Organizzazione dei dati in forma tabellare e grafica

Insegnamento: Statistica

Corso di Laurea Triennale in Economia

Dipartimento di Economia e Management

Università di Ferrara

Docenti: Prof. S. Bonnini, Dott.ssa V. Mini, Dott.ssa Brunori

Argomenti

- Il dataset
- Il dotplot
- La frequenza
- Tabelle e grafici di frequenza
- Il diagramma di Pareto
- La tabella di contingenza

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

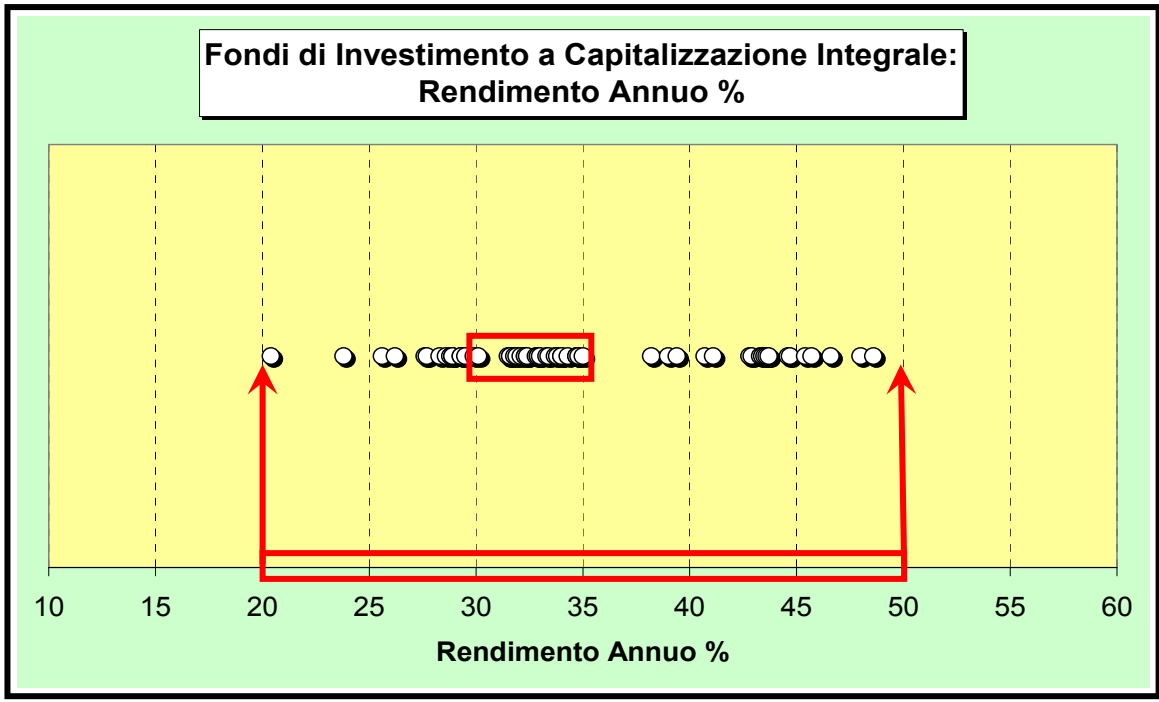
Unità statistica

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 per rappresentare adeguatamente il fenomeno diventa necessario utilizzare degli strumenti grafici.

Se raffiguriamo in un grafico (denominato *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 ⇒ *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

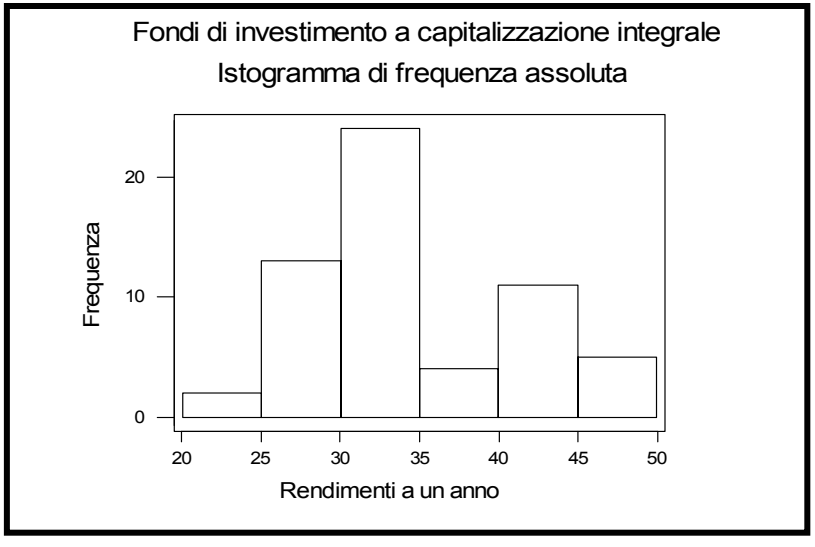
Tabella e Istogramma di frequenza assoluta e relativa

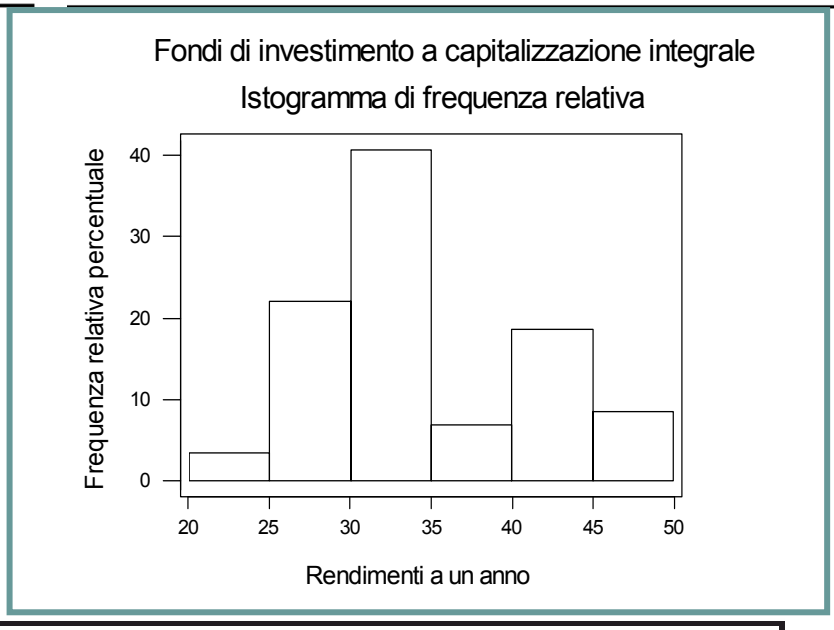
Tabella 2.2 *Distribuzione delle frequenze dei rendimenti percentuali a un anno realizzati dai 59 fondi a capitalizzazione integrale*

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

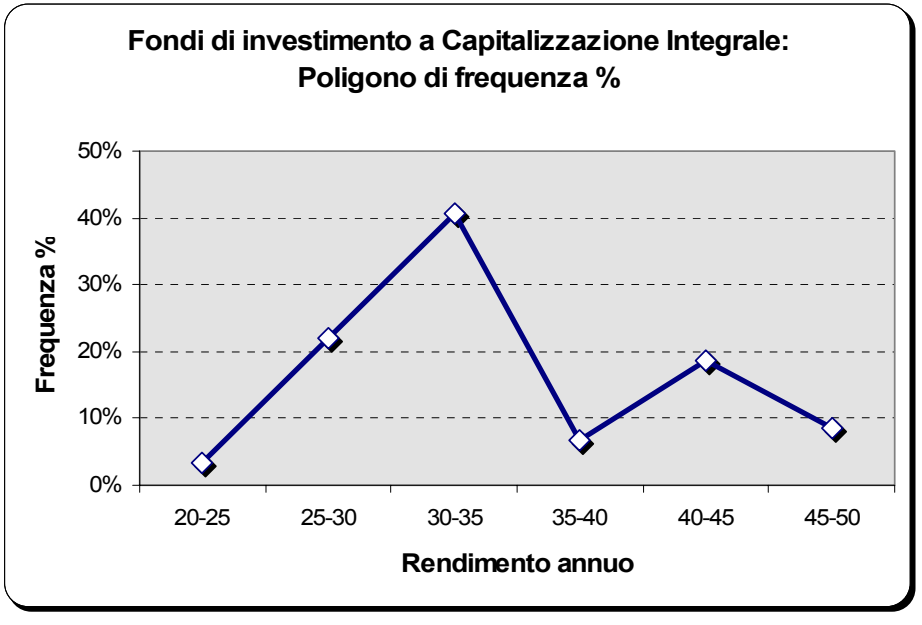




L'**istogramma** è un diagramma a barre verticali in cui le barre rettangolari hanno come base gli intervalli in cui sono state raggruppate le osservazioni

Il poligono: un'alternativa all'istogramma di frequenza

Anche nel caso del poligono l'asse orizzontale rappresenta il fenomeno oggetto dell'analisi, mentre sull'asse verticale viene indicato il numero, la percentuale o la frequenza relativa di osservazioni per ogni intervallo di raggruppamento.



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

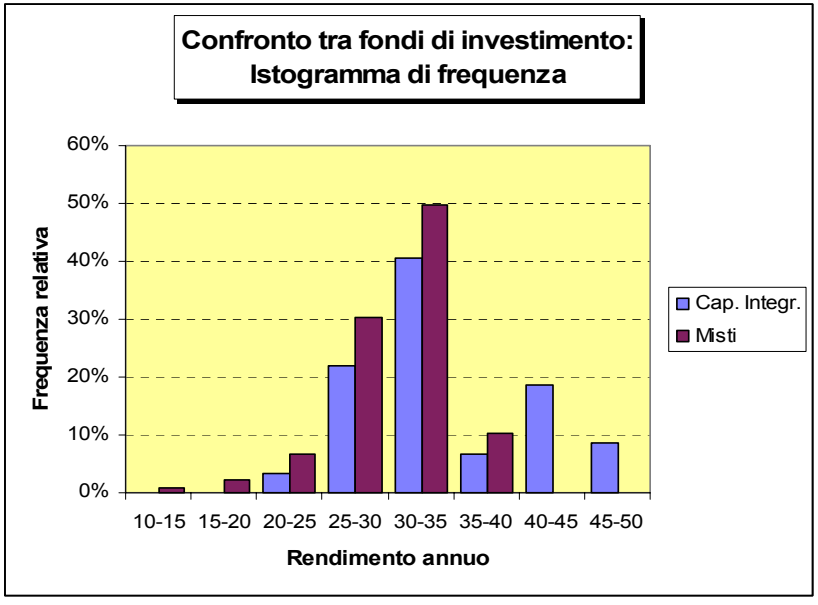
Il **punto medio di una classe**, valore a metà strada fra gli estremi, rappresenta convenzionalmente tutti i valori compresi nell'intervallo.

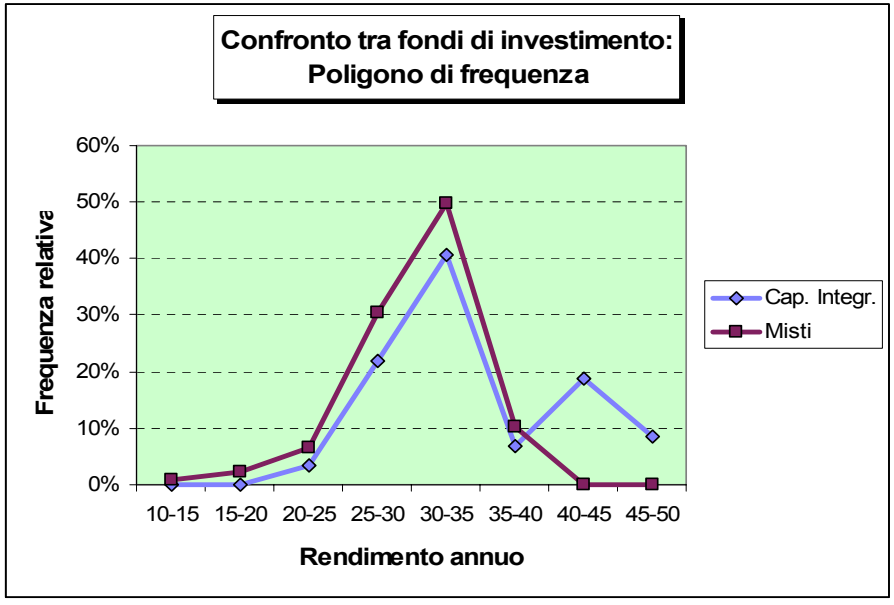
Tabella e Istogramma (o Poligono) di frequenza per il confronto tra due gruppi

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.



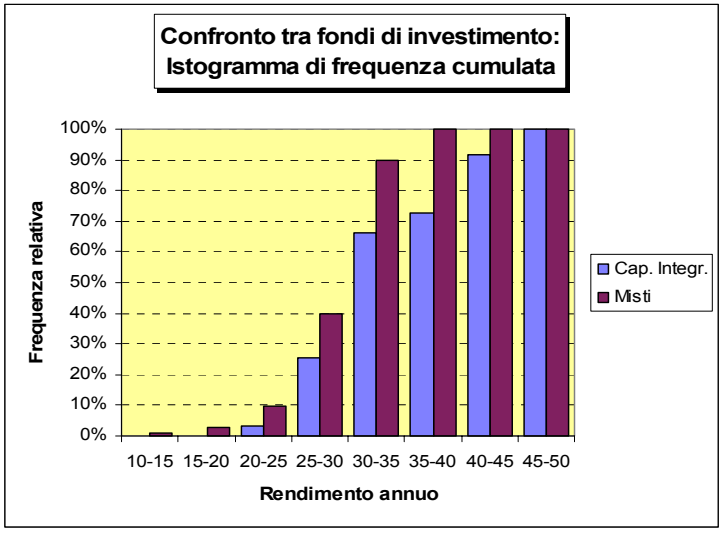


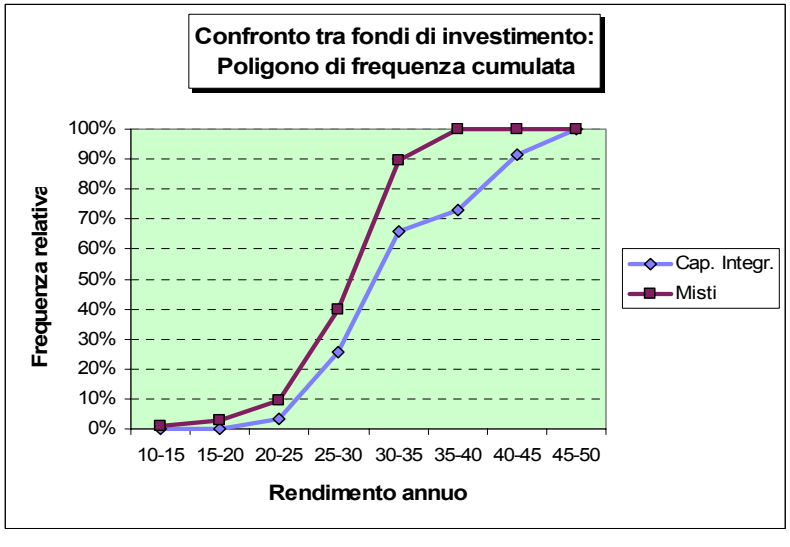
La frequenza cumulata

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

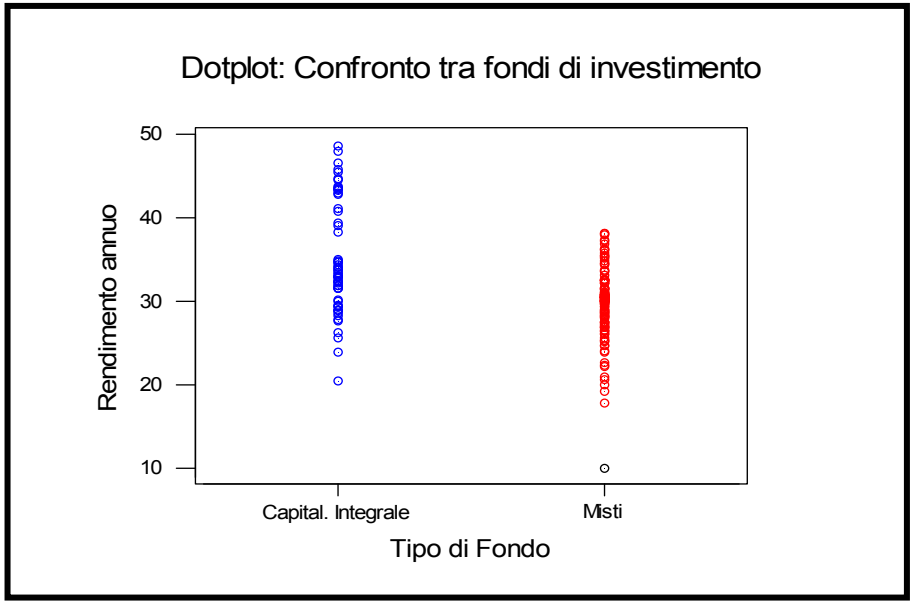
Rendimento Annuo	Formato della Frequenza Cumuta			
	Assoluta		Relativa	
	Tipo di Fondo		Tipo di Fondo	
	Cap. Integr.	Misti	Cap. Integr.	Misti
10-15		1	0.0%	0.7%
15-20		4	0.0%	3.0%
20-25	2	13	3.4%	9.6%
25-30	15	54	25.4%	40.0%
30-35	39	121	66.1%	89.6%
35-40	43	135	72.9%	100.0%
40-45	54	135	91.5%	100.0%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



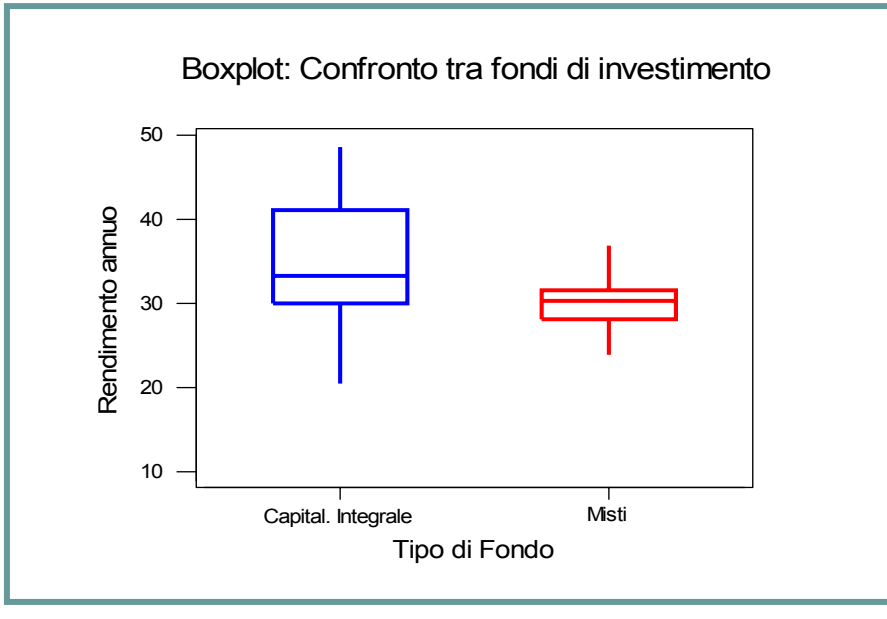


Dotplot e Boxplot: un'alternativa per il confronto tra due gruppi



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 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.

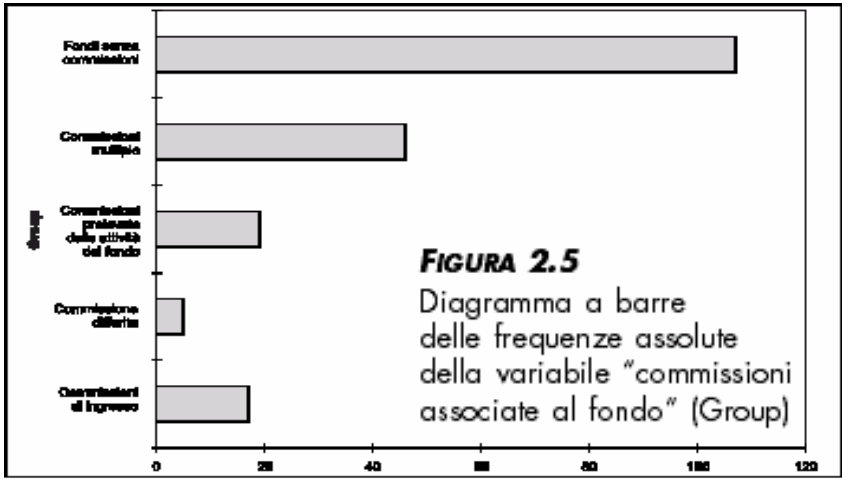


FIGURA 2.5
Diagramma a barre delle frequenze assolute della variabile "commissioni associate al fondo" (Group)

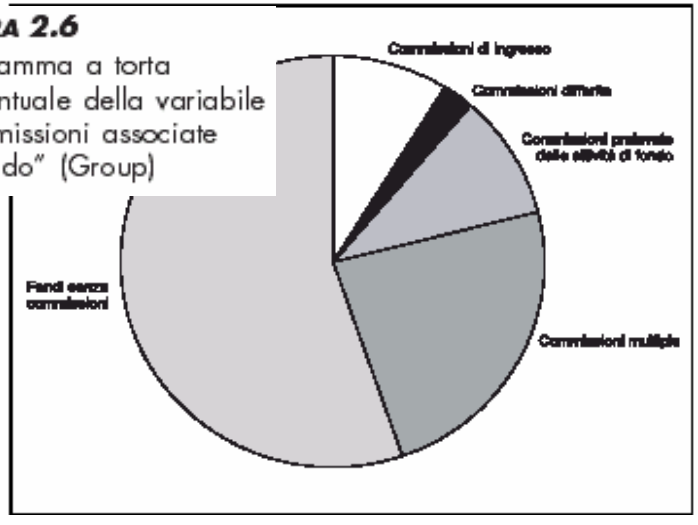


FIGURA 2.6
Diagramma a torta percentuale della variabile "commissioni associate al fondo" (Group)

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.

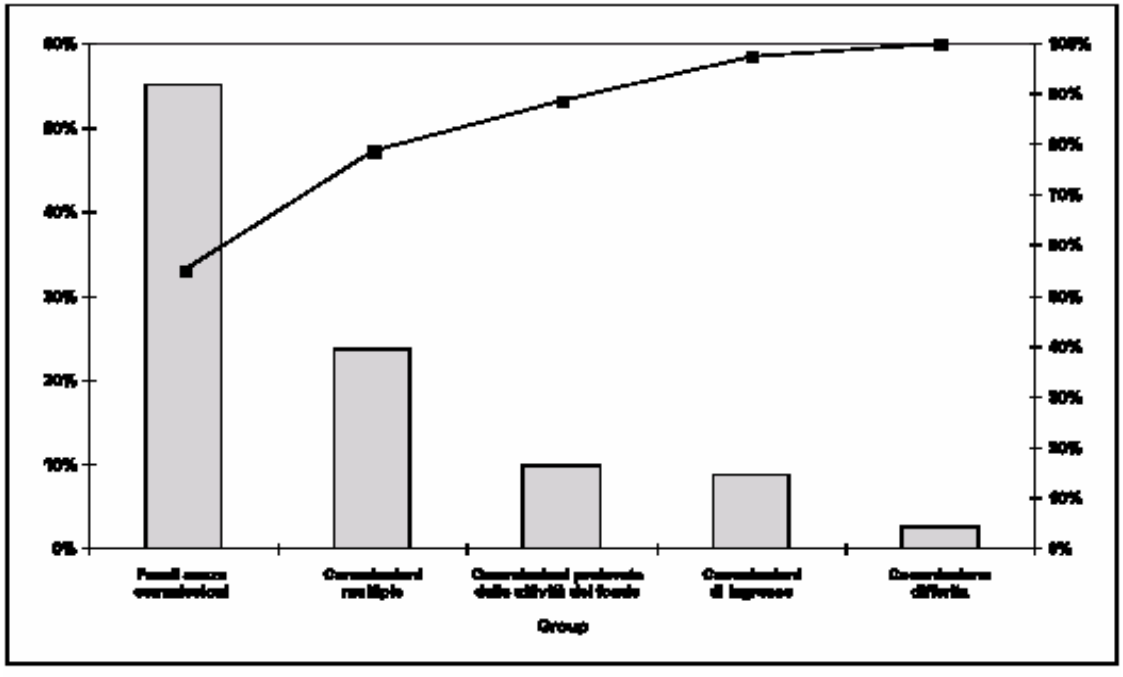


FIGURA 2.7
Diagramma di Pareto della variabile "commissioni associate al fondo" (Group)

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)*

	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	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

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.

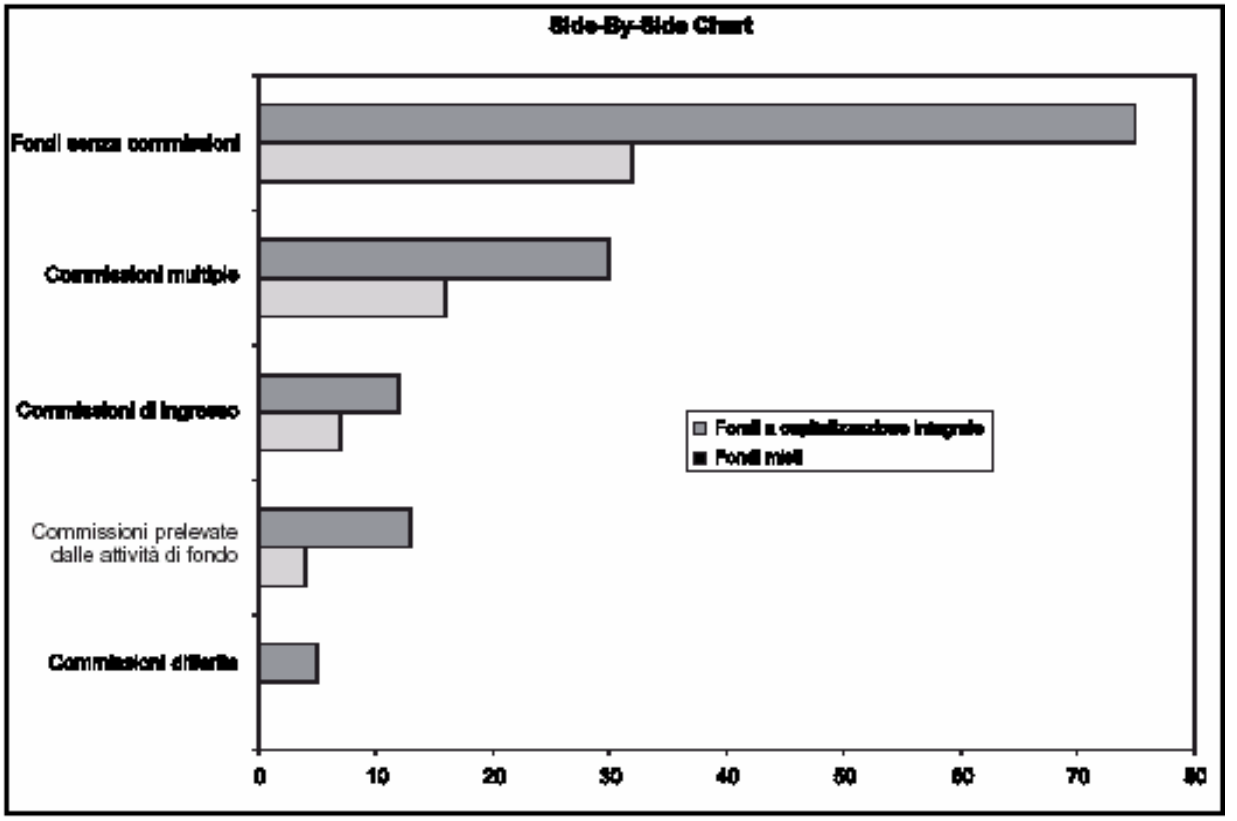


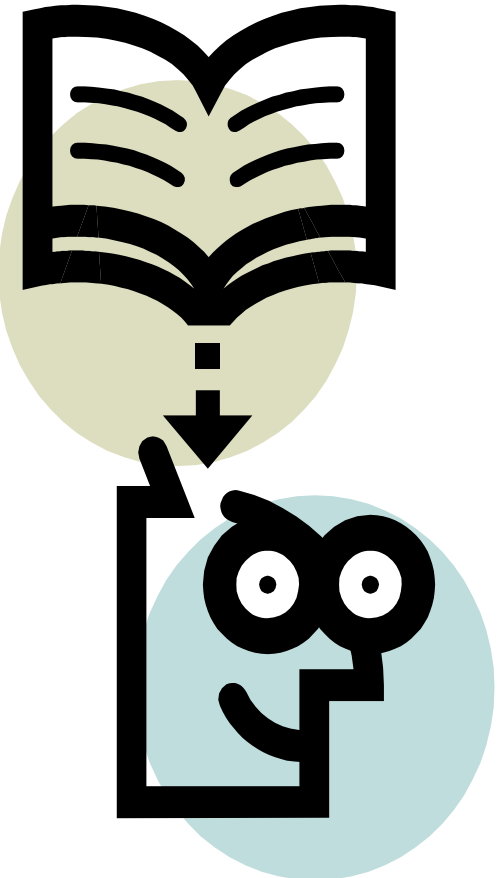
FIGURA 2.8
Diagramma a barre non in pila della variabile "commissioni associate al fondo" (Group) rispetto alla variabile "obiettivo del fondo" (Object)

Concetti principali della lezione

- Il dataset è la forma ottimale di organizzazione dei dati
- Il dotplot è un'utile rappresentazione grafica per visualizzare il range e l'intervallo di maggiore concentrazione dei dati
- La **frequenza** è il concetto fondamentale per lo studio della distribuzione dei valori della variabile di interesse
- La frequenza è rappresentabile in forma tabellare o grafica, in formato assoluto o percentuale ed anche nella variante di frequenza cumulata
- Per i dati qualitativi abbiamo degli strumenti analoghi a quelli dei dati quantitativi e un'ulteriore strumento dato dal diagramma di Pareto
- In presenza di due variabili categoriali possiamo avvalerci della tabella di contingenza e di diagrammi a barre

ESERCIZI

- Argomento 1
- 2
- 3
- 4
- 5



Esercizio 2.11

2.11 I prezzi dei monolocali di Queens, un distretto amministrativo di New York, variano da 103 000 a 295 000 dollari

a) Range =(295000-103000)=192000
Range/10=192000/10=19200

100000-120000	120000-140000	140000-160000	160000-180000	180000-200000
200000-220000	220000-240000	240000-260000	260000-280000	280000-300000

b) Ampiezza intervallo=19200≅20000
approssimazione per facilitare l'interpretazione

c) Punti medi:
110000, 130000, 150000, 170000, 190000
210000, 230000, 250000, 270000, 290000

Esercizio 2.18

Le società A e B hanno condotto un esperimento su 40 lampadine da 100 watt, sulle quali hanno misurato la durata di vita in ore.

- Dati grezzi

Società A					Società B				
684	697	720	773	821	819	836	888	897	903
831	835	848	852	852	907	912	918	942	943
859	860	868	870	876	952	959	962	986	992
893	899	905	909	911	994	1004	1005	1007	1015
922	924	926	926	938	1016	1018	1020	1022	1034
939	943	946	954	971	1038	1072	1077	1077	1082
972	977	984	1005	1014	1096	1100	1113	1113	1116
1016	1041	1052	1080	1093	1153	1154	1174	1188	1230

Società	Lampadina	Durata in ore
A	1	684
A	2	697
...	...	...
A	40	1093
B	1	819
B	2	836
...	...	...
B	39	1188
B	40	1230

- Corretta struttura del dataset:

a) Costruire la distribuzione di frequenza per ciascuna società:

Conteggio di Durata in ore	Società	
Durata in ore	A	B
650-749	3	
750-849	5	2
850-949	20	8
950-1049	9	16
1050-1149	3	9
1150-1250		5
Totale complessivo	40	40

Esercizio 2.18

b) Come cambia la distribuzione riducendo l'ampiezza degli intervalli da 100 a 50?

Conteggio di Durata in ore	Società	
Durata in ore	A	B
650-749	3	
750-849	5	2
850-949	20	8
950-1049	9	16
1050-1149	3	9
1150-1250		5
Totale complessivo	40	40

La distribuzione non sembra cambiare in maniera rilevante.

Conteggio di Durata in ore	Società	
Durata in ore	A	B
650-699	2	
700-749	1	
750-799	1	
800-849	4	2
850-899	9	2
900-949	11	6
950-999	5	6
1000-1049	4	10
1050-1099	3	5
1100-1149		4
1150-1199		4
1200-1250		1
Totale complessivo	40	40

Con l'ampiezza = 100 $\Rightarrow$ per A, la classe più frequente è 850-949; per B, la classe più frequente è 950-1049; sia per A che B, le classi a destra e sinistra della classe più frequente degradano gradualmente in maniera simmetrica.

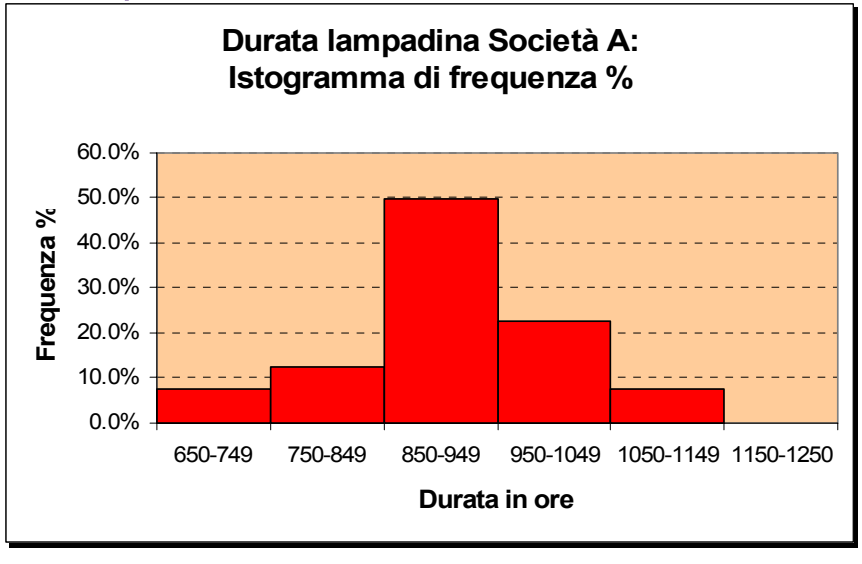
Con l'ampiezza = 50 $\Rightarrow$ per A, la classe più frequente è 900-949; per B, la classe più frequente è 1000-1049. Nei due casi la differenza tra classi + frequenti è = 100.

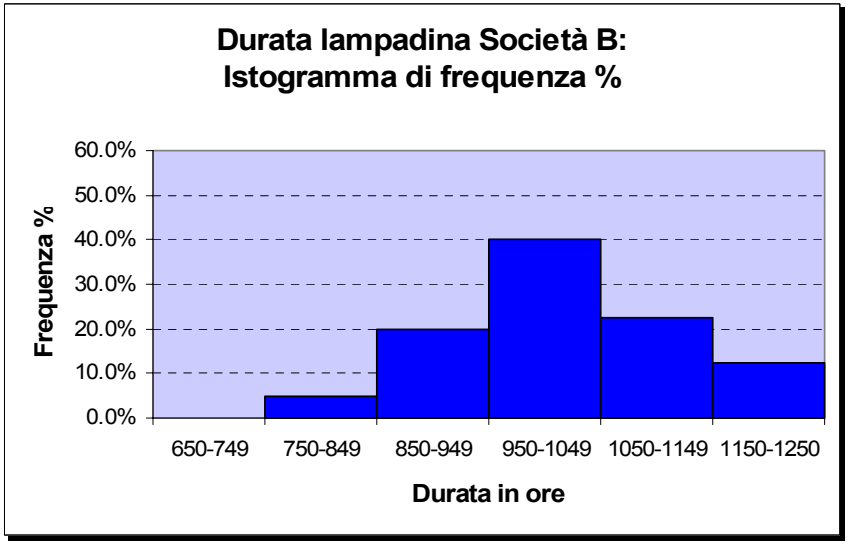
Esercizio 2.18

c) Costruite le due distribuzioni percentuali.

Conteggio di Durata in ore	Società	
Durata in ore	A	B
650-749	7.5%	0.0%
750-849	12.5%	5.0%
850-949	50.0%	20.0%
950-1049	22.5%	40.0%
1050-1149	7.5%	22.5%
1150-1250	0.0%	12.5%
Totale complessivo	100.0%	100.0%

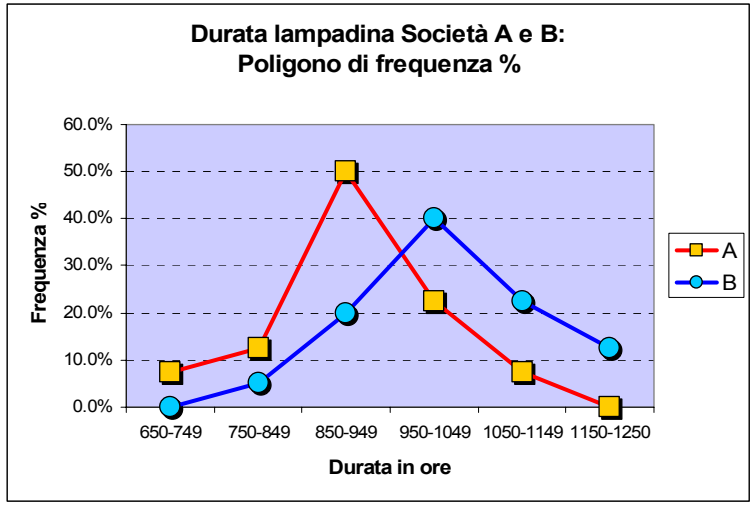
d) Rappresentate i due istogrammi di frequenza.





Esercizio 2.18

e) Disegnate i due poligoni in un unico grafico.

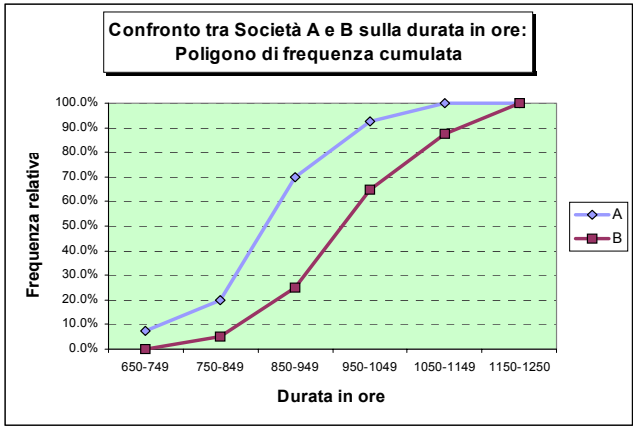


f) Calcolate le frequenze e le percentuali cumulate.

Durata in ore	Formato della Frequenza Cumuta			
	Assoluta		*Relativa*	
	Società		Società	
	A	B	A	B
650-749	3		7.5%	0.0%
750-849	8	2	20.0%	5.0%
850-949	28	10	70.0%	25.0%
950-1049	37	26	92.5%	65.0%
1050-1149	40	35	100.0%	87.5%
1150-1250	40	40	100.0%	100.0%

Esercizio 2.18

g) Rappresentate le due ogive in unico grafico.



h) Quali delle due società costruisce lampadine con una vita più lunga? Commentate i risultati.

In base ai dati relativi al campione di 80 lampadine possiamo affermare che

- la società B costruisce lampadine che si distribuiscono più frequente-mente su valori più elevati di durata in ore;
- esiste una forte indicazione che le lampadine della società B possano avere una durata maggiore rispetto alle lampadine della società A.

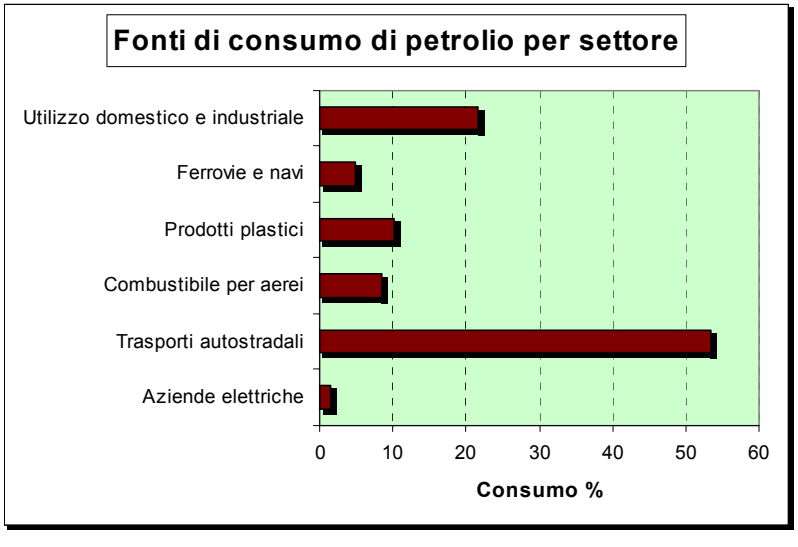
Esercizio 2.28

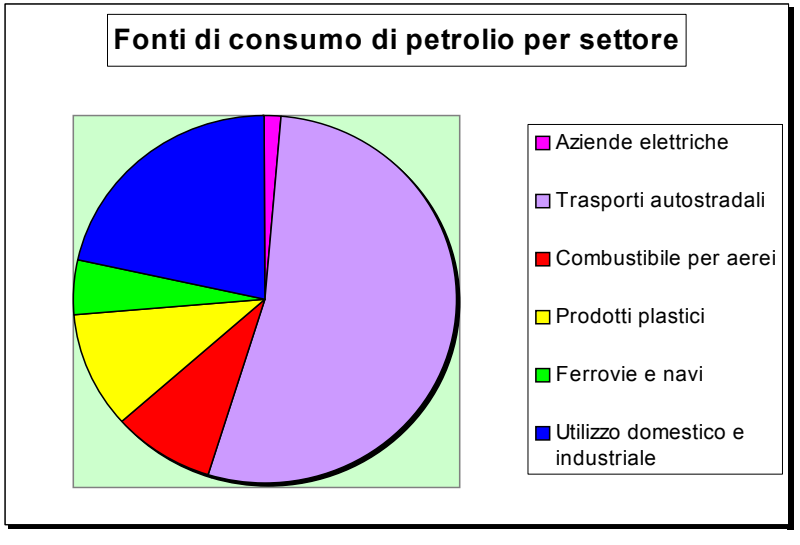
2.28 Nel 1995 in tutti gli Stati Uniti sono stati consumati 17.7 milioni di barili di petrolio al giorno. Nella seguente tabella sono elencate le percentuali di consumo per settore.

a) Costruire il diagramma a barre.

b) Costruire il diagramma a torta.

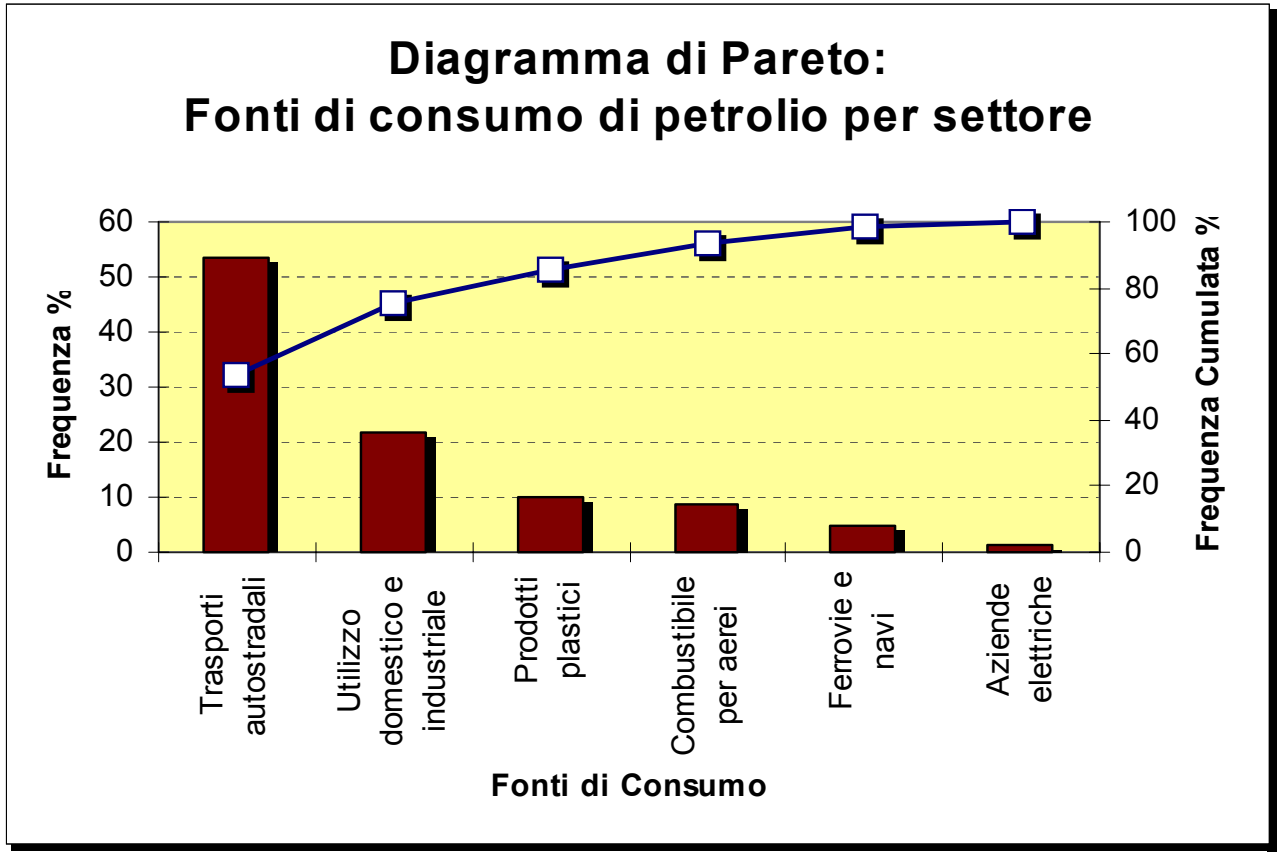
Fonte di consumo	Percentuale
Aziende elettriche	1.4
Trasporti autostradali	53.4
Combustibile per aerei	8.5
Prodotti plastici	10.2
Ferrovie e navi	4.8
Utilizzo domestico e industriale	21.7
Totale	100.0





Esercizio 2.28

c) Costruire il diagramma di Pareto.



Esercizio 2.28

c) Quale di questi grafici secondo voi è più adatto alla rappresentazione dei dati oggetto dell'analisi?

Il diagramma di Pareto è il più dettagliato: fornisce sia la frequenza che la frequenza cumulata.

d) Quali fra i settori elencati sono i maggiori consumatori di petrolio? Commentate i risultati ottenuti.

Trasporti e utilizzo domestico industriale coprono oltre il 70% del consumo, mentre aerei e prodotti plastici consumano globalmente un altro 20% circa. Il resto non conta quasi nulla, dal momento che vale appena poco più del 6% del totale.