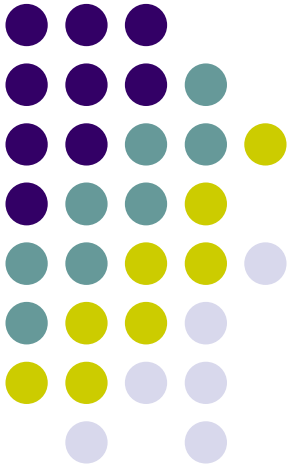
Levine, Krehbiel, Berenson

Statistica

Capitolo 3

Sintesi e descrizione dei dati quantitativi



Insegnamento: Statistica (gruppo C)

Corso di Laurea Triennale in Economia

Università degli Studi di Ferrara

Docente: Dott.ssa A. Grassi

Si ringrazia il Prof. S. Bonnini per aver condiviso le slide del suo corso

Argomenti

- Gli indici statistici di sintesi
  - Misure di tendenza centrale
  - Misure di dispersione
  - Misure di forma
- Misure di sintesi descrittive per una popolazione
- Analisi esplorativa dei dati: il diagramma a “scatola e baffi” (o boxplot)
- La covarianza ed il coefficiente di correlazione

Indici statistici di sintesi

Per trarre delle indicazioni adeguate quando si considerano dati quantitativi, non è sufficiente rappresentare i dati mediante tabelle e grafici di frequenza.

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

Tipi di indici:

- Misure di **tendenza centrale**
- Misure di **variabilità**
- Misure di **forma**

Misure di Tendenza Centrale

Nella maggior parte degli insiemi di dati, le osservazioni mostrano una tendenza a raggrupparsi attorno a un valore centrale.

Risulta in genere quindi possibile selezionare un valore tipico per descrivere un intero insieme di dati.

Tale valore descrittivo è una misura di posizione o di tendenza centrale.

Tipi di misure di tendenza centrale:

- Media
- Mediana
- Moda

Misure di Tendenza Centrale: la Media

La **media aritmetica** (anche chiamata semplicemente **media**) è la misura di posizione più comune. Si calcola dividendo la somma dei valori osservati per il numero totale di osservazioni.

La media aritmetica

La media aritmetica è la somma dei valori divisa per il numero dei valori.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (3.1)$$

dove

$\overline{X}$ = media aritmetica campionaria

n = ampiezza del campione

X_i = i–esima osservazione della variabile casuale X

$\sum_{i=1}^{n} X_i$ = somma di tutti i valori X_i del campione

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \cdots + X_n$$

Misure di Tendenza Centrale: la Media

Un esempio: studiamo i 17 fondi comuni azionari che prelevano le commissioni di commercializzazione direttamente dalle attività del fondo (Group = 1).

Tabella 3.1 *Rendimenti percentuali a un anno per i fondi comuni azionari le cui commissioni sono prelevate dalle attività del fondo*

FONDO	RENDIMENTI A DODICI MESI (IN %)
Amcore Vintage Equity	32.2
Baron Funds Asset	29.5
Berger SmCoGrow	29.9
Chicago Trust GrowInc	32.4
Dodge & Cox DominiSo	30.5
Federated Institut MaxCapSvc	30.1
First Funds GroInc III	32.1
Harris Insight Inst Haven	35.2
Mentor Merger	10.0
Rainler Reich Tang	20.6
Robertson Stephens ValGrow	28.6
SSgA S&P500Idx	30.5
SSgA SmallCap	38.0
1784 GrowInc	33.0
Stagecoach CorpStk	29.4
Westwood Eq R	37.1
Wright Yacktman	28.6

La media aritmetica per questo campione è calcolata come segue:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{32.2 + 29.5 + 29.9 + \cdots + 28.6}{17} = 29.86$$

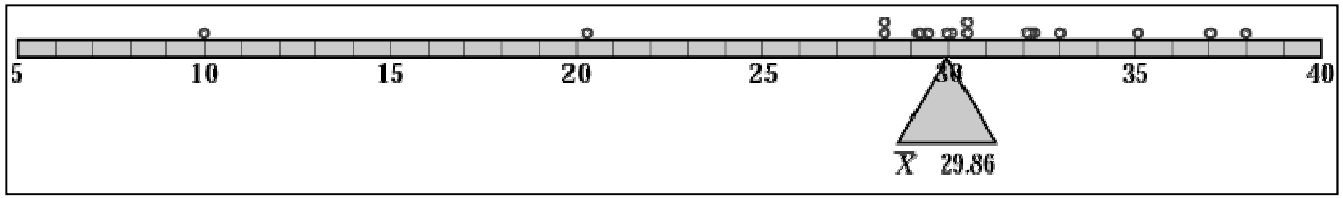
- La media si presenta come un “punto di equilibrio” tale che le osservazioni più piccole bilanciano quelle più grandi.
- Il calcolo della media si basa su tutte le osservazioni (X_1, X_2, X_3, . . . , X_n) dell’insieme di dati, proprietà questa che non è presentata da nessun’altra misura di posizione comunemente usata.

Misure di Tendenza Centrale: la Media

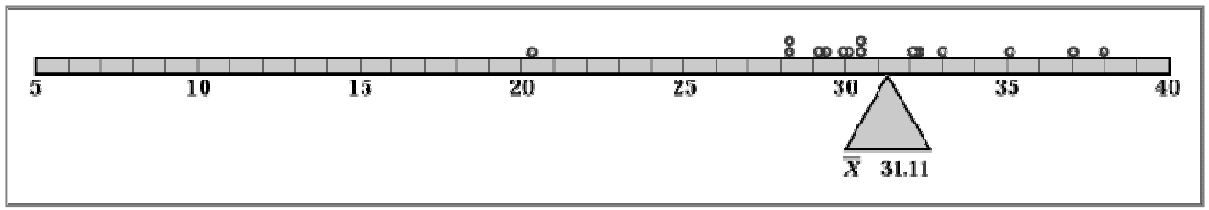
***Commento**: quando usare la media aritmetica*

Proprio perché il calcolo della media si basa su tutte le osservazioni, tale misura di posizione risulta influenzata da valori estremi.

In presenza di valori estremi, la media aritmetica fornisce una rappresentazione distorta dei dati ed è pertanto opportuno in questi casi ricorrere ad altre misure di posizione.



Se dal campione rimuoviamo il fondo Mentor Merger (rendimento=10.0), che possiamo considerare come un *outlier* (dato anomalo), ricalcolando la media otteniamo un valore pari a 31.11.



Misure di Tendenza Centrale: la Mediana

La **mediana** è il valore centrale in un insieme di dati ordinati dal valore più piccolo al più grande (cioè in ordine non decrescente).

La mediana

La mediana è l'osservazione che, nella serie ordinata dei dati, si lascia alla destra il 50% delle osservazioni e a sinistra il 50% delle osservazioni. Quindi, il 50% delle osservazioni risulteranno maggiori della mediana e il 50% risulteranno minori della mediana.

$$\text{Mediana} = \text{osservazione di posto } \frac{n+1}{2} \text{ nella serie ordinata} \qquad (3.2)$$

***Commento**:* La mediana non è influenzata dalle osservazioni estreme di un insieme di dati. Nel caso di osservazioni estreme è quindi opportuno descrivere l'insieme di dati con la mediana piuttosto che con la media.

Misure di Tendenza Centrale: la Mediana

Per trovare la posizione occupata dal valore mediano nella serie ordinata delle osservazioni si usa l'equazione (3.2) secondo una delle due regole seguenti:

REGOLA 1. Se l'ampiezza del campione è un numero **dispari**, la mediana coincide con il valore centrale, vale a dire con l'osservazione che occupa la posizione (n+1)/2 nella serie ordinata delle osservazioni.

REGOLA 2. Se l'ampiezza del campione è un numero **pari**, la mediana allora coincide con la media dei valori corrispondenti alle due osservazioni centrali.

Misure di Tendenza Centrale: la Mediana

Esempio 3.3 *Il calcolo della mediana in un campione di ampiezza dispari*

Nel nostro esempio del rendimento percentuale a un anno conseguito dai fondi comuni azionari che prelevano le commissioni di commercializzazione direttamente dalle attività del fondo, i dati grezzi sono:

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

Calcolate la mediana.

SOLUZIONE

La serie ordinata è:

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

↑
Mediana

Posizione ↑

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Mediana = 30.5

Per questi dati il valore centrale coincide con la nona osservazione nella serie ordinata [ossia, $(n + 1)/2 = (17 + 1)/2 = 9$]. Pertanto la mediana è 30.5.

Misure di Tendenza Centrale: la Moda

La **moda** è il valore più frequente in un insieme di dati.

- A differenza della media, la moda non è influenzata dagli outlier.
- Tuttavia tale misura di posizione viene usata solo per scopi descrittivi, poiché è caratterizzata da maggiore variabilità rispetto alle altre misure di posizione (piccole variazioni in un insieme di dati possono far variare in modo consistente la moda).

Esempio 3.5 *Il calcolo della moda*

Calcolate la moda dei rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo utilizzando la serie ordinata nell'esempio 3.3.

Misure di Tendenza Centrale: la Moda

Esempio 3.5 *Il calcolo della moda*

Calcolate la moda dei rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo utilizzando la serie ordinata nell'esempio 3.3.

SOLUZIONE

La serie ordinata per questi dati è la seguente:

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

Possiamo osservare che ci sono due valori "più tipici" o due mode: 28.6 e 30.5. Questo insieme di dati si dice *bimodale*.

NOTA: un insieme di dati può non avere moda, se nessuno valore è "più tipico".

Misure di Tendenza Centrale: i Quartili

Mentre la mediana è un valore che divide a metà la serie ordinata delle osservazioni, i **quartili** sono misure descrittive che dividono i dati ordinati in quattro parti.

Il primo quartile, Q_1

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

$$Q_1 = \text{osservazioni di posto } \frac{(n+1)}{4} \text{ nella serie ordinata} \qquad (3.4)$$

Il terzo quartile, Q_3

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

$$Q_3 = \text{osservazioni di posto } \frac{3(n+1)}{4} \text{ nella serie ordinata} \qquad (3.5)$$

Misure di Tendenza Centrale: i Quartili

Tre sono le regole usate per il calcolo dei quartili.

- *REGOLA 1*. Se il punto di posizionamento è un numero intero, si sceglie come quartile il valore dell'osservazione corrispondente.
- *REGOLA 2*. Se il punto di posizionamento è a metà tra due numeri interi, si sceglie come quartile la media delle osservazioni corrispondenti.
- *REGOLA 3.* Se il punto di posizionamento non è né un intero né a metà tra due numeri interi, una regola semplice consiste nell'approssimarlo per eccesso o per difetto all'intero più vicino e scegliere come quartile il valore numerico dell'osservazione corrispondente.

Misure di Tendenza Centrale: i Quartili

Esempio 3.8 *Il calcolo dei quartili*

Calcolate i quartili dei rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni dalle attività del fondo considerati nell'esempio 3.3.

SOLUZIONE

La serie ordinata è

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 abbiamo

$$Q_1 = \frac{n+1}{4} \text{–esima osservazione ordinata}$$

$$= \frac{17+1}{4} = 4.5\text{–esima osservazione ordinata}$$

Pertanto Q_1, usando la regola 2, può essere approssimato con la media tra la quarta e la quinta osservazione nella serie ordinata.

$$Q_1 = \frac{28.6 + 29.4}{2} = 29.0$$

$$Q_3 = \frac{3(n+1)}{4} \text{–esima osservazione ordinata}$$

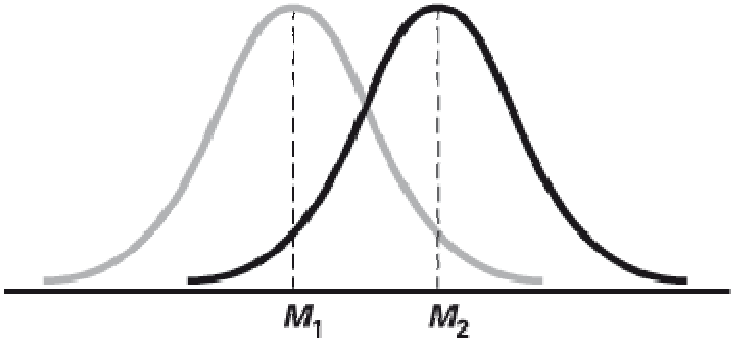
$$= \frac{3(17+1)}{4} = 13.5\text{–esima osservazione ordinata.}$$

Pertanto Q_3, usando la regola 2, può essere approssimato con la media tra la tredicesima e la quattordicesima osservazione nella serie ordinata.

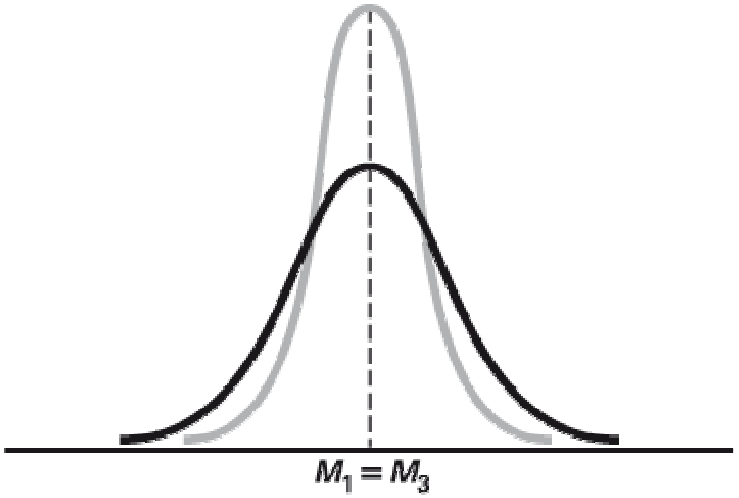
$$Q_3 = \frac{32.4 + 33.0}{2} = 32.7$$

Misure di dispersione

Una seconda caratteristica importante di un insieme di dati è la **variabilità**. La variabilità è la quantità di dispersione presente nei dati. Due insiemi di dati possono differire sia nella posizione che nella variabilità; oppure possono essere caratterizzati dalla stessa variabilità, ma da diversa misura di posizione; o ancora, possono essere dotati della stessa misura di posizione, ma differire notevolmente in termini di variabilità.



(*a*) Due distribuzioni simmetriche a forma campanulare che differiscono solo nella posizione



(*b*) Due distribuzioni simmetriche a forma campanulare che differiscono solo nella variabilità

Misure di dispersione

Le misure più utilizzate per misurare la dispersione sono:

- Intervallo di variazione (range)
- Intervallo di variazione interquartile (interquartile range)
- Varianza
- Scarto Quadratico Medio (o deviazione standard)
- Coefficiente di Variazione

Misure di dispersione: il Range

Il **range** (o **intervallo di variazione**) è la differenza tra l'osservazione più grande e quella più piccola in un insieme di dati.

Il range

Il range è uguale all'osservazione più grande meno quella più piccola.
È importante sottolineare che il range deve assumere sempre valori maggiori di zero. Quindi se la quantità $(X_{\text{più grande}} - X_{\text{più piccola}})$ risulta minore di zero, dobbiamo considerarne l'opposto, $-(X_{\text{più grande}} - X_{\text{più piccola}})$. In definitiva quindi:

$$\text{Range} = | X_{\text{più grande}} - X_{\text{più piccola}} | \qquad (3.7)$$

(Quando inseriamo un certo valore tra le due barrette, |.|, significa che stiamo considerando il *valore assoluto* della quantità considerata; ad esempio, $|3| = 3$ e $|-3| = 3$).

La serie ordinata è

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 $38.0 - 10.0 = 28.0$.

NOTA: un limite del range consiste nel fatto che non tiene conto di come i dati si distribuiscono effettivamente tra il valore più piccolo e quello più grande.

Per questo motivo, in presenza di osservazioni estreme, risulta una misura inadeguata della variabilità.

Misure di dispersione: il Range Interquartile

Il **range** (o **intervallo**) **interquartile** è la differenza tra il terzo e il primo quartile in un insieme di dati.

Il range interquartile

Il range interquartile si ottiene sottraendo al terzo quartile il primo quartile. Anche il range interquartile deve essere sempre maggiore di zero. Quindi:

$$\text{Range interquartile} = |Q_3 - Q_1| \qquad (3.8)$$

NOTA: Questa misura di variabilità sintetizza la dispersione del 50% delle osservazioni che occupano le posizioni centrali, e non è pertanto influenzata da valori estremi.

La serie ordinata è

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 sappiamo già dall'esempio 3.8 che $Q_1 = 29.0$ e $Q_3 = 32.7$. In base all'equazione (3.8) abbiamo:

$$\text{Range interquartile} = 32.7 - 29.0 = 3.7$$

L'intervallo compreso tra i due quartili 29 e 32.7 racchiude il 50% delle osservazioni centrali. L'ampiezza di tale intervallo, 3.7, racchiude i rendimenti percentuali annui conseguiti dal *gruppo centrale* dei 17 fondi comuni azionari che prelevano le commissioni dalle attività del fondo

Misure di dispersione: la Varianza

Sebbene il range sia una misura della dispersione totale e il range interquartile della dispersione centrale, nessuna di queste due misure tiene conto di come le osservazioni si distribuiscano o si concentrino intorno a una misura di tendenza centrale, come ad esempio la media.

Due misure della variabilità che forniscono questo tipo di informazione sono la **varianza** e la sua radice quadrata, lo **scarto quadratico medio**. Queste misure sintetizzano la dispersione di valori osservati attorno alla loro media.

Misure di dispersione: la Varianza

La **varianza** e la sua radice quadrata, lo **scarto quadratico medio**, sintetizzano la dispersione dei valori osservati attorno alla loro media.

La varianza campionaria

La varianza campionaria è la somma dei quadrati delle differenze dalla media divisa per $(n - 1)$:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$$

dove

$\overline{X}$ = media aritmetica campionaria

n = ampiezza del campione

X_i = i–esima osservazione della variabile casuale X

$\sum_{i=1}^{n}(X_i - \overline{X})^2$ = somma dei quadrati delle differenze tra i valori X_i e $\overline{X}$

Misure di dispersione: la Deviazione Standard

Lo scarto quadratico medio (o deviazione standard)

Lo scarto quadratico medio campionario (detto anche deviazione standard) è la radice quadrata della varianza campionaria:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \qquad (3.10)$$

Esempio 3.12 *Il calcolo della varianza campionaria e dello scarto quadratico medio campionario*

Per il campione contenente i 17 fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo, i dati grezzi relativi ai rendimenti percentuali annui 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

La media aritmetica per questo campione è pari a $\overline{X} = 29.86$. Calcolate la varianza campionaria, S^2, e lo scarto quadratico medio campionario, S.

Soluzione

Per calcolare S^2 seguiamo la procedura indicata nel Riquadro 3.1, riportata nella tabella a pagina seguente.

Utilizzando la formula (3.9), si ottiene che la varianza campionaria è:

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

$$= \frac{(32.2 - 29.86)^2 + (29.5 - 29.86)^2 + (29.9 - 29.86)^2 + \cdots + (28.6 - 29.86)^2}{17 - 1}$$

$$= \frac{658.5592}{16}$$

$$= 41.15995$$

Dall'equazione (3.10), lo scarto quadratico medio, S, risulta pari a

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} = \sqrt{41.15995} = 6.42$$

Interpretazione della Varianza e dello Scarto Quadratico Medio

- La varianza e lo scarto quadratico medio misurano la dispersione “media” attorno alla media: sono ottenute “valutando’’ come le osservazioni più grandi oscillano sopra la media e come le osservazioni più piccole si distribuiscono al di sotto della media.

- La varianza possiede alcune importanti proprietà matematiche; tuttavia, la sua unità di misura coincide con il quadrato dell’unità di misura dei dati (euro al quadrato, metri al quadrato e così via). Mentre lo scarto quadratico medio è espresso nell’unità di misura originaria dei dati (euro, metri, …).

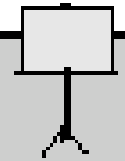
Interpretazione della Varianza e dello Scarto Quadratico Medio

- Lo scarto quadratico medio ci aiuta a stabilire se e quanto i dati sono concentrati o dispersi intorno alla loro media.
- Per quasi tutti gli insiemi di dati, la maggior parte dei valori osservati si trova nell'intervallo centrato sulla media e i cui estremi distano dalla media 1 scarto quadratico medio.

***COMMENTO:** Cosa indica lo scarto quadratico medio*

Nel campione dei 17 fondi comuni azionari che prelevano le commissioni dalle attività del fondo, lo scarto quadratico medio del rendimento percentuale a un anno è 6.42. Ci aspettiamo allora che i rendimenti percentuali annui della maggioranza dei fondi nel campione si raggruppino nel raggio di 6.42 punti dalla media (vale a dire che si raggruppano tra $\bar{X} - 1S = 23.44$ e $\bar{X} + 1S = 36.28$). In effetti, possiamo osservare che i rendimenti del 76.5% dei fondi (13 su 17) cadono in questo intervallo.

Capire la variabilità dei dati



Riquadro 3.2 Capire la variabilità dei dati

✓ **1.** Quanto più i dati sono dispersi, tanto maggiori saranno il range, il range interquartile, la varianza e lo scarto quadratico medio.

✓ **2.** Quanto 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 eguali (in modo che non vi è variabilità nei dati) il range, il range interquartile, la varianza e lo scarto quadratico medio sono tutti eguali a zero.

✓ **4.** Nessuna delle misure di variabilità (il range, il range interquartile, la varianza e lo scarto quadratico medio) può essere negativa.

Misure di dispersione: il Coefficiente di Variazione

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**, indicato con CV, misura la dispersione nell'insieme di dati relativamente alla media.

Il coefficiente di variazione

Il coefficiente di variazione è uguale allo scarto quadratico medio diviso per la media aritmetica, moltiplicato per 100%.

$$CV = \left(\frac{S}{|\bar{X}|}\right)100\% \qquad (3.11)$$

dove

S = scarto quadratico medio

$\bar{X}$ = valore assoluto della media aritmetica nell'insieme dei dati

Esempio 3.13 *Il calcolo del coefficiente di variazione*

Per questi dati, la media del rendimento percentuale a un anno $\bar{X}$ è 29.86 e lo scarto quadratico medio S è 6.42. Usando l'equazione (3.11) il coefficiente di variazione è dato da:

$$CV = \left(\frac{S}{\bar{X}}\right)100\% = \left(\frac{6.42}{29.86}\right)100\% = 21.5\%$$

Per questo campione, la "diffusione media attorno alla media" è pari al 21.5%.

Misure di dispersione: il Coefficiente di Variazione

NOTA: Il coefficiente di variazione è particolarmente utile quando si confrontano le variabilità di due o più insiemi di dati che sono espressi in unità di misura diverse.

Esempio 3.13 *Il calcolo del coefficiente di variazione*

Per questi dati, la media del rendimento percentuale a un anno $\bar{X}$ è 29.86 e lo scarto quadratico medio S è 6.42. Usando l'equazione (3.11) il coefficiente di variazione è dato da:

$$CV = \left(\frac{S}{\bar{X}}\right)100\% = \left(\frac{6.42}{29.86}\right)100\% = 21.5\%$$

Per questo campione, la "diffusione media attorno alla media" è pari al 21.5%.

Misure di forma

La terza caratteristica dei dati che prendiamo in considerazione è la forma della loro distribuzione, cioè il modo in cui si distribuiscono.

La distribuzione dei dati può essere simmetrica o meno.

Se la distribuzione dei dati non è simmetrica, si dice asimmetrica oppure obliqua.

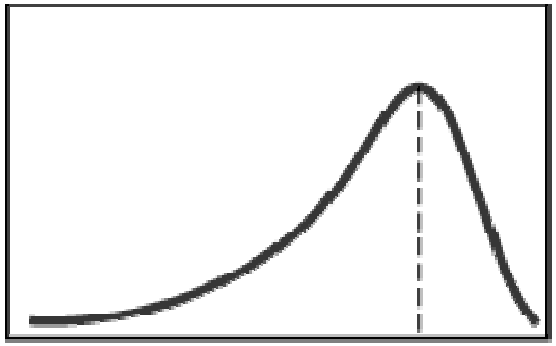
Tipi di misure di forma:

- Asimmetria

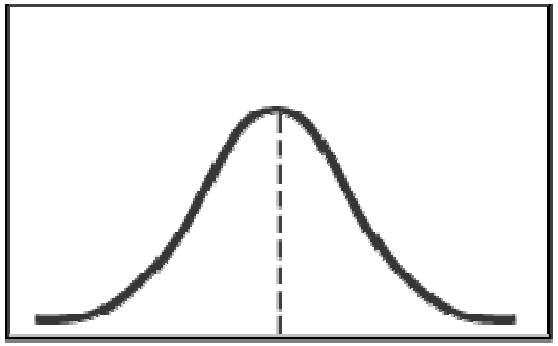
Forma della distribuzione: Simmetrica o Obliqua

Per descrivere la forma della distribuzione è sufficiente confrontare la media con la mediana. Se queste due misure sono uguali, la distribuzione è considerata simmetrica.

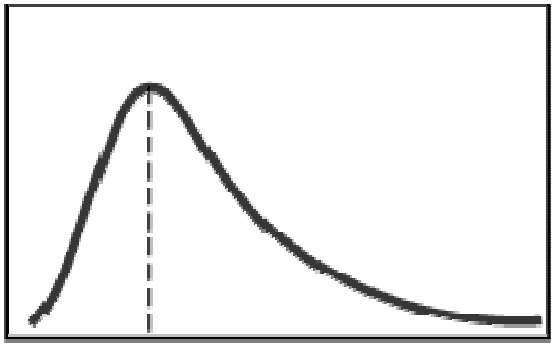
- media < mediana:
  asimmetria negativa o **distribuzione obliqua a sinistra**
- media = mediana:
  simmetria
- media > mediana:
  asimmetria positiva o **distribuzione obliqua a destra**



(a) Obliqua a sinistra



(b) Simmetrica



(c) Obliqua a destra

Misure di sintesi descrittive per una popolazione

- Finora abbiamo preso in considerazioni diverse statistiche che sintetizzano le informazioni contenute in un campione. In particolar modo, abbiamo usato queste statistiche per descrivere le caratteristiche di posizione, di variabilità e di forma.
- Supponiamo ora che l'insieme di dati che abbiamo a disposizione non sia un campione, ma una raccolta di misurazioni numeriche da una intera **popolazione**.
- Quando si considera un'intera popolazione, le misure di sintesi descrittive corrispondenti alla media aritmetica, alla varianza e allo scarto quadratico medio sono i **parametri** della popolazione.

Misure di sintesi descrittive per una popolazione

La media della popolazione viene indicata con il simbolo μ, la lettera minuscola dell'alfabeto greco mu.

Media della popolazione

La media della popolazione è data dalla somma dei valori della popolazione divisa per la dimensione della popolazione

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (3.12)$$

dove

N = dimensione della popolazione

X_i = i–esima osservazione della variabile casuale X

$\sum_{i=1}^{N} X_i$ = somma di tutti i valori X_i della popolazione

Misure di sintesi descrittive per una popolazione

La varianza della popolazione si indica con il simbolo σ^2, la lettera minuscola dell'alfabeto greco sigma elevato al quadrato (si legge "sigma quadro") e lo scarto quadratico medio della popolazione si indica con il simbolo σ.

Varianza della popolazione

La varianza della popolazione è la somma dei quadrati delle differenze dalla media della popolazione divisa per la dimensione della popolazione

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \qquad (3.13)$$

dove

N = dimensione della popolazione

X_i = i–esima osservazione della variabile casuale X

$\sum_{i=1}^{n} (X_i - \mu)^2$ = somma dei quadrati delle differenze tra i valori X_i e μ

Scarto quadratico medio della popolazione

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \qquad (3.14)$$

NOTA: Osserviamo che le formule per il calcolo della varianza e dello scarto quadratico medio della popolazione differiscono da quelle della varianza e dello scarto quadratico medio campionari per il termine (n - 1) che compare al denominatore di S^2 e S che è sostituito da N per il calcolo di σ^2 e σ.

Misure di sintesi descrittive per una popolazione

Quando la distribuzione dei dati non è caratterizzata da una forte asimmetria e le osservazioni sono concentrate intorno a media e mediana, possiamo usare la cosiddetta regola empirica per esaminare la variabilità dei dati e per analizzare più approfonditamente il significato dello scarto quadratico medio.

La regola empirica

La regola empirica afferma che, nella maggior parte degli insiemi di dati, circa due osservazioni su tre (il 67%) si trovano ad una distanza dalla media pari ad una volta lo scarto quadratico medio, e che una percentuale tra il 90% e il 95% circa delle osservazioni si trova ad una distanza dalla media pari a due volte lo scarto quadratico medio.

NOTA: Pertanto lo scarto quadratico medio ci aiuta a capire come le osservazioni si distribuiscono al di sotto e al di sopra della media, e ad individuare e segnalare osservazioni anomale (gli outlier).

Misure di sintesi descrittive per una popolazione

Tabella 3.6

La dispersione dei valori intorno alla media

	% di valori che appartengono a intervalli centrati sulla media	
Intervallo	Regola di Chebyshev (per qualunque distribuzione)	Regola empirica (per distribuzioni a forma campanulare)
$(\mu - \sigma, \mu + \sigma)$	Almeno 0%	Approssimativamente 68%
$(\mu - 2\sigma, \mu + 2\sigma)$	Almeno 75%	Approssimativamente 95%
$(\mu - 3\sigma, \mu + 3\sigma)$	Almeno 88.89%	Approssimativamente 99.7%

Analisi esplorativa dei dati

Tabella 3.7

Relazioni tra i cinque numeri di sintesi e la forma della distribuzione

	Forma della distribuzione		
Confronto	Obliqua a sinistra	Simmetrica	Obliqua a destra
Distanza tra X_{min} e la mediana e distanza tra la mediana e Q_{max}	La distanza tra X_{min} e la mediana è maggiore della distanza tra la mediana e Q_{max}	Le due distanze sono approssimativamente uguali	La distanza tra X_{min} e la mediana è minore della distanza tra la mediana e Q_{max}
Distanza tra X_{min} e Q_1 e distanza tra Q_3 e X_{max}	La distanza tra X_{min} e Q_1 è maggiore della distanza tra Q_3 e X_{max}.	Le due distanze sono approssimativamente uguali	La distanza tra X_{min} e Q_1 è minore della distanza tra Q_3 e X_{max}.
Distanza tra Q_1 e la mediana e distanza tra la mediana e Q_3	La distanza tra Q_1 la mediana è maggiore della distanza tra la mediana e Q_3	Le due distanze sono approssimativamente uguali	La distanza tra Q_1 e la mediana è minore della distanza tra la mediana e Q_3

I cinque numeri di sintesi

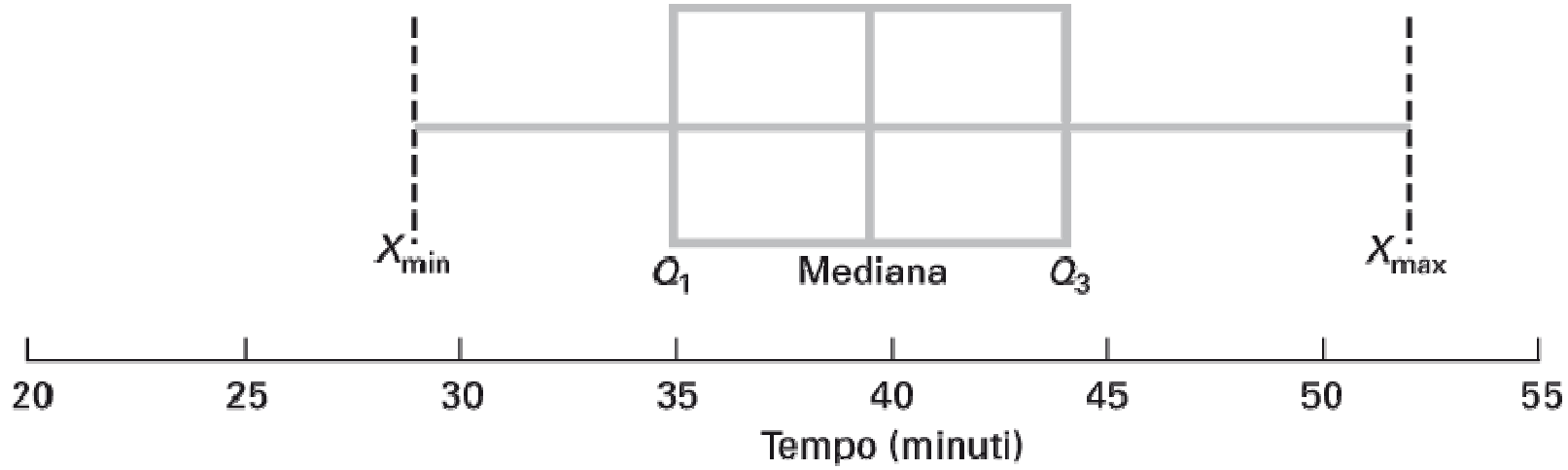
Un approccio che mira a sintetizzare opportunamente le diverse caratteristiche dei dati consiste nel considerare i **cinque numeri di sintesi**:

X_{min} Q_1 Mediana Q_3 X_{max}

A partire da questi numeri è possibile ottenere due misure di posizione (la mediana, la media interquartile) e due misure di variabilità (il range interquartile e il range), che consentono di effettuare “un’analisi esplorativa dei dati” per avere un’idea più precisa della forma della distribuzione.

Il diagramma a “Scatola e Baffi” (o Boxplot)

Il **diagramma scatola e baffi o (o boxplot)** fornisce una rappresentazione grafica dei dati sulla base dei cinque numeri di sintesi.

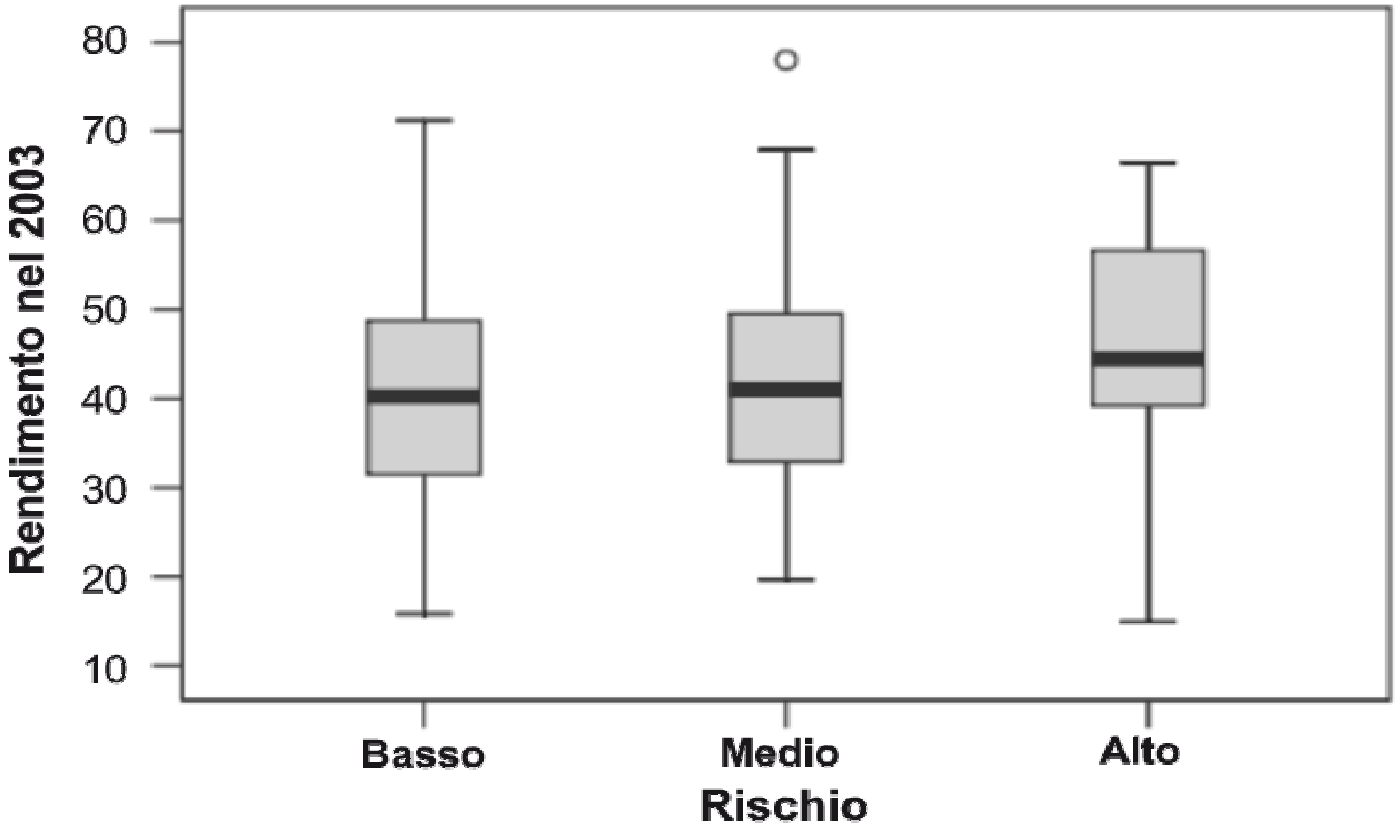


Linea verticale al centro della scatola $\Rightarrow$ mediana	Linea verticale a sinistra della scatola $\Rightarrow$ Q_1	Linea verticale a destra della scatola $\Rightarrow$ Q_3
Linea tratteggiata a sinistra $\Rightarrow$ minimo		Linea tratteggiata a destra $\Rightarrow$ massimo

Il diagramma a “Scatola e Baffi” (o Boxplot)

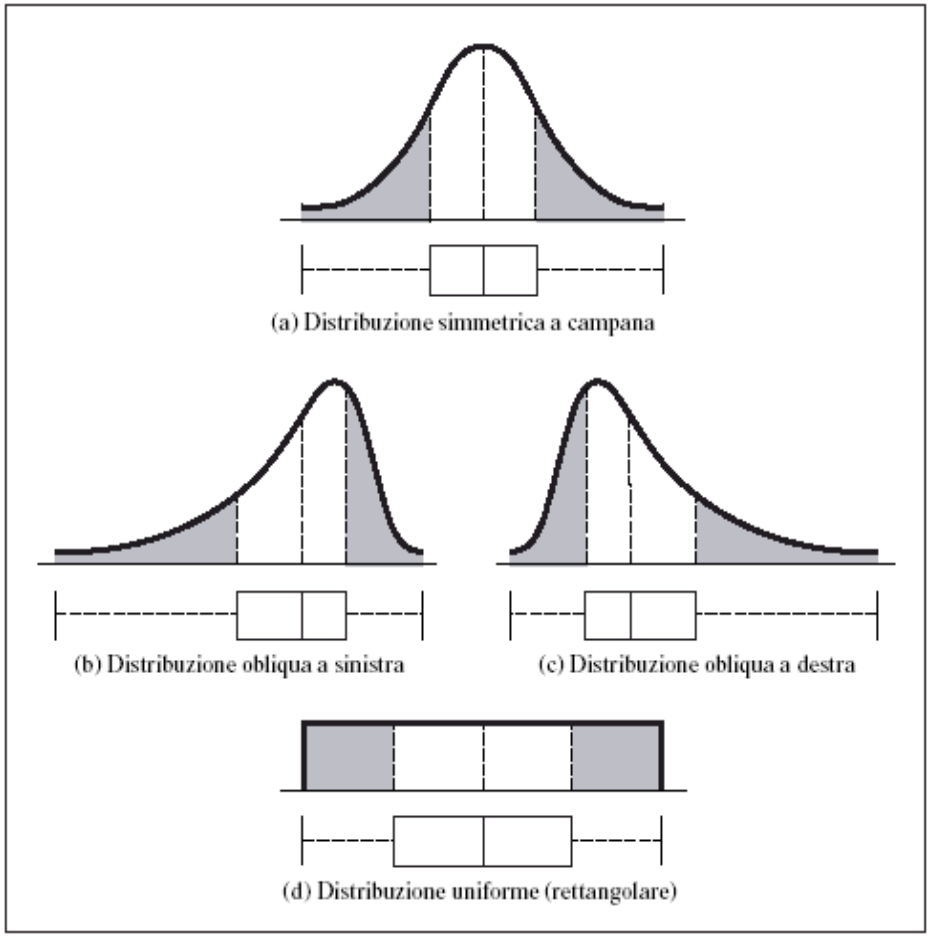
Figura 3.7

Diagrammi a scatola e baffi relativi ai rendimenti nel 2003 dei fondi a basso, medio e alto rischio



Il diagramma a “Scatola e Baffi” (o Boxplot)

Per valutare la relazione che sussiste tra i metodi di analisi esplorativa dei dati, come il diagramma scatola e baffi, e le rappresentazioni grafiche, come i poligoni, consideriamo la Figura, nella quale sono riportati i diagrammi scatola e baffi e i poligoni relativi a quattro ipotetiche distribuzioni.



NOTA: l’area sottostante a ciascuna curva è divisa nei quartili corrispondenti ai cinque numeri di sintesi su cui si basa il diagramma scatola e baffi.

La covarianza ed il coefficiente di correlazione

Nel paragrafo 2.5 abbiamo introdotto il diagramma di dispersione come strumento grafico atto a visualizzare la relazione tra due variabili numeriche. In questo paragrafo introduciamo la covarianza ed il coefficiente di correlazione, che misurano la forza della relazione lineare tra due variabili.

La **covarianza** è una misura che sintetizza la forza della relazione lineare tra due variabili numeriche (*X* e *Y*).

La covarianza campionaria

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \qquad \textbf{(3.16)}$$

La covarianza ed il coefficiente di correlazione

Ad esempio, si può calcolare la covarianza tra il rendimento nel 2003 e le spese dei fondi comuni ad alto rischio specializzati in aziende di piccole dimensioni.

Spese	Rendimento 2003
1.25	37.3
0.72	39.2
1.57	44.2
1.40	44.5
1.33	53.8
1.61	56.6
1.68	59.3
1.42	62.4
1.20	66.5

La covarianza ed il coefficiente di correlazione

Figura 3.9

Foglio di Microsoft Excel con i calcoli necessari per ottenere la covarianza tra le spese e i rendimenti 2003 dei fondi ad alto rischio specializzati in aziende di piccole dimensioni

	A	B	C	D	E	F	G
1	Spese (X)	Rendimento 2003 (Y)	(X-XBar)(Y-YBar)				
2	1,25	37,3	1,47078		=(A2-C13)*(B2-C14)		
3	0,72	39,2	7,81111		=(A3-C13)*(B3-C14)		
4	1,57	44,2	-1,58889		=(A4-C13)*(B4-C14)		
5	1,4	44,5	-0,32822		=(A5-C13)*(B5-C14)		
6	1,33	53,8	-0,05289		=(A6-C13)*(B6-C14)		
7	1,61	56,6	1,30044		=(A7-C13)*(B7-C14)		
8	1,68	59,3	2,53711		=(A8-C13)*(B8-C14)		
9	1,42	62,4	0,72444		=(A9-C13)*(B9-C14)		
10	1,2	66,5	-2,29489		=(A10-C13)*(B10-C14)		
11							
12		Calcoli					
13		XBar	1,353333333		=MEDIA(A2:A10)		
14		YBar	51,53333333		=MEDIA(B2:B10)		
15		*n*-1	8		=CONTA.NUMERI(A2:A10)-1		
16		Sum	9,57900		=SOMMA(C2:C10)		
17		Covarianza	1,19738		=C16/C15		

$$\text{cov}(X,Y) = \frac{9.579}{9} = 1.19738$$

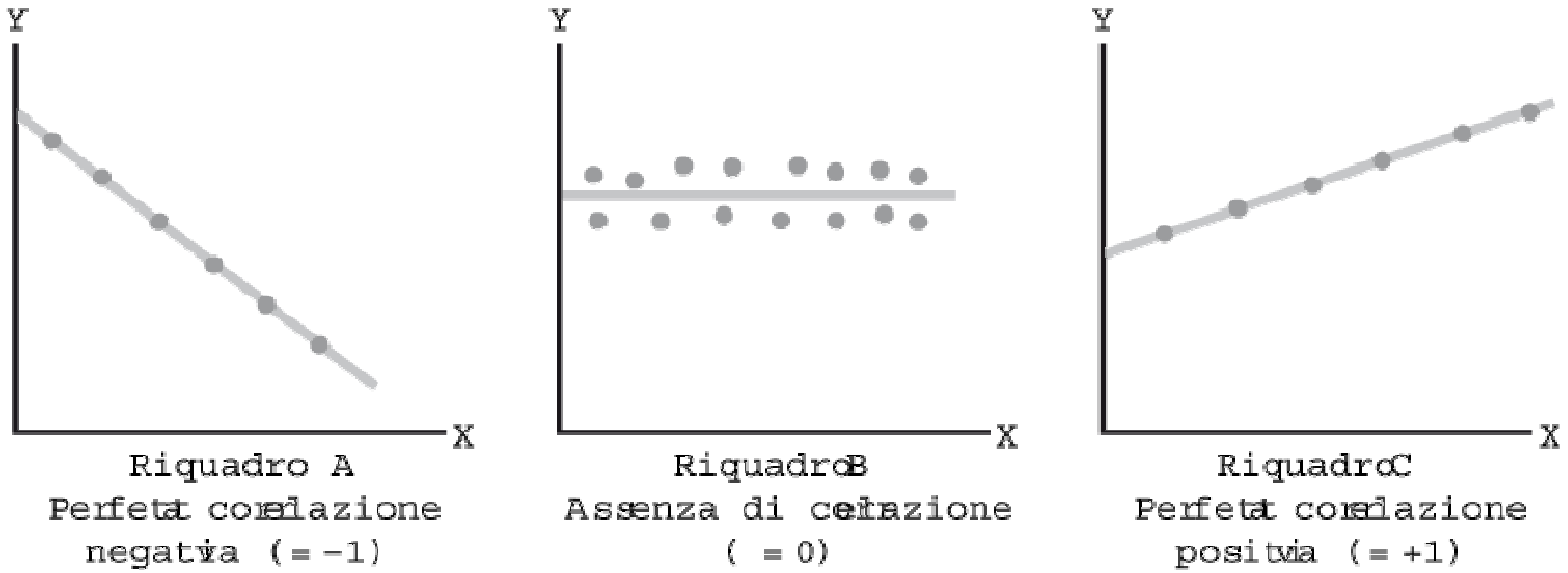
La covarianza ed il coefficiente di correlazione

Un limite della covarianza è che questa misura può assumere un qualunque valore, e non è quindi possibile fare riferimento ad essa per valutare la forza della relazione lineare in termini relativi. A questo scopo bisogna riferirsi al coefficiente di correlazione.

Il **coefficiente di correlazione** misura l'intensità della relazione lineare tra due variabili quantitative. I valori del coefficiente di correlazione variano tra −1 nel caso di una relazione lineare inversa (o negativa) perfetta e +1 nel caso di una relazione lineare diretta (o positiva) perfetta.

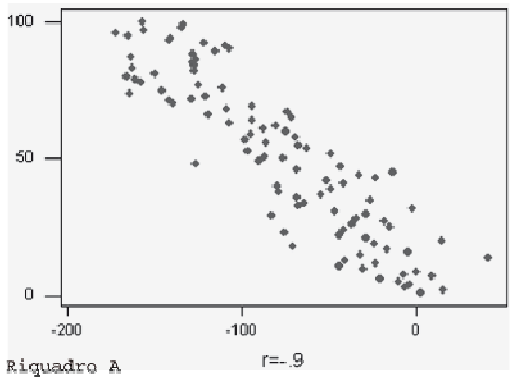
Una relazione lineare si dice perfetta quando i punti del diagramma di dispersione si trovano tutti sulla stessa retta.

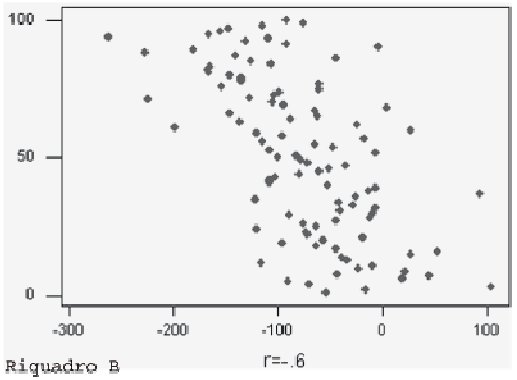
La covarianza ed il coefficiente di correlazione

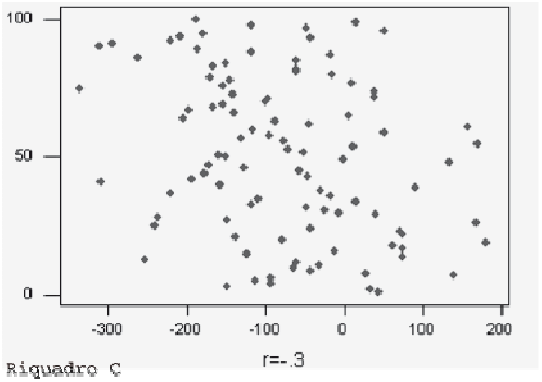


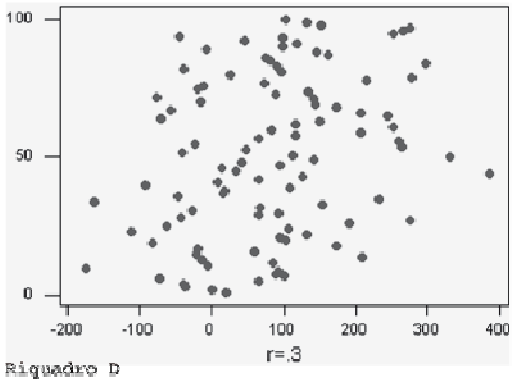
Ovviamente, nel caso si abbiano dati campionari, si calcolerà il coefficiente di correlazione campionaria, che viene indicato con *r*. Con dati campionari è praticamente impossibile osservare coefficienti di correlazioni uguali a −1 o a +1.

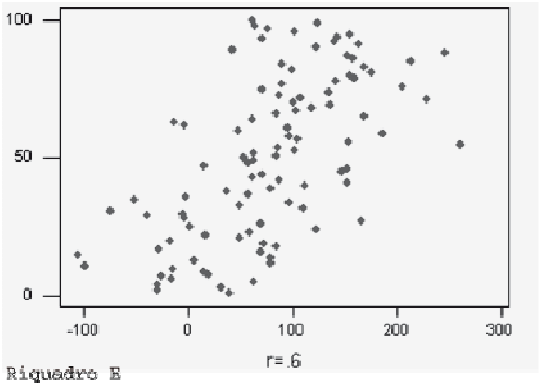
La covarianza ed il coefficiente di correlazione

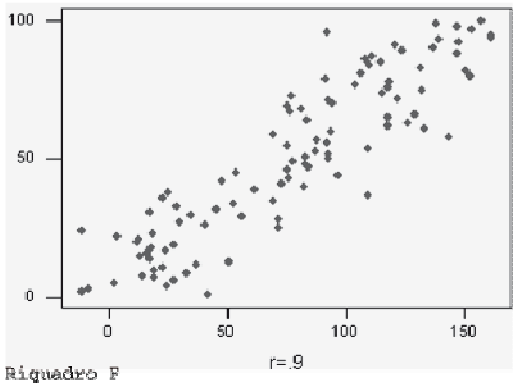












La covarianza ed il coefficiente di correlazione

E' importante sottolineare che la correlazione di per sé non può essere utilizzata per trarre conclusioni sull'esistenza di un nesso causa-effetto tra due variabili.

Il coefficiente di correlazione campionario

$$r = \frac{\mathrm{cov}(X,Y)}{S_X S_Y} \qquad \textbf{(3.17)}$$

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

La covarianza ed il coefficiente di correlazione

Figura 3.12

Foglio di Microsoft Excel con i calcoli necessari per ottenere le correlazioni tra le spese e i rendimenti 2003 dei fondi ad alto rischio specializzati in aziende di piccole dimensioni

	A	B	C	D	E
1	Spese	Rendimento 2003	$(X-XBar)^2$	$(Y-YBar)^2$	(X-XBar)(Y-YBar)
2	1,25	37,3	0,0107	202,5878	1,4708
3	0,72	39,2	0,4011	152,1111	7,8111
4	1,57	44,2	0,0469	53,7778	-1,5889
5	1,4	44,5	0,0022	49,4678	-0,3282
6	1,33	53,8	0,0005	5,1378	-0,0529
7	1,61	56,6	0,0659	25,6711	1,3004
8	1,68	59,3	0,1067	60,3211	2,5371
9	1,42	62,4	0,0044	118,0844	0,7244
10	1,2	66,5	0,0235	224,0011	-2,2949
11		Somme:	0,662	891,16	9,5790
12		:			
13				Calcoli	
14				XBar	1,353333333
15				YBar	51,53333333
16				*n*-1	8
17				Covarianza	1,19738
18				S_X	0,287662997
19				S_Y	10,55438298
20				*r*	0,394378596
21					
22					
23					
24					

(non si riportano le formule per i calcoli nelle celle C2:E11)

=MEDIA(A2:A10)
=MEDIA(B2:B10)
=CONTA.NUMERI(A2:A10)-1
=E11/E16
=RADQ(C11/E16)
=RADQ(D11/E16)
=CORRELAZIONE(A2:A10;B2:B10)
o
E17/(E18*E19)

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y} = \frac{1.19738}{(0.287663)(10.554383)} = 0.3943786$$