Università di Ferrara

Informatica applicata alla EBP e alla ricerca in fisioterapia

Carlo Giannelli
carlo.giannelli@unife.it

http://docente.unife.it/carlo.giannelli

Si ringrazia il dott. Valerio Muzzioli per il materiale didattico

A.A. 2019/20

Informazioni generali

- Sito Web del corso: http://www.unife.it/medicina/fisioterapia/minisiti-fe/evidence-based-practice-e-metodologia-della-ricerca/

- Materiale didattico
  - slide presentate a lezione
  - file Excel di esempio

- Prova d'esame: esercizi di analisi statistica al PC utilizzando Microsoft Excel
  - ad esempio, dato un file di testo contenente dati quantitativi o qualitativi, importarlo in un foglio di lavoro Excel per analizzarne i contenuti: formattazione condizionale e filtraggio dei dati, suddivisione in classi, **frequenza assoluta e relativa**, misure di dispersione come range e varianza, rappresentazioni grafiche dei dati, **organizzazioni stem and leaf**, analisi basata su **tabelle pivot**.

Fogli elettronici

- Consentono di eseguire calcoli di tipo tabellare con visualizzazione immediata dei risultati
- Esempi: consuntivi, preventivi, budget, valutazione di investimento, piani di ammortamento, etc.
- Esigenza di eseguire calcoli ripetuti considerando diverse ipotesi.
- Calcoli lunghi se non automatizzati.
- Alcuni fogli elettronici: Microsoft Office Excel, LibreOffice Calc

Foglio elettronico

- È una matrice composta da un numero elevato di caselle contenuta nella memoria del calcolatore.
- Ogni casella può contenere qualche decina di caratteri
- La matrice eccede, in genere, la dimensione del video
- Il video si comporta come una finestra che mostra una parte della matrice
- La finestra può essere fatta scorrere sulla matrice

Tabelle e celle

È un programma che permette di effettuare calcoli, elaborare dati e tracciare efficaci rappresentazioni grafiche. Il principio su cui si basa il foglio di calcolo è semplice: fornire una **tabella**, detta anche foglio di lavoro, **formata da celle** in cui si possono inserire dati, numeri o formule.

Le celle sono la base fondamentale del foglio di calcolo. Le **celle sono identificabili da una lettera (la colonna) e da un numero (la riga)**. Es: la prima cella in alto a sinistra sarà A1, quella accanto a destra B1 e così via. Quelle invece sotto alla cella A1 saranno A2, A3, A4 e così via.

È possibile effettuare operazioni o applicare funzioni ai contenuti delle celle, specificandone l'indirizzo tramite la lettera identificativa della casella ed il numero identificativo della riga.

Pianificazione crediti accademici

Laurea in lettere ad indirizzo storico-musicale

RIEPILOGO SEMESTRI

	CREDITI	CORSI
Semestre 1	46	12
Semestre 2	20	8
Semestre 3	9	4
Semestre 4	4	2
Semestre 5	2	1

Per aggiornare il grafico pivot precedente, fare clic con il pulsante destro del mouse in un'area vuota del grafico e scegliere **Aggiorna dati**.

CREDITI RICHIESTI	TOTALE	GUADAGNATI	NECESSARI	
Corso accademico principale	54	22	32	32
Corso accademico secondario	NA	0		
Corso facoltativo	4	4	0	✓ 0
Studio generale	66	42	24	24
TOTALI	**124**	**68**	**56**	

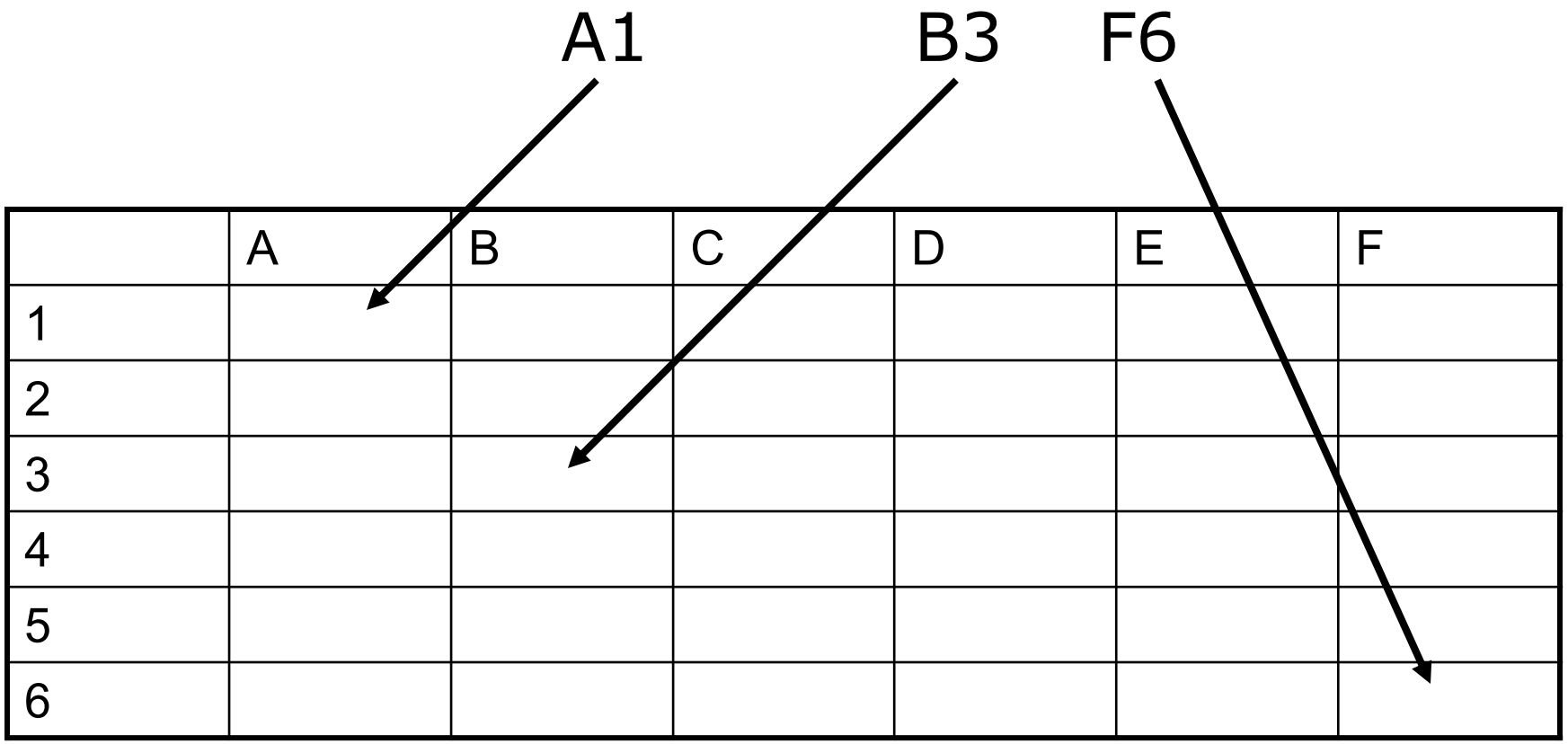
STATO DI AVANZAMENTO GENERALE: 55% COMPLETATO.

Hai raggiunto più della metà del tuo obiettivo.

Corsi accademici

TITOLO CORSO	N. CORSO	TITOLO ACCADEMICO RICHIESTO	CREDITI	COMPLETATO?	SEMESTRE
Antropologia	GEN 108	Studio generale	20	Sì	Semestre 1
Musica applicata	MUS 215	Corso accademico principale	3		Semestre 3
Storia dell'arte	ART 101	Studio generale	2	Sì	Semestre 1
Storia dell'arte	ART 201	Studio generale	2	Sì	Semestre 2
Ear training I	MUS 113	Corso accademico principale	2	Sì	Semestre 1
Ear training I II	MUS 213	Corso accademico principale	2	Sì	Semestre 2
Ear training I III	MUS 313	Corso accademico principale	2		Semestre 3
Ear training I IV	MUS 413	Corso accademico principale	2		Semestre 4

Indirizzamento



Riferimenti a celle

È possibile **copiare il contenuto di celle semplicemente trascinando nella direzione desiderata la selezione**.

Dato che tale procedura è valida sia per dati che per operazioni o formule su tali dati, è possibile specificare due tipi di indirizzi di cella:

- Riferimenti relativi, ad esempio A2
- Riferimenti assoluti, ad esempio $A$2
- Riferimenti misti, ad esempio $A2 e A$2

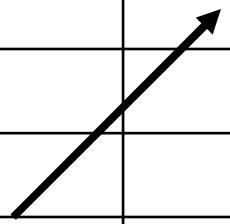
Nei **riferimenti relativi**, durante la copia per trascinamento di una cella, vengono memorizzate e copiate le distanze tra coordinate relative tra le celle.

Nei **riferimenti assoluti** il simbolo $ specifica che tale colonna o riga non deve essere modificata durante la copia.

Riferimenti misti: sono sintassi accettate anche $A1 (nel quale la copia modifica il numero di riga, ma non la colonna) e A$1 (nel quale la copia modifica la lettera di colonna, ma non la riga).

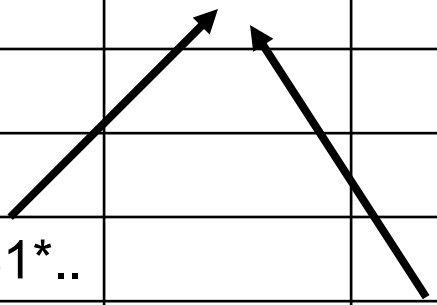
Riferimento assoluto

	A	B	C	D	E	F
1						
2						
3						
4	=$B$1*..					
5						
6						



Riferimento assoluto

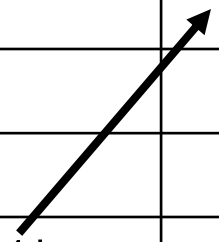
	A	B	C	D	E	F
1						
2						
3						
4	=$B$1*..					
5			=$B$1*..			
6						



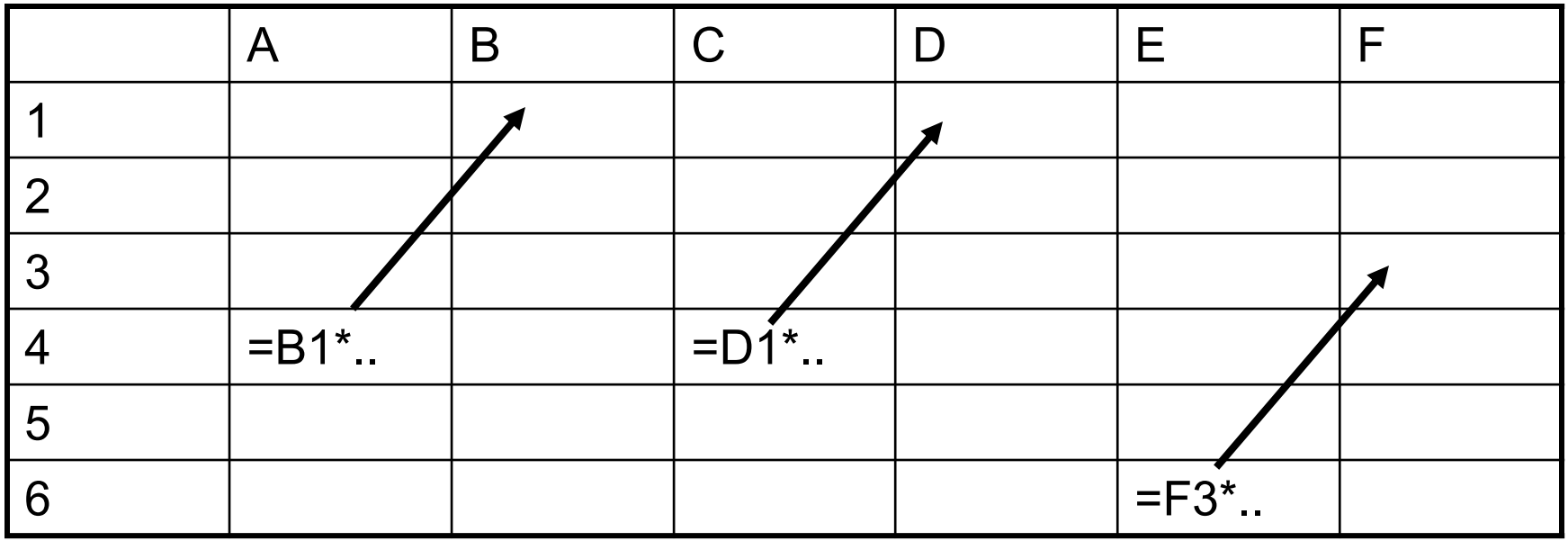
Copia espressione senza modifica del riferimento alla cella

Riferimento relativo

	A	B	C	D	E	F
1						
2						
3						
4	=B1*..					
5						
6						



Riferimento relativo



Copia espressione con modifica del riferimento alla cella, dipendentemente dalla distanza tra cella sorgente e destinazione

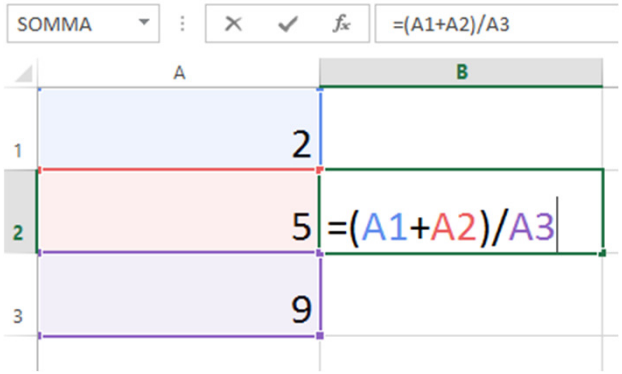
Funzioni e operazioni

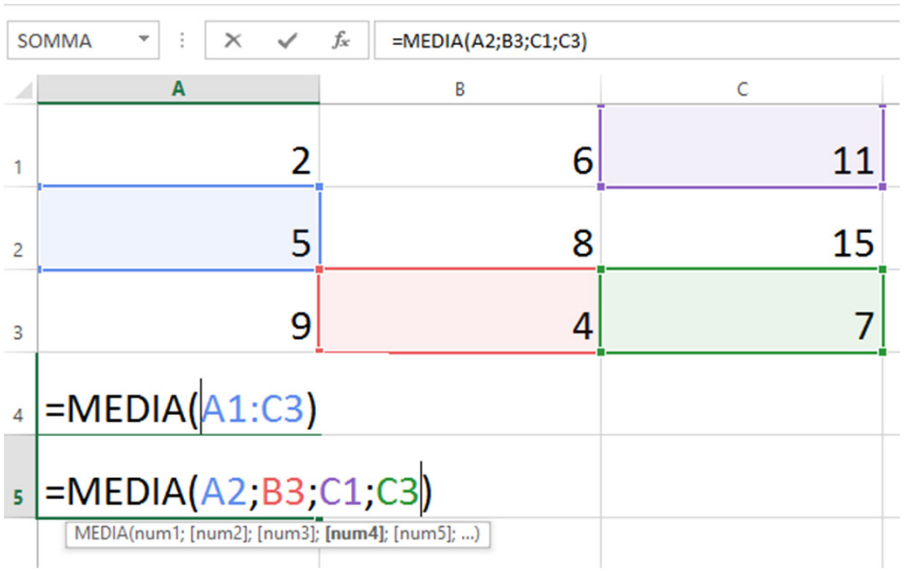
È possibile **effettuare operazioni o applicare funzioni ai contenuti delle celle**, specificandone l'indirizzo tramite la lettera identificativa della casella e il numero identificativo della riga.

Le funzioni seguono la sintassi =NOMEFUNZIONE(arg1, …, argN)

Per identificare un blocco di celle adiacenti, se A1 è la prima cella in alto a sinistra e C3 è l'ultima cella in basso a destra, la sintassi è (A1:C3).

Se al contrario si vuole identificare una sequenza di celle non adiacenti, è necessario separarle con il simbolo ;





Calcolo interessi

- Esempio: si consideri il caso di un investimento che richiede il versamento annuale di una quota fissa pari a 1000 per 4 anni. Sul capitale versato ogni anno viene riconosciuto un tasso di interesse (supponiamo 5%)

Tasso di interesse		0.05
Quota		1000
Capitale alla fine del		
anno	1	=$C$2*(1+$C$1)
anno	2	=(C4+$C$2)*(1+$C$1)
anno	3	=(C5+$C$2)*(1+$C$1)
anno	4	=(C6+$C$2)*(1+$C$1)

Calcolo interessi

- Nel caso di tasso di interesse del 5%, al termine dei quattro anni avremmo € 4.525.631

Tasso di interesse		5%
Quota		EUR 1,000.00
Capitale alla fine del		
anno	1	EUR 1,050.00
anno	2	EUR 2,152.50
anno	3	EUR 3,310.13
anno	4	EUR 4,525.63

Esercizio

- Per un'impresa ceramica si prevede un **incremento costante mensile dello 0.8% dei costi**, supponete di considerare tre diverse voci di costo (materie prime, personale, macchinari)

- Due obiettivi:
  - per ciascuna voce, calcolare il costo per ciascun mese dell'anno e il totale annuale, conoscendo i valori dei costi nel mese di Gennaio
  - per ciascun mese, si calcoli il totale delle spese

Suggerimenti

- Inserire nelle righe i nome dei mesi e nelle colonne le varie voci di costo
- Occorre, dopo aver introdotto i dati per il mese di Gennaio, calcolare quelli dei mesi successivi utilizzando una formula del tipo (CostoMesePrecedente *1.008)

Soluzione: formule

	materie prime	personale	macchinari	Totali mensili
Gennaio	400000	600000	350000	=SUM(B2:D2)
Febbraio	=B2*1.008	=C2*1.008	=D2*1.008	=SUM(B3:D3)
Marzo	=B3*1.008	=C3*1.008	=D3*1.008	=SUM(B4:D4)
Aprile	=B4*1.008	=C4*1.008	=D4*1.008	=SUM(B5:D5)
Maggio	=B5*1.008	=C5*1.008	=D5*1.008	=SUM(B6:D6)
Giugno	=B6*1.008	=C6*1.008	=D6*1.008	=SUM(B7:D7)
Luglio	=B7*1.008	=C7*1.008	=D7*1.008	=SUM(B8:D8)
Agosto	=B8*1.008	=C8*1.008	=D8*1.008	=SUM(B9:D9)
Settembre	=B9*1.008	=C9*1.008	=D9*1.008	=SUM(B10:D10)
Ottobre	=B10*1.008	=C10*1.008	=D10*1.008	=SUM(B11:D11)
Novembre	=B11*1.008	=C11*1.008	=D11*1.008	=SUM(B12:D12)
Dicembre	=B12*1.008	=C12*1.008	=D12*1.008	=SUM(B13:D13)
Totali annuali	=SUM(B2:B13)	=SUM(C2:C13)	=SUM(D2:D13)	=SUM(E2:E13)

Soluzione: valori ottenuti

	materie prime	personale	macchinari	Totali mensili
Gennaio	EUR 400,000	EUR 600,000	EUR 350,000	EUR 1,350,000
Febbraio	EUR 403,200	EUR 604,800	EUR 352,800	EUR 1,360,800
Marzo	EUR 406,426	EUR 609,638	EUR 355,622	EUR 1,371,686
Aprile	EUR 409,677	EUR 614,516	EUR 358,467	EUR 1,382,660
Maggio	EUR 412,954	EUR 619,432	EUR 361,335	EUR 1,393,721
Giugno	EUR 416,258	EUR 624,387	EUR 364,226	EUR 1,404,871
Luglio	EUR 419,588	EUR 629,382	EUR 367,140	EUR 1,416,110
Agosto	EUR 422,945	EUR 634,417	EUR 370,077	EUR 1,427,439
Settembre	EUR 426,328	EUR 639,493	EUR 373,037	EUR 1,438,858
Ottobre	EUR 429,739	EUR 644,609	EUR 376,022	EUR 1,450,369
Novembre	EUR 433,177	EUR 649,765	EUR 379,030	EUR 1,461,972
Dicembre	EUR 436,642	EUR 654,964	EUR 382,062	EUR 1,473,668
Totali annuali	EUR 5,016,935	EUR 7,525,402	EUR 4,389,818	EUR 16,932,155

Excel: elementi di base

Tabella: insieme di celle disposte secondo righe (identificate da numeri) e colonne (identificate da lettere).
Costituisce un **foglio di lavoro.**
Cartella di lavoro: insieme di fogli di lavoro
Formati dei dati: tipi di formato e gestione del formato

→ riferimenti di cella

- **relativo**: viene aggiornato se la formula è copiata in un'altra cella (es. A1)
- **assoluto**: NON viene aggiornato se la formula viene copiata in un'altra (es. $A$1)
- **misto**: indica un riferimento assoluto solo per la riga o la colonna scelta (es. A$1)

→ inserimento di dati

Inserimento dei dati in una cella
Importazione dei dati da un file di testo (TXT)

→ formattazione dei dati in tabella

- La **formattazione condizionale** dei dati
- Assegnazione di un **Nome delle celle**
- Sintassi per l'accesso ai dati

Excel: elementi di base

→ funzioni di conteggio e statistiche

Funzioni di conteggio:

- **CONTA.NUMERI**
- **CONTA.VALORI**
- **CONTA.VUOTE**

Alcune funzioni statistiche:

MAX (num1, num2,…) valore massimo
MIN (num1, num2,…) valore minimo
MEDIA (num1, num2,…) media
MEDIANA (num1, num2,…) mediana
MODA (num1, num2,…) moda
DEV. ST (num1, num2,…) deviazione standard
VAR (num1, num2,…) varianza
QUARTILE (matrice, quarto) quartile
PERCENTILE (matrice, k) percentile

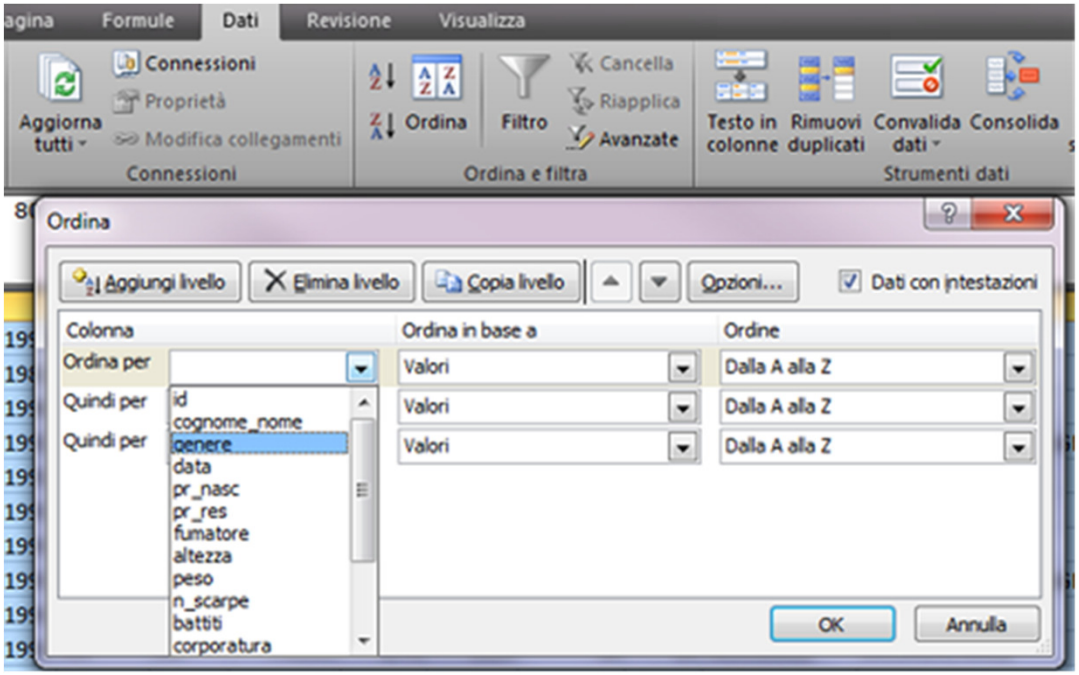
Excel: operazioni avanzate

ORDINAMENTO DATI

Per ordinare i dati, selezionare i dati che si vogliono ordinare e dal menu DATI cliccare ORDINA; **attenzione: CRITERI DI ORDINAMENTO e opzioni**

Quando si ordina un elenco (ovvero una serie di righe contenenti dati correlati), **le righe sono ridisposte** in base al contenuto di una colonna specificata. Distinguiamo due tipi di ordinamento: **crescente e decrescente**.

- I dati devono essere in colonne e righe contigue
- È possibile specificare se i dati hanno o meno una riga di intestazione che in caso sia presente, viene mostrare nel campo ORDINA PER.
- L'operazione di ordinamento, se bene seguita, non devo modificare l'integrità dei dati, ma solo riordinare le righe in base ai valori della colonna (o colonne) scelte come criterio di ordinamento



Excel: operazioni avanzate

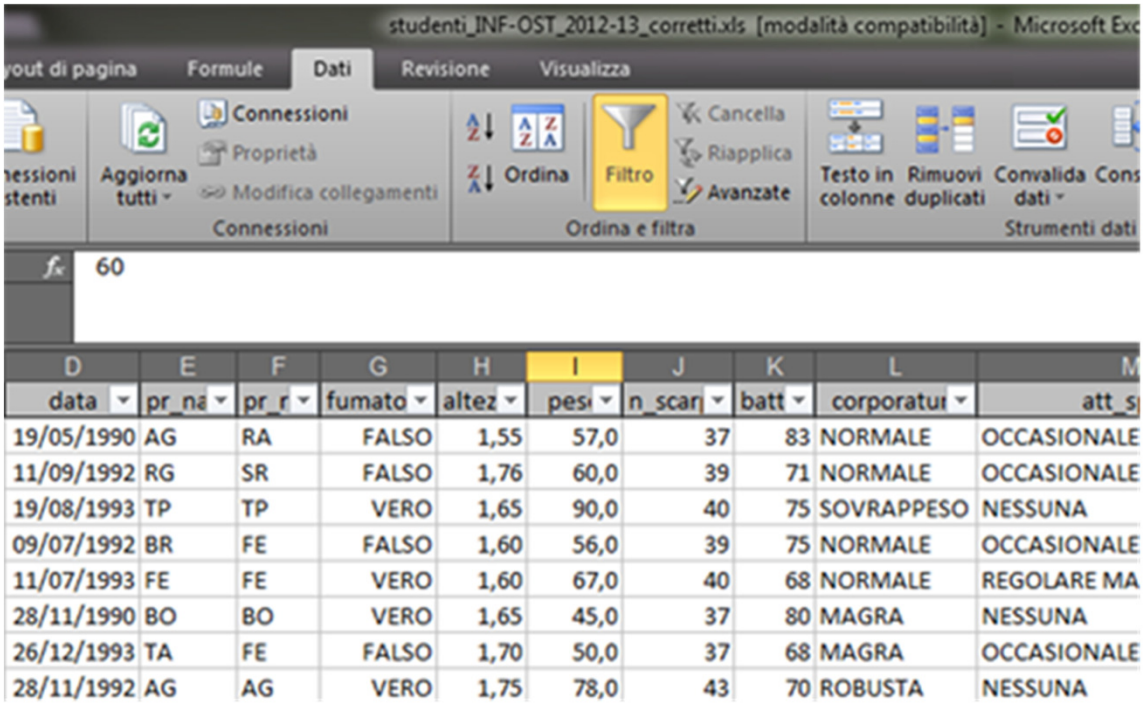
FILTRARE I DATI

Un archivio tabellare è una tabella in cui **la prima riga ospita le intestazioni dei campi** in cui si articolano i record dell'archivio stesso, mentre **le righe sottostanti contengono i suddetti record.**

In un prospetto così impostato si possono eseguire velocemente ricerche anche complesse, utilizzando il **Filtro automatico**, che si attiva selezionando l'intera tabella, aprendo il menu Dati, scegliendo la voce Filtro, e optando per Filtro automatico nel corrispondente sottomenu**.**

Così facendo, nella parte destra delle celle della prima riga dell'archivio compare un pulsante associato ad un menu a tendina che visualizza tutti i valori che il campo assume.

Selezionando uno dei valori Excel mostra solo le righe che hanno il valore selezionato in quel campo: le restanti righe non vengono visualizzate.

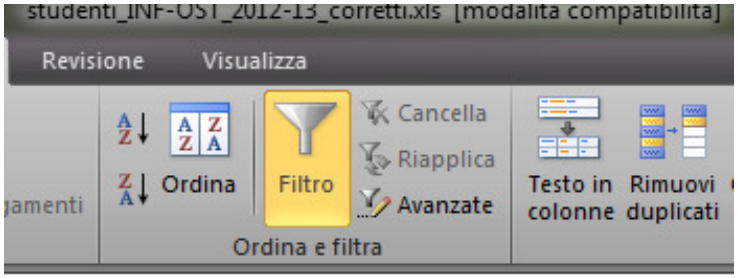


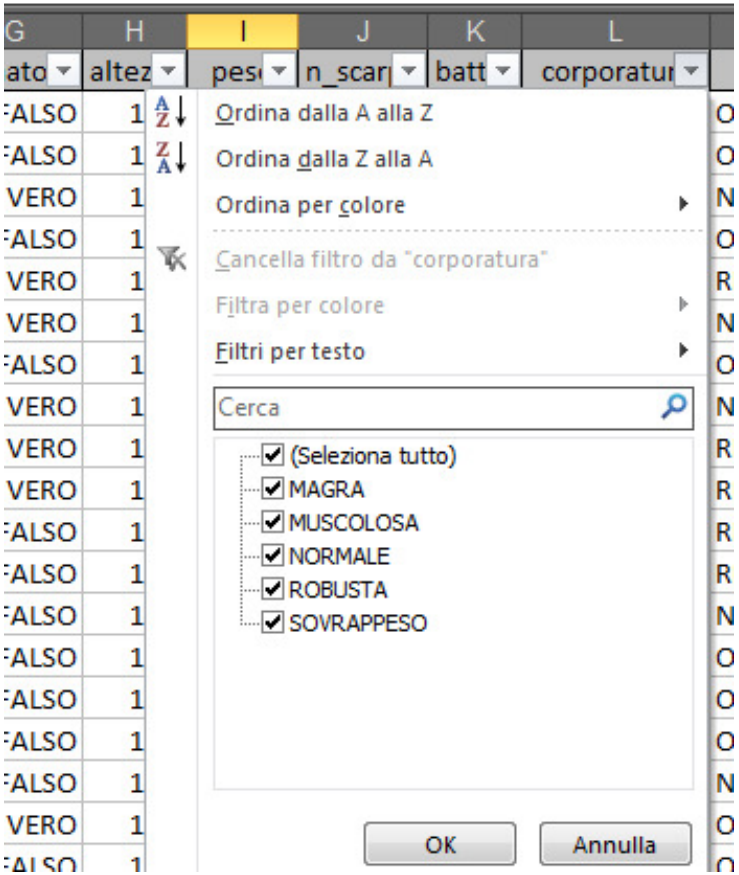
data	pr_na	pr_r	fumato	altez	pes	n_scar	batt	corporatur	att_s
19/05/1990	AG	RA	FALSO	1,55	57,0	37	83	NORMALE	OCCASIONALE
11/09/1992	RG	SR	FALSO	1,76	60,0	39	71	NORMALE	OCCASIONALE
19/08/1993	TP	TP	VERO	1,65	90,0	40	75	SOVRAPPESO	NESSUNA
09/07/1992	BR	FE	FALSO	1,60	56,0	39	75	NORMALE	OCCASIONALE
11/07/1993	FE	FE	VERO	1,60	67,0	40	68	NORMALE	REGOLARE MA
28/11/1990	BO	BO	VERO	1,65	45,0	37	80	MAGRA	NESSUNA
26/12/1993	TA	FE	FALSO	1,70	50,0	37	68	MAGRA	OCCASIONALE
28/11/1992	AG	AG	VERO	1,75	78,0	43	70	ROBUSTA	NESSUNA

Excel: operazioni avanzate

FILTRARE I DATI

Più filtri possono essere applicati in sequenza senza disattivare i precedenti: in tal modo Excel mostrerà le sole righe che soddisfano i criteri di selezione impostati nei vari filtri.





Excel: operazioni avanzate

TABELLE PIVOT

La tabella pivot deve il suo nome al fatto che le intestazioni di riga e colonna possono essere ruotate intorno all'area dati principale per offrire diverse visualizzazioni dei dati di origine.

Una tabella pivot può essere costruita partendo da un elenco di dati di Excel, da una tabella pivot già esistente nella cartella di lavoro o da un'origine esterna, come una tabella di database.

Attenzione: l'elenco non deve contenere righe o colonne vuote e deve essere fornito di intestazioni di colonna.

Una tabella pivot serve per riepilogare i dati provenienti da elenchi o database esistenti utilizzando i metodi di calcolo come somma e media.

Nella tabella pivot vengono eseguiti rapporti che ti permettono di migliorare l'analisi dei dati: non vengono inseriti dati ma vengono solamente presentati in un modo diverso quelli già esistenti.

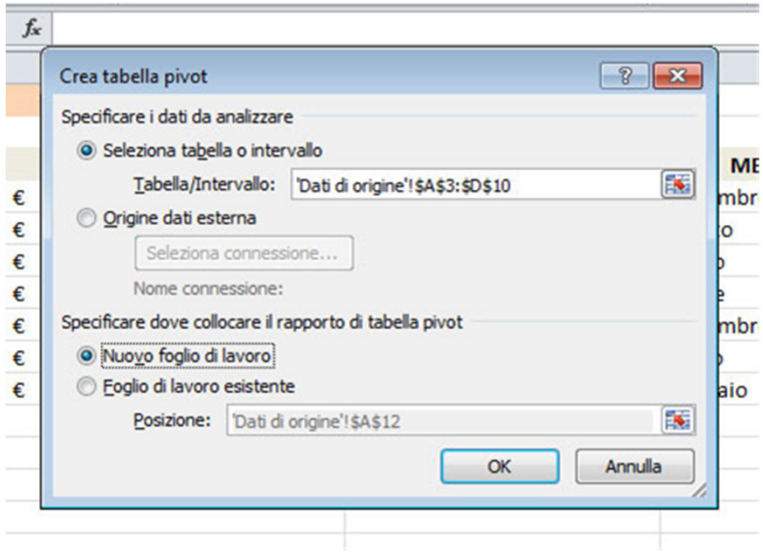
STRUTTURA DI UNA TABELLA PIVOT
Una tabella pivot è composta da:
Campi Riga
Campi Colonna
Campi Filtro
Dati

Trascinando i nomi degli elementi dal "ELENCO CAMPI" è possibile costruire la tabella sulla base delle analisi dati da effettuare.

Excel: operazioni avanzate

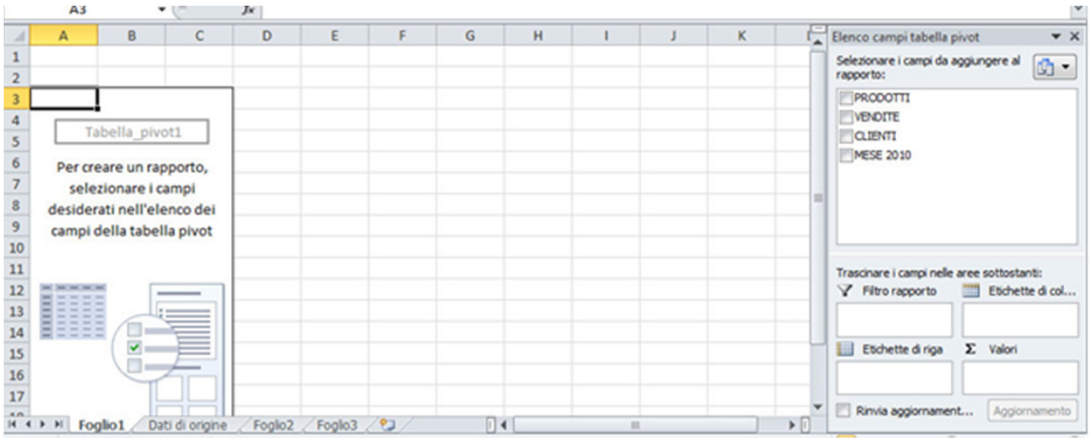
CREARE UNA TABELLA PIVOT

1 Aprire il file XLS e posizionare la cella attiva sui dati. Controllare che i dati siano in righe e colonne contigue, ovvero NON ci siano intere righe e/o colonne interamente vuote tra i dati.



2 Al menu Inserisci clicca su "tabella Pivot".
Si apre la finestra per la creazione guidata della tabella Pivot.
Come prima cosa ti viene chiesto dove si trovano i dati da analizzare e quale tipo di rapporto vuoi creare.

3 Seleziona (se necessario) le caselle di opzione "Elenco o database Excel" e "Tabella pivot" e indicare dove si trovano i dati da utilizzare. Solitamente viene proposto l'intera area dei dati già inseriti.



Excel: operazioni avanzate

CREARE UNA TABELLA PIVOT

La struttura della tabella pivot è composta da quattro aree in cui potete inserire i dati:

- **Filtro rapporto**: contiene l'elemento che deve essere posto sulla terza dimensione. Si tratta, cioè, dei dati che verranno visualizzati uno alla volta (di solito si inserisce in questo campo quella variabile che ha un elevato numero di valori osservati).

- **Campi/Etichette di riga**: contiene i dati che saranno utilizzati come etichette per le righe della tabella pivot. Excel collocherà in OGNI RIGA un valore assunto dalla variabile.

- **Campi/Etichette di colonna**: contiene i dati che saranno utilizzati come etichette per le colonne della tabella pivot. Excel collocherà in OGNI COLONNA un valore assunto dalla variabile.

- **Valori**: contiene i dati da riepilogare nella tabella pivot sulla base di una funzione matematica**.**

STATISTICA

DESCRITTIVA: consiste di metodi per l'organizzazione, la visualizzazione e la descrizione di dati tramite tabelle, grafici (metodi di esposizione, sintesi e rappresentazione dei dati) e misurazioni.

INFERENZIALE: consiste di metodi che utilizzano risultati di "campioni" per aiutare nel prendere decisioni relative ad una "popolazione"

Popolazione: consiste di elementi le cui caratteristiche sono di particolare interesse per lo studio statistico che si va a compiere.
Esempio:
- La % di tutti i votanti che voterà per un particolare candidato
- Il prezzo di tutte le case a Roma
- Il fatturato del 2017 di tutte le aziende di Ferrara

Si è interessati a "TUTTI"

Decisione basata su una porzione della popolazione → Campione
Campione: è una porzione della popolazione selezionata per compiere uno studio statistico.

Esempio: exit poll elettorale

TERMINI SPECIFICI

Raw Data: sono i dati registrati nella sequenza in cui sono stati raccolti e devono ancora essere processati.

Età di 50 studenti universitari

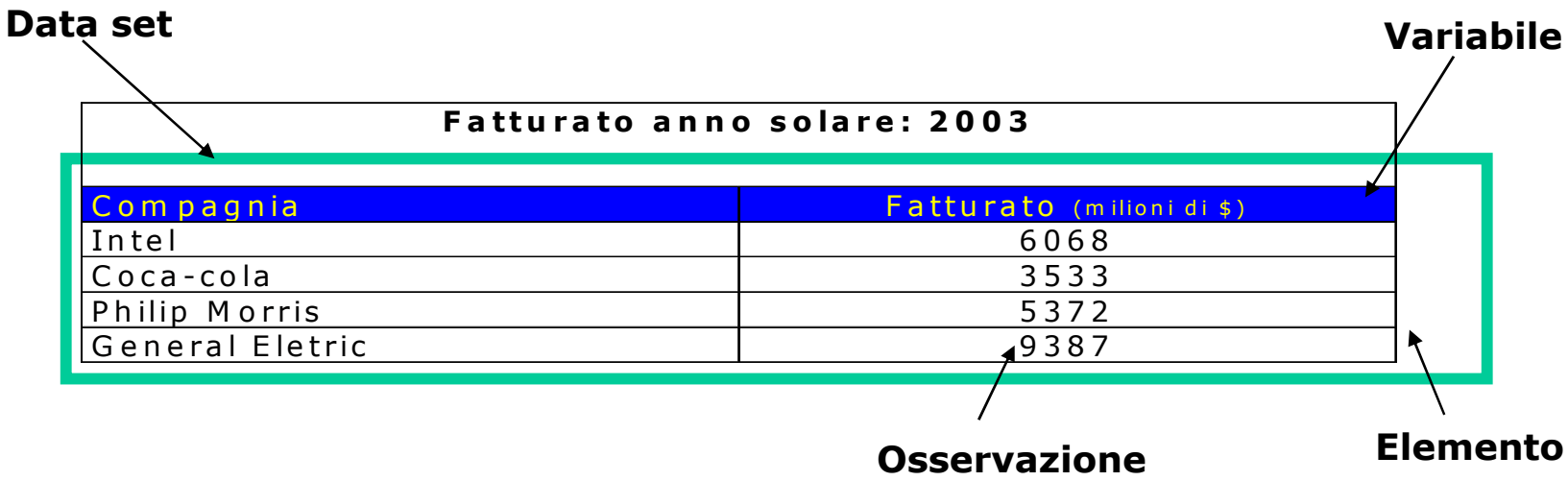
24	27	20	22	20	19	27	23	22	27
26	25	26	18	28	19	24	27	19	18
22	28	26	28	28	28	22	18	26	27
27	25	22	18	25	26	19	20	20	21
19	19	25	22	18	26	20	23	18	25

Data Set: è una collezione di osservazioni di una o più variabili.

Elemento: è uno specifico oggetto o soggetto appartenente al campione o alla popolazione.

Variabile: caratteristica del dato che assume diversi valori per diversi elementi. (tipi di variabili)

Osservazione: il valore che una variabile assume per un elemento,



TIPI DI VARIABILI

VARIABILI QUALITATIVE (CATEGORICHE)
VARIABILI QUANTITATIVE (NUMERICHE)

VARIABILI QUALITATIVE: descrivono caratteristiche che non possono essere misurate con un numero, ma consentono di inserire gli elementi di un campione in una categoria o un gruppo.
Si dividono in:

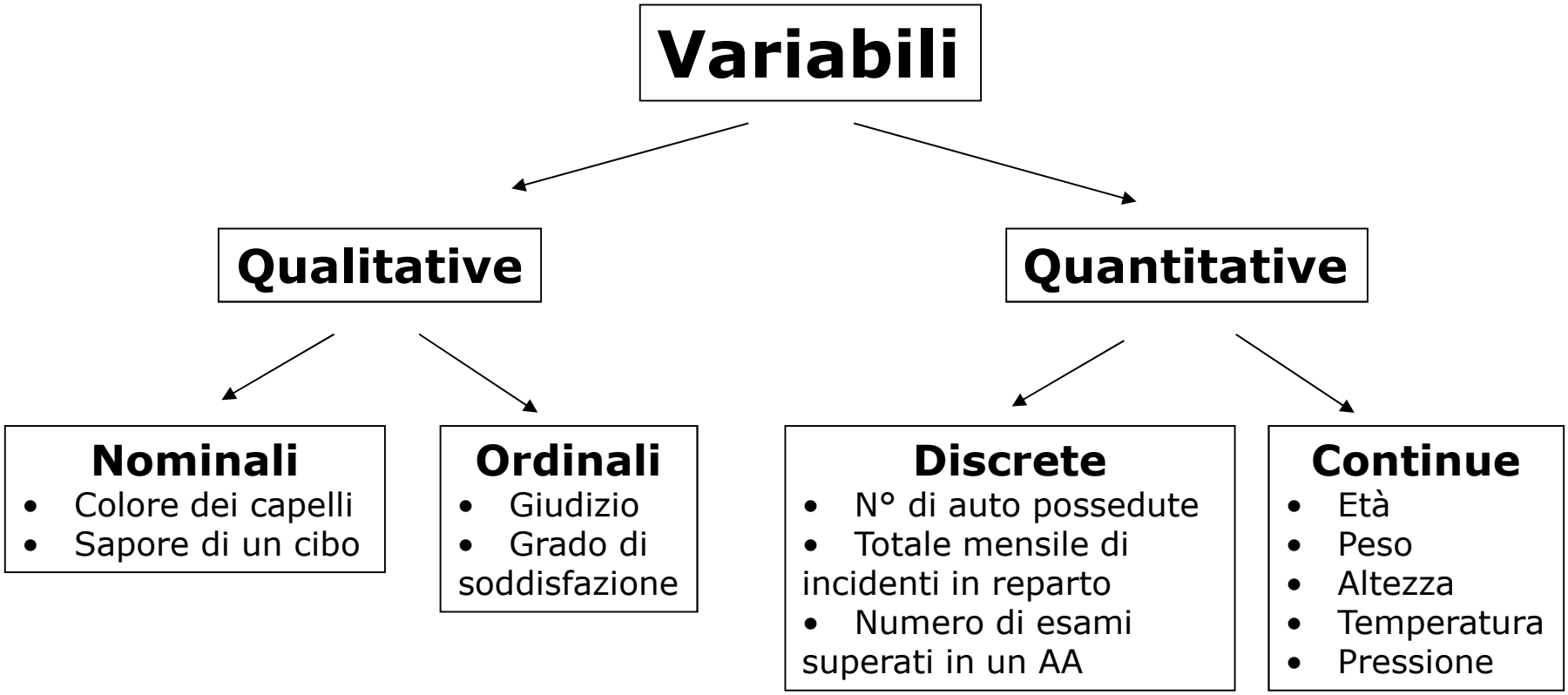
- **NOMINALI**, se la differenza di categoria non ha un ordine intrinseco, ma solo un nome (es. colore dei capelli)
- **ORDINALI**, se i valori possono essere ordinati nonostante non sia rappresentabili su una scala numerica (es. giudizio da insufficiente a ottimo)

VARIABILI QUANTITATIVE: esprimono caratteristiche attraverso un valore su una scala numerica. Sono variabili numeriche quelle che rappresentano conteggi, percentuali, tassi, dimensioni, ecc.
Si dividono in:

- **CONTINUE**, se possono assumere qualsiasi valore numero reale in un certo intervallo: pertanto tra 2 valori qualsiasi ne possono esistere infiniti altri. I valori vengono arrotondati (valori di temperatura, altezza, ecc.).
- **DISCRETE,** se i valori sono unità indivisibili (numero di figli, sigarette fumate in un giorno, ecc).

Attenzione: si possono usare i numeri anche per denominare categorie,
così come dati numerici possono essere ridotti a categorici.

TIPI DI VARIABILI



STRUMENTI per

- ORGANIZZARE

- Stem and leaf (per variabili quantitative)

- ANALIZZARE

- Frequenza (assoluta, relativa, percentuale, cumulativa)
- Misure di tendenza centrale (media, mediana, moda)
- Misure di dispersione (range, varianza e deviazione standard)
- Misure di posizione (quartile, percentile)

- RAPPRESENTARE

- Grafici a barre, a torta, poligoni
- Box and Whisker

Attenzione!

I metodi di organizzazione, analisi e rappresentazione delle variabili sono strettamente legati al TIPO DI VARIABILE

DISTRIBUZIONI DI FREQUENZA di DATI QUALITATIVI

Nella **TABELLA DI DISTRIBUZIONE DELLE FREQUENZE**, vengono indicate le **CATEGORIE** e le **FREQUENZE (assolute)** di ogni categoria.

variabile → | Tipo di impiego | n° di studenti |
categoria → | industria | 44 |
| pubblica amm. | 16 | ← **frequenza**
forze armate	23
artigiani	17
totale	**100**

Una distribuzione di frequenza per dati qualitativi elenca tutte le categorie e il numero di elementi appartenenti ad ogni categoria.

FREQUENZA

ASSOLUTA = n° di elementi per ogni categoria
RELATIVA = n° elementi per categoria rispetto al totale
PERCENTUALE = frequenza relativa * 100

Tipo di diploma di maturità di 110 studenti

Categoria	Freq. Ass.	Freq. Rel.	Freq. %
Classico	22	0,20	20,00
Commericiale	17	0,15	15,45
Scientifico	33	0,30	30,00
Tecnico	17	0,15	15,45
Altri	21	0,19	19,09
tot.	**110**	**1**	**100**

RAPPRENSENTAZIONE GRAFICA di DATI

5 aspetti fondamentali di un grafico:

- **Accuratezza** precisione dei dettagli
- **Semplicità** uso essenziale degli elementi grafici
- **Chiarezza** comunicazione non ambigua del significato
- **Aspetto** estetico (dimensioni e tratti proporzionati)
- **Struttura** deve essere definita (elementi grafici posti in maniera gerarchica e distinti)

le componenti di un grafico Excel:

- Tipo di grafico
- **Serie di dati**
- Area del grafico
- Area del tracciato
- Opzioni del grafico

GRAFICO A BARRE

Asse delle X: categorie (in Excel → asse delle CATEGORIE)
Asse delle Y: frequenze (in Excel → asse dei VALORI)

GRAFICO A TORTA

Usato per visualizzare le frequenze relative o percentuali.

- Il cerchio rappresenta l'intero campione.
- Le fette rappresentano le frequenze relative o percentuali.

Come si crea?
ANGOLO = 360° * frequenza relativa di ogni categoria

Livello di stress	Freq. Relativa	Angolo
Elevato	0,333	119,88
Medio	0,467	168,12
Basso	0,2	72

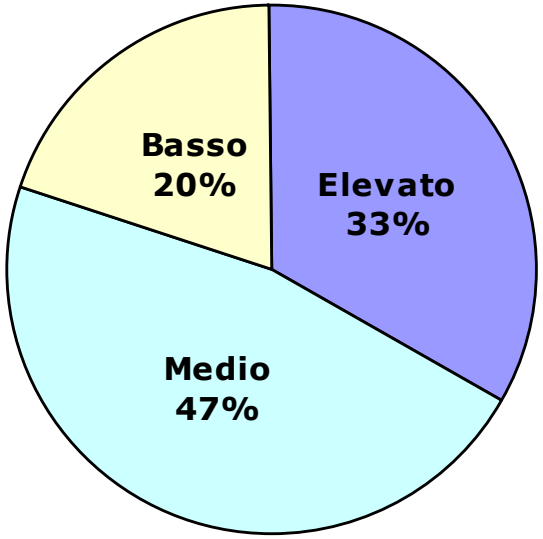


Grafico Stem-leaf

Rappresentazione condensata di dati quantitativi, in grado di conservare le informazioni sulle osservazioni individuali

75	52	80	96	65	79	71	87	93	95
69	72	81	61	76	86	79	68	50	92
83	84	77	64	71	87	72	92	57	98

Punteggio conseguito al test di ammissione

Separazione di ciascun valore:
la prima parte contiene la prima cifra; la seconda parte contiene la seconda cifra.
75: 7 → stem; 5 → leaf
123: 1→ stem; 23 → leaf
1250: 125 → stem; 0 →leaf

5	2	0	7						
6	5	9	1	8	4				
7	5	9	1	2	6	9	7	1	2
8	0	7	1	6	3	4	7		
9	6	3	5	2	2	8			

Si deduce che:
- il "tronco" 5 ha frequenza meno elevata
- il "tronco" 7 ha frequenza più elevata

5	0	2	7						
6	1	4	5	8	9				
7	1	1	2	2	4	6	7	9	9
8	0	1	3	4	6	7	7		
9	2	2	3	5	6	8			

Stem-leaf ordinato
Conservazione delle informazioni iniziali
Esempio: un solo studente ha ottenuto 98 come punteggio

Data Set A		**Data Set B**
Leaf	**Stem**	**Leaf**
3 2 0	4	1 5 6 7

I numeri 40, 42, 43 appartengono al Data Set A.
I numeri 41, 45, 46, 47 appartengono al Data Set B.

DISTRIBUZIONE DI FREQUENZA di DATI QUANTITATIVI

L'elaborazione di dati QUANTITATIVI avviene dopo averli suddivisi in **CLASSI**.

CLASSE: intervallo che include tutti i valori che cadono all'interno di un limite inferiore ed un limite superiore

- Le classi **RAPPRESENTANO SEMPRE LE VARIABILI**
- Le classi **NON SONO MAI SOVRAPPOSTE**: ogni valore appartiene ad 1 ed 1 sola classe

La frequenza di una classe rappresenta il numero di valori del data set che cade in quella classe.

Una distribuzione di frequenza per dati quantitativi è un **elenco di tutte le classi e del numero di valori** che appartiene ad ogni classe.

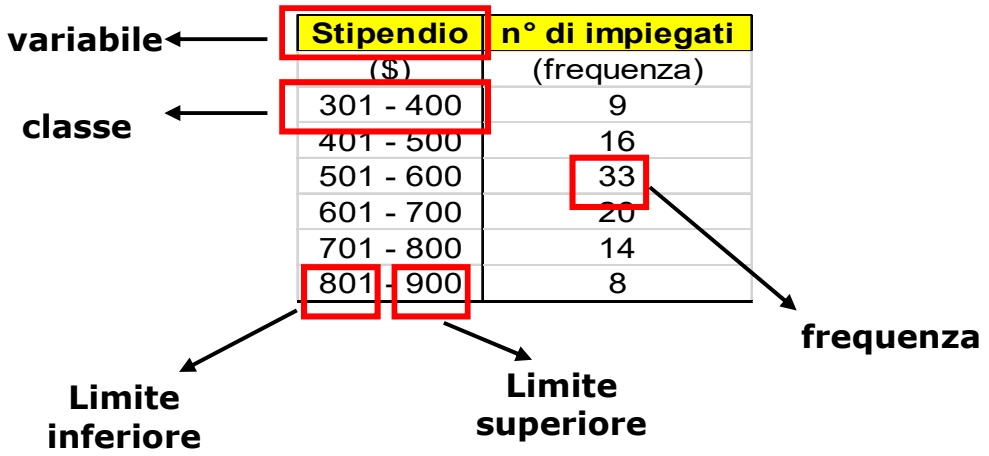


TABELLA DI DISTRIBUZIONE DELLE FREQUENZE

I dati sono rappresentati in modo raggruppato

Stipendio ($)	n° di impiegati (frequenza)
301 - 400	9
401 - 500	16
501 - 600	33
601 - 700	20
701 - 800	14
801 - 900	8

Limite inferiore: 301, 401, 501, 601, 701, 801
Limite superiore: 400, 500, 600, 700, 800, 900

Confine fra le classi: punto intermedio fra l'estremo superiore di una classe e l'estremo inferiore della classe successiva

- (400+401)/2= 400.5 è il confine superiore della prima classe e il confine inferiore della seconda classe
- (500+501)/2= 500.5 è il confine superiore delle seconda classe e il confine inferiore della terza classe

Larghezza di una classe: differenza tra gli estremi superiori ed inferiori di ciascuna classe

- 500.5-400.5=100

Punto centrale (mark) di classe: somma dei limiti di ogni classe diviso 2

- (301+400)/2=350.5

Classe	Confine di classe	Larghezza di classe	Mark
301 - 400	da 300.5 a < 400,5	100	350,5
401 - 500	da 400.5 a < 500,5	100	450,5
501 - 600	da 500.5 a < 600,5	100	550,5
601 - 700	da 600.5 a < 700,5	100	650,5
701 - 800	da 700.5 a < 800,5	100	750,5
801 - 900	da 800.5 a < 900,5	100	850,5

Costruzione della Tabella di Distribuzione di Frequenza

N° delle classi: da 5 a 20 (radice quadrata del numero della osservazioni)

Larghezza delle classi (arrotondato)=**(Valore Max -Valore Min)/N° classi**

- È preferibile avere classi delle stessa larghezza, anche se può succedere di avere classi di diversa dimensione.
- Arrotondando questo numero, può variare il numero della classi precedentemente stabilito.

La tabella di distribuzione delle frequenze non conserva alcuna informazione sulle osservazioni individuali presente in un campione di dati raggruppati.

$$\textbf{Frequenza relativa} = \frac{\text{Frequenza di ogni classe}}{\text{Somma di tutte le frequenze}}$$

Frequenza percentuale = Frequenza relativa * 100

Excel: FUNZIONE FREQUENZA

Calcola la frequenza di occorrenza dei valori di un intervallo e restituisce una matrice verticale di numeri. Dal momento che FREQUENZA restituisce una matrice, deve essere immessa come formula in forma di matrice.

Sintassi: FREQUENZA(**matrice_dati**;**matrice_classi**)

- **Matrice_dati** è una matrice o un insieme di valori di cui si desidera calcolare la frequenza.
- **Matrice_classi** è una matrice o un riferimento agli intervalli in cui si desidera raggruppare i valori contenuti in matrice_dati.

Il numero di elementi, contenuti nella matrice restituita, è maggiore di una unità rispetto al numero di elementi contenuti in matrice_classi (viene aggiunta una classe denominata ALTRO).

L'elemento in più nella matrice restituirà il conteggio di qualsiasi valore superiore all'intervallo più alto. FREQUENZA ignora le celle vuote e il testo.

Immettere una formula in forma di matrice

Quando si immette una formula in forma di matrice, essa verrà automaticamente racchiusa tra parentesi graffe { }.

FREQUENZA viene immessa come formula matrice dopo aver selezionato un intervallo di celle adiacenti nel quale dovrà apparire il risultato: le celle selezionate sono tante quante sono le classi definite.

Digitare la formula in forma di matrice. **Premere CTRL+SHIFT+INVIO**

Distribuzione della frequenza cumulativa

Dalla tabella di distribuzione della frequenza si ricavano **informazioni puntali, cioè specifiche per ogni classe.**

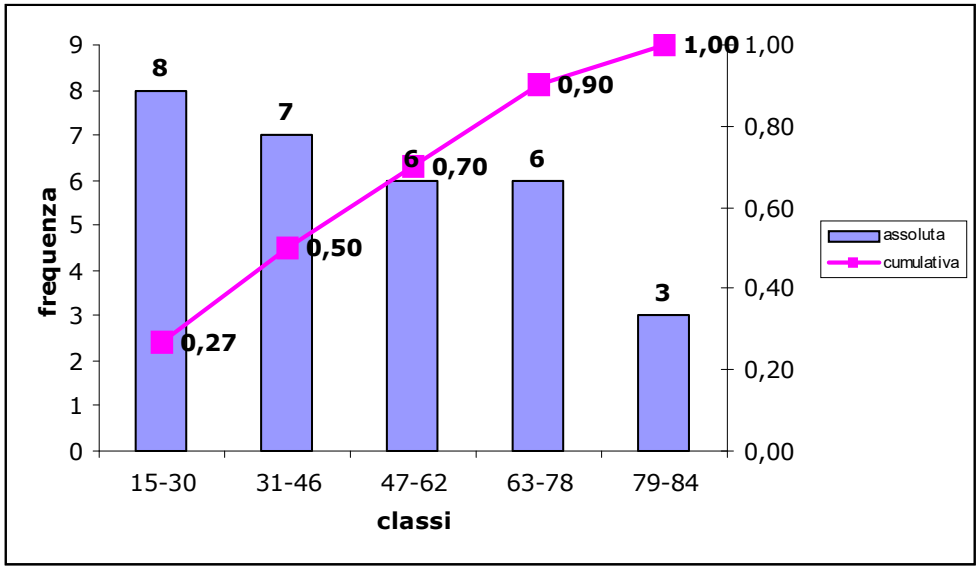
La frequenza cumulativa esprime il totale di valori delle frequenze che stanno al di sotto di un preciso valore.
La frequenza cumulativa è calcolata solo per le variabili quantitative

Frequenza cumulativa

classe	frequenza
15-30	8
31-46	7
47-62	6
63-78	6
79-94	3

Classe	intevallo di classe	Freq.cumulativa
15-30	da 14,5 a meno di 30,5	8
31-46	da 14,5 a meno di 46,5	8+7=15
47-62	da 14,5 a meno di 62,5	8+7+6=21
63-78	da 14,5 a meno di 78,5	8+7+6+6=27
79-94	da 14,5 a meno di 94,5	8+7+6+6+3=30

classe	freq.cumulat. Rel.	Cumulat. %
15-30	8/30=,267	26,7
31-46	15/30=,500	50,0
47-62	21/30=,700	70,0
63-78	27/30=,900	90,0
79-94	30/30=1	100,0



Ogiva: è la curva che rappresenta la frequenza cumulativa. E' una linea spezzata che unisce gli estremi superiori di ogni classe (asse X) con il relativo valore della frequenza (asse Y)

EXCEL STRUMENTO ANALISI DEI DATI - ISTOGRAMMA

Lo strumento di analisi Istogramma consente di calcolare le frequenze individuali e cumulative per un intervallo di celle e di classi di dati.

- **Intervallo di input:** riferimento di cella per l'intervallo di dati da analizzare.
- **Intervallo della classe:** intervallo di celle contenente un insieme di valori limite che definiscano gli intervalli delle classi. **I valori devono essere disposti in ordine crescente**.

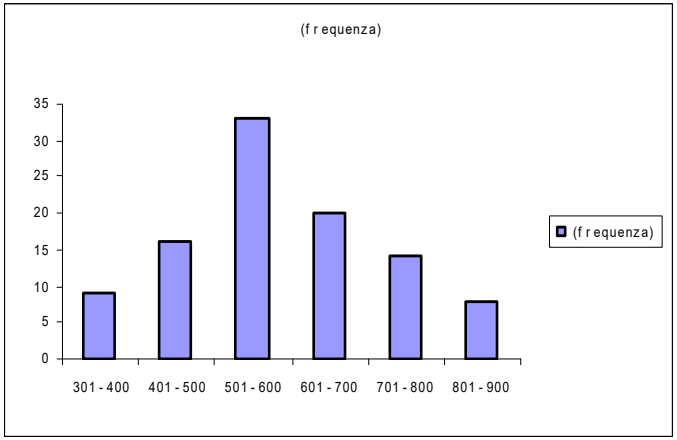
Un numero viene conteggiato in una particolare classe se è uguale o minore al numero di classe. Vengono conteggiati insieme tutti i valori inferiori al primo valore della classe e tutti i valori superiori all'ultimo valore della classe.
Se non si specifica l'intervallo di classe, verrà automaticamente creato un insieme di classi distribuite uniformemente tra il valore minimo e il valore massimo dei dati.

- **Etichette:** Selezionare questa casella se la prima riga dell'intervallo di input contiene etichette.
- **Intervallo di output:** riferimento della cella superiore sinistra della tabella di output.
- **Nuovo foglio di lavoro:** consente di inserire un nuovo foglio di lavoro nella cartella di lavoro corrente e incollare i risultati a partire dalla cella A1 del nuovo foglio di lavoro.
- **Nuova cartella di lavoro :** consente di creare una nuova cartella di lavoro e incollare i risultati in un nuovo foglio della nuova cartella di lavoro.
- **Pareto (istogramma ordinato):** rappresenta i dati nella tabella di output in ordine di frequenza decrescente. Se questa casella è deselezionata, i dati verranno presentati in ordine crescente
- **Percentuale cumulativa:** genera una tabella di output una colonna per le percentuali cumulative e per includere nel grafico di istogramma una riga per la percentuale cumulativa.
- **Grafico in output:** Selezionare questa opzione per generare un istogramma incorporato nella tabella di output.

RAPPRESENTAZIONE GRAFICA DEI DATI RAGGRUPPATI

I dati raggruppati possono essere visualizzati tramite un ISTOGRAMMA o un POLIGONO

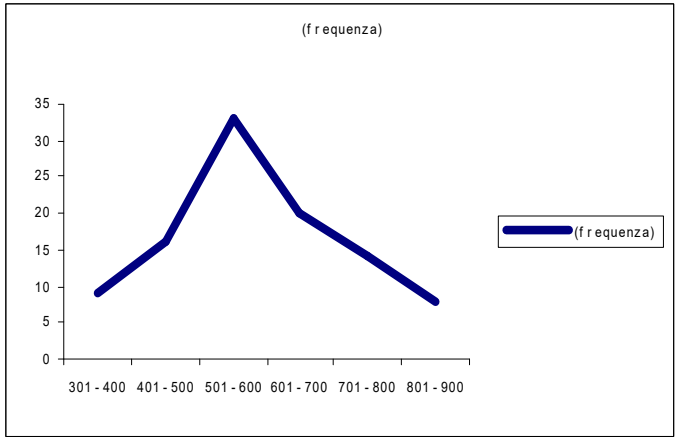
ISTOGRAMMA Asse X → CLASSI



POLIGONO

Si ottiene costruendo una spezzata passante per il punto di mezzo (mark) di ciascuna delle barre che indicano la frequenza di una classe.

Se il dataset è di grandi dimensioni e il dati sono suddivisi in tante classi di ampiezza ridotta, allora la spezzata si trasforma in una curva ed il grafico viene chiamato CURVA DI DISTRIBUZIONE.



PRINCIPALI INDICI STATISTICI

INDICI

- **DI POSIZIONE (TENDENZA CENTRALE)**
  - MEDIA
  - MEDIANA
  - MODA
- **DI DISPERSIONE**
  - RANGE
  - VARIANZA
  - DEVIAZIONE STANDARD
- **DI FORMA**
  - ASIMMETRIA (SKEWNESS)
  - CURTOSI (KURTOSIS)

MISURE DI TENDENZA CENTRALE per dati non raggruppati

Definiscono un "centro di insieme dei dati" ovvero un valore attorno al quale si forma la rosa dei dati.

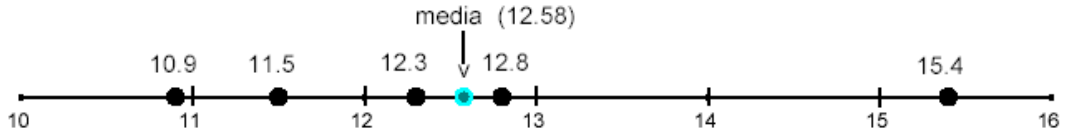
Non esiste un modo univoco di intendere questa dicitura, ma si utilizzano 3 parametri specifici, ognuno dei quali fornisce una particolare caratteristica che dai dati in esame:

- **MEDIA**
- **MEDIANA**
- **MODA**

MEDIA per dati non raggruppati

La **MEDIA (valore medio)** è il parametro più utilizzato nel calcolo di misura di tendenza centrale. Per dati non raggruppati, è calcolata dividendo la somma di tutti i valori per il numero di valori del data set.

$$\text{MEDIA} = \frac{\text{Somma di tutti i valori}}{\text{N° valori}}$$



MEDIA per dati raggruppati

Non si hanno informazioni sui valori individuali e quindi non è possibile calcolare la somma dei valori.

Calcolo della MEDIA:

- Calcolare il punto medio di ogni classe
- Moltiplicare il punto medio di ogni classe per la relativa frequenza
- Sommare i prodotti ottenuti
- Dividere la somma per il n° totale di osservazioni

N° ordini ricevuti giornalmente nei passati 50 giorni

numero ordini	frequenze (f)	punto medio (m)	f*m (mf)
10-12	4	11	44
13-15	12	14	168
16-18	20	17	340
19-21	14	20	280
	N=50		**S=832**

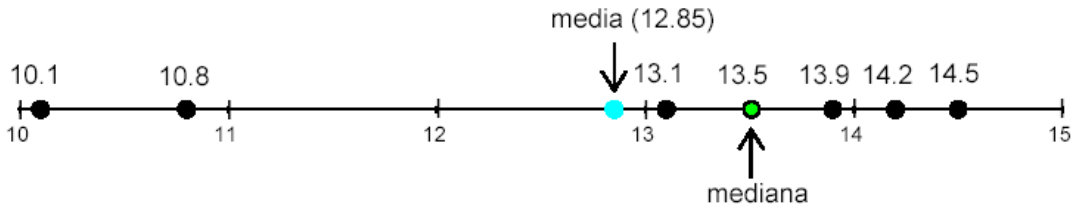
Media= 832/50= 16.64 ordini

La **MEDIANA** è **il valore di mezzo** di un dataset di N dati **ordinato in maniera crescente: la mediana divide il data set in 2 parti uguali.**

- Se N è dispari → Mediana = (N+1)/2
- Se N è pari → Mediana = **media** tra N/2 e (N/2)+1

Dataset:
10 5 19 8 3 → Dataset ordinato: **DISPARI**
3 5 8 10 19 → Calcolo della mediana
(5+1)/2=3 → 8 è la mediana

Dataset:
10 5 8 3 → Dataset ordinato: **PARI**
3 5 8 10 → Calcolo della mediana
(5+8)/2=6,5 → 6,5 è la mediana

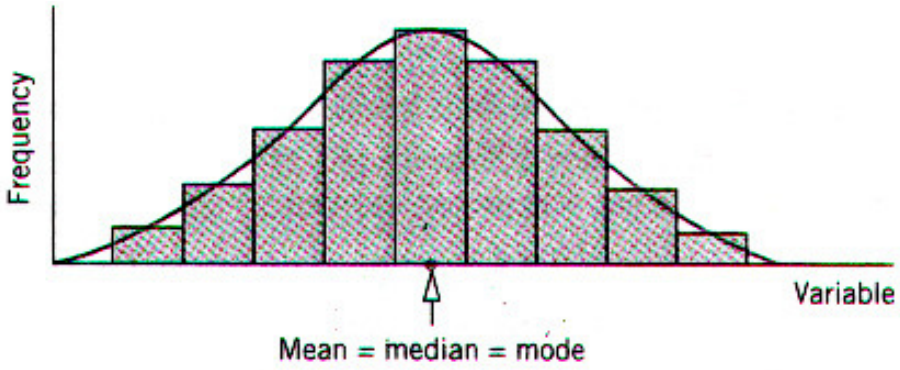


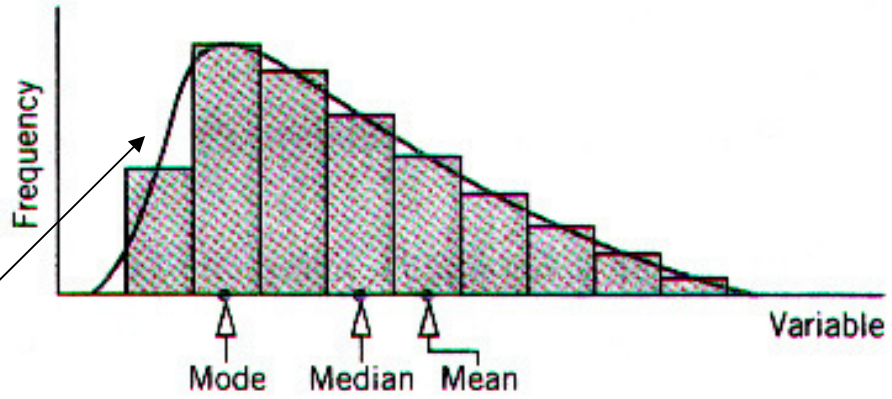
La **MODA** di un dataset è (se esiste) **il valore che si presenta con maggior frequenza.** Il dataset può essere unimodale, bimodale, multimodale.

Esempio:
77 69 74 81 74 62 74 73 → Moda = 74 20 15 45 21 69 41 5 → Moda = ??

- Gli argomenti devono essere numeri, nomi, matrici o riferimenti che contengono numeri.
- Se una matrice o un riferimento contiene testo, valori logici o celle vuote, tali valori verranno ignorati. Le celle contenenti il valore zero verranno invece incluse nel calcolo.
- Se l'insieme dei dati non contiene valori duplici, MODA restituirà il valore di errore #N/D.

RELAZIONE TRA MEDIA - MEDIANA – MODA

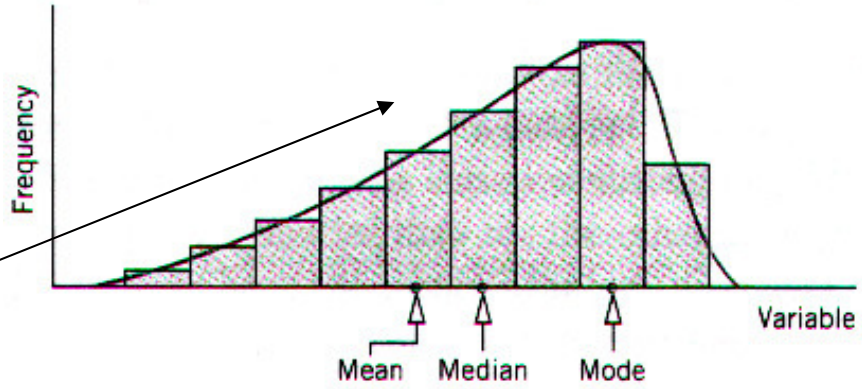




positiva

Skewness
(asimmetria)

negativa



OUTLIER

Outlier: valori troppo grandi o troppo piccoli rispetto alla maggior parte dei valori del data set.

Stato	Popolazione (migliaia di abitanti)
Washington	5894
Oregon	3421
Alaska	627
Hawaii	1212
California	33872

OUTLIER

La media è una misura molto sensibile alla presenza di outlier.

MISURE DI DISPERSIONE per dati non raggruppati

Media, Moda e Mediana non danno alcuna informazione sulla distribuzione dei dati all'interno di un data set.

Pertanto, due data set con stesso valore medio, possono avere un **intervallo di variabilità** completamente diverso

Esempio: età degli impiegati di due uffici
Ufficio 1: 47 38 35 40 36 45 39
Ufficio 2: 70 33 18 52 27

Hanno lo stesso valore medio,
ma i dati hanno dispersione diversa

Le misure di dispersione permettono di definire un valore che indica quanto i dati sia concentrati o dispersi attorno ai valori tipici.

Il più semplice indice di dispersione è il **RANGE**

RANGE = VALORE MASSIMO (DATASET) - VALORE MINIMO (DATASET)

VARIANZA (σ^2) E DEVIAZIONE STANDARD (σ)

La deviazione standard è un **indice di dispersione dei dati di un dataset attorno al valor medio**.
Generalmente un basso valore di D.S. di un data set indica che i valore del data set sono sparsi in un'area relativamente piccola attorno al valor medio.

Al contrario, un alto valore di D.S. denota che i valore dei dataset sono sparsi in un'area abbastanza grande attorno al valore medio.

La D.S. (σ) è la radice quadrata della varianza (σ^2)

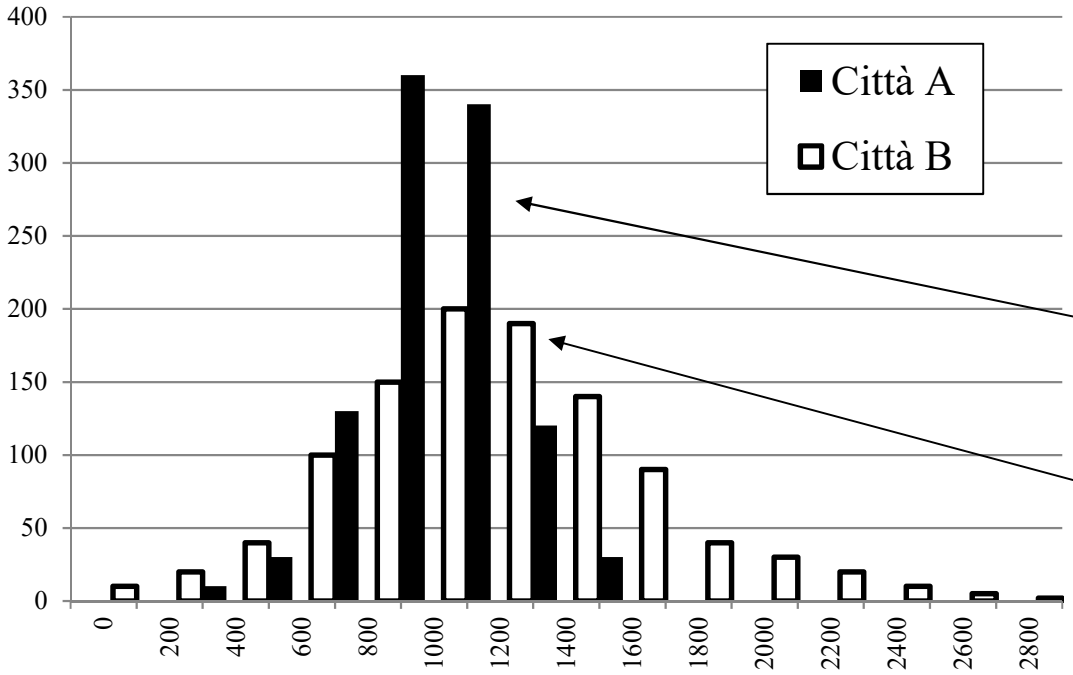
$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} \qquad \sigma = \text{sqrt}(\sigma^2)$$

Si osserva che:

σ^2 e σ non sono MAI < 0

Se N = valore costante → $\sigma=0$

MISURE DI DISPERSIONE



	Media	Mediana	Moda
Città A	1188,2	1000	1000
Città B	1227,8	1000	1000

Kurtosis positiva

Kurtosis negativa

- Gli istogrammi sono quasi simmetrici: media, mediana e moda sono quasi coincidenti
- Il valore degli indici centrali è approssimativamente lo stesso per i due istogrammi
- Le due distribuzioni sono centrate attorno allo stesso valore

Differente dispersione dei dati

EXCEL STRUMENTO ANALISI DEI DATI – STATISTICA DESCRITTIVA

Lo strumento di analisi Statistica descrittiva genera un rapporto di statistica univariata per i dati dell'intervallo di input, fornendo informazioni sulla tendenza centrale e la variabilità dei dati.

Intervallo di input: riferimento di cella per l'intervallo di dati da analizzare che deve consistere in due o più intervalli di dati adiacenti disposti in colonne o righe.

Dati raggruppati per: indicare se i dati nell'intervallo di input sono disposti in righe o in colonne

Etichette nella prima riga/Etichette nella prima colonna:
Se la prima riga dell'intervallo di input contiene etichette, selezionare la casella di controllo **Etichette nella prima riga**.
Se le etichette si trovano invece nella prima colonna dell'intervallo di input, selezionare la casella di controllo **Etichette nella prima colonna**.
Se l'intervallo di input non contiene etichette, queste caselle di controllo dovranno essere deselezionate.

Intervallo di output: immettere il riferimento della cella superiore sinistra della tabella di output.
Questo strumento genera due colonne (etichette di statistica + statistiche) di informazioni per ciascun insieme di dati. Verrà scritta una tabella di statistiche a due colonne per ciascuna colonna o riga dell'intervallo di input, a seconda dell'opzione selezionata nella casella **Dati raggruppati per**.

Nuovo foglio di lavoro: consenti di inserire un nuovo foglio di lavoro nella cartella di lavoro corrente e incollare i risultati a partire dalla cella A1 del nuovo foglio di lavoro.

Nuova cartella di lavoro: consente di creare una nuova cartella di lavoro e incollare i risultati in un nuovo foglio della nuova cartella di lavoro.

Riepilogo statistiche: genera un campo nella tabella di output per ciascuna delle seguenti statistiche: Media, Errore standard (della media), Mediana, Modalità, Deviazione standard, Varianza, Curtosi, Asimmetria, Intervallo, Minimo, Massimo, Somma, Conteggio, Più grande (#), Più piccolo (#) e Livello di confidenza.

MISURE DI POSIZIONE

Una misura di posizione determina la posizione di un singolo valore in relazione agli altri, appartenenti ad un campione o ad una popolazione di dati.

Gli indici di posizione più utilizzati sono:

- **Quantile**
- **Quartile**
- **Percentile**

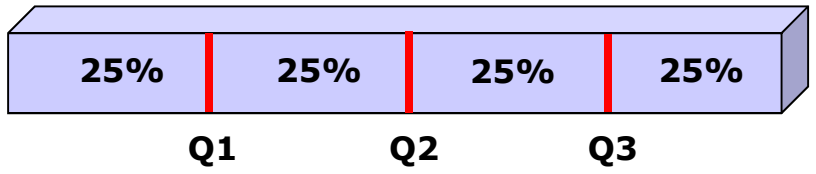
Un quantile è solitamente quel valore x_q per il quale la somma di tutte le frequenze è uguale al valore q (compreso tra zero e uno).
Qualora il quantile venga espresso in termini percentuali, allora si parla pure di **percentile.**
Alcuni quantili vengono indicati in modo proprio:
quartile, quando q assume valori pari a 0,25 o 0,5 o 0,75
mediana, quando q=0,5 (pari al secondo quartile)

I quartili ripartiscono il dataset ordinato, in 4 parti di pari frequenza.

Il primo quartile (Q1) è il valore (o l'insieme di valori) di una distribuzione X per cui la frequenza cumulativa vale 0,25.

Il secondo quartile (Q2) è la mediana.

Il terzo quartile (Q3)è il valore (o l'insieme di valori) di una distribuzione X per cui la frequenza cumulativa vale 0,75.



I percentili dividono un dataset ordinato in 100 parti uguali.
Ogni dataset ordinato ha 99 percentili che lo dividono in 100 parti uguali.