Elementi di statistica descrittiva

Organizzazione dei dati

I risultati numerici di una ricerca dovrebbero essere rappresentati in forma chiara e coincisa in modo da fornire una rapida idea generale delle caratteristiche globali. Esistono varie tecniche tabellari e grafiche che si sono sviluppate nel corso degli anni.

È possibile rappresentare la distribuzione di un campione di taglia *n* di dati tramite un grafico delle frequenze relative. In particolare, si possono usare grafici *a bastoncini, a barre e a linee*.

Il *diagramma di Pareto* è un diagramma a barre verticali (in cui le modalità compaiono in ordine decrescente rispetto alle frequenze di ciascuna) combinato con un poligono cumulativo nella stessa scala. Il vantaggio di questo tipo di grafico sta nel fatto che si possono separare le poche modalità alle quali è associata una frequenza più alta da quelle meno rappresentate nei dati, consentendo così di concentrarsi sulle modalità più importanti. Il diagramma di Pareto si costruisce a partire dal diagramma a barre semplicemente scambiando l'ordine delle modalità in modo che risultino in ordine decrescente delle frequenze. Sono presenti due assi verticali: su quello di sinistra sono riportate le frequenze o le percentuali, su quello di destra le frequenze cumulate o le percentuali cumulate. Naturalmente, questi grafici sono analoghi a quelli delle frequenze assolute ma con valori delle ordinate scalati di un fattore *1/n*.

Un altro tipo di rappresentazione grafica è il *grafico a torta* molto utile soprattutto quando i dati non sono numerici ma categorici. Si traccia un cerchio e lo si suddivide in tanti settori quante sono le categorie distinte di dati, ogni settore è caratterizzato da un angolo al centro proporzionale alla frequenza, che può essere relativa o assoluta, della categoria.

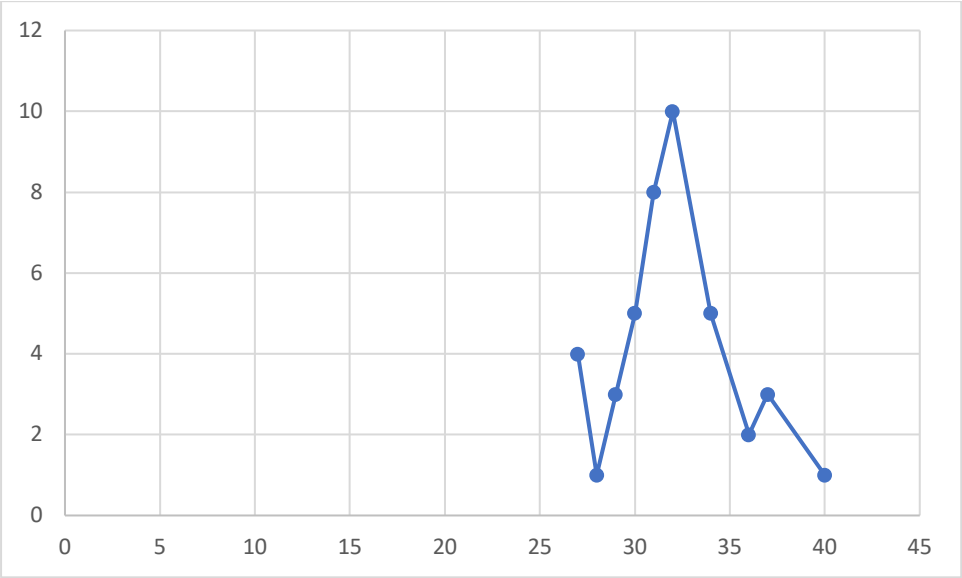
Quando i dati si sviluppano in numero relativamente basso di valori distinti possono essere rappresentati in una tabella tramite le loro frequenze. Come mostrato nel seguente esempio:

Esempio 1 In Tabella 1 sono riportati i dati relativi alla retribuzione annua iniziale di 42 neolaureati. Dai dati si evince che lo stipendio minimo annuale è stato di 27000 euro ed è stato corrisposto a 4 neolaureati, mentre lo stipendio massimo, di 40000 euro, è stato corrisposto ad una sola persona. La cifra più frequente è stata di 32000 euro.

Stpendio iniziale	Frequenza
27	4
28	1
29	3
30	5
31	8
32	10
34	5
36	2
37	3
40	1

Tabella 1: Stipendi iniziali annui in migliaia di euro

Figura 1



Poligono (grafico a linee)

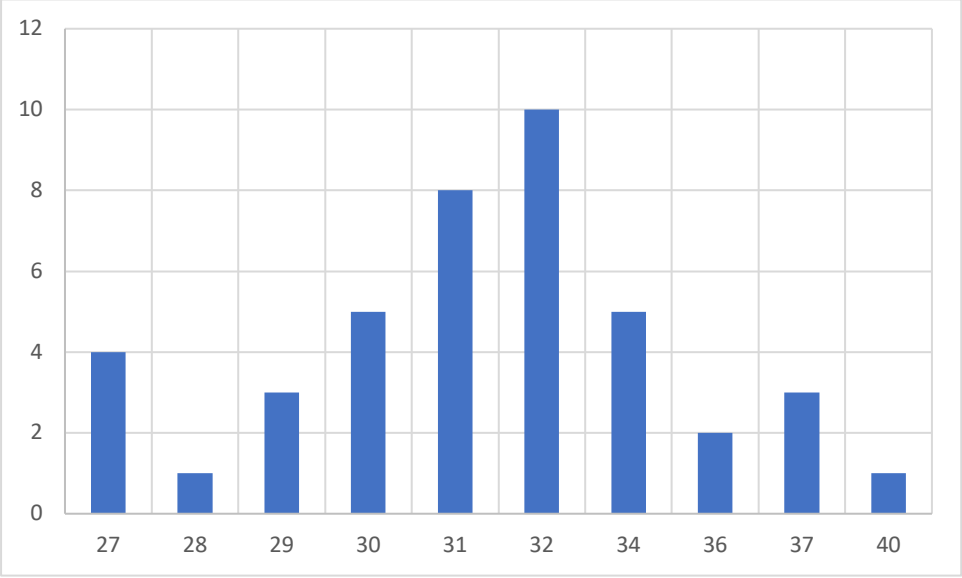


Diagramma a colonne

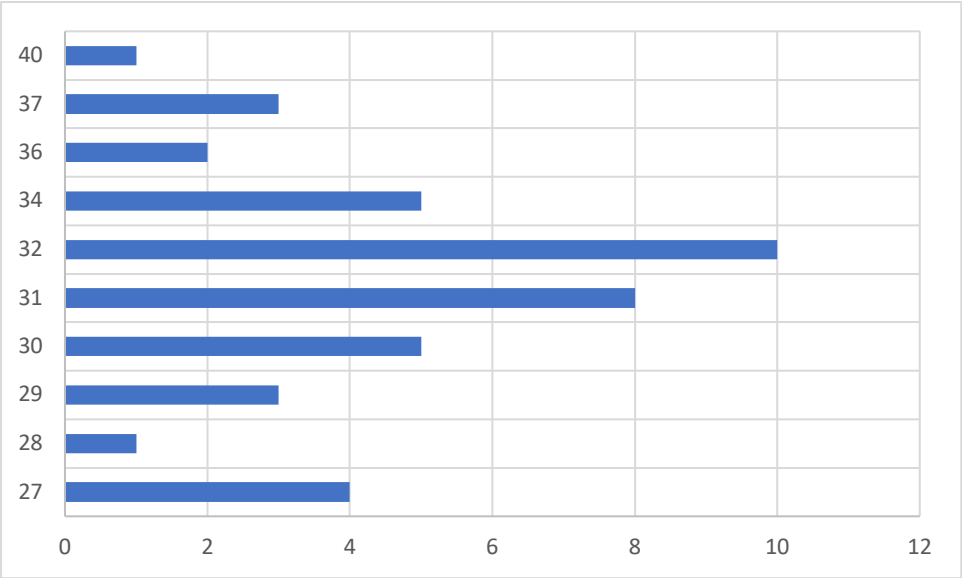


Diagramma a barre

Esempio 2. In Tabella 2 sono riportate le frequenze assolute e relative con cui sono stati riscontrati alcuni tipi di tumore su un campione di 200 ammalati.

Tipo di tumore	Numero di casi	Frequenza relativa
Polmoni	42	0.210
Seno	50	0.250
Colon	32	0.160
Prostata	55	0.275
Melanoma	9	0.045
Vescica	12	0.060
Totale	200	1

Tabella 2: Frequenze relative e assolute di tumori riscontrati in 200 ammalati

Figura 2

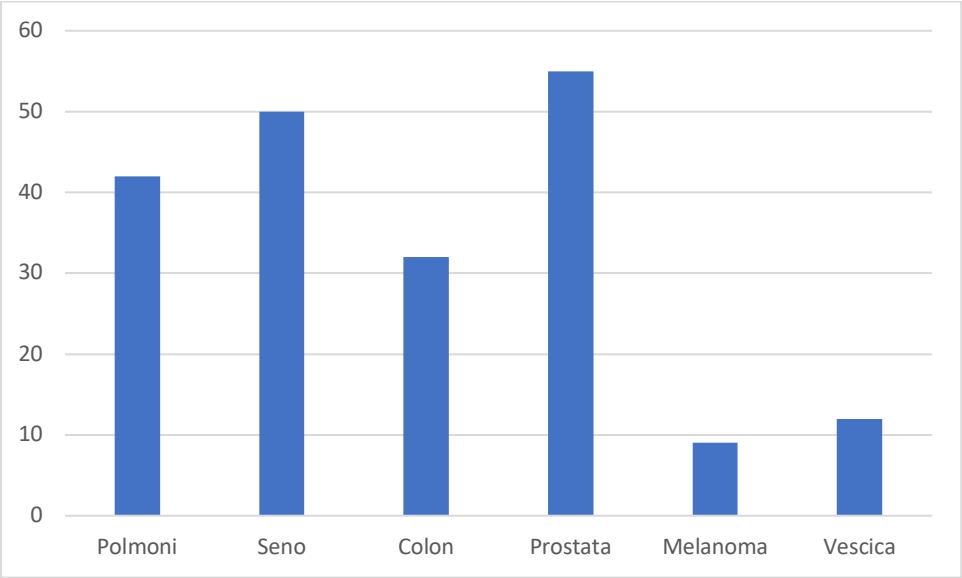


Diagramma a colonne

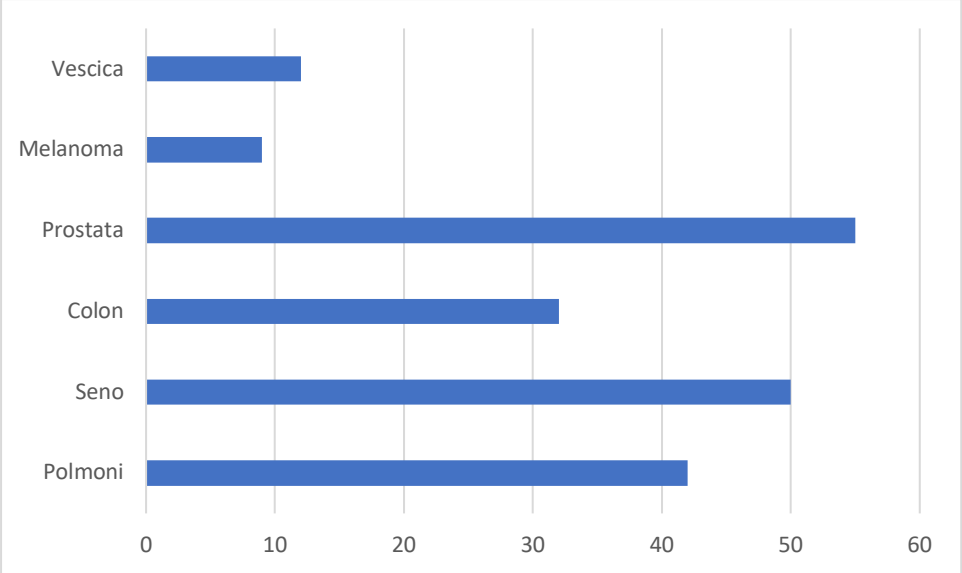


Diagramma a barre

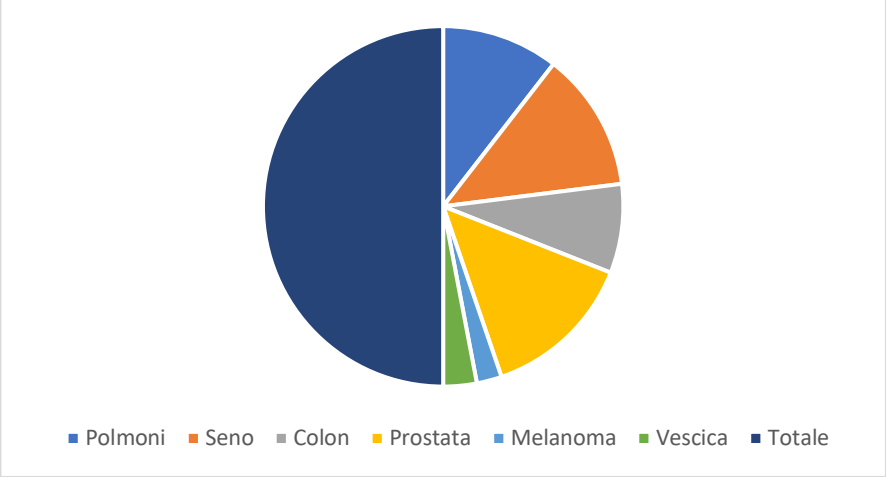


Diagramma a torte

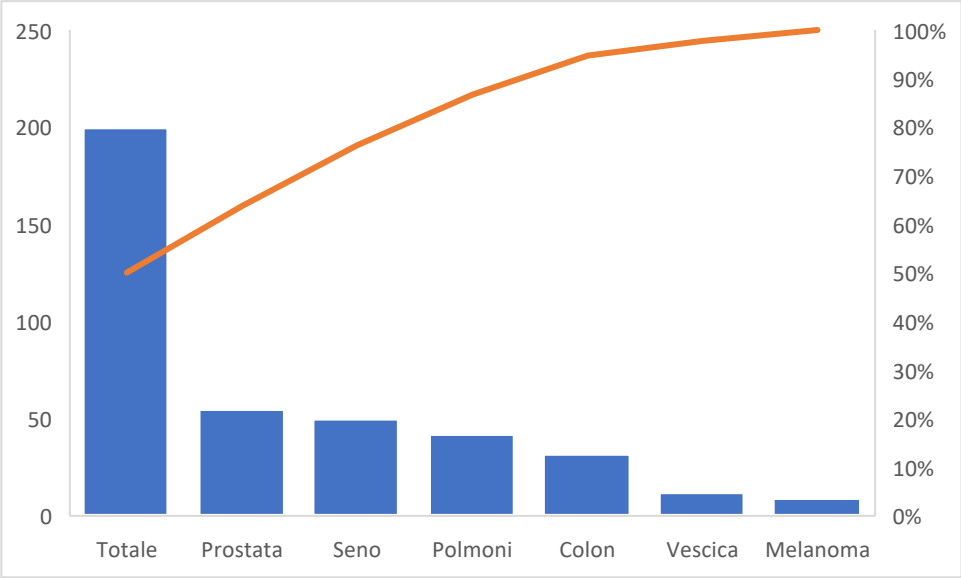


Diagramma di Pareto

All'aumentare del numero di osservazioni è opportuno dividere i dati in classi di raggruppamento o in categorie e riportarli successivamente sotto forma di tabelle. In questo modo si ottiene la distribuzione delle frequenze delle osservazioni. Aumentando le osservazioni disponibili è opportuno aumentare il numero delle classi che dovrebbe, comunque, variare da un minimo di 5 ad un massimo di 15. Se il numero delle classi non è adeguato diminuisce la quantità delle informazioni che si può ottenere. Spesso è opportuno che le classi abbiano la stessa ampiezza determinata come

$$\textit{Ampiezza delle classi} = \textit{range/ numero delle classi}$$

dove il range rappresenta l'intervallo di valori che la variabile assume (massimo – minimo). Per semplicità il rapporto è approssimato all'intero più vicino. Gli estremi delle classi devono essere individuati in modo da evitare sovrapposizioni e da coprire tutto il range delle osservazioni. Ad esempio, su un range che varia tra 21 e 48 si possono definire 6 classi di ampiezza pari al 5% visto che $48 - 21 = 27$, ed essendo $27/6 \approx 5\ (5\%)$.

Esempio 3 Supponiamo che i rendimenti ad un anno di 59 fondi a capitalizzazione siano quelli riportati nella Tabella 3. I dati sono stati ordinati in modo crescente per semplicità.

20,4	23,8	25,6	26,2	27,6	27,7	28,3	28,6	28,8	28,9
28,9	29,3	29,3	29,5	29,9	30,1	31,5	31,6	31,6	31,8
31,9	32,1	32,3	32,3	32,4	32,8	32,9	32,9	33,0	33,9
33,4	33,7	33,8	34,0	34,0	34,3	34,7	34,7	34,8	35,0
38,2	39,0	39,4	40,7	41,1	42,8	42,9	43,3	43,4	43,5
43,6	43,7	44,6	44,7	45,4	45,7	46,6	48,0	48,6	

Tabella 3: Ordinamento dei rendimenti percentuali ad un anno fatti registrare dai 59 fondi a capitalizzazione integrale.

In questo caso un numero di classi pari a 6 è sicuramente sufficiente a raggruppare i 59 fondi a capitalizzazione integrale. Risulta:

range = 48,6 – 20,4 = 28, *ampiezza dell'intervallo = 28,2/6 = 4,7*

Nella costruzione della distribuzione delle frequenze è necessario stabilire gli estremi delle classi per evitare sovrapposizioni. Così, se fissiamo al 5% l'ampiezza di ciascuna classe, gli estremi devono essere calcolati in modo da coprire l'intero range delle osservazioni. Quando possibile, la scelta degli estremi deve essere finalizzata a facilitare la lettura e l'interpretazione dei dati. Procedendo in questo modo, la prima classe andrà dal 20% al 25%, la seconda dal 25% al 30% e così via secondo lo schema qui sotto riportato (Tabella 4)

Analizzando la prima colonna della Tabella 2 si può osservare che il range approssimato dei valori corrisponde all'intervallo *20 – 50* e le osservazioni tendono ad accentrarsi tra i valori *30* e *35*. Comunque, non si riesce ad individuare come i valori si distribuiscono all'interno delle classi.

Generalmente si sceglie il punto medio come valore rappresentativo della classe.

In Tabella 2 sono riportate anche due varianti della distribuzione di frequenze:

la **distribuzione delle frequenze relative** che si ottiene rapportando le frequenze assolute della distribuzione delle frequenze al numero delle osservazioni,

la **distribuzione delle percentuali** si ottiene, poi, moltiplicando per cento ciascuna frequenza relativa.

Rendimenti percentuali a un anno	Numero di fondi	Distribuzione delle frequenze relative	Distribuzione delle frequenze percentuali
Da 20,0 a 25,0	2	0,034	3,4
Da 25,0 a 30,0	13	0,220	22,0
Da 30,0 a 35,5	24	0,407	40,7
Da 35,5 a 40,0	4	0,068	6,8
Da 40,0 a 45,5	11	0,186	18,6
Da 45,5 a 50,0	5	0,085	8,5
Totale	59	1,000	100,0

Tabella 4: Distribuzione delle frequenze dei rendimenti a un anno fatti registrare dai 59 fondi a capitalizzazione integrale

Con riferimento alla Tabella 4, la proporzione (ossia la frequenza relativa) di fondi a capitalizzazione integrale che hanno ottenuto un rendimento percentuale ad un anno compreso tra il 35,0 e il 40,0 è pari 0,068, questo equivale a dire che il 6,8% dei fondi ha avuto una tale performance.

Un altro strumento per rappresentare i dati è fornito dalla **distribuzione cumulativa**. Questa può essere ottenuta a partire dalle frequenze assolute, da quelle relative e da quelle percentuali,

semplicemente considerando l'estremo inferiore della classe e sommando tutte le frequenze minori a tale estremo. Nella Tabella 5 è riportata la distribuzione cumulata per le frequenze percentuali:

Rendimenti percentuali a un anno	Numero di fondi	Distribuzione delle frequenze percentuali	Rendimenti percentuali a una anno	Percentuale cumulata
Da 20,0 a 25,0	2	3,4	20,0	0
Da 25,0 a 30,0	13	22,0	25,0	3,4
Da 30,0 a 35,5	24	40,7	30,0	25,4
Da 35,5 a 40,0	4	6,8	35,0	66,1
Da 40,0 a 45,5	11	18,6	40,0	72,9
Da 45,5 a 50,0	5	8,5	45,0	91,5
Totale	59	100,0	50,0	100,0

Tabella 5: Distribuzione cumulativa delle percentuali dei rendimenti a un anno fatti registrare dai 59 fondi a capitalizzazione integrale

Molto spesso però si preferiscono i grafici alle tabelle. I grafici adatti alla rappresentazione di dati numerici sintetizzati con le distribuzioni di frequenze (assolute o relative) sono gli **istogrammi.**
L'istogramma è un diagramma a barre verticali in cui le barre rettangolari hanno come base gli intervalli in cui sono state raggruppate le osservazioni.
Nel grafico di un istogramma l'asse orizzontale corrisponde ai valori assunti dalle variabili in esame, mentre l'asse verticale rappresenta la proporzione o la percentuale delle osservazioni che cade in ciascuna classe.

Figura 3

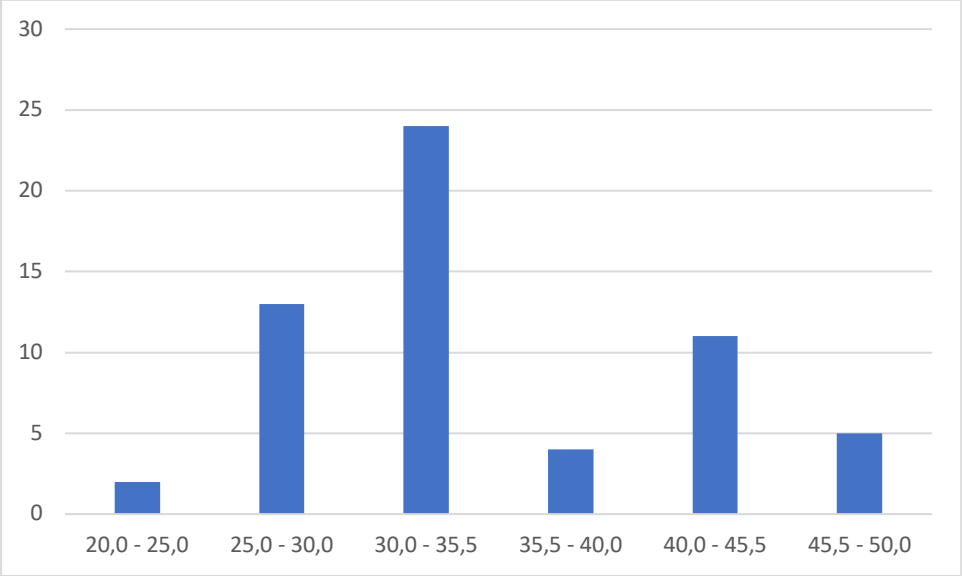


Grafico a colonne raggruppate

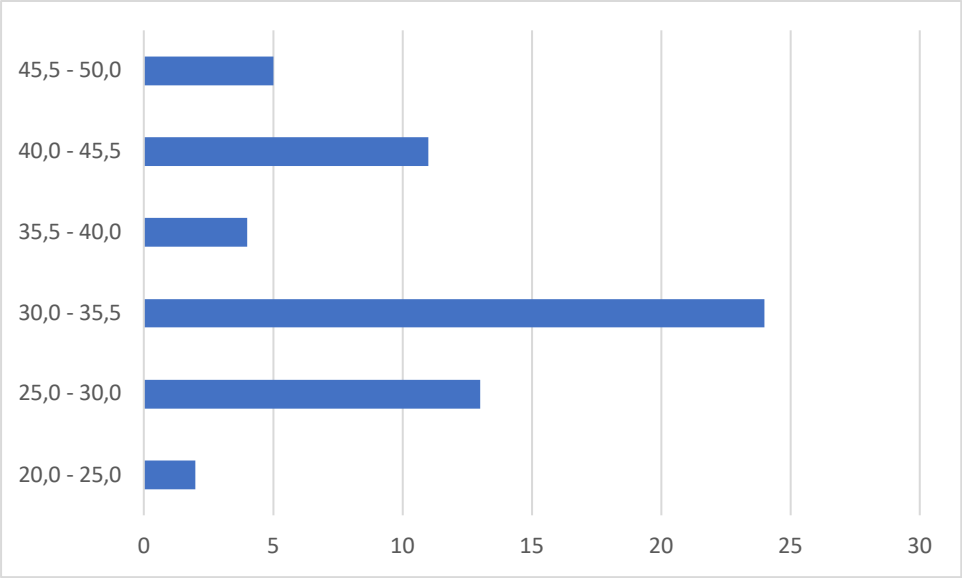


Grafico a barre raggruppate

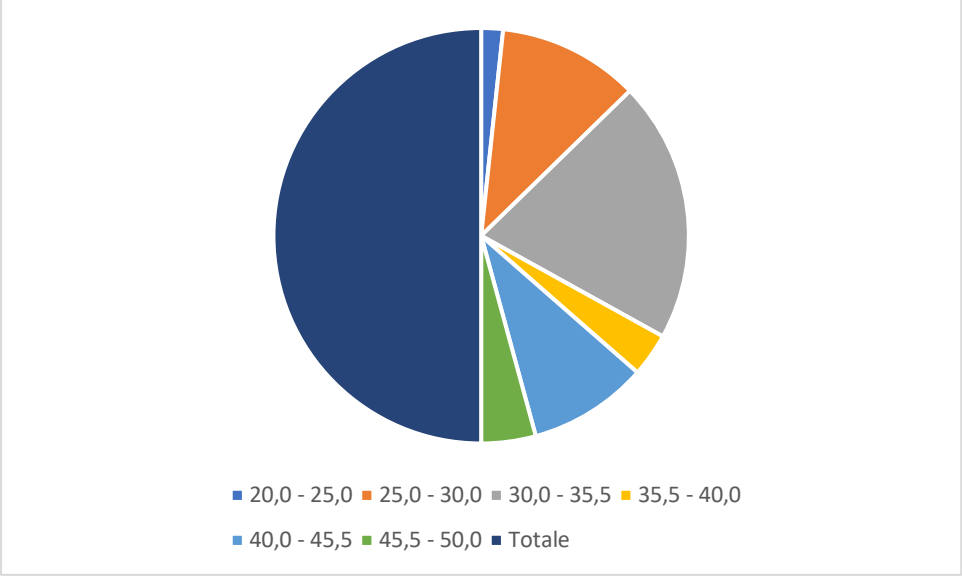


Grafico a torta

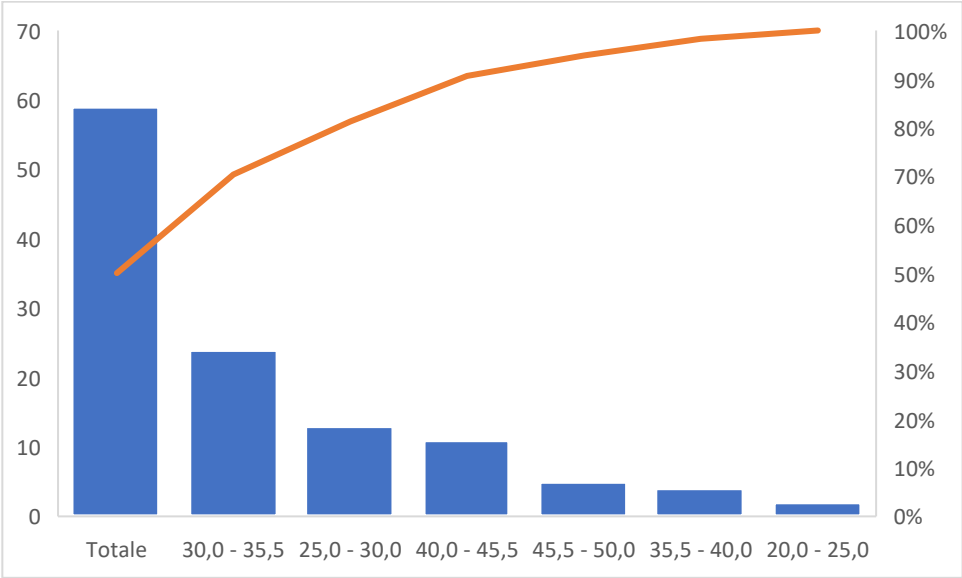


Grafico di Pareto

Il *poligono* si costruisce scegliendo il punto medio di ciascuna classe per rappresentare tutte le osservazioni che cadono nella classe stessa, congiungendo poi la sequenza dei punti medi alla percentuale di osservazioni nella classe corrispondente.

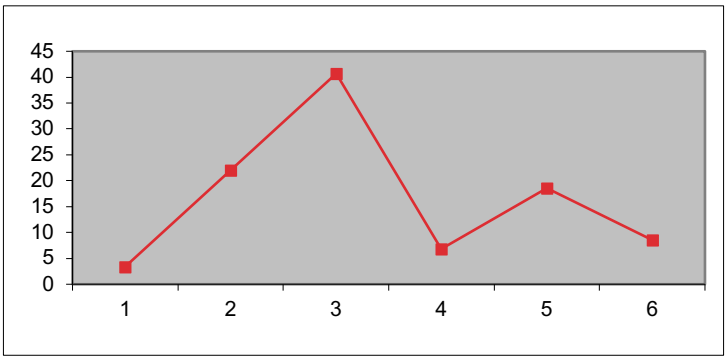


Figura 4: Poligono per i rendimenti ad un anno per i 59 fondi, le percentuali usate sono date in Tabella 2.

Il *poligono cumulativo* o *ogiva* è un rappresentazione grafica della tabella delle frequenze cumulative. Sull'asse orizzontale si rappresenta il fenomeno di interesse mentre su quello verticale si rappresenta la proporzione o la percentuale delle osservazioni cumulate.

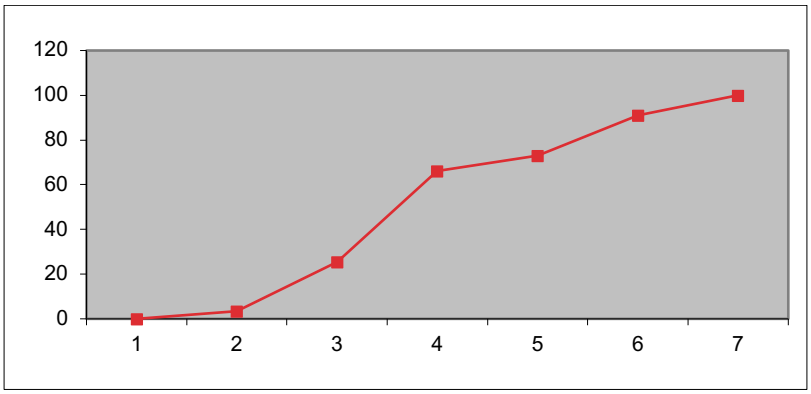


Figura 5: Ogiva per i rendimenti ad un anno per i 59, le percentuali cumulate usate sono date in Tabella 3.

È possibile sintetizzare anche le osservazioni di carattere quantitativo attraverso grafici e tabelle. Il primo passo verso questa direzione consiste nel sintetizzare le osservazioni *non numeriche* in forma tabellare, poi si fornisce una rappresentazione grafica. Tra tali strumenti grafici quelli più usati sono il diagramma a barre, il diagramma a torta e il diagramma di Pareto.

La **tabella di sintesi** per i dati qualitativi presenta le stesse caratteristiche della tabella delle frequenze utilizzata per i dati quantitativi. L'esempio che tratteremo in questa sezione riguarda la variabile relativa alle commissioni associate al fondo (Group).

Consideriamo un campione di *194* fondi azionari di cui *17* hanno commissioni prelevate dalle attività del fondo, *5* prevedono commissioni differite, *19* prevedono commissioni di ingresso, *46* hanno

commissioni multiple, mentre i restanti *107* fondi non prevedono alcuna commissione. Nella Tabella 4 sono riportate le frequenze assolute e le percentuali.

Commissioni	**Frequenze assolute**	**Percentuali**
Commissioni prelevate dalle attività del fondo	17	8,8
Commissioni differite	5	2,6
Commissioni di ingresso	19	9,8
Commissioni multiple	46	23,7
Fondi senza commissione	107	55,2
Totale	194	100,1*

Tabella 6: Tabella di sintesi e delle percentuali della variabile "commissioni associate al fondo" per i 194 fondi azionari del campione

In Figura 4 sono mostrati vari tipi di grafici basati per l'esempio in questione

Figura 4: Grafici relativi all'Esempio 4.

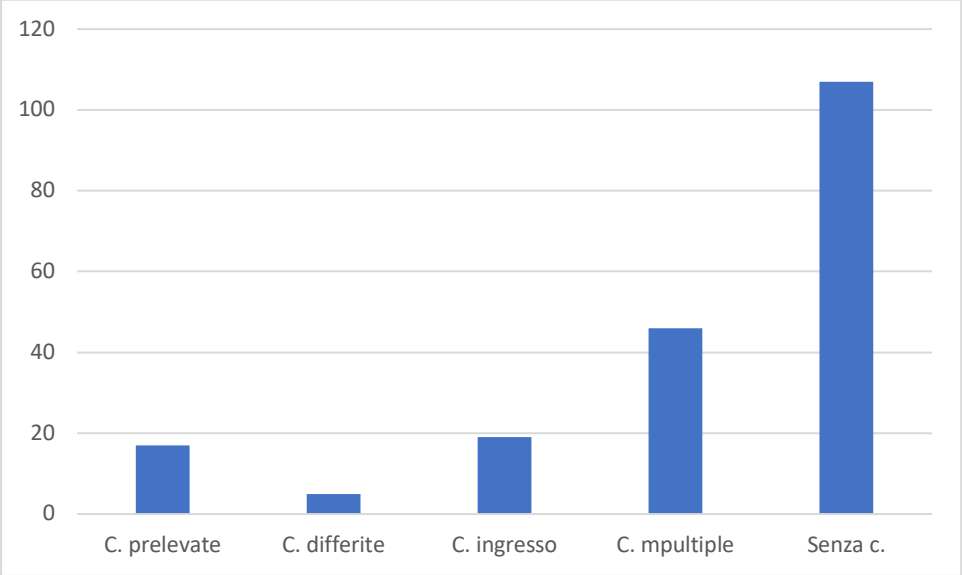


Grafico a colonne raggruppate

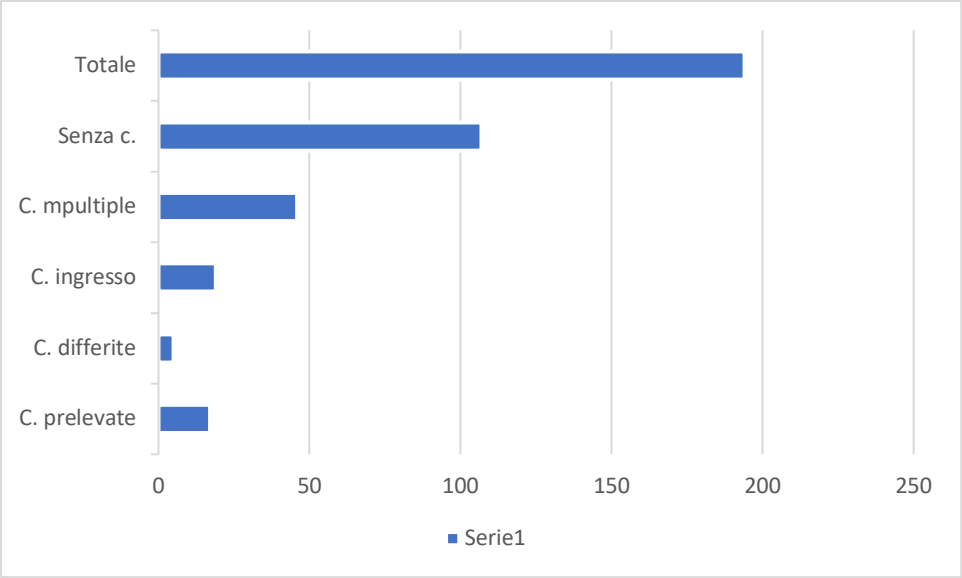


Grafico a barre raggruppate

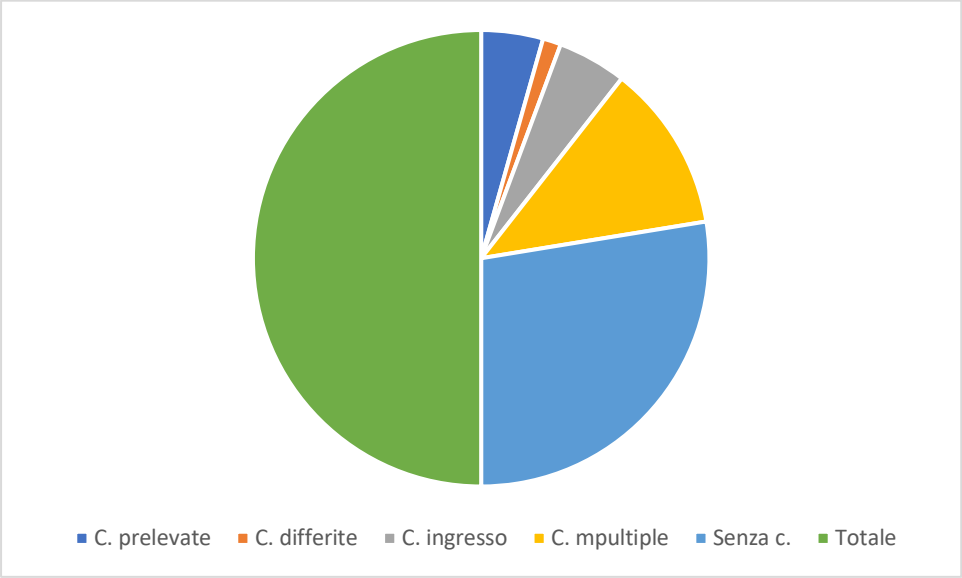


Grafico a torta

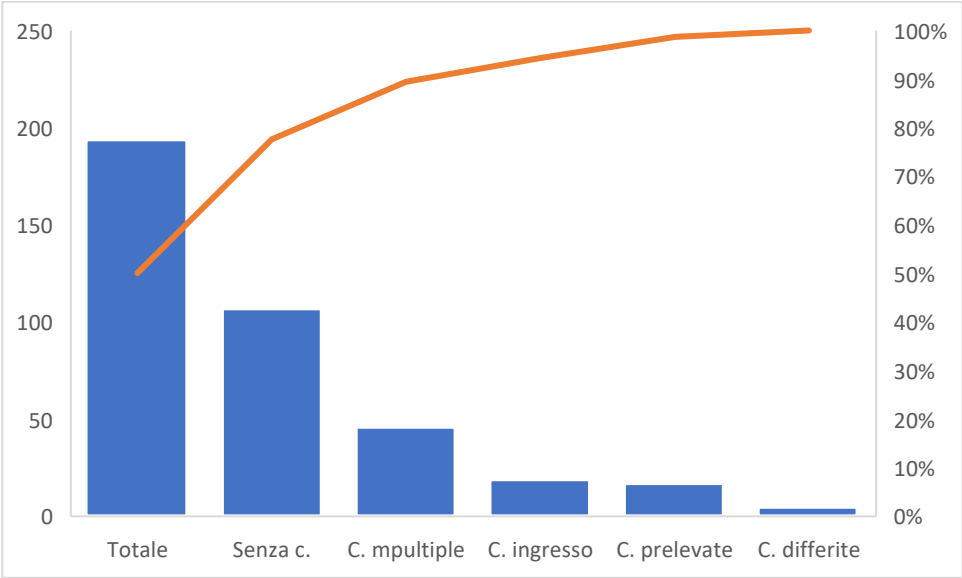


Grafico di Pareto

Sintesi e descrizione dei dati

Una buona analisi dei dati richiede anche che le caratteristiche principali delle osservazioni siano sintetizzate con opportune misure e che tali misure siano adeguatamente analizzate e interpretate.

Misure di posizione

Nella maggior parte dei dati, le osservazioni mostrano una tendenza a raggrupparsi intorno ad un valore centrale. Così, in generale, risulta possibile selezionare un valore tipico per descrivere un intero insieme dei dati. Il valore descrittivo così ottenuto è una misura di posizione o di *tendenza centrale*. Le principali misure di posizione sono: la media aritmetica, la mediana, la moda, il midrange e la media interquartile.

La media aritmetica, o semplicemente media, è sicuramente la misura di posizione più comune. Si calcola dividendo per il numero delle osservazioni la somma delle osservazioni stesse:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

dove n rappresenta la taglia del campione, X_i rappresenta la i-esima osservazione della variabile aleatoria X.

Notiamo che poiché il calcolo della media si basa su tutte le osservazioni, tale misura è influenzata da valori estremi. La presenza di tali valori può condurre ad una rappresentazione distorta dei dati. In questi casi è opportuno ricorrere ad altre misure di posizione. A volte, rimuovendo qualche "valore estremo" e ricalcolando la media su un campione di taglia inferiore si può ottenere un valore più significativo.

Esempio 3.1 Si desidera calcolare la media dei rendimenti percentuali a un anno dei 17 fondi azionari della Tabella 3.1. In questo caso risulta

$$\overline{X} = 29,86$$

Se rimuoviamo l'outlier Mentor Merger risulta che

$$\overline{X} = 31,11$$

In questo caso, l'eliminazione dell'outlier comporta un aumento della media aritmetica.

Fondo	Rendimenti a 12 mesi (in %)	
Amcore Vintage Equità	32,2	
Baron Funds Asset	29,5	
Berger SmCoGrow	29,9	
Chicago Trust GrowInc	32,4	
Dodge and Cox DominiSo	30,5	
Federated Institut axCapSve	30,1	
First Funds GroInc III	32,1	
Harris Insight Inst Haven	35,2	
Mentor Merger	*10,0*	
Rainler Reich Tang	20,6	
Robertsson Stephens ValGrow	28,6	
SSgA SandP500idx	30,5	
ssgA SmallCap	38,0	
1784 Growinc	33,0	
Stagecoach	29,4	
Westwood Eq R	37,1	
Wright Yacktman	28,6	

Media campionaria = 29,86 Range = 28
Midrange=(38,0+10,0)/2 = 24,0
Q_1 = 29,0
Mediana = 30,5
Q_3= 32,7
Media interquartile = (Q_1 + Q_3) = 30,85
Range interquartile = 3,7
S^2 = 41,16
S = 6,42
CV =S/media campionaria = 21,5%

Tabella 3.1 Rendimenti percentuali a un anno per i fondi comuni azionari le cui commissioni sono prelevate dalle attività del fondo. In questo caso la media campionaria è 29,86 mentre se si escludono gli outlier è 31,11.

La **mediana** è il valore centrale in una successione ordinata di dati. In assenza di ripetizioni nei dati, metà delle osservazioni cadranno a sinistra della mediana e l'altra metà a destra. A differenza della media aritmetica, questo parametro non è influenzato dalla presenza di valori estremi. Per calcolare la mediana è necessario in primo luogo disporre i dati in ordine crescente. Successivamente, si sceglierà il valore che lascia alla sua sinistra il 50% delle osservazioni. Ciò può essere formalizzato attraverso la seguente espressione:

$$Mediana = osservazione\ di\ posto\ \frac{n+1}{2}\ nella\ serie\ ordinata$$

Per trovare la posizione occupata dalla mediana si possono usare le seguenti due regole di carattere generale:

- Se la taglia del campione è un numero *dispari* la mediana coincide con il valore centrale, ossia con l'osservazione che occupa la posizione *(n+1)/2* nella serie ordinata
- Se la taglia del campione è un numero *pari* il valore centrale si trova a metà tra le due osservazioni centrali della serie ordinata. La mediana coincide con la media dei valori corrispondenti a queste due osservazioni centrali.

Esempio 3.2 Consideriamo i dati del rendimento percentuale annuale conseguito dai fondi riportati in Tabella 3.1 e calcoliamone la mediana. I dati originari (grezzi) sono i seguenti:

32,2 29,5 29,9 32,4 30,5 30,1 32,1 35,2 10,0 20,6 28,6 30,5 38,0 33,0 29,4 37,1 28,6.

Per calcolare la mediana devono essere ordinati:

10,0	20,6	28,6	28,6	29,4	29,5	29,9	30,1	**30,5**	30,5	32,1	32,2	32,4	33,0	35,2	37,1	38,0
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Essendo (17+1)/2 = 9, segue che la mediana è la nona osservazione nella serie ordinata ossia 30,5.
In questo esempio è evidente che la mediana non è influenzata dall'osservazione estrema 10,0 nel senso che la mediana rimane la stessa indipendentemente dal fatto che l'osservazione più piccola sia 1 o 10 oppure 20. Inoltre, nel calcolo della mediana non si tiene conto delle ripetizioni. Più in particolare, la mediana nell'esempio in considerazione coincide con la nona osservazione anche se nella serie di dati considerata la decima è ugualmente 30,5.
Esempio 3.3 Supponiamo che il nostro campione sia costituito dal valore del patrimonio netto di 14 fondi azionari misti a bassa capitalizzazione. I valori "grezzi" coincidenti con i valori in dollari del patrimonio netto, per ciascuna quota dei fondi in questione sono forniti in Tabella.
In questo caso la serie ordinata delle osservazioni è la seguente:

7,35	11,62	14,07	14,09	16,95	17,30	**18,26**	**18,60**	20,34	21,17	21,69	24,01	26,10	37,61
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Poiché risulta che *(14+1)/2=7,5*, la mediana cade tra il settimo e l'ottavo valore nella serie ordinata delle osservazioni. In questo caso, come detto precedentemente, si sceglie come mediana la media aritmetica dei valori 18,26 e 18,60 ossia 18,43 ovviamente misurato in dollari.

Fondo	**Rendimento**	
X_1	7,35	
X_2	17,30	
X_3	11,62	Media campionaria = 19.16
X_4	26,10	Range = 30,26
X_5	21,69	Midrange=(37,61+7,35)/2 = 22,48
X_6	21,17	$Q_1 = X_4^* = 14,09$
X_7	14,07	Mediana = 18,43
X_8	14,09	$Q_3 = X_{11,25}^* = X_{11}^* = 21,69$
X_9	24,01	Media interquartile = $(Q_1 + Q_3) = 17,89$
X_{10}	20,34	Range interquartile = 7,6
X_{11}	18,26	$S^2 = 52,41$
X_{12}	37,61	S = 7,24
X_{13}	18,60	CV =S/media campionaria = 0,38%
X_{14}	16,95	

Tabella 3.2 Patrimonio netto di 14 fondi azionari misti a bassa capitalizzazione. I valori riportati sono i valori del patrimonio netto misurati in dollari

La **moda** è il valore più frequente in un insieme di dati. Anche la moda, a differenza della media, non è influenzata dagli outlier. Tuttavia, questa misura di posizione viene usata solo per scopi descrittivi, poiché è affetta da maggiore variabilità rispetto alle altre misure di posizione.

Esempio 3.4 Si desidera calcolare la moda dei rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo utilizzando la seguente serie ordinata (Esempio 3.2):

10,0 20,6 **28,6 28,6** 29,4 29,5 29,9 30,1 **30,5 30,5** 32,1 32,2 32,4 33,0 35,2 37,1 38,0

In questa serie si possono osservare due valori "più tipici": 28,6 e 30,5. questo insieme di dati si dice **bimodale**.

Esempio 3.5 Analizziamo la moda dei patrimoni netti (in dollari) dei 14 fondi misti a bassa capitalizzazione dell'Esempio 3.3:

7,35 11,62 14,07 14,09 16,95 17,30 18,26 18,60 20,34 21,17 21,69 24,01 26,10 37,61

In questo insieme di dati nessun valore di patrimonio netto è "più tipico" degli altri. Pertanto, in questo caso la moda non esiste.

Il **midrange** è la media tra la più piccola e la più grande delle osservazioni di un insieme di dati:

$$Midrange = \frac{X_{minimo} + X_{massimo}}{2}$$

Esempio 3.6 Calcoliamo il Midrange dei rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo considerati nell'Esempio 3.2. I dati ordinati sono:

10,0 20,6 28,6 28,6 29,4 29,5 29,9 30,1 **30,5** 30,5 32,1 32,2 32,4 33,0 35,2 37,1 38,0

Così che risulta

$$Midrange = (10,0+38,0)/2 = 24,0$$

Il Midrange è spesso utilizzato dagli analisti finanziari e dai meteorologi in quanto è semplice da calcolare ed è una misura adeguata per caratterizzare l'intero insieme dei dati. Comunque il suo utilizzo è vincolato ad alcune considerazioni. Quando si lavora con dati quali i costi quotidiani di chiusura delle azioni o le rilevazioni orarie della temperatura è probabile che i valori estremi non vengano osservati. In altre occasioni, è opportuno tener conto del fatto che il Midrange è basato esclusivamente sui valori estremi così che, quando sono presenti valori anomali questo indice potrebbe essere poco significativo per la sintesi dei dati.

I **quartili** sono un caso particolare di una struttura più generale di misure dette **quantili** che rappresentano le misure di posizione "non centrale" più ampiamente usate. Vengono usate in particolar modo per sintetizzare o descrivere le caratteristiche di ampi insiemi di dati quantitativi.

Come abbiamo già visto la mediana è il valore che divide a metà la serie delle osservazioni, i quartili invece sono misure che dividono i dati ordinati in quattro parti. Altri quantili frequentemente usati sono i decili che dividono i dati ordinati in dieci parti e i percentili che dividono i dati ordinati in cento parti.

Il **primo quartile** Q_1 è il valore tale che il 25% delle osservazioni è più piccolo di Q_1 mentre il 75% è più grande di Q_1.

$$Q_1 = osservazione\ di\ posto\ \frac{n+1}{4}\ nella\ serie\ ordinata$$

Il **terzo quartile** Q_3 è il valore tale che il 75% delle osservazioni è più piccolo di Q_3 mentre il 25% è più grande di Q_3.

$$Q_3 = osservazione\ di\ posto\ \frac{3(n+1)}{4}\ nella\ serie\ ordinata$$

Osserviamo che

- se il punto di posizionamento corrispondente a Q_1 (Q_3) è un numero intero si sceglie come quartile il valore dell'osservazione corrispondente
- se il punto di posizionamento corrispondente a Q_1 (Q_3) è a metà tra due numeri interi si sceglie come quartile la media delle osservazioni corrispondenti
- se il punto di posizionamento corrispondente a Q_1 (Q_3) non è né un numero intero né cade a metà tra due interi allora, una semplice regola consiste nell'approssimarlo per eccesso o per difetto all'intero più vicino e scegliere come quartile il valore numerico dell'osservazione corrispondente.

Esempio 3.6 Calcoliamo i quartili dei rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo considerati nell'Esempio 3.2. I dati ordinati sono:

10,0 20,6 28,6 28,6 29,4 29,5 29,9 30,1 30,5 30,5 32,1 32,2 32,4 33,0 35,2 37,1 38,0

Per questi dati si ha

$(n+1)/4= (17+1)/4= 4,5,$ $3(n+1)/4=3\ (17+1)/4= 13,5$

Essendo 4,5 a metà tra la quarta e la quinta osservazione, si può usare la prima regola e scegliere

$Q_1 =(28,6+29,4)/2=29,0.$

Similmente, essendo 13,5 a metà tra due osservazioni consecutive si può usare la prima regola e scegliere

$Q_3 =(32,4+33,0)/2=32,7.$

La **media interquartile** è una misura di sintesi che viene utilizzata per evitare i problemi che possono sorgere in presenza di valori estremi. Essa è data dalla media tra il primo e il terzo quartile dell'insieme dei dati

$$Media\ interquartile = \frac{Q_1 + Q_3}{2}$$

Esempio 3.7 Calcoliamo la media interquartile dei rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo considerati nell'Esempio 3.2. I dati ordinati sono:

10,0 20,6 28,6 28,6 29,4 29,5 29,9 30,1 30,5 30,5 32,1 32,2 32,4 33,0 35,2 37,1 38,0

Come visto nell'Esempio 3.6, $Q_1 = 29{,}0$ e $Q_3 = 32{,}7$, pertanto risulta che

Media interquartile = (29,0+32,7)/2 = 30,85.

La media interquartile, in quanto media tra Q_1 e Q_3, non è influenzata dagli outlier infatti nel calcolo non si utilizza nessuna delle osservazioni più piccole di Q_1 e nessuna delle osservazioni più grandi di Q_3. Misure di sintesi come la mediana e la media interquartile sono dette **misure robuste**.

Misure di variabilità

Un'altra caratteristica importante di un insieme di dati è la variabilità. Questa è la quantità di dispersione presente nei dati. Due insiemi di dati possono differire sia per posizione che per variabilità o anche per una sola delle due caratteristiche.

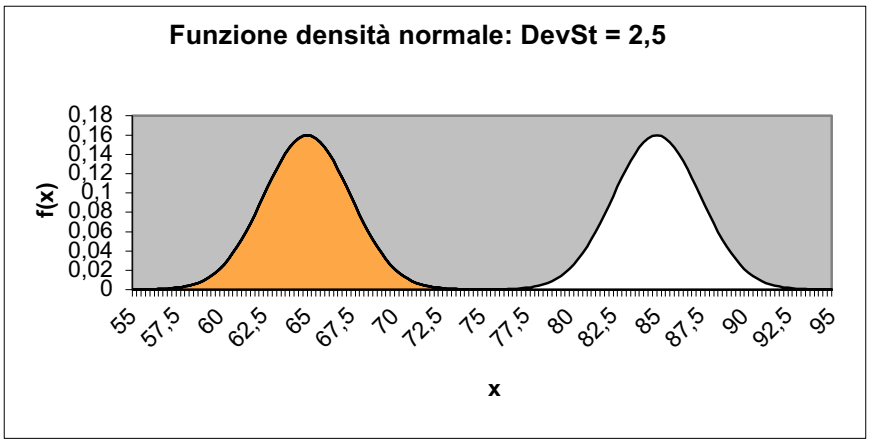


Figura 3.1 Due distribuzioni simmetriche a forma campanulare che differiscono solo nella posizione.

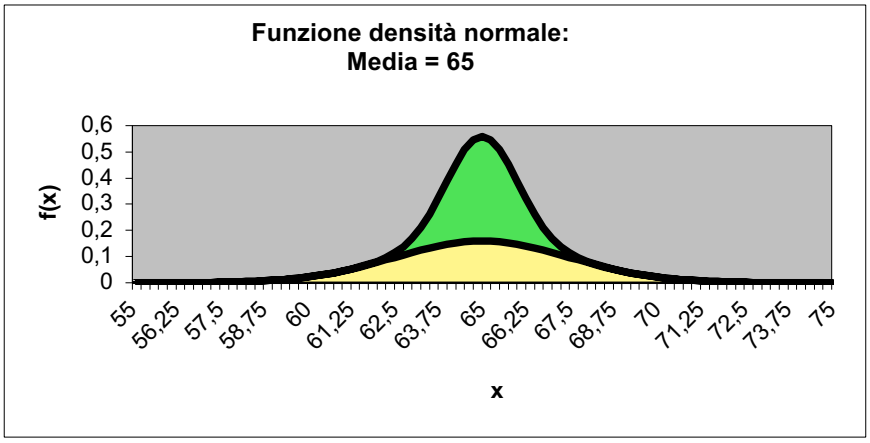


Figura 3.2 Due distribuzioni simmetriche a forma campanulare che differiscono solo nella variabilità.

Nel seguito prenderemo in esame cinque misure di dispersione: il *range*, il *range interquartile*, la *varianza*, lo *scarto quadratico medio* e il *coefficiente di variazione*.

Il **range** è la differenza tra l'osservazione più grande e quella più piccola in un insieme di dati. Il range deve assumere sempre valori positivi pertanto è così definito:

$$Range = | X_{più\ grande} - X_{più\ piccolo} |$$

Si tratta di una misura di dispersione totale nell'insieme dei dati. Un limite di tale misura consiste nel fatto che dipende da come i dati si distribuiscono effettivamente tra le due osservazioni estreme. Questo fatto spesso rende tale misura inadeguata ad esprimere la variabilità.

In presenza di osservazioni estreme non è opportuno considerare il range ma è più adeguato considerare il **range interquartile** così definito: Il range interquartile è, quindi, la differenza tra il terzo ed il primo quartile, misura la dispersione del 50% delle osservazioni che occupano le posizioni centrali e pertanto non è influenzato dai valori estremi.

Esempio 3.8 Consideriamo i rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni delle attività del fondo considerati nell'Esempio 3.2:

10,0 20,6 28,6 28,6 29,4 29,5 29,9 30,1 30,5 30,5 32,1 32,2 32,4 33,0 35,2 37,1 38,0

Per questi dati il range è *38 – 10 = 28* mentre, ricordando che nell'Esempio 3.6 si è calcolato $Q_1 = 29{,}0$ e $Q_3 = (32{,}4 + 33{,}0)/2 = 32{,}7$, il range interquartile risulta $32{,}7 - 29 = 3{,}7$. L'intervallo compreso tra i due quartili 29 e 32,7 comprende il 50% delle osservazioni centrali, l'ampiezza di tale intervallo è 3,7 e racchiude i rendimenti percentuali annui conseguiti dal gruppo centrale dei 17 fondi comuni azionari che prelevano le commissioni delle attività del fondo.

È opportuno osservare che né il range, né il range interquartile tengono conto di come le osservazioni si distribuiscono o si concentrano intorno ad una misura di tendenza centale, quale può essere ad esempio la media.

La **varianza** e lo **scarto quadratico medio** sintetizzano, invece, la dispersione dei valori osservati intorno alla media.

La varianza è *approssimativamente* la media dei quadrati degli scarti di ciascuna osservazione dalla media. Dato un campione di n osservazioni, $X_1, X_2, \ldots, X_n$, la **varianza campionaria** è

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Notiamo esplicitamente che se al denominatore di S^2 ci fosse stato n al posto di $n-1$ allora la varianza sarebbe stata coincidente con la media dei quadrati delle differenze intorno alla media. Si usa $n-1$ perché S^2gode di alcune proprietà che la rendono una misura adeguata nell'inferenza statistica. Ovviamente, aumentando il numero delle osservazioni diventa sempre meno rilevante la differenza tra n e $n-1$.

Lo scarto quadratico medio, denotato con S, è la radice quadrata della **varianza campionaria**:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Esempio 3.9 Consideriamo i rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni delle attività del fondo considerati nell'Esempio 3.2:

32,2 29,5 29,9 32,4 30,5 30,1 32,1 35,2 10,0 20,6 28,6 30,5 38,0 33,0 29,4 37,1 28,6.

La media campionaria per questo insieme di dati è 29,86 (vedi Esempio 3.1). Facendo uso della definizione si ottiene che

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$=[(32,2–29,86)^2+ (29,5–29,86)^2+ (29,9–29,86)^2+ (32,4–29,86)^2+ (30,5–29,86)^2 +(30,1–29,86)^2 + (32,1–29,86)^2+ (35,2–29,86)^2+ (10,0–29,86)^2+ (20,6–29,86)^2+ (28,6–29,86)^2+ (30,5–29,86)^2 +(38,0–29,86)^2+ (33,0–29,86)^2+ (29,4–29,86)^2+ (37,1–29,86)^2+ (28,6–29,86)^2]/(17–1)$

$=658,5592/16 = 41,15995$.

Inoltre, risulta che

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$=(41,15995)^{1/2}= 6,42.$

Nel calcolo della varianza campionaria le differenze tra ciascuna osservazione e la media sono elevate al quadrato, così che né la varianza né lo scarto quadratico medio possono essere negativi. Queste quantità possono essere nulle solo quando non c'è variabilità nei dati, cioè quando tutte le osservazioni nel campione sono uguali tra di loro e coincidono con la media campionaria. In questi casi eccezionali anche il range ed il range interquartile sono nulli. Generalmente, comunque, i dati sono intrinsecamente variabili, cioè non costanti: infatti, ogni fenomeno di interesse assume una varietà di valori. Proprio a causa di questa variabilità dei dati quantitativi diventa importante sintetizzare i dati non solo attraverso misure di posizione che ne sintetizzano i dati, ma anche attraverso misure di variabilità che ne sintetizzano la dispersione. La varianza e lo scarto quadratico medio misurano la dispersione media intorno alla media. La varianza soddisfa alcune importanti proprietà matematiche. Comunque la sua unità di misura coincide con il quadrato dell'unità di misura dei dati. Per questo motivo la misura di variabilità più usata è lo scarto quadratico medio che ha la stessa unità di misura dei dati.

Lo scarto quadratico medio aiuta a stabilire se e quando i dati sono concentrati intorno alla media. Molto spesso accade che la maggior parte dei dati osservati cade nell'intervallo centrato intorno alla media i cui estremi distano dalla media per uno scarto quadratico medio. In altri termini, l'intervallo $Media \pm S$ cattura almeno la maggior parte dei dati. In questo modo è evidente che la conoscenza della media aritmetica e dello scarto quadratico medio in genere aiuta a definire in quale intervallo si concentra la maggior parte dei dati osservati

Considerazioni circa le misure di variabilità

1. Quando più i dati sono dispersi, tanto maggiori sono il range, il range interquartile, la varianza e lo scarto quadratico medio.
2. Quando più i dati sono concentrati, o omogenei, tanto minori saranno il range, il range interquartile, la varianza e lo scarto quadratico medio.
3. Se le osservazioni sono tutte uguali (in modo che non vi sia variabilità nei dati) il range, il range interquartile, la varianza e lo scarto quadratico medio sono tutti nulli.
4. Nessuna delle misure di variabilità (il range, il range interquartile, la varianza e lo scarto quadratico medio) può essere negativa.

A differenza delle altre misure di variabilità, il *coefficiente di variazione* è una misura relativa, espressa come una percentuale e non nell'unità di misura dei dati.

Il **coefficiente di variazione**, denotato con CV è dato dal rapporto tra lo scarto quadratico medio e la media aritmetica il tutto moltiplicato per 100%:

$$CV = \left(\frac{S}{\overline{X}}\right) 100\%$$

Esempio 3.10 Consideriamo i rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni delle attività del fondo considerati nell'Esempio 3.2:

32,2 29,5 29,9 32,4 30,5 30,1 32,1 35,2 10,0 20,6 28,6 30,5 38,0 33,0 29,4 37,1 28,6.

La media campionaria per questo insieme di dati è 29,86 (vedi Esempio 3.1), mentre lo scarto quadratico medio è 6,42 (Esempio 3.9). Pertanto risulta:

$$CV = \left(\frac{S}{\overline{X}}\right) 100\% = \left(\frac{6,42}{29,86}\right) 100\% = 21,5\%$$

Per questo campione la "diffusione media intorno alla media" è pari al 21,5%.

Il coefficiente di variazione è particolarmente utile quando si misurano le variabilità di due o più insiemi di dati che sono espressi in unità di misura diversi. Qui di seguito sono mostrati alcuni esempi.

Esempio 3.11 Supponiamo che il direttore operativo di un'azienda di consegna pacchi stia valutando la possibilità di acquisto di un nuovo parco di autocarri. Quando i pacchi sono depositati negli autocarri si deve tener conto di due valori: il peso del pacco (misurato in chilogrammi) e il suo volume (misurare in metri cubi).

Supponiamo che in un campione di *200* pacchi il peso medio sia di *9 kg* con uno scarto quadratico di *1,5 kg* e che il volume medio dei pacchi sia di *2,7 mq* con uno scarto quadratico medio di *0,6 mq*. Si vuole confrontare la variabilità del peso e del volume.

Osserviamo in primo luogo che il peso e il volume sono espressi in unità di misura diverse, per cui il direttore operativo dovrebbe analizzare la variabilità relativa delle osservazioni.

Per il peso risulta *CV = (1,5/9) 100%= 16,67%*

Per il volume si ha *CV = (0,6/2,7) 100%= 22,22%*

Quindi, rispetto alla media, il volume è più variabile del peso.

Il coefficiente di variazione è molto utile anche quando si confrontano due o più insiemi di dati che sono misurati nella stessa unità di misura, ma che differiscono nella dimensione tanto da rendere poco significativo il confronto tra i rispettivi scarti quadratici medi. Questo è mostrato nel seguente esempio.

Esempio 3.12 Supponiamo che un potenziale investitore sia indeciso se acquistare le azioni della società *A* o quelle della società *B*. Se nessuna delle due società distribuisce dividendi ai suoi azionisti e se entrambe sono valutate molto positivamente (da varie società di servizi azionari) in termini di crescita potenziale, l'investitore potrebbe prendere in considerazione la *volatilità* (variabilità) dei due titoli per prendere una decisione.

Supponiamo che le azioni di *A* abbiano conseguito un prezzo medio di *50 $* nel corso degli ultimi mesi con un scarto quadratico medio di *10 $*. Immaginiamo, inoltre, che nello stesso periodo le azioni di *B* abbiano avuto un prezzo medio di *12 $* con uno scarto quadratico medio di *4 $*. Vogliamo stabilire quale dei due titoli è più variabile.

Considerando lo scarto quadratico medio, il prezzo delle azioni di *A* sembra più volatile di quello delle azioni di *B*, però i prezzi medi sono molto diversi. Per valutare la volatilità/stabilità dei due titoli è più appropriato analizzare la variabilità del prezzo rispetto al prezzo medio. Si ha:

Per la società *A* il $CV = (10\ \$/50\ \$)\ 100\% = 20{,}0\%$

Per la società *B* il $CV = (4\ \$/12\ \$)\ 100\% = 33{,}3\%$

Pertanto, si può concludere che relativamente alla media, il prezzo del titolo della società *B* è più variabile del prezzo del titolo della società *A*.

La forma della distribuzione

La terza caratteristica dei dati che prendiamo in considerazione è la **forma** della loro distribuzione. Il modo con cui i dati si distribuiscono può essere simmetrico o meno. Per descrivere la forma della distribuzione è sufficiente confrontare la media con la mediana. Se queste due misure sono uguali, la distribuzione è simmetrica. Se la media è maggiore della mediana la distribuzione è asimmetrica a destra, viceversa, se la media è minore della mediana allora la distribuzione è asimmetrica a sinistra. Lo schema qui sotto riportato riassume la situazione.

Media > *Mediana*: asimmetria positiva corrispondente ad una distribuzione obliqua a destra
Media = *Mediana*: simmetria
Media < *Mediana*: asimmetria negativa corrispondente ad una distribuzione obliqua a sinistra

Un'asimmetria positiva si verifica tutte le volte che la media supera la mediana a causa della presenza di valori eccezionalmente alti; in questo caso la distribuzione presenterà una coda molto lunga a destra e si dice anche obliqua a destra. Un'asimmetria negativa si presenta invece in corrispondenza di una media inferiore alla mediana per la presenza di valori estremamente bassi; in questo caso la distribuzione esibisce una coda più lunga a sinistra e si dice anche obliqua a sinistra. Invece, i dati sono distribuiti in modo simmetrico quando la media e la mediana sono uguali; in questo caso la distribuzione è perfettamente simmetrica rispetto alla media.

Esempio 3.13 Consideriamo i rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni delle attività del fondo considerati nell'Esempio 3.2:

32,2 29,5 29,9 32,4 30,5 30,1 32,1 35,2 10,0 20,6 28,6 30,5 38,0 33,0 29,4 37,1 28,6.

Cosa si può dire rispetto alla forma della distribuzione?

Osserviamo in primo luogo che in questo insieme di dati c'è un outlier e che i 17 fondi non si raggruppano intorno alla media che sappiamo essere pari a *29,82*. il rendimento percentuale annuo conseguito da Mentor Menger è *10,0* che risulta molto più piccolo della media. Inoltre, la mediana è stata calcolata pari a *30,5*. Pertanto, essendo la media minore della mediana, i dati si distribuiranno in modo asimmetrico con una coda più lunga a sinistra.

Osservazione L'opzione *statistica descrittiva* del componente aggiuntivo Analisi dati di Excel calcola tutte le misure che abbiamo descritto fino ad ora: la media, la mediana, la moda, la deviazione standard (lo scarto quadratico medio) la varianza, il range, il massimo, il minimo, il conteggio (l'ampiezza del campione). Excel calcola anche altre misure come l'*errore standard* che si ottiene dividendo la deviazione standard per la radice quadrata della dimensione del campione. L'*asimmetria* misura la mancanza di simmetria nella distribuzione e si basa su una funzione del cubo delle differenze dalla media. La *curtosi* è una misura della concentrazione relativa dei valori al centro della distribuzione rispetto alla concentrazione sulle code e si basa sulle differenze dalla media elevate alla quarta.

Consumi	*Consumi*	
27		
29	Media	23,471
33	Errore standard	2,235
21	Mediana	21
21	Moda	21
12	Deviazione standard	9,214
16	Varianza campionaria	84,890
25	Curtosi	–0,547
8	Asimmetria	0,361
17	Intervallo	33
24	Minimo	8
34	Massimo	41
38	Somma	399
15	Conteggio	17
19	Più grande(4)	33
19	Più piccolo(4)	16
41	Livello di confidenza(90,0%)	3,901

Tabella 3.3: Riepilogo di statistica descrittiva calcolato su un insieme di dati rappresentati i consumi di 17 autovetture.

32,2
29,5
29,9
32,4
30,5
30,1
32,1
35,2
10
20,6
28,6
30,5
38
33
29,4
37,1
28,6

Media	29,86470588
Errore standard	1,556011615
Mediana	30,5
Moda	30,5
Deviazione standard	6,415600242
Varianza campionaria	41,15992647
Curtosi	5,546604844
Asimmetria	− 2,011159015
Intervallo	28
Minimo	10
Massimo	38
Somma	507,7
Conteggio	17

Tabella 3.4: Riepilogo di statistica descrittiva calcolato sui rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni delle attività del fondo considerati nell'Esempio 3.2

L'analisi esplorativa dei dati

Vogliamo ora affrontare il problema di come sintetizzare opportunamente le caratteristiche dei dati. Un approccio a questa "analisi esplorativa dei dati" consiste nel calcolare i cinque numeri di sintesi e nel costruire il *diagramma scatola e baffi* (*box and whisker plot*).

I cinque numeri di sintesi sono

$$X_{min} \qquad Q_1 \qquad \text{Mediana} \qquad Q_3 \qquad X_{max}$$

A partire da questi valori è possibile ottenere tre misure di posizione (la mediana, la media interquartile e il midrange) e due misure di variabilità (il range, il range interquartile) che consentono di avere un'idea più precisa della forma della distribuzione. In particolare si ha:

Caso di dati simmetrici

1. La distanza tra Q_1 e la mediana è uguale alla distanza tra Q_3 e la mediana
2. La distanza tra X_{min} e Q_1 è uguale ala distanza tra Q_3 e X_{max}
3. La mediana, la media interquartile e il midrange sono tutti uguali e coincidono anche con la media dei dati

Caso di dati asimmetrici

1. Nelle distribuzioni oblique a destra la distanza tra Q_3 e X_{max} è maggiore della distanza tra X_{min} e Q_1
2. Nelle distribuzioni oblique a destra la mediana e la media interquartile sono minori del midrange
3. Nelle distribuzioni oblique a sinistra la distanza tra Q_3 e X_{max} è minore della distanza tra X_{min} e Q_1
4. Nelle distribuzioni oblique a sinistra la mediana e la media interquartile sono maggiori del midrange

Esempio 3.14 Consideriamo i rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni delle attività del fondo considerati nell'Esempio 3.2:

10,0 20,6 28,6 28,6 29,4 29,5 29,9 30,1 30,5 30,5 32,1 32,2 32,4 33,0 35,2 37,1 38,0

Vogliamo calcolare i cinque numeri di sintesi per questo insieme di dati.

Sappiamo che la mediana è 30,5 il primo quartine è 29 e il terzo è pari a 32,7. Pertanto, i cinque numeri di sintesi sono **X_{min} =10,0 Q_1=29 Mediana =30,5 Q_3=32,7 X_{max} =38.**

Usiamo questi valori per studiare la forma della distribuzione. In base alle tabelle di sintesi che abbiamo riportato risulta che la distribuzione del rendimento ad un anno è obliqua a sinistra essendo $Q_1 - X_{min} = 19 > X_{max} - Q_3 = 5,3$. Inoltre, confrontando la mediana con la media interquartile = 30,85 e il midrange = 24, notiamo che il midrange è influenzato dalla presenza dell'outlier 10 e risulta quindi molto più piccolo delle altre due misure di sintesi, mentre la mediana e la media interquartile che sono misure più robuste perché non influenzate degli outlier, assumono valori più vicini.

Il diagramma scatola e baffi (box-plot)

Il diagramma scatola e baffi fornisce una rappresentazione grafica dei dati sulla base dei cinque numeri di sintesi. Si tratta di riportare su un'asse graduata i cinque parametri. Come già detto, sulla base di questi valori si possono individuare alcune caratteristiche grafiche della distribuzione.

In Figura 3.1 è riportato il diagramma scatola e baffi per il rendimento percentuale ad un anno conseguito dai 17 fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo di cui all'Esempio 3.2. Per i dati in questione si ha (cf. Esempio 3.14)

X_{min} =10,0 Q_1=29 Mediana =30,5 Q_3=32,7 X_{max} =38.

In Figura 3.1 sono rappresentati $Q_1 = 29$ Mediana = 30,5 $Q_3 = 32,7$. La linea verticale all'interno della scatola rappresenta la mediana, la linea verticale a sinistra indica Q_1, la linea a destra Q_3. Pertanto, la scatola contiene il 50% delle osservazioni della distribuzione. Il 25% dei dati con valori più piccoli è rappresentato con una linea tratteggiata (un baffo) che collega il lato sinistro della scatola con il valore X_{min}; similmente, il 25% dei dati con valore più grande è rappresentato dall'altro baffo che collega il lato destro della scatola con X_{max}. Dal grafo si evince che, sebbene la mediana sia più

vicina al lato sinistro della scatola, la distribuzione è obliqua a sinistra essendo il baffo sinistro più lungo di quello destro. Ciò è dovuto alla presenza dell'outlier.

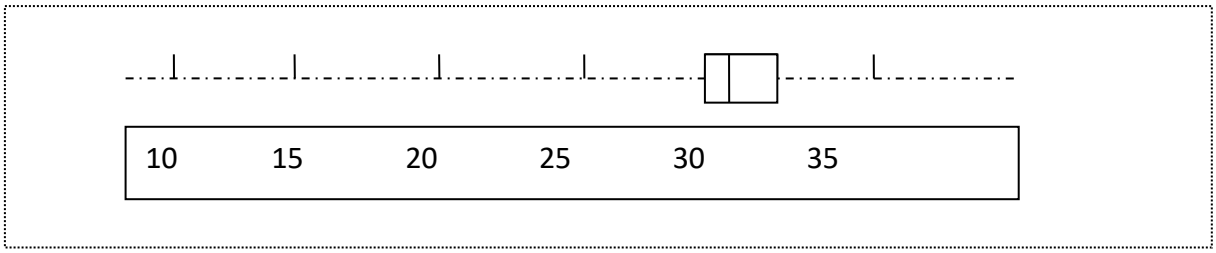


Figura 3.1: Il diagramma scatola e baffi del rendimento percentuale a un anno conseguito da 17 fondi comuni azionari le cui commissioni sono prelevate direttamente dall'attivo di gestione (Esempio 3.1). In questo caso $Q_1 = 29{,}0$, $Q_2 = 30{,}5$, $Q_3 = 32{,}7$.

Più in generale è possibile dire che

Se i dati sono perfettamente simmetrici i due baffi hanno uguale lunghezza e la linea della mediana divide a metà la scatola. Se ciò si verifica con buona approssimazione diremo che i dati sono "approssimativamente simmetrici".

Quando la distribuzione è obliqua a sinistra, le poche osservazioni piccole distorcono il midrange e la media verso sinistra. In questo caso si ha:

midrange < media < media interquartile < mediana < moda.

Ciò comporta una forte concentrazione delle osservazioni intorno ai valori più elevati sull'asse graduato visto che il 75% delle osservazioni cade tra il lato sinistro della scatola (Q_1) e la fine del baffo destro (X_{max}). Così che sul baffo sinistro cade il 25% delle osservazioni più piccole. Da ciò deriva che la distribuzione è obliqua a sinistra.

Se la distribuzione dei dati è obliqua a destra, allora sono le poche osservazioni grandi che provocano una distorsione dei midrange e della media verso destra. In questo caso si ha

moda < mediana < media interquartile< media < midrange

Ciò comporta una forte concentrazione delle osservazioni intorno ai valori più bassi sull'asse graduato visto che il 75% delle osservazioni cade tra l'inizio del baffo sinistro (X_{min}) e il lato destro della scatola (Q_3). Così che sul baffo destro cade il 25% delle osservazioni maggiori. Da ciò deriva che la distribuzione è obliqua a destra.

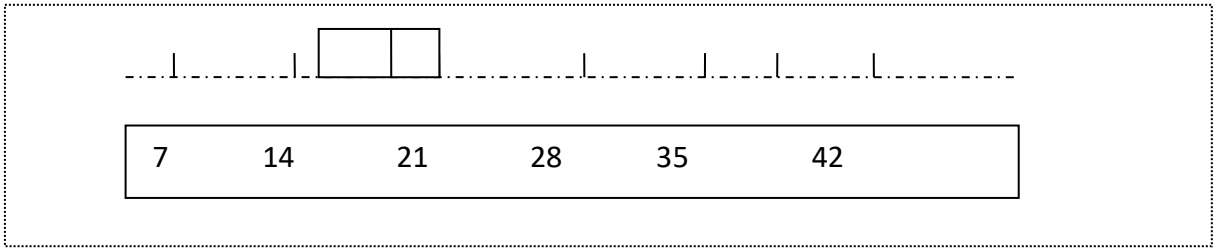


Figura 3.2: Il diagramma scatola e baffi del patrimonio netto di 14 fondi azionari misti a bassa capitalizzazione misurati in dollari (Esempio 3.3). In questo caso $Q_1 = 14{,}09$, $Q_2 = (18{,}26+18{,}60)/2 = 18{,}43$, $Q_3 = 21{,}69$.

Il diagramma illustrato in Figura 3.3 sintetizza quanto detto nel presente capitolo.

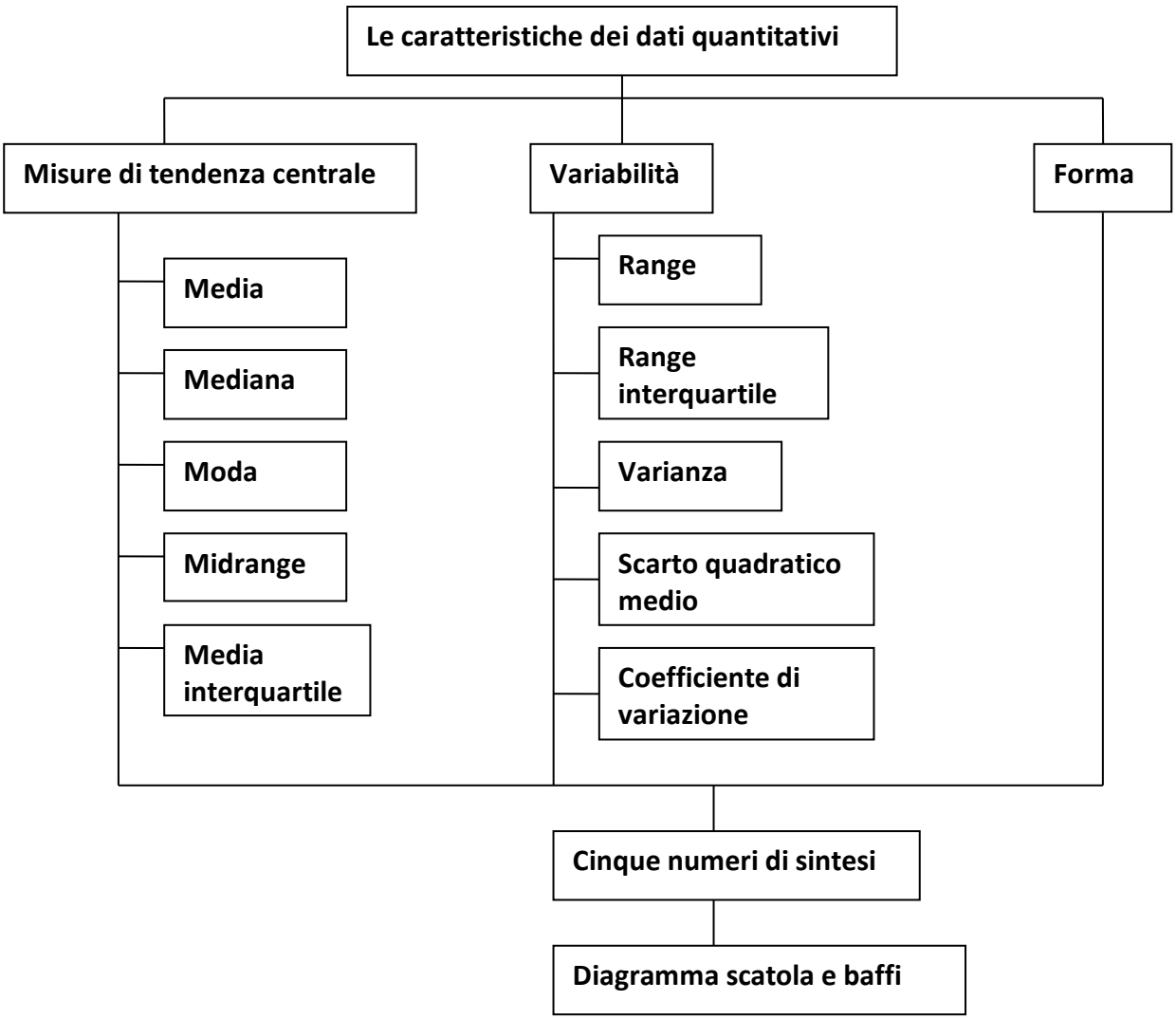


Figura 3.3: Sintesi del presente capitolo.