribuzioni asimmetriche a sinistra e valore positivo per quelle asimmetriche a destra e sarà quindi un utile strumento di descrizione sintetica della forma della distribuzione.

Tutte e tre le misure precedenti non sono analitiche, cioè la loro definizione non implica la utilizzazione di tutti i valori assunti dalla variabile studiata e che ne descrivono la distribuzione.

L'espressione che segue, invece, può essere considerata la migliore formulazione di un indice di asimmetria sia per i casi pratici sia per scopi di natura teorica. Essa è definita come la media delle potenze terze degli scostamenti dei singoli valori dalla loro media aritmetica, ossia non è altro che il *terzo momento rispetto alla media,* definito dalla relazione:

$$S_k = \frac{\sum_{i=1}^{i=n}(x_i - M)^3 . f_i}{N}, \text{ con } N = \sum_{i=1}^{i=n} f_i$$

E' chiaro che questa espressione si annulla per distribuzioni simmetriche, perché ad ogni scostamento positivo (x_i - M) di frequenza f_i, corrisponderà uno scostamento negativo di pari frequenza, cosicché le terze potenze avranno rispettivamente segno positivo e negativo e quindi i valori uguali, ma di segno contrario, si elideranno l'un l'altro. Nel caso in cui la distribuzione abbia una asimmetria positiva, la somma dei termini positivi che figurano nel sommatorio dell'indice sarà maggiore della somma dei termini negativi e quindi il risultato sarà positivo. Nel caso contrario la distribuzione avrà asimmetria negativa e l'indice stesso darà un risultato negativo.



Occorre, tuttavia, osservare che questa misura, per come è stata descritta, non rappresenta un numero puro, perché il suo valore dipende dall'unità di misura della variabile x . Per avere un indice indipendente dalla scala di misura adottata e dall'origine dei valori prescelta, l'espressione prima data viene divisa per il cubo dello scostamento quadratico medio ed indicata con la lettera greca α (alfa), come segue:

$$\alpha_3 = \frac{1}{\sigma^3 N}\sum_{i=1}^{i=n}(x_i - M)^3 . f_i = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

L'indice sopra detto può assumere valori negativi o positivi ed è uguale a zero per distribuzioni simmetriche, ma non è vero il viceversa, cioè ogni distribuzione per la quale l'indice suddetto sia zero non è necessariamente simmetrica. Malgrado tale difetto, questo indice di asimmetria è utilizzato in statistica con sufficiente soddisfazione per misurare il grado di simmetria di una distribuzione.

Un'altra misura dell'asimmetria spesso utilizzata è uguale al quadrato dell'indice or ora esaminato. Tale secondo indice è il *coefficiente β_1 (beta uno) del Pearson* definito dalla relazione:

$$\beta_1 = \alpha_3^2 = \frac{\mu_3^2}{\mu_2^3}$$

5.4.2. Appiattimento o disnormalità

Due o più distribuzioni simmetriche possono avere la *stessa media* e lo *stesso scostamento quadratico medio* e tuttavia avere una *forma della curva piuttosto differente* per quanto riguarda l'appiattimento (in inglese *Kurtosis*) più o meno accentuato dei loro grafici in prossimità della media. Per poterle differenziare allora vi è bisogno di una misura appropriata del grado di appiattimento rispetto ad un'altra curva che sia stata assunta come modello comparativo. Quando tale modello è la curva *normale* si parla di *disnormalità*.

Come misura di appiattimento è stata proposta la espressione seguente, che contiene le quarte potenze degli scostamenti dalla media aritmetica:

$$K_u = \frac{1}{N}\sum_{i=1}^{i=n}(x_i - M)^4 . f_i = m_4$$

Se si considerano due distribuzioni campanulari riferite allo stesso numero di casi e che abbiano lo stesso scostamento quadratico medio o deviazione standard, quella più appuntita, avente una maggiore percentuale dei suoi valori concentrati intorno alla sua media, in generale avrà le estremità più lunghe di quelle dell'altra curva meno appuntita per compensare la più elevata

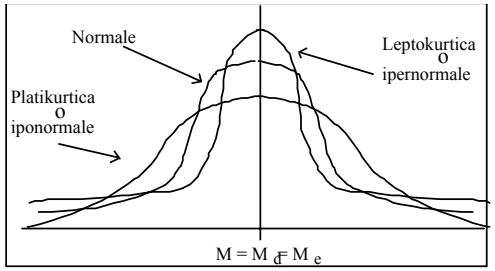
percentuale di scostamenti piccoli. Ora se si considerano le quarte potenze delle deviazioni più grandi in valore assoluto - che si riferiscono quindi ai valori più lontani dalla media - esse daranno alla somma sopra riportata un contributo molto maggiore di quello che potranno dare le quarte potenze degli scostamenti più piccoli; ne segue, che le distribuzioni più appuntite e che hanno le estremità della curva più allungate avranno un valore delle quarte potenze degli scostamenti superiore a quello delle curve più appiattite e con estremità più corte. Poichè questo indice dipende dall'unità di misura delle variabili, per evitare questo fatto si è soliti dividerlo per la quarta potenza della deviazione standard ed indicarlo come segue:

$$\alpha_4 = \frac{1}{\sigma^4 N} \sum_{i=1}^{i=n} (x_i - M)^4 . f_i$$

Questa quantità misura l'appiattimento o disnormalità della curva rappresentativa della distribuzione data, perché essa sarà tanto più bassa in valore quanto più appiattita sarà la curva di distribuzione della variabile.

L'indice di appiattimento o disnormalità talvolta è indicato con il simbolo β_2 ed è individuato come *coefficiente β_2 (beta due) del Pearson*; la sua espressione è definita dal rapporto tra il momento quarto rispetto alla media ed il quadrato del momento secondo rispetto alla media, come indicato qui di seguito:

$$\beta_2 = \alpha_4 = \frac{\mu_4}{\mu_2^2} .$$



Le curve di frequenza che, pur essendo anch'esse di tipo campanulare simmetrico, sono più appuntite della curva normale si chiamano *ipernormali o leptokurtiche*; quelle che sono più appiattite della curva normale si chiamano *iponormali o platikurtiche*. Il grafico sopra riportato mostra il confronto tra le curve ipernormali ed iponormali e la curva normale.

E' da osservare che per la curva normale l'indice di asimmetria β_1 è uguale a zero, mentre l'indice di appiattimento β_2 è uguale a 3.

6. Rappresentazione semigrafica degli indici delle distribuzioni

I diagrammi Box-and-Whisker (*scatola e baffi*), chiamati anche *boxplot*, sono un metodo grafico diffuso recentemente con i programmi informatici per rappresentare visivamente alcuni indici sintetici delle distribuzioni statistiche. Evidenziano tre caratteristiche: il *grado di dispersione* o variabilità dei dati, rispetto alla mediana e/o alla media; la *simmetria*; la presenza di *valori anomali*.

Sono utili per analizzare la distribuzione senza ricorrere obbligatoriamente alle misure della statistica parametrica, come la varianza e gli indici fondati sui momenti. *Servono per evidenziare la presenza di dati che si discostano in modo rilevante dagli altri*. Potrebbero essere valori reali; ma non raramente sono generati da errori di misura o di scrittura. E' sempre importante utilizzare tecniche che li evidenziano, per una ulteriore verifica della correttezza della rilevazione e della trascrizione.

Secondo il metodo originale proposto da Tukey, la costruzione di un diagramma Box-and-Whisker ha origine da una linea orizzontale, che rappresenta la *mediana*; vicino ad essa è collocata una croce, che rappresenta la *media*; attorno ad esse è costruita una scatola, i cui margini inferiore e superiore indicano rispettivamente il valore del *primo e* del *terzo quartile*; infine, all'esterno della scatola si estendono due linee verticali (una sopra e l'altra sotto), che

terminano con un breve segmento orizzontale (i baffi), per un valore uguale a *1,5 volte la distanza interquartilica*.

I valori che si discostano dalla mediana tra 1,5 e 3 volte la distanza interquartilica possono essere considerati nella norma; quelli che si discostano oltre 3 volte dovrebbero essere molto rari e meritano una verifica ulteriore, per escludere con sicurezza banali errori di misura o trascrizione.

Quando la media è distante dalla mediana e/o il primo ed il terzo quartile distano diversamente dalla mediana, la distribuzione è asimmetrica. La distanza del primo e del terzo quartile , inoltre, è una misura della dispersione o variabilità dei dati.

I diagrammi Box-and-Whisker hanno avuto una serie di adattamenti ed 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*;
- 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, presuppongono una distribuzione normale ed evidenziano sia la dispersione dei dati sia quella della media campionaria.

L'esempio III.10, riportato nelle pagine seguenti, utilizza la distribuzione dei dati sulle concentrazioni del sodio e di cloruri in 36 laghi degli Appennini. Esso può aiutare a chiarire l'uso e le potenzialità di questi diagrammi o boxplot. I confronti con la serie di statistiche dei dati, con gli indici che sono stati calcolati sulle medesime due distribuzioni e con la loro rappresentazione in istogrammi, di seguito riportate, offrono l'opportunità di meglio comprendere i rapporti tra le varie modalità di rappresentazione grafica e gli indici sintetici.

Nei primi 2 Box-and-Whisker il valore di riferimento centrale è la mediana, la scatola delimita il primo ed il terzo quartile, mentre i baffi in neretto individuano il valore minimo e quello massimo e quelli più chiari individuano la mediana più o meno 1,5 volte lo scarto interquartilico. 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.

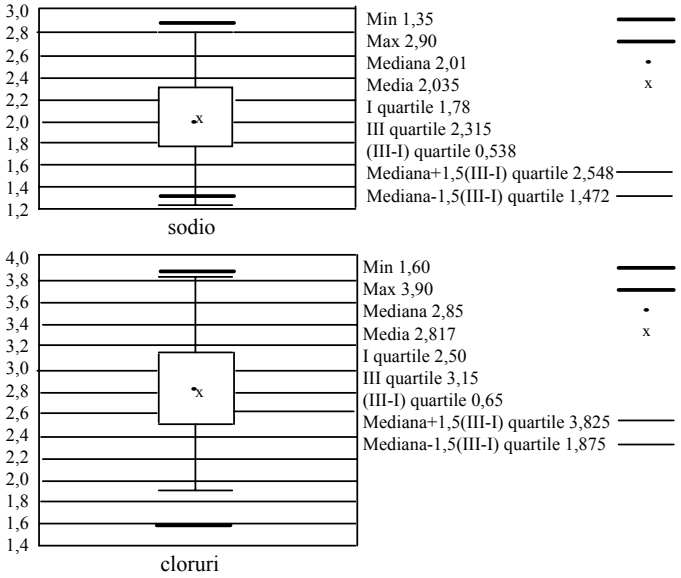


Figura III.15. Box-and-Whisker con misure non parametriche.

La distribuzione dei dati del sodio ha una asimmetria positiva o destra, mentre la distribuzione dei valori dei cloruri 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. Purtroppo sull'asimmetria, come sulla disnormalità o appiattimento, non esistono metodi oggettivi che conducano tutti i ricercatori alle medesime conclusioni. Fortunatamente, molti test parametrici sono robusti e permettono di ottenere risultati attendibili, anche in presenza di una distorsione sensibile dei dati rispetto ad una distribuzione normale.

Negli altri 2 boxplot (Fig. III.16), il valore di riferimento è la media, la scatola riporta la distanza di 1 errore standard ed i baffi una distanza di 1 deviazione standard. Sono misure parametriche di dispersione rispettivamente

della media e delle singole osservazioni. Anche in questo caso le linee in neretto si riferiscono ai valori minimo e massimo. I baffi (Whisker) esterni 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 e numerosità del campione raccolto.

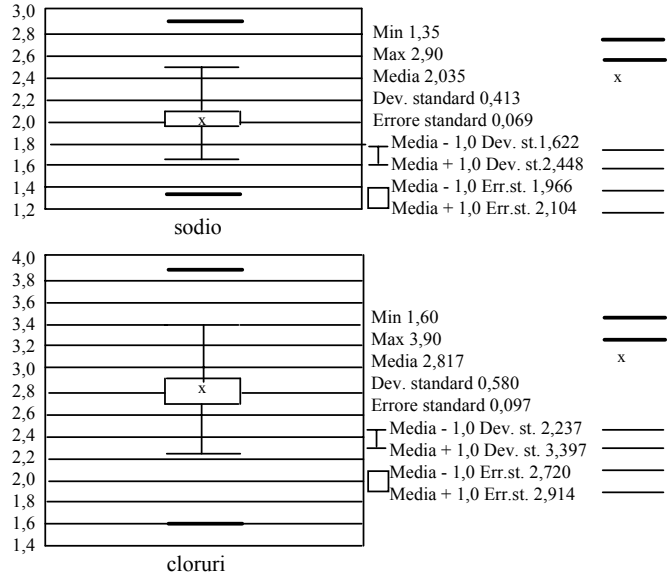


Figura III.16. Box-and-Whisker con misure parametriche.

Come sarà chiarito con l'uso della distribuzione normale, la frazione o percentuale di valori compresi nell'intervallo dipende da quante volte è riportato il valore della deviazione standard o dell'errore standard.

Si tratta di concetti che saranno sviluppati con l'uso della distribuzione normale, quando è nota la varianza della popolazione; si ricorrerà alla distribuzione t di Student, quando si utilizza la varianza del campione. I metodi servono per valutare sia la dispersione dei dati intorno alla media, sia la distribuzione delle medie di campioni con *n* osservazioni.

Esempio III.10

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. Vogliamo calcolare le misure della tendenza centrale, della variabilità e degli indici di forma e rappresentare graficamente in istogrammi i dati che sono riportate nella tabella seguente:

Lago	Sodio	Cloruri	Lago	Sodio	Cloruri
1	1,78	1,60	19	1,75	2,60
2	1,63	1,80	20	2,11	2,60
3	1,85	2,90	21	2,30	2,60
4	2,10	2,90	22	1,95	2,70
5	1,35	2,90	23	2,60	2,90
6	1,40	2,90	24	2,44	2,90
7	1,82	2,00	25	2,18	3,00
8	1,35	2,00	26	2,51	3,10
9	2,06	2,00	27	2,37	3,10
10	1,85	2,20	28	2,54	3,30
11	1,51	2,30	29	2,03	3,30
12	2,00	2,30	30	2,77	3,40
13	2,02	2,80	31	2,31	3,40
14	1,90	2,80	32	2,81	3,60
15	1,60	2,80	33	2,33	3,70
16	2,18	2,50	34	1,45	3,80
17	1,82	2,50	35	1,78	3,80
18	1,90	2,50	36	2,90	3,90

Le statistiche calcolate dai programmi informatici comprendono varie misure di tendenza centrale, di dispersione, di simmetria e di disnormalità o appiattimento o 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 primi 2 boxplot riportati in figura III.15.

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):

Statistiche	Sodio	Cloruri
Conteggio (Count; N. of data)	36	36
Minimo (Minimum)	1,35	1,60
Massimo (Maximum)	2,90	3,90
Intervallo (Range)	1,55	2,30
Somma totale (Total)	73,25	101,40
Moda (Mode)	1,78	2,90
I quartile (First quartile)	1,780	2,500
Media armonica (Harmonic mean)	1,953	2,692
Media geometrica (Geometric mean)	1,994	1,994
Mediana (Median)	2,010	2,850
Media (mean)	2,035	2,817
III quartile (Third quartile)	2,315	3,150
Errore standard (Standard error)	0,069	0,097
Deviazione standard (Standard deviation)	0,413	0,580
Devianza (Sum of squares)	5,981	11,770
Varianza campionaria (Sample variance)	0,171	0,336
Appiattimento o disnormalità (Kurtosis)	-0,521	-0,424
Asimmetria (Skewness)	0,266	-0,016

Diff. interquartilica	0,535	0,650

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 enunciati.

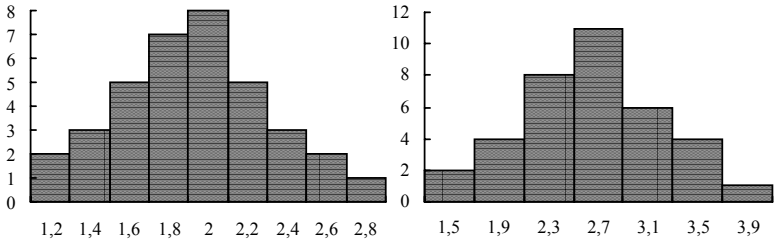


Figura III.17

Istogramma delle concentrazioni del Sodio Istogramma delle concentrazioni dei Cloruri

Le due serie di valori hanno una distribuzione molto vicino a quella 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.