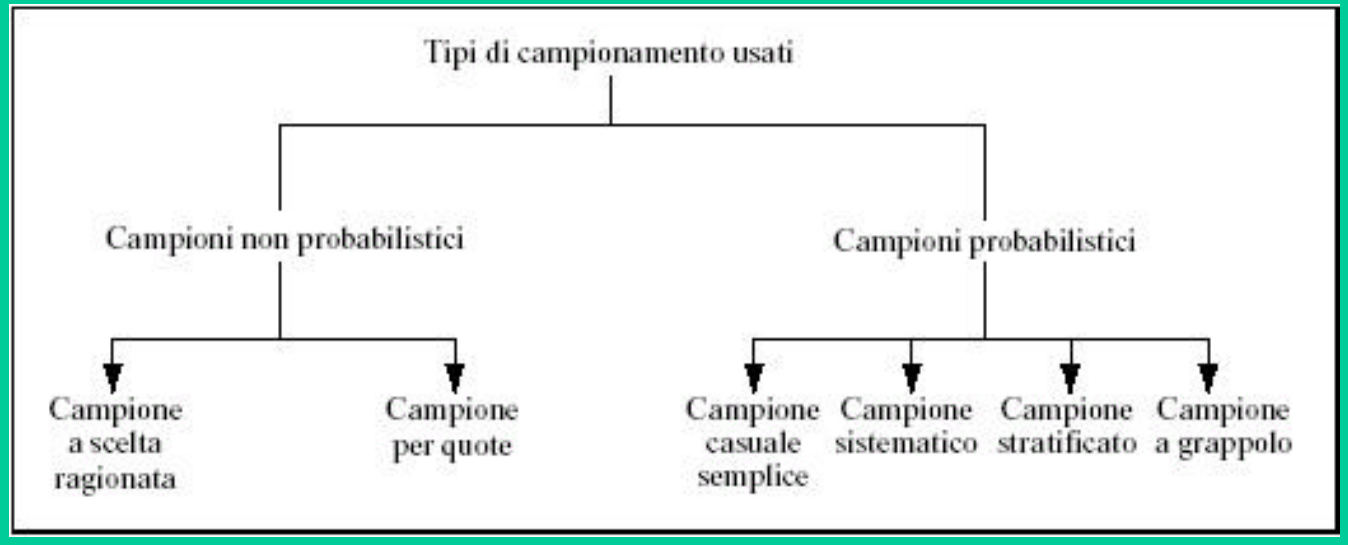
RICHIAMI DI STATISTICA DESCRITTIVA E DI INFERENZA:

RACCOLTA E ORGANIZZAZIONE DEI DATI IN FORMA TABELLARE E GRAFICA

I tipi di campionamento

Un **campione non probabilistico** è un campione in cui gli oggetti o gli individui sono inclusi senza tenere conto della loro probabilità di appartenere al campione.

Un **campione probabilistico** è un campione in cui i soggetti sono scelti sulla base delle probabilità note.



I tipi di campionamento

Un campione non probabilistico è un campione in cui gli oggetti o gli individui sono inclusi senza tenere conto della loro probabilità di appartenere al campione

Esempio: sondaggi proposti da aziende ai visitatori del loro sito web => autoselezione del campione

Vantaggi: comodità, velocità, costi bassi

Svantaggi: mancanza di accuratezza per la selezione distorta e impossibilità di generalizzare i risultati

I tipi di campionamento

Un campione probabilistico è un campione in cui gli oggetti o gli individui sono scelti sulla base delle probabilità note di appartenere al campione

Campione casuale semplice: ogni individuo o oggetto della popolazione ha la stessa probabilità di essere selezionato.

Assegnando ad ogni unità della popolazione un numero progressivo da 1 a N (numerosità totale) genero n numeri casuali compresi tra 1 e N per individuare le unità del campione (n = numerosità campionaria).

I tipi di campionamento

Campione sistematico: gli N individui o oggetti della popolazione sono ripartiti in n gruppi e si calcola:

$$k = \frac{N}{n}$$

dove k è arrotondato all'intero più vicino. Il primo individuo è scelto casualmente tra i k individui o oggetti del primo gruppo. Il resto del campione si ottiene scegliendo da quel punto in poi ogni k-esimo elemento successivo dell'intera lista della popolazione.

Vantaggi: velocità

I tipi di campionamento

Campione stratificato: gli N elementi della popolazione sono suddivisi in distinte sottopopolazioni o strati, sulla base di una caratteristica comune.
Si conduce un campionamento casuale semplice in ogni strato e i risultati dei singoli campionamenti sono poi messi assieme.

Vantaggi: più efficiente del campionamento casuale semplice e del campionamento sistematico perchè assicura che gli individui o oggetti della popolazione siano rappresentati adeguatamente nel campione.

I tipi di campionamento

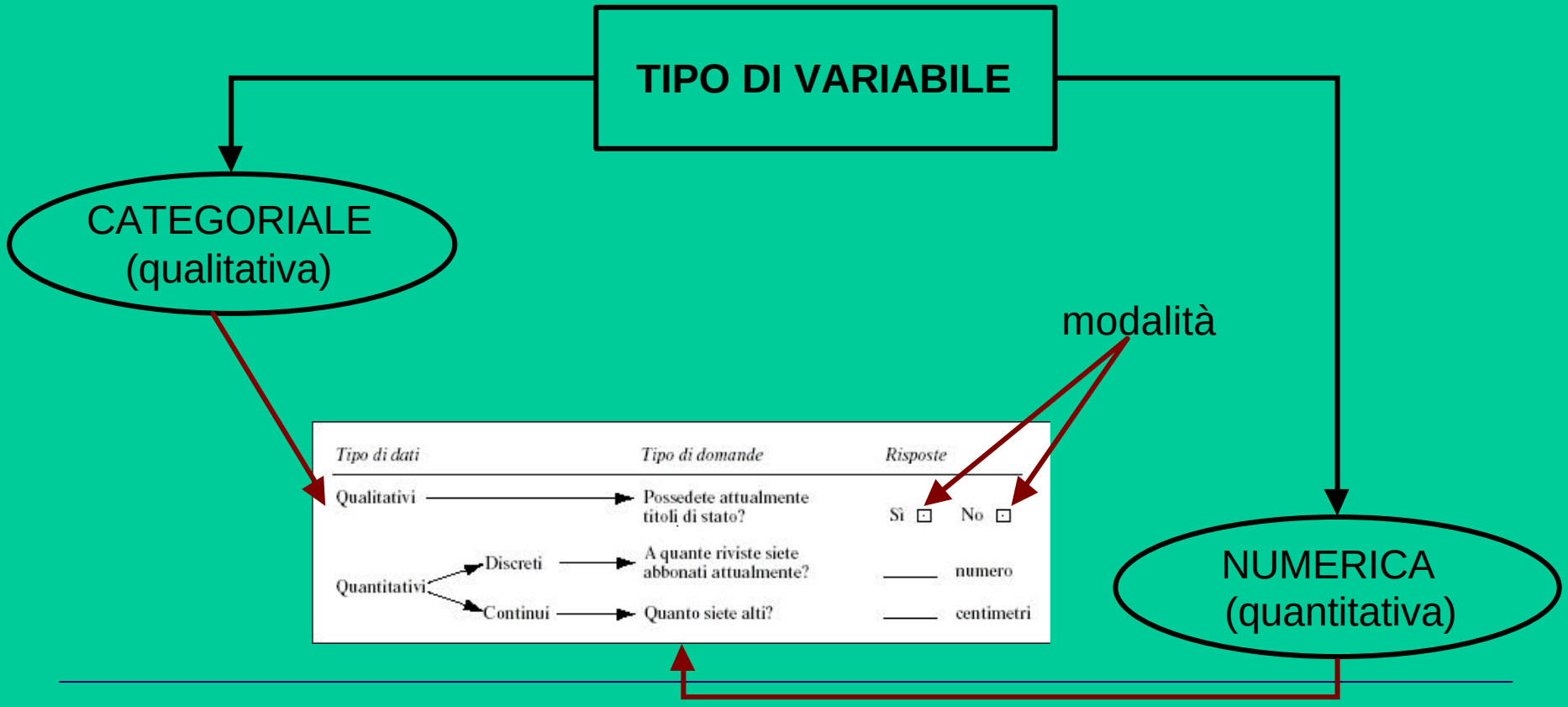
Campione a grappolo: gli N elementi della popolazione sono suddivisi in molti gruppi, detti grappoli, in maniera tale che ogni grappolo sia rappresentativo dell'intera popolazione. Si estrae poi un campione casuale di grappoli e tutti gli elementi dei grappoli selezionati sono inclusi nel campione.

Vantaggi: meno costoso del campionamento casuale semplice soprattutto se i grappoli sono circoscrizioni o aree geografiche

Svantaggi: è necessaria una dimensione complessiva del campione più grande per ottenere risultati precisi

I tipi di variabile

Per variabile si intende un aspetto del fenomeno di interesse oggetto di studio, del quale è disponibile una serie di misurazioni.



Statistica descrittiva e statistica inferenziale

La **statistica descrittiva** si può definire come un complesso di metodi che comprendono la raccolta, la presentazione e la caratterizzazione di un insieme di dati con lo scopo di descriverne le varie caratteristiche in maniera appropriata.

La **statistica inferenziale** può essere definita come il complesso dei metodi che consentono di stimare una caratteristica di una popolazione, oppure di prendere una decisione che concerne l'intera popolazione, sulla base dei soli risultati campionari.

Una **popolazione** (o **universo**) è l'insieme degli elementi o delle "cose" che si prendono in considerazione.
Un **campione** è la porzione della popolazione che si seleziona per l'analisi.
Un **parametro** è una misura di sintesi che descrive una caratteristica dell'intera popolazione.
Una **statistica** è una misura di sintesi che si calcola per descrivere una caratteristica soltanto sulla base di un campione della popolazione.

Statistica descrittiva

Una prima descrizione e sintesi dei dati si ottiene mediante una serie di strumenti tabellari e grafici

L'analisi dei dati con i grafici è semplice e ricca di informazioni

Gli svantaggi rispetto ai metodi numerici sono:

- **Anche se le conclusioni finali dell'interpretazione sono univoche le informazioni ricavabili sono soggettive**
- **La precisione delle informazioni è minore, soprattutto per certi tipi di analisi (stima intervallare e puntuale, verifiche d'ipotesi, ecc.)**

Il *dataset*: la corretta organizzazione dei dati

Esempio: sono stati raccolti i dati relativi alla performance (1Yr$Ret=rendimento percentuale a un anno) di un campione di 194 fondi di investimento, suddivisi in 59 a capitalizzazione integrale (Object=1) e 135 misti (Object=2).

Per una corretta ed efficace analisi statistica dei dati, essi devono essere strutturati secondo il seguente schema:

N	Fund	1Yr$Ret	Object
1	Alliance Capital A GrowInc	30.8	2
2	Berger SmCoGrow	29.9	1
3	Jurika & Voyles Kaufmann	28.9	1
4	Baron Funds BanRosSC	35.5	2
...	....	...	...
192	MainStay Inst MainPwrGr	36.1	2
193	Vanguard Index Inst	30.9	2
194	Vanguard Index 500	30.8	2

Nome Variabili

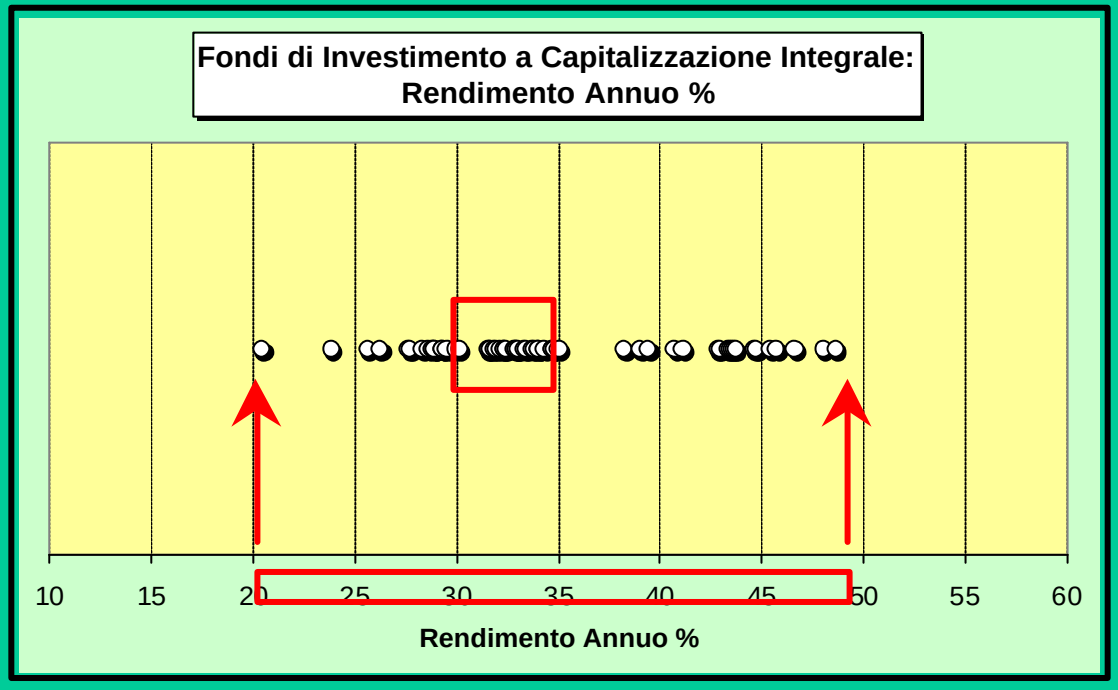
Unità statistica

Non devono esserci né righe né colonne completamente vuote. Se ci sono dei dati mancati essi vanno codificati in maniera appropriata (in Excel, cella vuota).

Una prima rappresentazione grafica: il *dotplot*

All'aumentare del numero di osservazioni tanto l'ordinamento quanto il diagramma ramo-foglia si rivelano inadeguati a rappresentare il fenomeno: diventa necessario utilizzare degli strumenti grafici.

Se raffiguriamo in un 75 *dotplot*) i 59 valori della performance dei fondi a capitalizzazione integrale otteniamo la seguente rappresentazione ...



L'informazione che risulta dal grafico è che la performance dei fondi a capitalizzazione varia tra 20 e 50 (*range*=30) e che la maggior parte dei valori si concentra tra 30 e 35.

La frequenza: definizione e motivazione

Sarebbe interessante conoscere esattamente quanti fondi cadono tra il valore 30 e 35 ed, in modo analogo, quanti cadono in una serie di intervalli, opportunamente definiti, in modo da coprire l'intero intervallo di variazione che va da 20 a 50.

DEFINIZIONE (per le variabili numeriche)

Frequenza: conteggio del numero di unità statistiche che cadano in un certo intervallo di valori, detto classe.

DEFINIZIONE (per le variabili categoriali)

Frequenza: conteggio del numero di unità statistiche che assumono una data modalità.

Lo studio della frequenza ci fornisce una fondamentale informazione sulla **distribuzione** della variabile di interesse: il modo in cui (ossia dove e come) i valori della variabile si distribuiscono nell'intervallo di variazione (variabili numeriche) o tra le diverse modalità (variabili categoriali).

La frequenza: caratteristiche

Numero di classi: da un minimo di 5 ad un massimo di 15.

Estremi delle classi: devono facilitare la lettura e l'interpretazione dei dati.

Ampiezza delle classi: si calcolano secondo la seguente formula:

Determinazione dell'ampiezza di una classe di raggruppamento

$$\text{Ampiezza dell'intervallo} \cong \frac{\text{range}}{\text{numero delle classi}} \qquad (2.1)$$

NOTA BENE Þ *Elementi di soggettività nel calcolo della frequenza*

Una diversa definizioni del numero e/o degli estremi e/o dell'ampiezza delle classi genera una differente espressione della frequenza, che può essere anche sensibile se la numerosità dei dati è scarsa.

Rappresentazione della frequenza: la frequenza può essere rappresentata

FORMA	FORMATO
• **Tabella** • **Grafico**	• **Frequenza assoluta** • **Frequenza relativa**

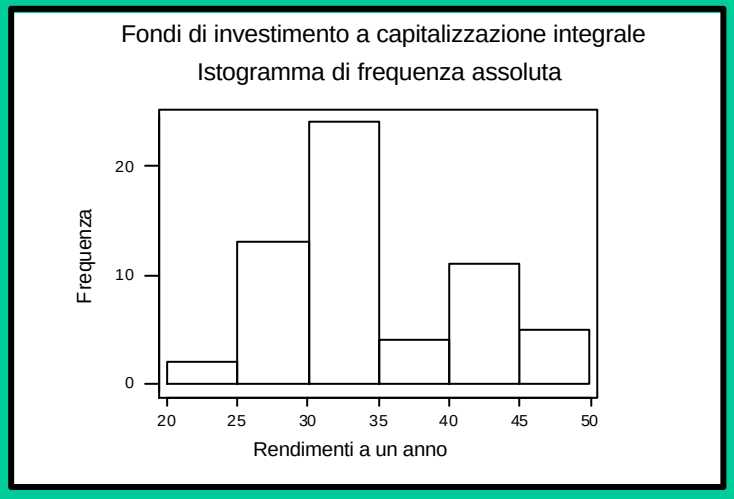
Tabella e Istogramma di frequenza assoluta e relativa

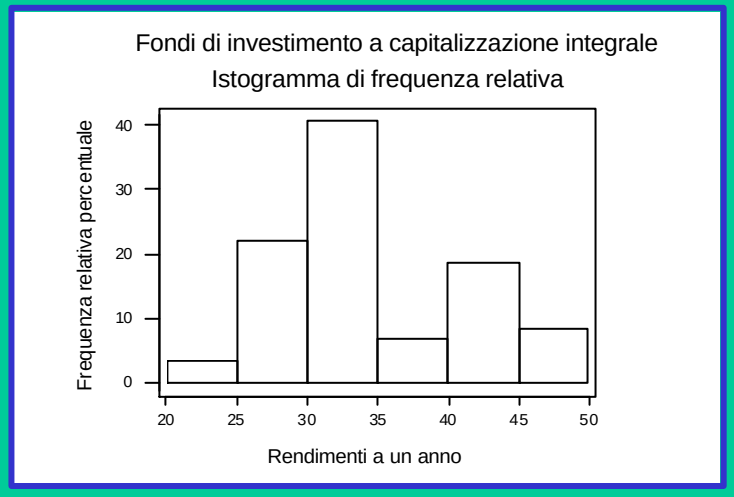
Tabella 2.2 *Distribuzione delle frequenze dei rendimenti percentuali a un anno realizzati dai 59 fondi a capitalizzazione integrale*

RENDIMENTI PERCENTUALI A UN ANNO	NUMERO DI FONDI
da 20.0 a 25.0	2
da 25.0 a 30.0	13
da 30.0 a 35.0	24
da 35.0 a 40.0	4
da 40.0 a 45.0	11
da 45.0 a 50.0	5
Totale	59

Tabella 2.3 *Distribuzione delle frequenze relative e delle percentuali dei rendimenti a un anno fatti registrare dai 59 fondi a capitalizzazione integrale*

RENDIMENTI PERCENTUALI A UN ANNO	PROPORZIONE DI FONDI	PERCENTUALE DI FONDI
da 20.0 a 25.0	0.034	3.4
da 25.0 a 30.0	0.220	22.0
da 30.0 a 35.0	0.407	40.7
da 35.0 a 40.0	0.068	6.8
da 40.0 a 45.0	0.186	18.6
da 45.0 a 50.0	0.085	8.5
Totale	1.000	100.0





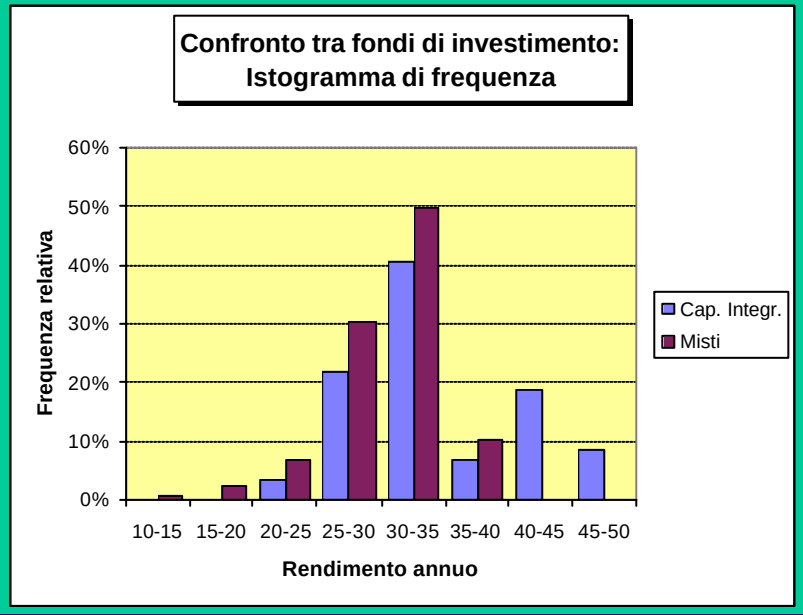
L'**istogramma** è un diagramma a barre verticali in cui le barre rettangolari hanno come base gli intervalli in cui sono state raggruppate le osservazioni

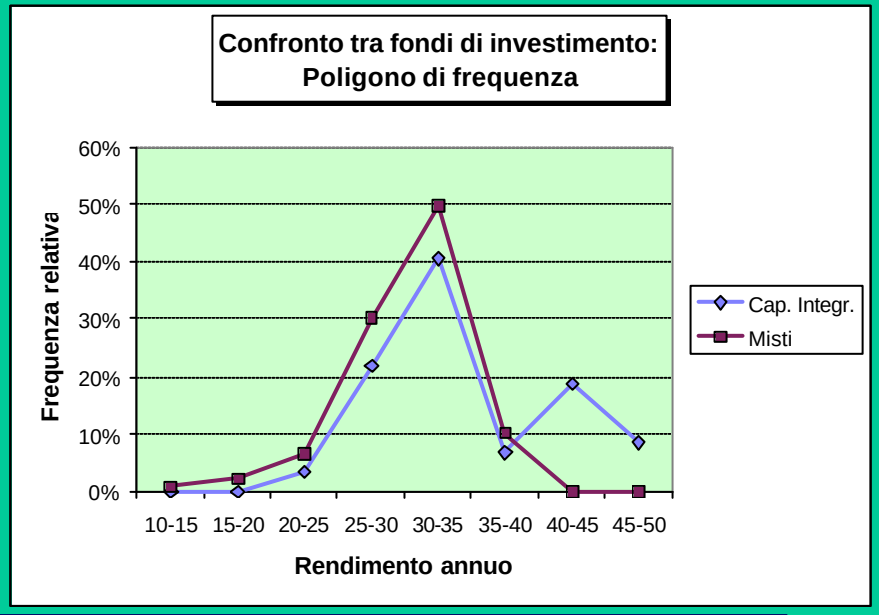
Tabella e Istogramma di frequenza per il confronto tra due gruppi

Rendimento Annuo	Formato della Frequenza			
	Assoluta		Relativa	
	Tipo di Fondo		Tipo di Fondo	
	Cap. Integr.	Misti	Cap. Integr.	Misti
10-15		1	0%	1%
15-20		3	0%	2%
20-25	2	9	3%	7%
25-30	13	41	22%	30%
30-35	24	67	41%	50%
35-40	4	14	7%	10%
40-45	11		19%	0%
45-50	5		8%	0%
Totale	59	135	100%	100%

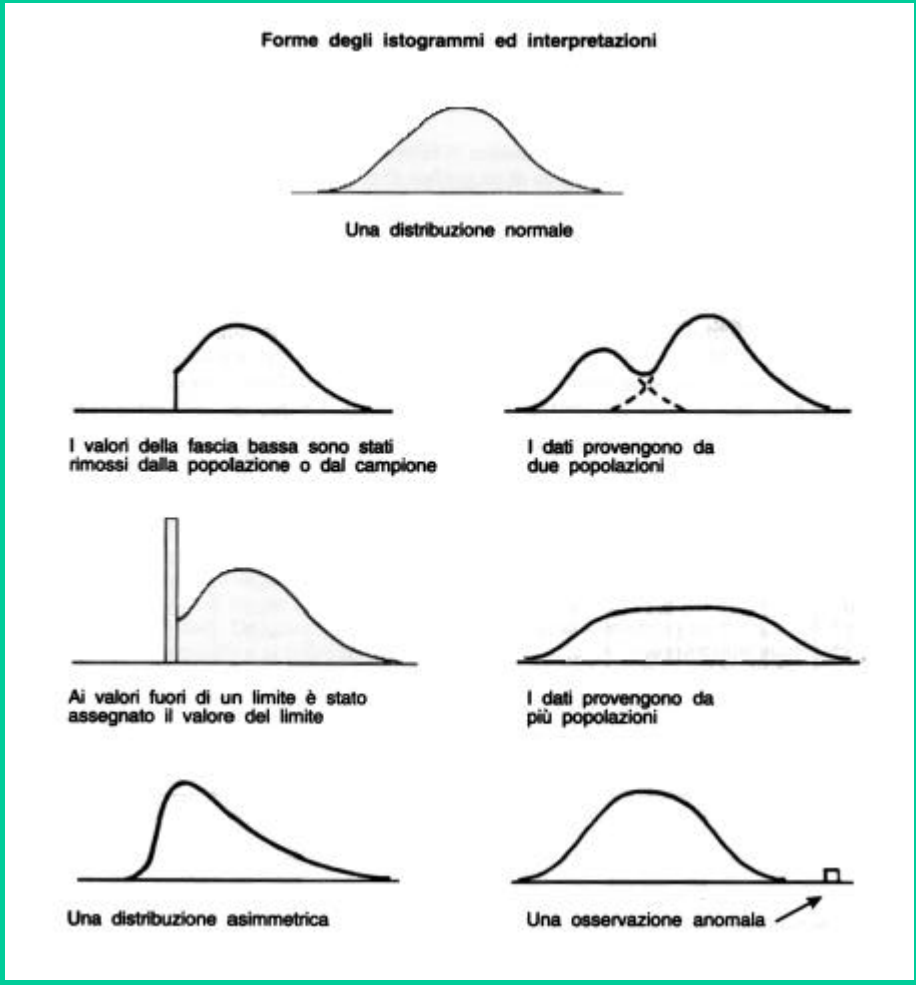
Ai fini del confronto tra due (o più) gruppi

- la frequenza relativa è più efficace di quella assoluta;
- graficamente, il poligono è più idoneo dell'istogramma.





Forme degli istogrammi ed interpretazioni

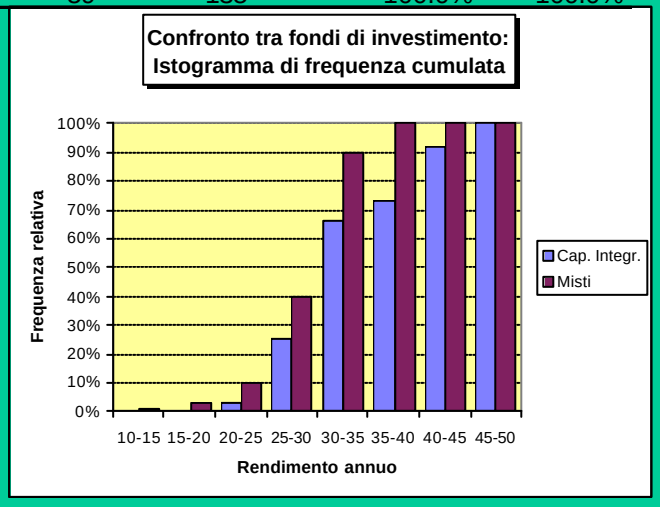


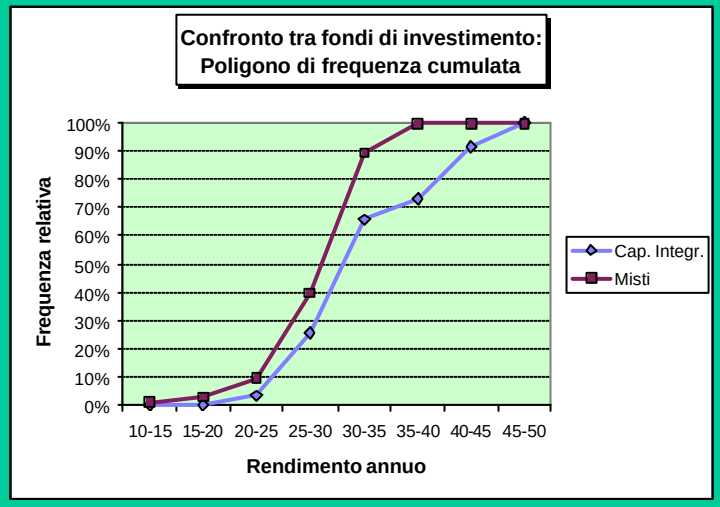
La frequenza cumulata

Se, a partire dalla seconda classe di intervallo, si sommano recursivamente le frequenze si ottiene la cosiddetta frequenza cumulata, sia assoluta che relativa.

Rendimento Annuo	Formato della Frequenza Cumuta			
	Assoluta		Relativa	
	Tipo di Fondo		Tipo di Fondo	
	Cap. Integr.	Misti	Cap. Integr.	Misti
10-15		1	0.0%	0.7%
15-20		4	0.0%	3.0%
20-25	2	13	3.4%	9.6%
25-30	15	54	25.4%	40.0%
30-35	39	121	66.1%	89.6%
35-40	43	135	72.9%	100.0%
40-45	54	135	91.5%	100.0%
45-50	59	135	100.0%	100.0%

RENDIMENTI PERCENTUALI A UN ANNO	PERCENTUALE DI FONDI NELL'INTERVALLO	PERCENTUALE CUMULATIVA DI FONDI FINO AL LIMITE INFERIORE DELL'INTERVALLO
da 20.0 a 25.0	3.4	0.0
da 25.0 a 30.0	22.0	3.4
da 30.0 a 35.0	40.7	25.4 = 3.4 + 22.0
da 35.0 a 40.0	6.8	66.1 = 3.4 + 22.0 + 40.7
da 40.0 a 45.0	18.6	72.9 = 3.4 + 22.0 + 40.7 + 6.8
da 45.0 a 50.0	8.5	91.5 = 3.4 + 22.0 + 40.7 + 6.8 + 18.6
da 50.0 a 55.0	0.0	100.0 = 3.4 + 22.0 + 40.7 + 6.8 + 18.6 + 8.5





Grafici di dispersione

Un'azienda chimica che produce detersivi effettua delle prove di lavaggio con diversi prodotti rilevando strumentalmente le variabili riflettanza (efficacia pulente) e scolorimento. Interessa valutare la relazione tra le due variabili

Prodotto	Riflettanza	Scolorimento
A	60	1.1
B	71	2.4
C	54	1.5
D	47	1.8
E	76	2.5
F	89	3.1
G	58	1.4
H	56	1.7
I	45	0.9
J	75	2.3

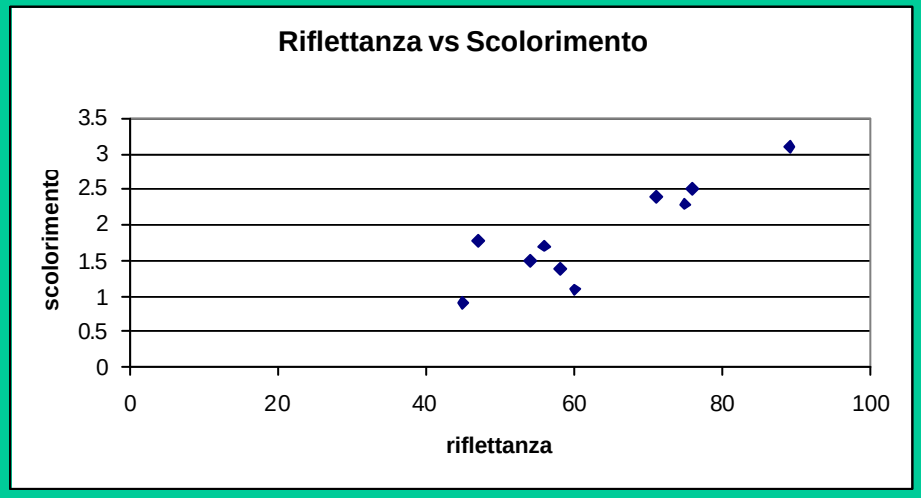
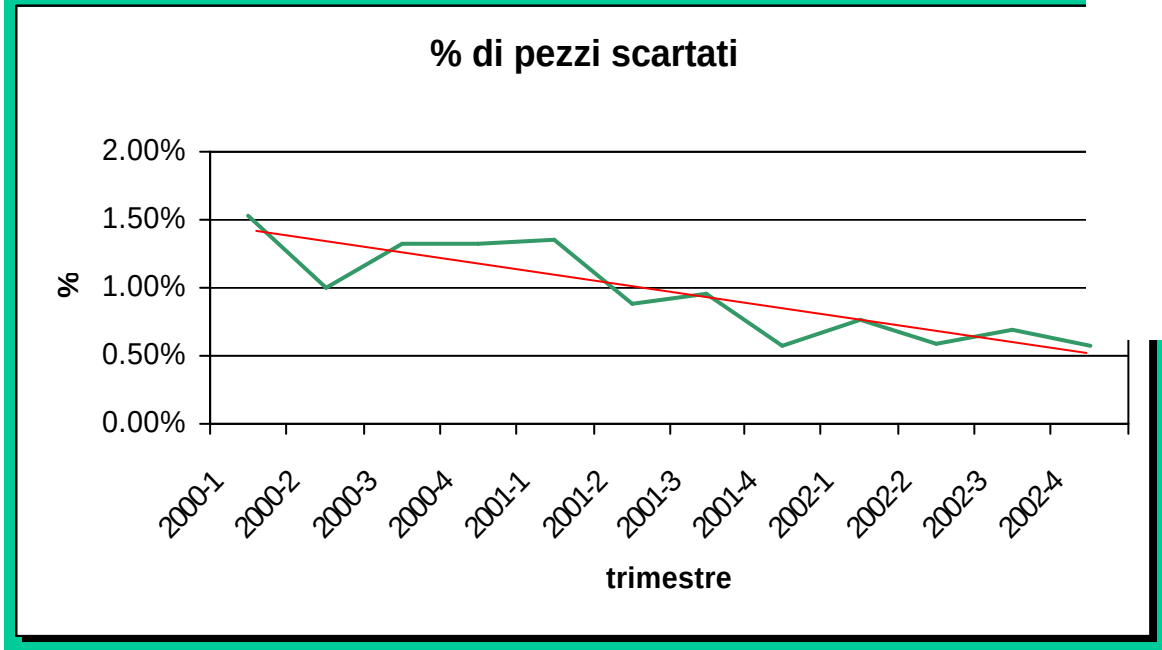


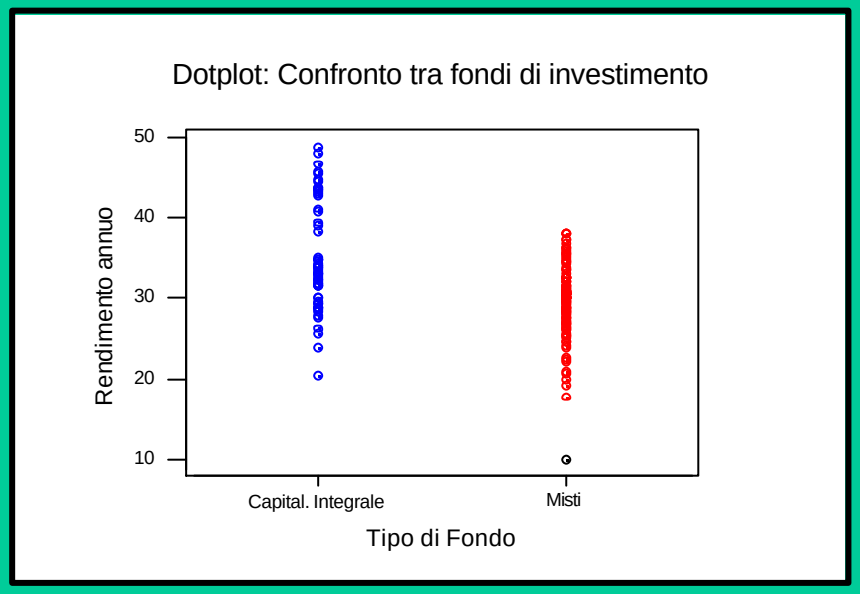
Diagramma in coordinate cartesiane rispetto al tempo

Un'azienda che produce componenti elettrici controlla periodicamente una parte della produzione, rilevando la percentuale di pezzi scartati

Progressivo	Trimestre	Anno	Pezzi scartati	Pezzi controllati	%
1	2000-1	2000	34	2200	1.55%
2	2000-2	2000	20	2000	1.00%
3	2000-3	2000	31	2310	1.34%
4	2000-4	2000	28	2100	1.33%
5	2001-1	2001	27	1998	1.35%
6	2001-2	2001	15	1700	0.88%
7	2001-3	2001	23	2400	0.96%
8	2001-4	2001	13	2300	0.57%
9	2002-1	2002	17	2250	0.76%
10	2002-2	2002	16	2700	0.59%
11	2002-3	2002	14	2070	0.68%
12	2002-4	2002	11	1925	0.57%

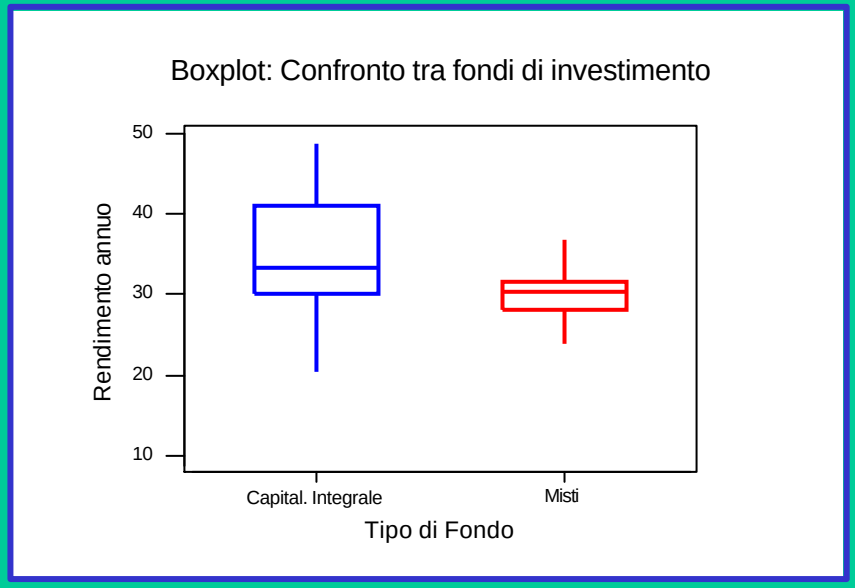


Dotplot e Boxplot: confronto tra due gruppi



Il Dotplot ci conferma che i fondi a capitalizzazione integrale ottengono tendenzialmente un rendimento annuo più alto rispetto ai fondi misti.

Il Boxplot suggerisce anche che i fondi a capitalizzazione integrale sono più variabili rispetto ai fondi misti.



Variabili categoriali: frequenza e frequenza cumulata

Anche i dati qualitativi possono essere sintetizzati utilizzando appropriati strumenti analoghi a quelli dei dati quantitativi.

Consideriamo un'estensione del dataset relativo ai fondi di investimento,

N	Fund	1Yr$Ret	Group	Object
1	Alliance Capital A GrowInc	30.8	4	2
2	Berger SmCoGrow	29.9	1	1
3	Jurika & Voyles Kaufmann	28.9	4	1
4	Baron Funds BanRosSC	35.5	2	2
...	...	...	...	...
192	MainStay Inst MainPwrGr	36.1	5	2
193	Vanguard Index Inst	30.9	5	2
194	Vanguard Index 500	30.8	5	2

includendo (oltre ad Object) anche la 2ª variabile categoriale Group="Tipo di commissione sul fondo", che può assumere 5 modalità (o livelli).

La **tabella di sintesi** per dati qualitativi presenta le stesse caratteristiche della tabella delle frequenze già vista in relazione ai dati quantitativi

Tabella 2.7 *Tabella di sintesi e tabella delle percentuali della variabile "commissioni associate al fondo" (Group) per i 194 fondi azionari del campione*

Commissione	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 commissioni	107	55.2
Totale	194	100.1[a]

Variabili categoriali: diagramma a barre e a torta

Il **diagramma a barre** è un grafico analogo all'istogramma di frequenza. Ciascuna barra del diagramma rappresenta una modalità della variabile, e la lunghezza della barra è proporzionale alla frequenza della barra è proporzionale alla frequenza dalla modalità considerata.

Il **diagramma a torta** si ottiene dividendo l'angolo di 360° in "fette" la cui dimensione è proporzionale alla percentuale di osservazioni che cadono in ciascuna categoria.

FIGURA 2.5
Diagramma a barre delle frequenze assolute della variabile "commissioni associate al fondo" (Group)

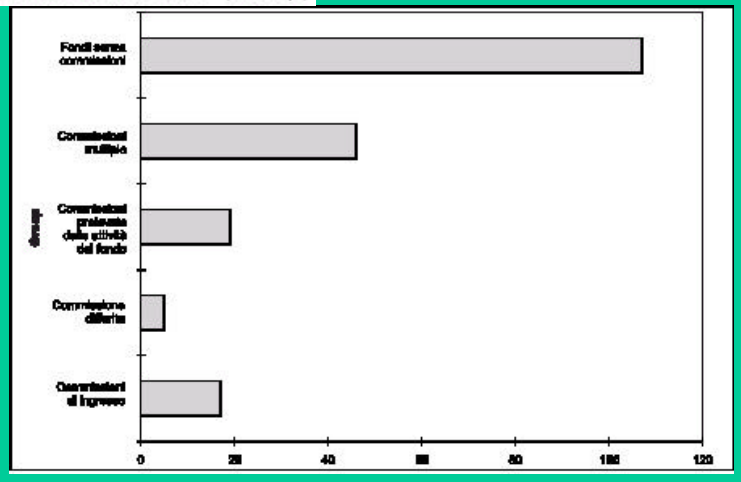
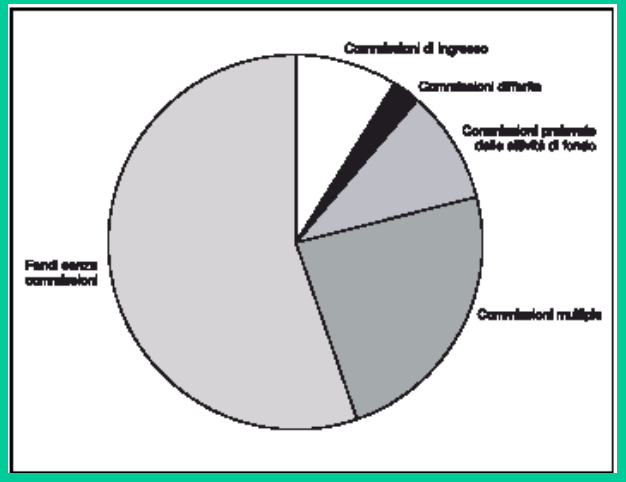


FIGURA 2.6
Diagramma a torta percentuale della variabile "commissioni associate al fondo" (Group)



Il diagramma di Pareto

Il diagramma di Pareto è un diagramma a barre verticali in cui le modalità compaiono in ordine decrescente rispetto alle frequenze di ciascuna e combinate con un poligono cumulativo nella stessa scala.

Il diagramma di Pareto diventa particolarmente utile quando le modalità della variabile di interesse sono molte.

Infatti il vantaggio di questo grafico consiste nella sua capacità di separare le poche modalità cui è associata una frequenza più alta da quelle meno rappresentate nei dati, permettendo al lettore di concentrarsi sulle modalità più importanti.

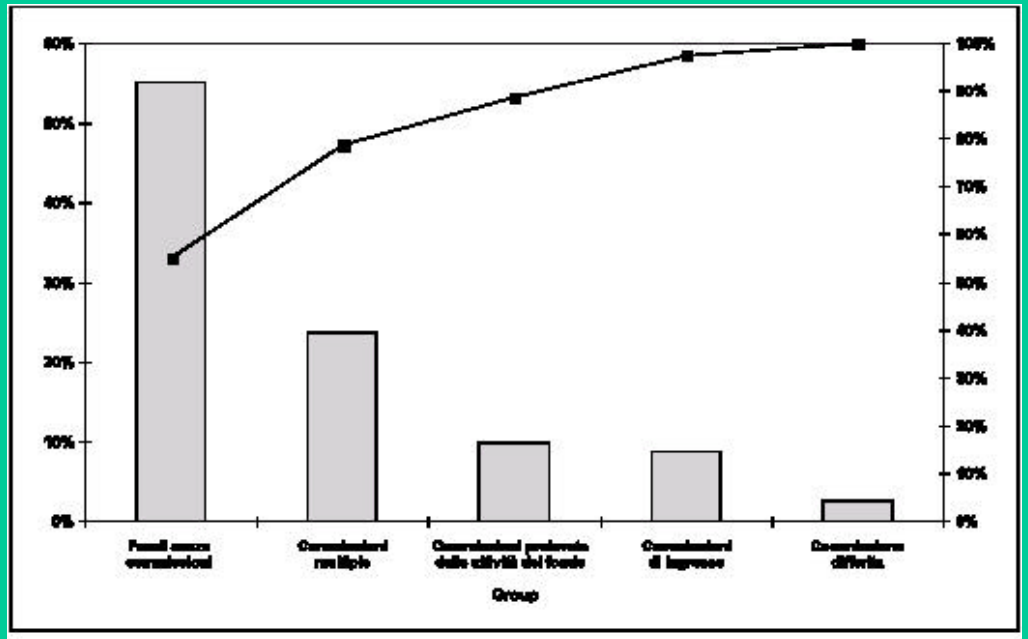


FIGURA 2.7
Diagramma di Pareto della variabile "commissioni associate al fondo" (Group)

1/2

Due variabili categoriali: la tabella di contingenza

In un'analisi statistica siamo spesso interessati a esaminare il comportamento simultaneo di due variabili qualitative: per esempio ci possiamo chiedere se esiste un legame fra il tipo di fondo (a capitalizzazione integrale o misto) e la particolare forma di commissione cui il fondo è assoggettato.

La **tabella di contingenza** è una tabella a doppia entrata in cui le osservazioni relative a due variabili categoriche vengono rappresentate simultaneamente.

Tabella 2.8 *Tabella di contingenza per le variabili "obiettivo del fondo" (Object) e "commissioni associate al fondo" (Group)*

	Commissioni sul fondo					
Obiettivo del fondo	Commissioni prelevate dalle attività del fondo	Commissioni differite	Commissioni di ingresso	Commissioni multiple	Fondi senza commissioni	Totale
Fondo a capitalizzazione integrale	4	0	7	16	32	59
Fondo misto	13	5	12	30	75	135
Totale	17	5	19	46	107	194

Due variabili categoriali: la tabella di contingenza

Al fine di analizzare tutte le possibili relazioni esistenti fra le due variabili, è utile convertire le frequenze congiunte assolute in frequenze percentuali rispetto:

1. Al totale complessivo (rappresentato nel nostro caso dai 194 fondi azionari dal campione)
2. Al totale per riga (rispetto al numero di fondi a capitalizzazione integrale e al numero di fondi misti)
3. Al totale per colonna (rispetto alle cinque tipologie di commissione)

Tabella 2.9 *Tabella di contingenza per le variabili "obiettivo del fondo" (Object) e "commissioni associate al fondo" (Group)(percentuali sul totale)*

Obiettivo del fondo	Commissioni sul fondo: Commissioni prelevate dalle attività del fondo	Commissioni differite	Commissioni di ingresso	Commissioni multiple	Fondi senza commissioni	Totale
Fondo a capitalizzazione integrale	2.1	0.0	3.6	8.2	16.5	30.4
Fondo misto	6.7	2.6	6.2	15.5	38.7	69.6
Totale	8.8	2.6	9.8	23.7	55.2	100.0

Tabella 2.10 *Tabella di contingenza per le variabili "obiettivo del fondo" (Object) e "commissioni associate al fondo" (Group) (percentuali di riga)*

Obiettivo del fondo	Commissioni sul fondo: Commissioni prelevate dalle attività del fondo	Commissioni differite	Commissioni di ingresso	Commissioni multiple	Fondi senza commissioni	Totale
Fondo a capitalizzazione integrale	6.8	0.0	11.9	27.1	54.2	100.0
Fondo misto	9.6	3.7	8.9	22.2	55.6	100.0
Totale	8.8	2.6	9.8	23.7	55.2	100.0

Tabella 2.11 *Tabella di contingenza per le variabili "obiettivo del fondo" (Object) e "commissioni associate al fondo" (Group) (percentuali di colonna)*

Obiettivo del fondo	Commissioni sul fondo: Commissioni prelevate dalle attività del fondo	Commissioni differite	Commissioni di ingresso	Commissioni multiple	Fondi senza commissioni	Totale
Fondo a capitalizzazione integrale	23.5	0.0	36.8	34.8	29.9	30.4
Fondo misto	76.5	100.0	63.2	65.2	70.1	69.6
Totale	100.0	100.0	100.0	100.0	100.0	100.0

Due variabili categoriali: diagrammi a barre

Una rappresentazione grafica delle tabelle di contingenza può essere fornita dal diagramma a barre non in pila, che qui sotto viene visualizzato nella forma della frequenza assoluta.

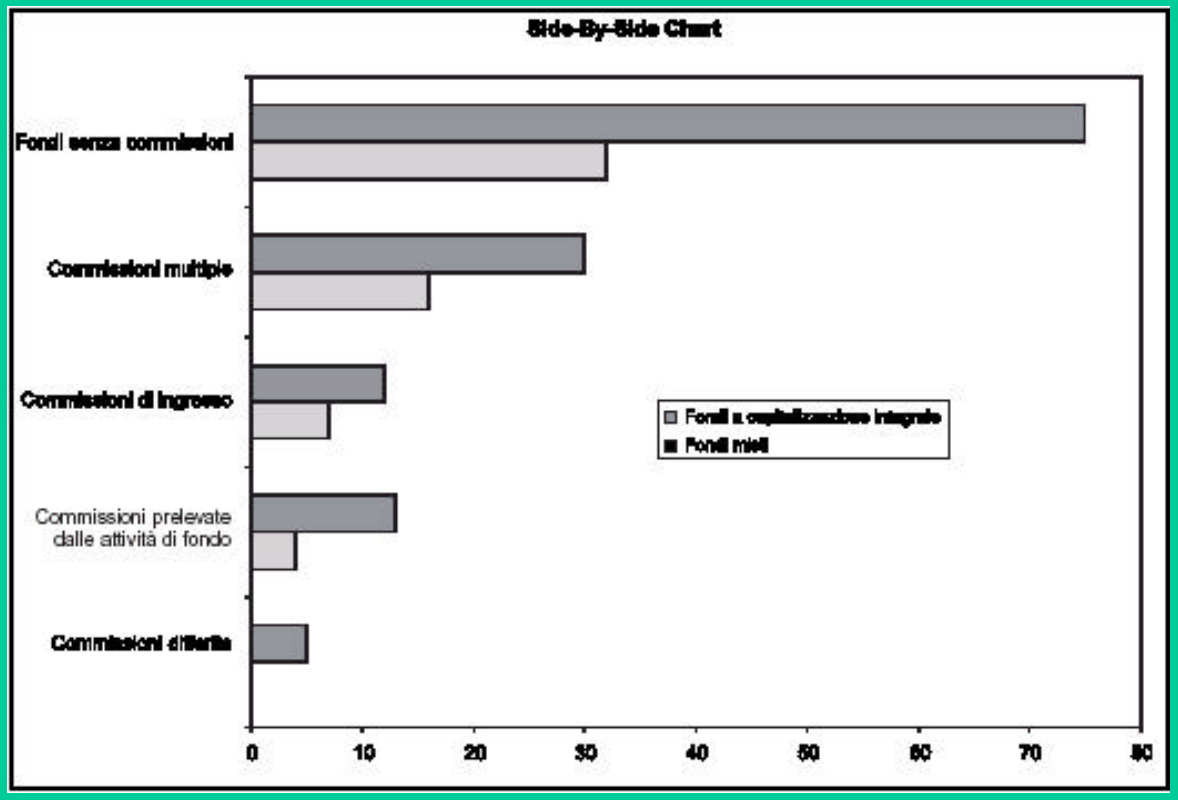


FIGURA 2.8
Diagramma a barre non in pila della variabile "commissioni associate al fondo" (Group) rispetto alla variabile "obiettivo del fondo" (Object)

RICHIAMI DI STATISTICA DESCRITTIVA E DI INFERENZA:

SINTESI E DESCRIZIONE DEI DATI QUANTITATIVI

Misure di Posizione (o 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 posizione:

- Media
  - Mediana
    - ✓ Moda
      - – Quartili

Misure di posizione: 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 (vedi Appendice B)

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

Misure di posizione: 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:

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

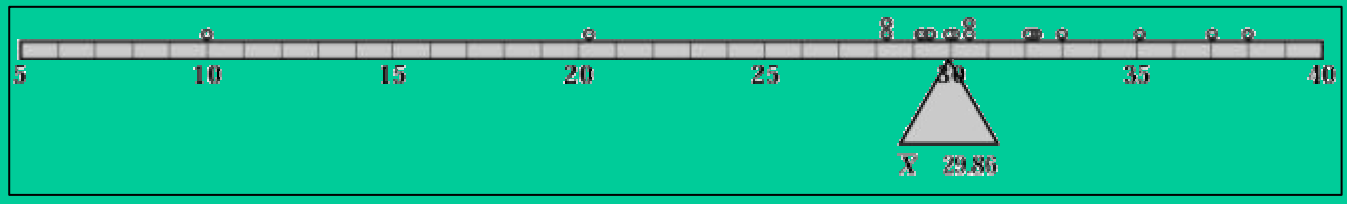
- La media si presenta come un “punto di equilibrio” tale che le osservazioni più piccole bilanciano quelle più grandi.
- Il calcolo della media si basi su tutte le osservazioni (X_1, X_2, X_3, . . . , X_n) dell’insieme di dati, proprietà questa che non è presentata da nessun’altra misura di posizione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

Misure di posizione: la Media

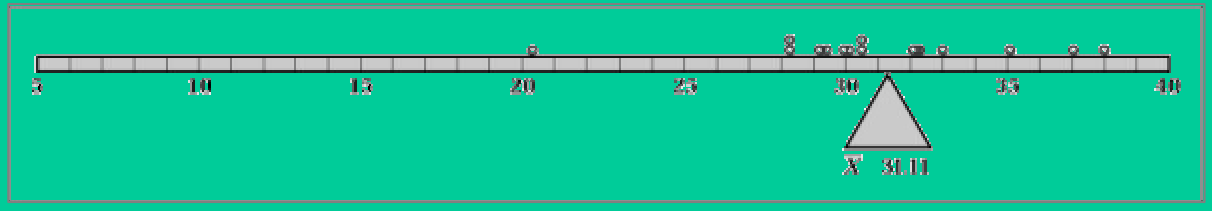
Commento: quando usare la Media Aritmetica

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 posizione: la Mediana

La **mediana** è il valore centrale in una successione ordinata di dati.

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.

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 posizione: 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 posizione: 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.

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 posizione: 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.

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

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 posizione: 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 variabilità

Una seconda caratteristica importante di un insieme di dati è la variabilità: la quantità di dispersione presente nei dati.

Due insiemi di dati possono differire o nella posizione o nella variabilità oppure sia nella posizione che nella variabilità.

FIGURA 3.3
Due distribuzioni simmetriche a forma campanulare che differiscono solo nella posizione

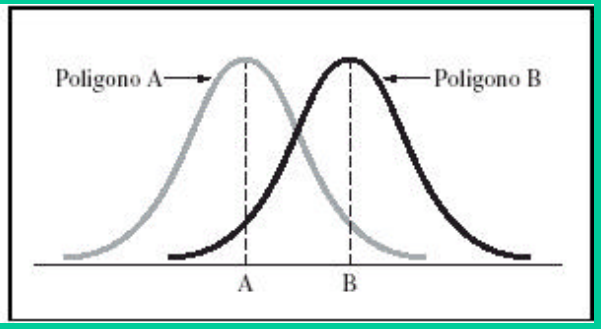
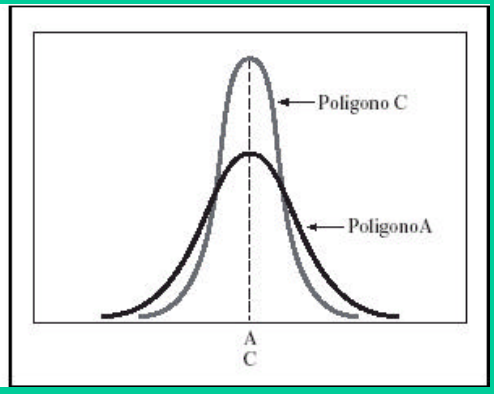


FIGURA 3.4
Due distribuzioni simmetriche a forma campanulare che differiscono solo nella variabilità



Tipi di misure di variabilità:

- ✓ Varianza
  - – Scarto Quadratico Medio
    - » Coefficiente di variazione

Misure di variabilità: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.

Varianza e la sua radice quadrata, lo scarto quadratico medio, invece 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 variabilità: lo Scarto Quadratico Medio

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

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

Misure di variabilità: 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 il simbolo 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:

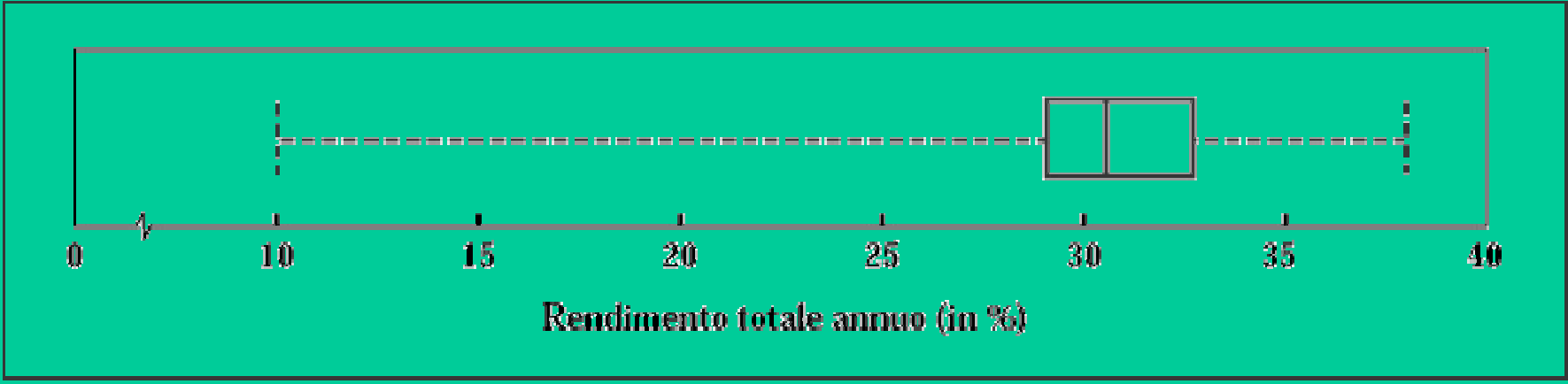
$$CV = \left(\frac{S}{\bar{X}}\right)100\% = \left(\frac{6.42}{29.86}\right)100\% = 21.5\%$$

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

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.

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

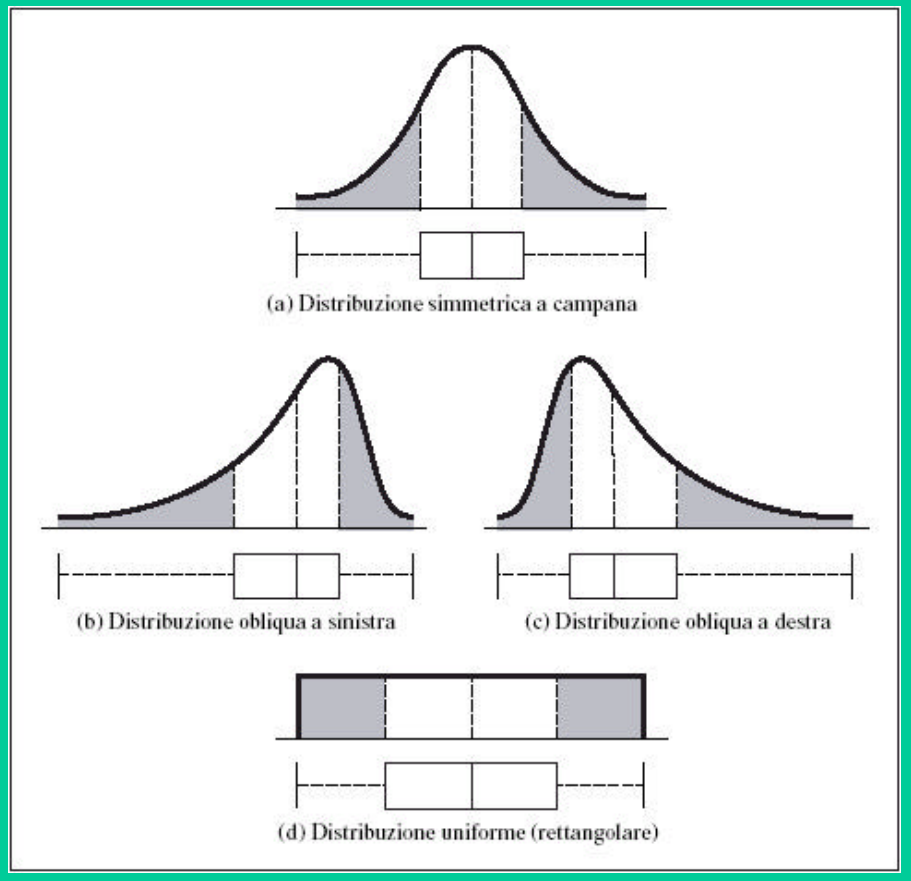
Il diagramma scatola e baffi o (o **boxplot**) fornisce una rappresentazione grafica dei dati sulla base dei cinque numeri di sintesi.



Linea verticale al centro della scatola Þ **mediana**	**Linea verticale a sinistra della scatola** Þ **Q_1**	**Linea verticale a destra della scatola** Þ **Q_3**
Linea tratteggiata a sinistra Þ **minimo**	**Linea tratteggiata a destra** Þ **massimo**	

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.**

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 a individuare e segnalare osservazioni anomale (gli outlier).

RICHIAMI DI STATISTICA DESCRITTIVA E DI INFERENZA:

LA PROBABILITA’

La probabilità

La probabilità rappresenta uno strumento indispensabile per poter utilizzare l'informazione contenuta nel campione al fine di fare inferenza su una popolazione più ampia.

Probabilità: definizione e tipi di approccio

La probabilità **può essere definita come il grado di verosimiglianza con cui un evento è destinato a verificarsi.**

La probabilità è una proporzione o frazione che varia tra i valori 0 e 1, estremi inclusi. Associamo il valore zero a un evento che non ha nessuna possibilità di verificarsi (*evento impossibile*) e il valore uno a un evento che si verificherà sicuramente (*evento certo*).

Secondo l'approccio classico, nel semplice caso che ciascun risultato sia ugualmente probabile, la probabilità che un evento si verifichi è definita nel seguente modo:

Probabilità del verificarsi di un evento

$$\text{Probabilità del verificarsi di un evento} = \frac{X}{T} \qquad (4.1)$$

dove

X = numero di risultati favorevoli all'evento

T = numero di risultati possibili

Distribuzione di probabilità di una variabile aleatoria discreta

Una **variabile aleatoria discreta** è una variabile quantitativa tale che ad ogni valore (modalità) che essa può assumere è associata una certa probabilità.

Il numero di valori di una variabile aleatoria discreta a cui è associata probabilità non nulla è finito o al più numerabile.

La distribuzione di probabilità di una variabile aleatoria discreta è data dall'elenco delle modalità che la variabile assume, a ciascuna delle quali è associata la relativa probabilità.

La somma di tutte le probabilità di una data distribuzione di probabilità deve essere uguale a uno.

Distribuzione di probabilità di una variabile aleatoria discreta

Il valore atteso di una variabile aleatoria discreta è una media ponderata delle modalità assunte dalla variabile, dove i coefficienti di ponderazione sono le probabilità associate a ciascun valore.

In genere si indica con m oppure con E(X), dove X è la variabile casuale.

Indicando con X_i l' i-esimo valore di X e con $P(X_i)$ la probabilità associata a quel valore, formalmente si ha:

$$m = E(X) = \sum_{i=1}^{N} X_i P(X_i)$$

Distribuzione di probabilità di una variabile aleatoria discreta

La varianza di una variabile aleatoria discreta è una media ponderata dei quadrati delle differenze tra ciascun valore e il valore atteso delle variabile dove i coefficienti di ponderazione sono rappresentati dalle probabilità associate a ciascuna modalità.

Il simbolo usato per rappresentare la varianza è s^2.

Indicando con X_i l' i-esimo valore di X e con $P(X_i)$ la probabilità associata a quel valore, formalmente si ha:

$$s^2 = \sum_{i=1}^{N} [X_i - m]^2 P(X_i)$$

Lo scarto quadratico medio di una variabile aleatoria discreta è:

$$s = \sqrt{\sum_{i=1}^{N} [X_i - m]^2 P(X_i)}$$

La distribuzione di probabilità di una variabile aleatoria discreta

Esempio: ad ognuno dei due impianti produttivi A e B è associata una distribuzione di probabilità per la variabile che misura il numero di unità produttive non conformi in un mese.

	Impianto A			Impianto B		
X	P(X)	XP(X)	$(X-\mu)^2P(X)$	P(X)	XP(X)	$(X-\mu)^2P(X)$
0	0.32	0	0.52	0.21	0	1.18
1	0.35	0.35	0.03	0.23	0.23	0.43
2	0.18	0.36	0.10	0.14	0.28	0.02
3	0.08	0.24	0.24	0.12	0.36	0.05
4	0.04	0.16	0.30	0.1	0.4	0.27
5	0.02	0.1	0.28	0.1	0.5	0.69
6	0.01	0.06	0.22	0.1	0.6	1.32
TOTALE	1	1.27	1.68	1	2.37	3.95
	m =	1.27		m =	2.37	
	s =	1.30		s =	1.99	

XXX

La distribuzione di probabilità di una variabile aleatoria discreta

La distribuzione di probabilità è un modello matematico tramite il quale è possibile rappresentare schematicamente un fenomeno.

Per esempio la distribuzione di probabilità dei possibili risultati della prima estrazione nel gioco del lotto è detta distribuzione di probabilità uniforme in quanto assegna una probabilità costante pari a 1/90 a tutti i 90 possibili risultati.

Altri tipi di modelli matematici sono stati sviluppati per rappresentare diversi fenomeni discreti tipici delle scienze sociali, naturali, ingegneristiche ed economiche.

In particolare prenderemo in considerazione i seguenti modelli:

- **Modello binomiale**
- **Modello di Poisson**

La distribuzione binomiale

Praticamente la distribuzione binomiale è la legge della variabile aleatoria che rappresenta il numero di successi ottenuti in un campione di n osservazioni.

Su n osservazioni il numero di successi è un intero compreso tra 0 ed n.

Distribuzione binomiale:

$$P(X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X}$$

Con n! = n · (n-1) · (n-2) · … · 2 · 1

La distribuzione binomiale

Il valore atteso di una distribuzione binomiale è uguale al prodotto tra l'ampiezza del campione n e la probabilità di successo p:

$$m = E(X) = np$$

Lo scarto quadratico medio di una distribuzione binomiale è dato da:

$$s = \sqrt{np(1-p)}$$

Esempio: se prelevo un campione di 4 prodotti finiti da un processo produttivo essendo 0.1 la probabilità che ogni prodotto risulti non conforme, la probabilità di avere meno di 3 prodotti non conformi è data da

P(X<3) = P(X=0)+P(X=1)+P(X=2) = 0.6561+0.2916+0.0486 = 0.9963

In media avrò (4)·(0.1)=0.4 prodotti non conformi e una variabilità (scarto quadratico medio) pari a 0.6

Distribuzione di probabilità di una variabile aleatoria continua

Una variabile aleatoria continua è una variabile quantitativa continua a cui è associata una funzione di densità di probabilità f(x) tale che la probabilità che la variabile aleatoria X assuma valori compresi in un dato intervallo (a,b) è data da

$$P(a < X < b) = \int_a^b f(x)dx$$

Per le variabili aleatoria continue la probabilità che X assuma un particolare valore è pari a zero.

L'integrale definito della funzione di densità di probabilità su tutta la retta reale deve essere pari a 1 cioè

$$P(-\infty < X < +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

La distribuzione Normale

La distribuzione normale è la distribuzione continua più usata in assoluto.

Tra i motivi del suo grande successo ne citiamo due:

- **Diversi fenomeni continui sembrano seguire, almeno approssimativamente, una distribuzione normale**
- **La distribuzione normale può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete.**

Le principali proprietà sono:

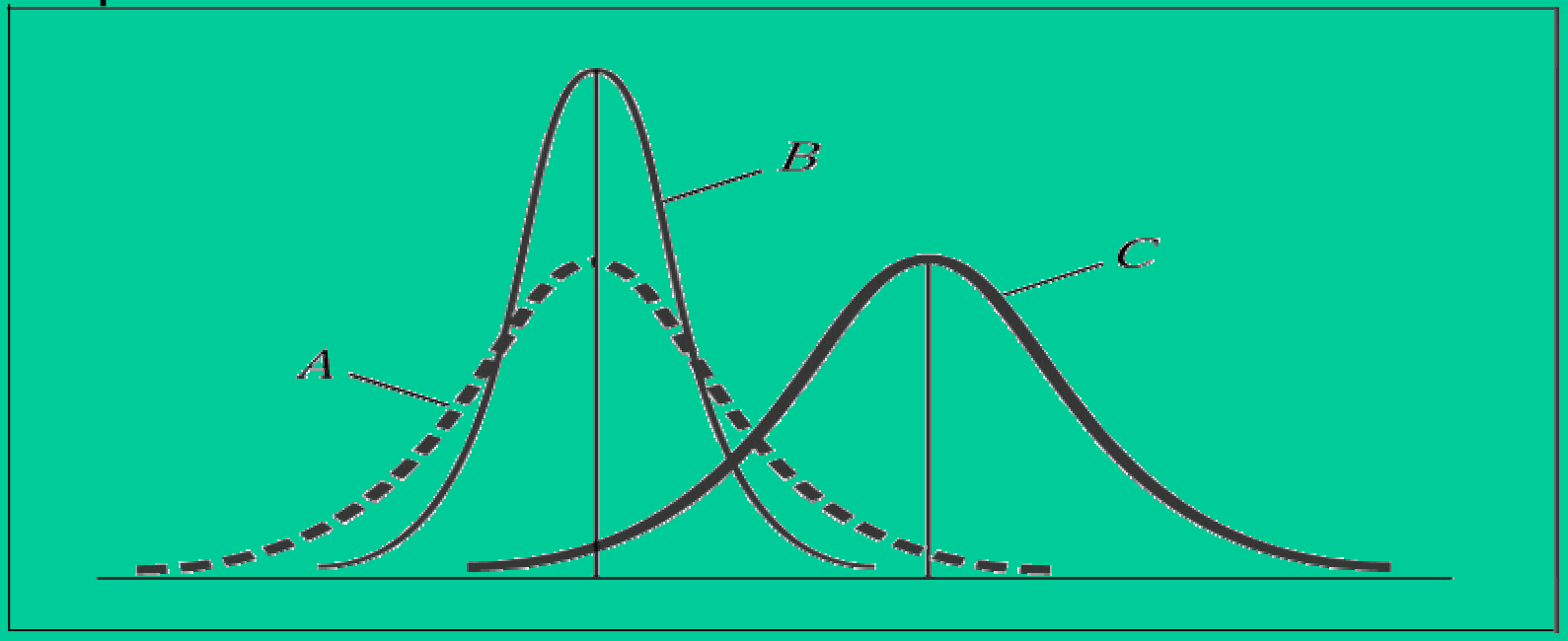
- **La distribuzione normale ha una forma campanulare e simmetrica**
- **Le sue misure di posizione centrale (valore atteso, moda, mediana, midrange) coincidono**
- **Il suo range interquartile è pari a 1.33 volte lo scarto quadratico medio, cioè copre un intervallo compreso tra $\mu - 2/3\sigma$ e $\mu + 2/3\sigma$**
- **La variabile aleatoria normale assume valori compresi tra $-\infty$ e $+\infty$.**

La distribuzione Normale

La funzione di densità di probabilità della distribuzione normale è data da:

$$f(X) = \frac{1}{\sqrt{2ps}} e^{-\frac{1}{2}(\frac{X-m}{s})^2}$$

La funzione di densità di probabilità ha una forma tipica campanulare che dipende dai parametri m **e** s**.**



La distribuzione Normale

La distribuzione normale cumulativa F(z) è data da P(X<z):

$$\int_{-\infty}^{z} f(x)dx$$

Essa coincide con l'area compresa tra la curva della funzione di densità, l'asse delle x e la retta perpendicolare all'asse x passante per il punto (z;0).

Sottraendo ad X la media e dividendo per lo scarto quadratico medio otteniamo la variabile aleatoria normale standardizzata, che è distribuita come una normale con media 0 e varianza 1:

$$Z = \frac{(X - m)}{s}$$

La distribuzione Normale

Conoscendo media e scarto quadratico medio di una variabile casuale normale X è possibile, ricorrendo alle tavole della distribuzione cumulativa della normale standardizzata, calcolare la funzione cumulativa di X in un certo punto x_1. Infatti:

$$P(X< x_1)=P[(X-\mu)/\sigma)< (x_1 -\mu)/\sigma] = P(Z<z_1) = F(z_1)$$

Viceversa, conoscendo media e scarto quadratico medio di una variabile casuale normale X è possibile, ricorrendo alle tavole della distribuzione cumulativa della normale standardizzata, calcolare il valore x_1 corrispondente ad un valore noto β della distribuzione cumulata. Infatti:

ricavando il valore z_1 tale che $F(z_1)= \beta$, dalla relazione $(x_1-\mu)/\sigma) = z_1$ ricavo $x_1 = \sigma z_1+\mu$

La distribuzione Normale

Esempio: il responsabile di un processo di assemblaggio in una fabbrica di automobili ha stabilito che il tempo necessario per assemblare un certo pezzo può essere considerato come una variabile aleatoria normale di parametri μ=75 (secondi) e σ=6 (secondi).

> Qual è la probabilità che un addetto scelto a caso impieghi un tempo superiore a 81 secondi ad assemblare un pezzo?

P(X>81) = 1-P(X<81) = 1- F **[(81-75)/6] = 1-** $F(1) = 1 - 0.8413 = 0.1587$

> Qual è la probabilità che un addetto scelto a caso impieghi un tempo compreso tra 69 e 81 secondi ad assemblare un pezzo?

P(69<X<81)=P(X<81)-P(X<69)= F **[(81-75)/6]-** F **[(69-75)/6]=** $F(1) - F(-1) = 0.8413 - 0.1587 = 0.6826$

> Qual'è il valore di X la cui probabilità cumulata è pari a 0.10?

F **(Z)=0.10 => Z= -1.2 => X = 75 + 6 (-1.28) = 67.32**

RICHIAMI DI STATISTICA DESCRITTIVA E DI INFERENZA:

DISTRIBUZIONI CAMPIONARIE E INTERVALLI DI CONFIDENZA

Le distribuzioni campionarie

Uno degli scopi principali dell'analisi dei dati consiste nell'uso delle statistiche, come la media campionaria e la proporzione campionaria, per stimare i corrispondenti parametri delle rispettive popolazioni.

Lo scopo dell'INFERENZA è di trarre conclusioni sulla popolazione e non sul campione.

Nella pratica, da una popolazione viene estratto a caso un solo campione, di ampiezza prestabilita.

Per usare le statistiche campionarie allo scopo di stimare i parametri della popolazione, dovremmo prendere in considerazione la distribuzione campionaria, cioè la distribuzione di tutti i possibili campioni che possono essere estratti dalla popolazione.

La distribuzione della media campionaria

Se la variabile casuale X è distribuita come una normale di media m e scarto quadratico medio s, in simboli X ~ N(m;s), allora la media campionaria è distribuita anch'essa come una normale di parametri m e $s / \sqrt{n}$

Per standardizzare la media campionaria, cioè trasformarla in una normale di media nulla e scarto quadratico medio unitario, sarà perciò sufficiente operare come segue:

Calcolo di Z per la distribuzione della media campionaria

Il valore Z è uguale alla differenza tra la media campionaria $\bar{X}$ e la media della popolazione μ, divisa per l'errore standard della media della media $\sigma_{\bar{X}}$.

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (5.4)$$

Esempio: se X ~ N(368;15), trovare la probabilità che la media di un campione casuale di numerosità 25 assuma valori inferiori a 365. Trovare la probabilità che una singola osservazione assuma un valore inferiore a 365.

La distribuzione della media campionaria

Si presentano spesso casi in cui la distribuzione della popolazione non è normale.

In questi casi è utile fare riferimento ad un importante teorema della statistica.

Teorema del limite centrale: quando l'ampiezza del campione diventa sufficientemente grande, la distribuzione della media campionaria può essere approssimata dalla distribuzione normale. Questo vale indipendentemente dalla distribuzione dei singoli valori della popolazione.

Riquadro 5.2 La distribuzione normale e la distribuzione della media campionaria

✓ 1. Per la maggior parte delle popolazioni, indipendentemente dalla forma della loro distribuzione, la distribuzione della media campionaria è approssimativamente normale, purché si considerino campioni di almeno 30 osservazioni.

✓ 2. Se la distribuzione della popolazione è abbastanza simmetrica, la distribuzione della media campionaria è approssimativamente una normale, purché si considerino campioni di almeno 15 osservazioni.

✓ 3. Se la popolazione ha una distribuzione normale, la media campionaria è distribuita secondo la legge normale, indipendentemente dall'ampiezza del campione

Stima puntuale e stima intervallare

Esistono due tipi fondamentali di stimatori:

- **Stimatore puntuale**
- **Stimatore intervallare**

Stimatore puntuale: singola statistica che viene usata per stimare il vero valore di un parametro della popolazione. Ad esempio la media campionaria è uno stimatore puntuale della media della popolazione m, la varianza campionaria è uno stimatore puntuale della varianza della popolazione s^2, ecc.

Stima puntuale e stima intervallare

Stimatore intervallare: **intervallo di valori che ha una certa probabilità o confidenza di comprendere il vero valore del parametro della popolazione.**

Esempio: se X ~ N(m;15), trovare un intervallo di confidenza del 95% per la media della popolazione sapendo che ho estratto un campione di 25 osservazioni con media campionaria pari a 362,12.

Sappiamo che l'intervallo in cui cade il 95% dei valori della media campionaria ha estremi:

$$(m - Zs/\sqrt{n}) e (m + Zs/\sqrt{n})$$

Sostituendo al parametro (ignoto) m della popolazione, il valore della media campionaria, otteniamo l'intervallo di confidenza cercato:

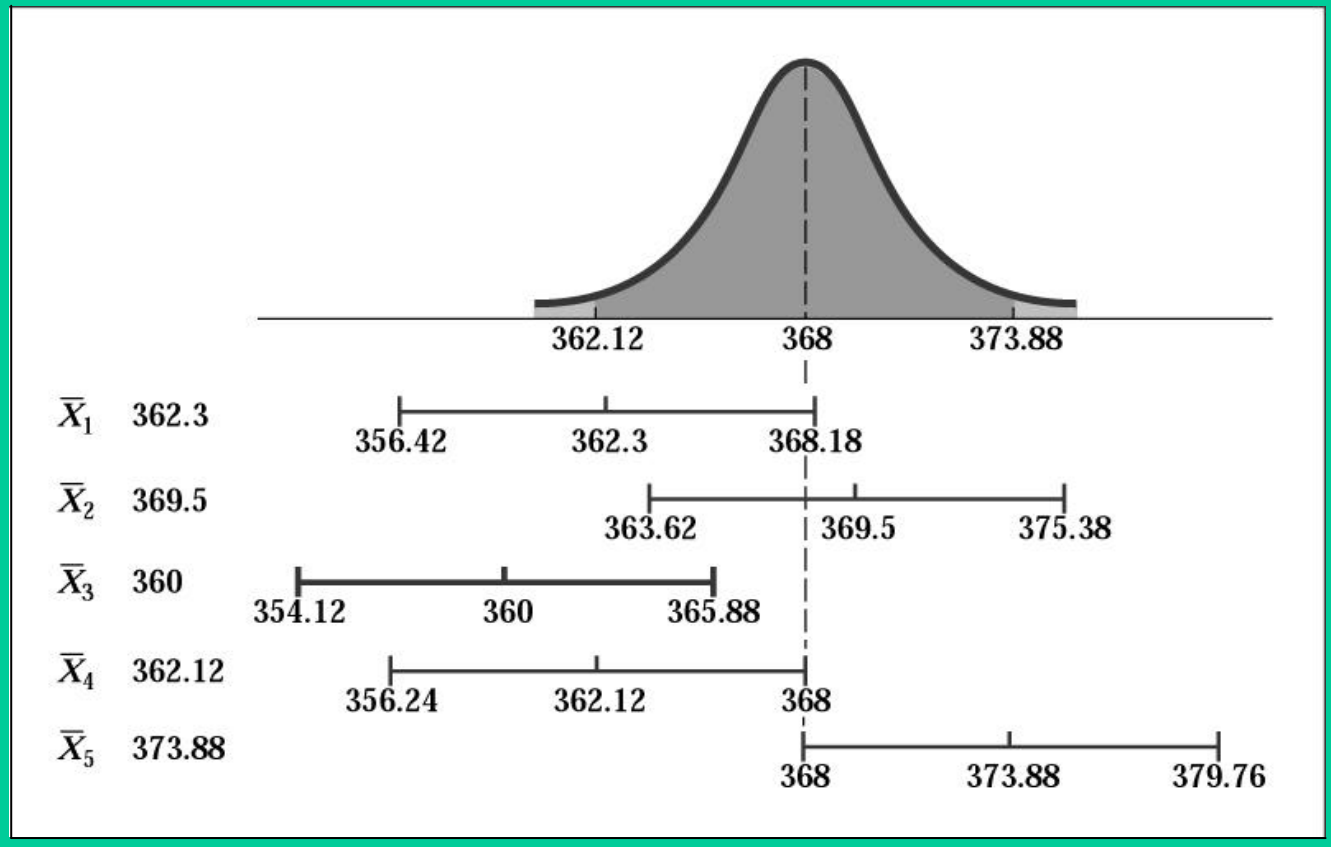
$$(\overline{X} - Zs/\sqrt{n}) e (\overline{X} + Zs/\sqrt{n})$$

La soluzione al problema è (362,12)-(1,96)(15)(5) e (362,12)+(1,96)(15)(5), cioè 356,24 e 368,00.

In generale il livello di confidenza è indicato con (1-a)% dove a è la probabilità che si trova nelle code della distribuzione, al di fuori dell'intervallo di confidenza (la probabilità della coda sinistra e della coda destra coincidono e sono pari a a/2).

Intervalli di confidenza

Intervalli di confidenza per cinque diversi campioni di ampiezza n=25, estratti da una popolazione normale con μ = 368 e s = 15



Intervalli di confidenza

Generalizzando la formula per la costruzione degli intervalli di confidenza conoscendo il valore dello scarto quadratico medio della popolazione si ottiene:

Intervallo di confidenza per la media (σ noto)

$$\bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

oppure

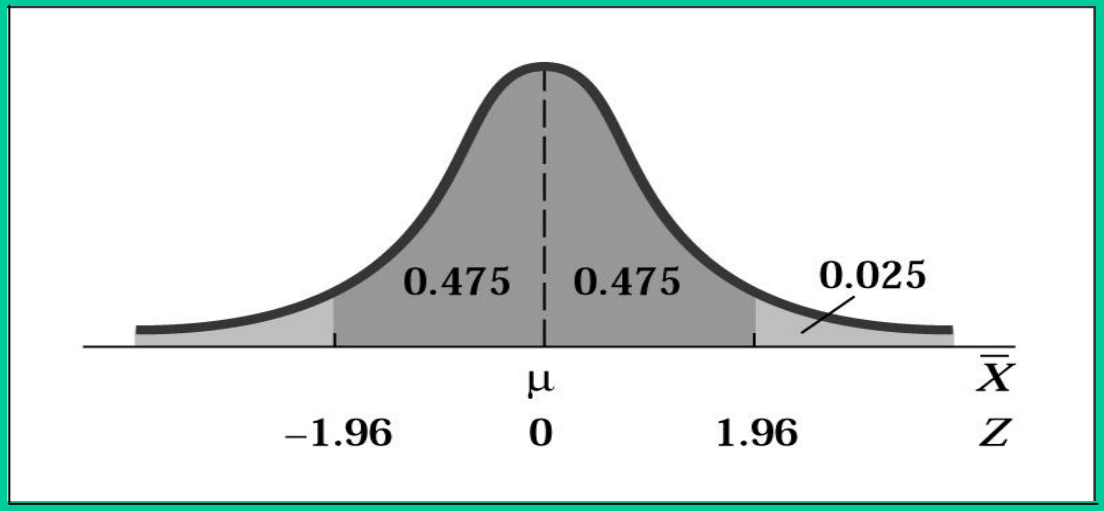
$$\bar{X} - Z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z\frac{\sigma}{\sqrt{n}} \qquad (5.7)$$

dove

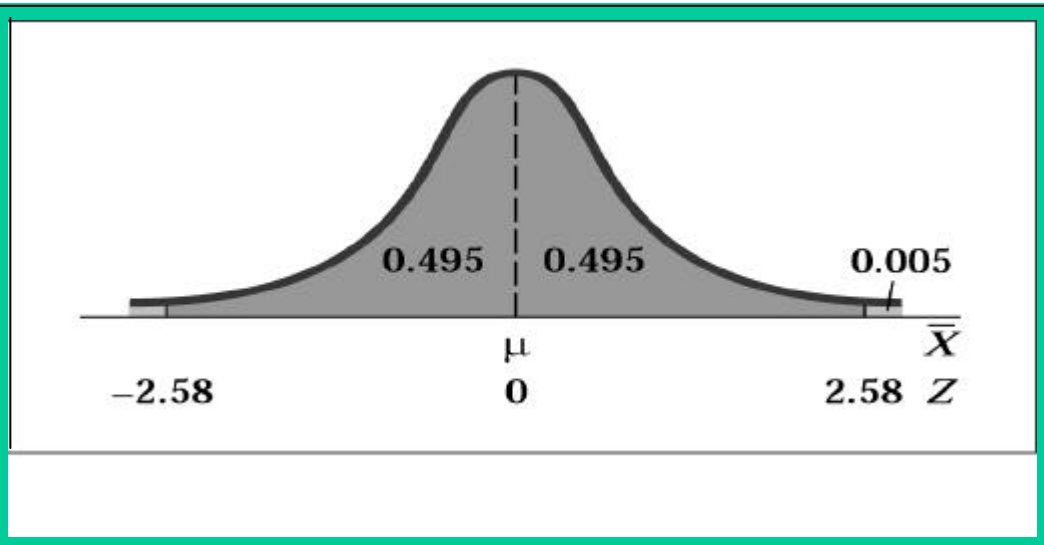
Z è quel valore tale che l'area sottesa alla curva normale standardizzata tra $-Z$ e Z è pari a $(1 - \alpha)$. Notiamo inoltre che Z si lascia alla destra un'area pari a $\alpha/2$ e che l'area sottesa alla curva normale standardizzata tra 0 e Z è pari a $(1 - \alpha)/2$.

Intervalli di confidenza

Curva normale per determinare il valore di Z necessario per un livello di confidenza del 95%



Curva normale per determinare il valore di Z necessario per un livello di confidenza del 99%



Intervalli di confidenza

In genere lo scarto quadratico medio della popolazione s, al pari della media m, non è noto. Pertanto, per ottenere un intervallo di confidenza per la media della popolazione possiamo basarci sulle sole statistiche campionarie.

La statistica utile per costruire intervalli di confidenza per la media è

$$t = \frac{\overline{X} - m}{\frac{S}{\sqrt{n}}}$$

Se la variabile casuale X ha una distribuzione normale allora la statistica t ha una distribuzione t di Student con n-1 gradi di libertà.

La distribuzione t di Student ha una forma molto simile a quella della normale standardizzata. Tuttavia il grafico risulta più appiattito e l'area sottesa sulle code è maggiore di quella della normale a causa del fatto che s non è noto e viene stimato da S. L'incertezza su s causa la maggior variabilità di t.

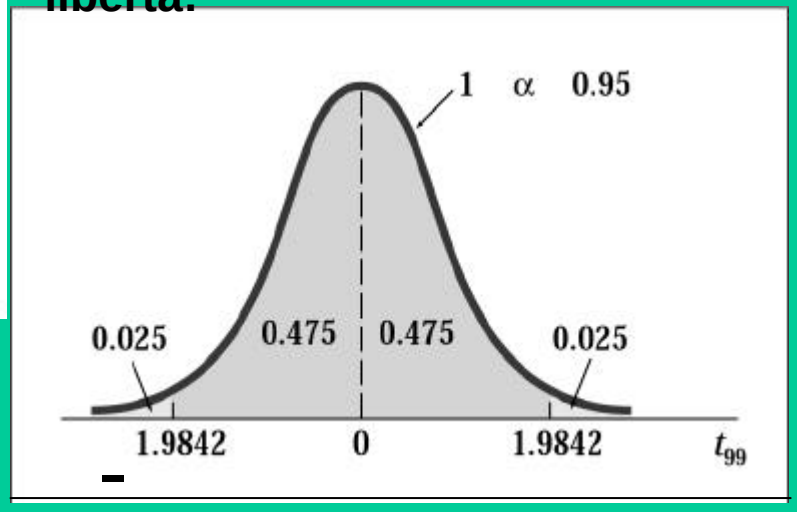
Intervalli di confidenza

Le tavole della distribuzione t di Student forniscono la probabilità (l'area sottesa) a destra del valore indicato.

Tabella 5.4 *Determinazione del valore critico della t con 99 gradi di libertà necessario per un livello di confidenza del 95%*

	Area nella coda di destra					
Gradi di libertà	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259

Distribuzione t con 99 gradi di libertà:



Intervalli di confidenza

L'intervallo di confidenza di livello (1-a)% per la media con s ignoto è definito come segue:

Intervallo di confidenza per la media (σ non noto)

$$\bar{X} \pm t_{n-1}\frac{S}{\sqrt{n}}$$

oppure

$$\bar{X} - t_{n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1}\frac{S}{\sqrt{n}} \qquad (5.8)$$

dove

t_{n-1} 1 è il valore critico della distribuzione t con $n - 1$ gradi di libertà che si lascia alla destra un'area pari a $\alpha/2$.

Intervalli di confidenza

Per ricavare l'intervallo di confidenza per la proporzione della popolazione p, che ha una certa caratteristica, si utilizza la proporzione campionaria p_s.

Se sia np che n(1-p) sono uguali almeno a 5, la distribuzione di p_s può essere approssimata alla distribuzione normale.

L'errore standard della proporzione è dato da $s_p = \sqrt{\frac{p(1-p)}{n}}$

L'intervallo di confidenza di livello (1-a)% per la proporzione p si ricava come segue:

Intervallo di confidenza per la proporzione

$$p_s \pm Z\sqrt{\frac{p_s(1-p_s)}{n}}$$

oppure (5.9)

$$p_s - Z\sqrt{\frac{p_s(1-p_s)}{n}} \leq p \leq p_s + Z\sqrt{\frac{p_s(1-p_s)}{n}}$$

dove

p_s = proporzione campionaria = $\frac{\text{numero di successi nel campione}}{\text{ampiezza campionaria}}$

p = proporzione della popolazione

Z = valore critico della normale standardizzata

n = ampiezza del campione