Informatica

Modulo III – Analisi statistica dei dati con Excel

Dr. Ing. Cristian Zambelli

a.a. 2019/2020

Corso di Laurea Triennale in Economia

DIPARTIMENTO
DI ECONOMIA
E MANAGEMENT

Analisi statistica dei dati (…come mai Excel?)

- La **statistica** è l'arte e la scienza di raccogliere, analizzare, interpretare e presentare **dati** nell'intento di trasformarli in informazione utile (…ve lo dirà il Prof. Bonnini nel suo libro «Statistica per le scienze economiche e aziendali» e nel corso di Statistica nel vostro piano degli studi)
- L'avere a che fare con **dati** che provengono da elementi di una stessa famiglia (**popolazione**) impone di usare adeguate metodologie di analisi che prevedono descrittori e misure implementabili come formule o grafici

- Mmmhhh… Mi pare che ci sia un applicativo che dovrebbe aiutare in questa serie di procedure… è qualcosa che usa formule e grafici…
- È proprio lui: Excel!

Alcune definizioni utili nella statistica descrittiva

- La **statistica descrittiva** si occupa direttamente di riassumere e presentare l'informazione contenuta nei **dati** (cit. Prof. Bonnini)
- Le tecniche di **statistica descrittiva** consistono ad esempio in **misure di sintesi** che permettano di riassumere l'informazione essenziale attraverso pochi **numeri** che "comunichino" qualcosa che non era evidente con la semplice osservazione dei **dati** raccolti (sempre dal Prof. Bonnini)
- Quando si analizza un **dato** è necessario individuare una caratteristica di interesse su cui portare l'osservazione **statistica** → si parla di **variabile**
- Esempi di variabile:
  - Soddisfazione di un cliente rispetto ad un investimento eseguito
  - Stipendio medio di un CFO (Chief Financial Officer)
  - Valore sul mercato di una commodity (es. petrolio, cereali, carbone, ecc.)

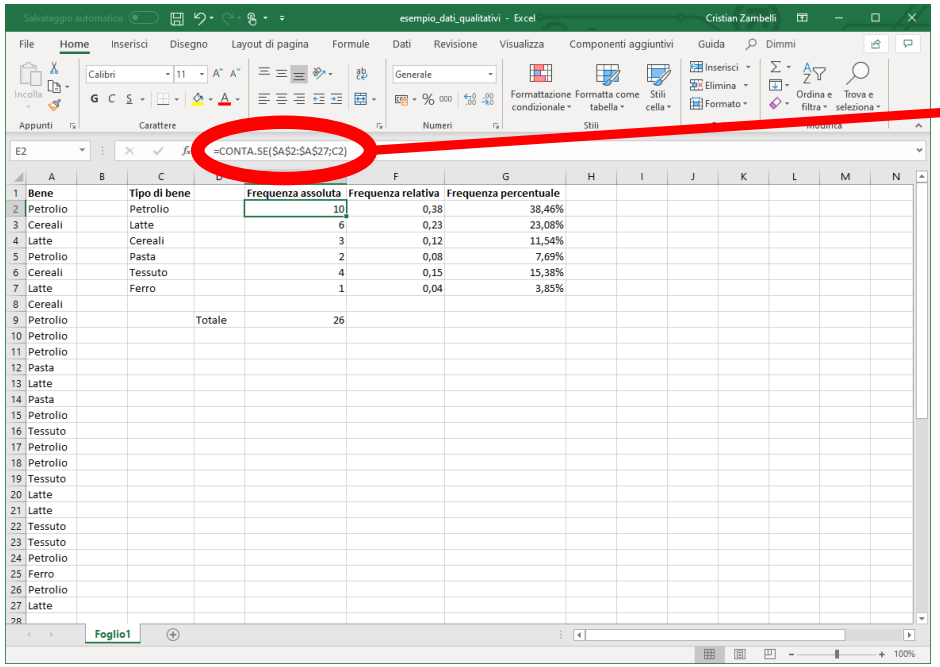
Alcune definizioni utili nella statistica descrittiva – Variabili

- Una **variabile** viene identificata come **qualitativa** se le sue **modalità** (i valori che può assumere) sono valori numerici e non misurabili (qualità appunto…)
  - Si dice che la **variabile qualitativa** è **ordinale** se le sue **modalità** sono passibili di un **ordinamento** (es. la soddisfazione di un cliente → poco, abbastanza, tanto)
  - Si dice che la **variabile qualitativa** è **nominale** se invece non è possibile stabilire un **ordinamento** delle sue **modalità** (es. nome di un bene → benzina, pane, latte)
- Una **variabile** viene identificata come **quantitativa** se le sue **modalità** sono **numeri** e quindi quantità misurabili
  - Se la **variabile quantitativa** prevede operazioni di conteggio su un **numero** finito di **modalità** allora si definisce **discreta** (es. numero di dipendenti di una piccola industria)
  - Se la **variabile quantitativa** può assumere un **numero** qualsiasi all'interno di un **intervallo** si definisce **continua** (es. stipendio medio di un impiegato)
- Ogni **variabile** viene quindi osservata su un insieme di unità **statistiche** e pertanto genera una **distribuzione** di valori o di **modalità**

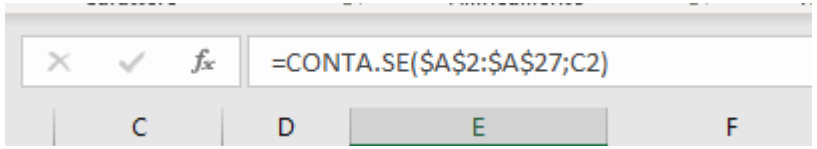
Analisi dei dati qualitativi – Frequenze

- L'analisi di una **variabile qualitativa** per rappresentare un certo fenomeno inizia sempre con la valutazione della sua **distribuzione**
- Nella pratica si avrà a che fare con una **tabella** in cui la maggior parte delle **modalità** assunte dalla **variabile** si ripetono per una o più volte
- Per definizione, la **distribuzione di frequenza** di una data **variabile** è una **tabella** di sintesi che indica il numero di osservazioni per ciascuna **modalità** della variabile → **Frequenza assoluta**
- Tuttavia, per favorire l'interpretazione dei dati a volte conviene prendere in esame anche la distribuzione delle **frequenze relative** e **percentuali**
- Le prima indica la proporzione di osservazioni ascrivibili a ciascuna **modalità**, la seconda ne indica la percentuale

Analisi di dati qualitativi in Excel – La funzione CONTA.SE



Bene		Tipo di bene		Frequenza assoluta	Frequenza relativa	Frequenza percentuale
Petrolio		Petrolio		10	0,38	38,46%
Cereali		Latte		6	0,23	23,08%
Latte		Cereali		3	0,12	11,54%
Petrolio		Pasta		2	0,08	7,69%
Cereali		Tessuto		4	0,15	15,38%
Latte		Ferro		1	0,04	3,85%
Cereali						
Petrolio			Totale	26		
Petrolio						
Petrolio						
Pasta						
Latte						
Pasta						
Petrolio						
Tessuto						
Petrolio						
Petrolio						
Tessuto						
Latte						
Latte						
Tessuto						
Tessuto						
Petrolio						
Ferro						
Petrolio						
Latte						



- Per definizione la somma delle **frequenze assolute** deve essere uguale al totale delle osservazioni
- Quando si usa la **funzione** *CONTA.SE* fate sempre attenzione ai **riferimenti**

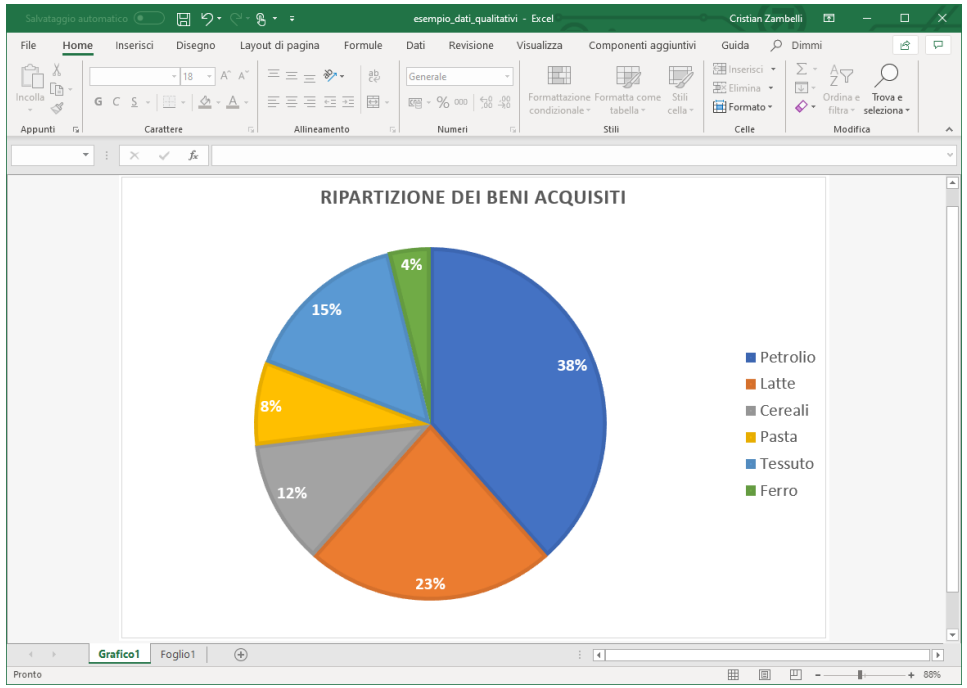
- Per costruire la **distribuzione di frequenza assoluta** in Excel si utilizza la **funzione** *CONTA.SE(intervallo_dati;criterio_di_conteggio)*
- *Intervallo_dati* contiene le osservazioni della **variabile** studiata e *criterio_di_conteggio* (scritto fra doppi apici) il criterio di selezione delle sue **modalità**

Analisi di dati qualitativi in Excel – La funzione CONTA.SE

- Le **frequenze relative** possono essere calcolate agevolmente dal risultato della **funzione** *CONTA.SE* applicata su ogni **modalità**
- Basta dividere le **frequenze assolute** ottenute dalla **funzione** per il numero totale delle **osservazioni** (…che ottenete con la **funzione** *SOMMA* applicata alle **celle** che contengono le **frequenze assolute** ad esempio…)
- Le **frequenze percentuali** si possono ottenere moltiplicando per un fattore *100* le **frequenze relative** oppure formattando le **celle** (in questo caso sarà non un calcolo ma una visualizzazione!) con il **tipo di dato numerico «percentuale»**

- In casi eccezionali si può usare la **funzione** *CONTA.PIU'.SE* che funziona come *CONTA.SE*, ma è applicata su più criteri di conteggio (utile quando si vogliono aggregare e confrontare più **modalità**). Tuttavia si ritiene questa una funzionalità avanzata per questo corso

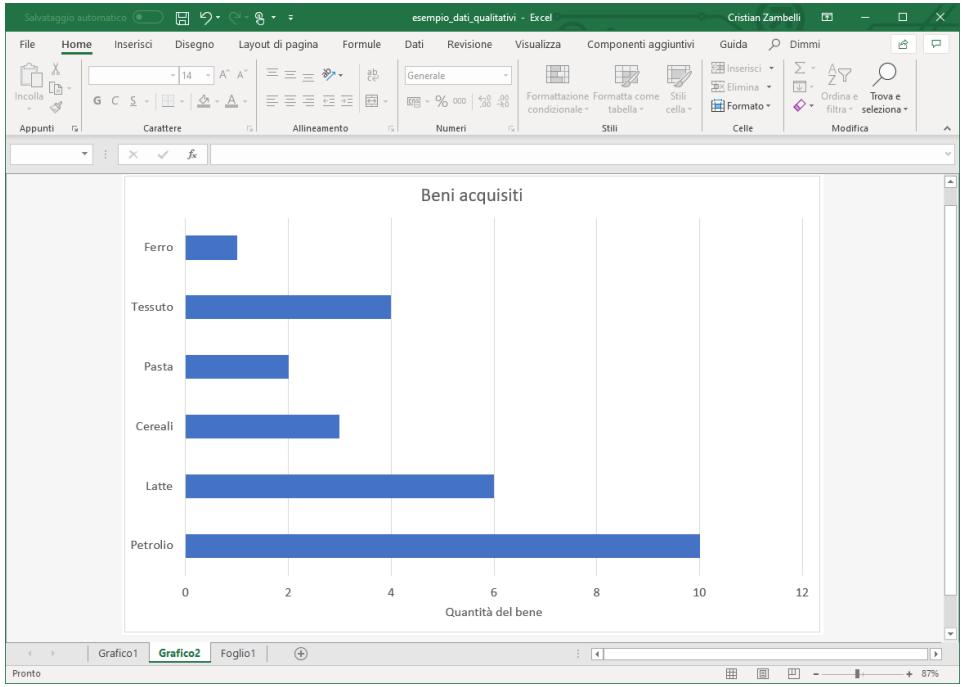
Grafici per l'analisi di dati qualitativi in Excel – Grafico a torta



- Il **grafico a torta** è un ottimo strumento per la rappresentazione dei **dati qualitativi**
- Facendo uso delle **frequenze relative** (o **percentuale**) si suddivide «la torta» in spicchi la cui grandezza è proporzionale al «peso» della **modalità** osservata
- Si noti che con il **grafico a torta** queste ultime due **frequenze** vengono calcolate automaticamente dai dati delle **frequenze assolute**

- Basta selezionare la colonna delle **frequenze assolute** come insieme di **dati** prima di usare la funzionalità *Inserisci grafico*
- Quando si utilizza un **grafico a torta** è molto importante che i dati da rappresentare costituiscano effettivamente il 100% delle **osservazioni**

Grafici per l'analisi di dati qualitativi in Excel – Grafico a barre



- Il **grafico a barre** è un'alternativa al grafico a torta per visualizzare **dati qualitativi**
- In questo **grafico** la **frequenza** di ciascuna modalità della variabile osservata è rappresentata mediante una barra
- La lunghezza della barra è proporzionale alla **frequenza** stessa (sia essa **assoluta**, **relativa** o **percentuale**)

- I valori numerici sull'ascissa identificano la scala delle **frequenze** adottata
- Sull'ordinata le barre sono equispaziate per evidenziare la caratteristica **qualitativa** delle **modalità**. Inoltre si nota che le barre non hanno un ordine di comparizione preciso sul **grafico**

Analisi di dati quantitativi in Excel – La funzione FREQUENZA

- Per i **dati** di tipo **quantitativo** è necessario organizzare molto spesso le **modalità** di una **variabile** in **gruppi di frequenza**
- In Excel la **funzione** che sintetizza le **frequenze** per **variabili quantitative discrete** è *FREQUENZA(celle_osservazioni;celle_modalità)*
- Il parametro *celle_osservazioni* indica il **riferimento** alle **celle** contenenti le **osservazioni** della **variabile** studiata, mentre il parametro *celle_modalità* indica il **riferimento** alle **celle** contenenti le **modalità** assunte dalla **variabile**
- ATTENZIONE! La **funzione** *FREQUENZA* è una **funzione** speciale in Excel che appartiene ad una famiglia di **funzioni** particolari chiamate «matrice»
- Per il loro inserimento si usa la seguente procedura:
  - Scegliere l'**intervallo di celle** in cui si vuole far apparire il risultato della **funzione**
  - Digitare la **formula** nella **barra della formula** senza inserirla
  - Premere contemporaneamente *CTRL+Maiusc+Invio* sulla tastiera

Leggibilità dei dati usando la funzione FREQUENZA

	A	B	C	D	E
1	Codice Titolo	Rating		Rating	Frequenza
2	1	5		1	2
3	2	4		2	3
4	3	2		3	2
5	4	7		4	4
6	5	4		5	4
7	6	5		6	3
8	7	4		7	3
9	8	6		8	1
10	9	7		9	1
11	10	1			
12	11	3			
13	12	2			
14	13	6			
15	14	5			
16	15	4			
17	16	7			
18	17	6			
19	18	3			
20	19	1			
21	20	2			
22	21	8			
23	22	9			
24	23	5			

In questo esempio di catalogazione del rating da 1 a 9 di 23 titoli bancari potrebbe essere scomodo usare la **funzione** *FREQUENZA* su tutte le **modalità** (troppi valori…)

G	H	I
Classi di rating	Limite superiore	
1-3	3	
4-6	6	
7-9	9	

- Quando la **variabile quantitativa discreta** esaminata assume un numero molto elevato di **modalità** è necessario rendere «leggibili» le **frequenze**
- In questo caso più che organizzare la **tabella delle frequenze** per singolo valore di occorrenza della **variabile** si preferisce costruire delle **classi di modalità** → è necessario stabilirle però…

Classi di modalità (così interpreto meglio i dati…)

G	H	I
Classi di rating	**Limite superiore**	**Frequenza**
1-3	3	7
4-6	6	11
7-9	9	5

- Per costruire una **tabella di frequenza** per **classi di modalità** in Excel si utilizza sempre la **funzione** *FREQUENZA* come visto precedentemente
- In questo caso però, il secondo argomento della **funzione** deve contenere il limite superiore (il valore più elevato) di ogni **classe di modalità** stabilite
- Prendendo l'esempio della slide precedente si può suddividere il rating dei titoli secondo le **classi** 1-3, 4-6, e 7-9
- I limiti superiori delle **classi** saranno quindi 3, 6, e 9, rispettivamente
- A questo punto si utilizza la stessa procedura di calcolo delle frequenze

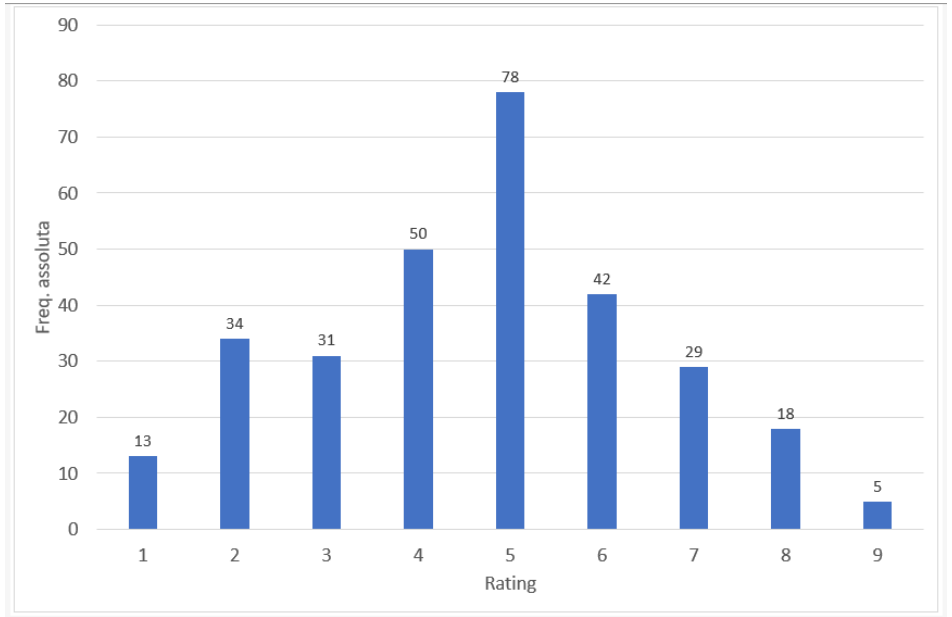
Frequenza cumulativa – Impariamola con un esempio

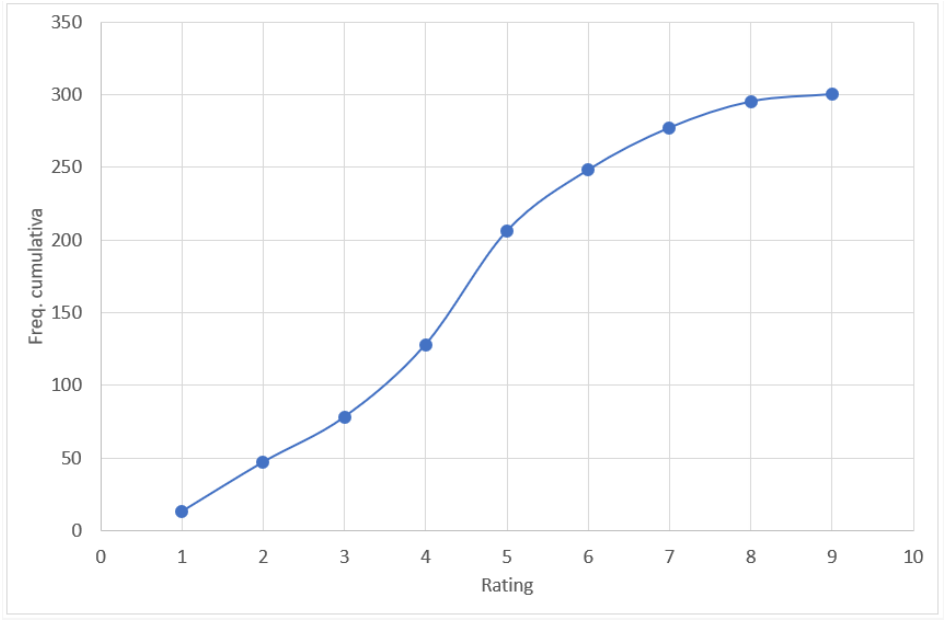
E19

	A	B	C	D	E	F
1	Rating	Freq. Assoluta	Freq. Relativa	Freq. Percentuale	Freq. Cumulativa	Freq. Cumulativa Percentuale
2	1	13	0,043	4,33%	13	4,33%
3	2	34	0,113	11,33%	47	15,67%
4	3	31	0,103	10,33%	78	26,00%
5	4	50	0,167	16,67%	128	42,67%
6	5	78	0,260	26,00%	206	68,67%
7	6	42	0,140	14,00%	248	82,67%
8	7	29	0,097	9,67%	277	92,33%
9	8	18	0,060	6,00%	295	98,33%
10	9	5	0,017	1,67%	300	100,00%
11						
12						
13	Totale	300				
14						

- Quando si ha a che fare con **variabili quantitative**, potrebbe essere necessario contare la **frequenza cumulativa**
- Essa rappresenta il **numero di occorrenze minori o uguali** alla **modalità** in esame. Molto utile per capire ad esempio quante volte un titolo ha un rating inferiore ad un certo valore
- Dalla **frequenza cumulativa** si può passare alla **frequenza cumulativa percentuale** dividendo la prima per il **totale delle osservazioni** e moltiplicando per *100*

Grafici per l'analisi di dati quantitativi in Excel



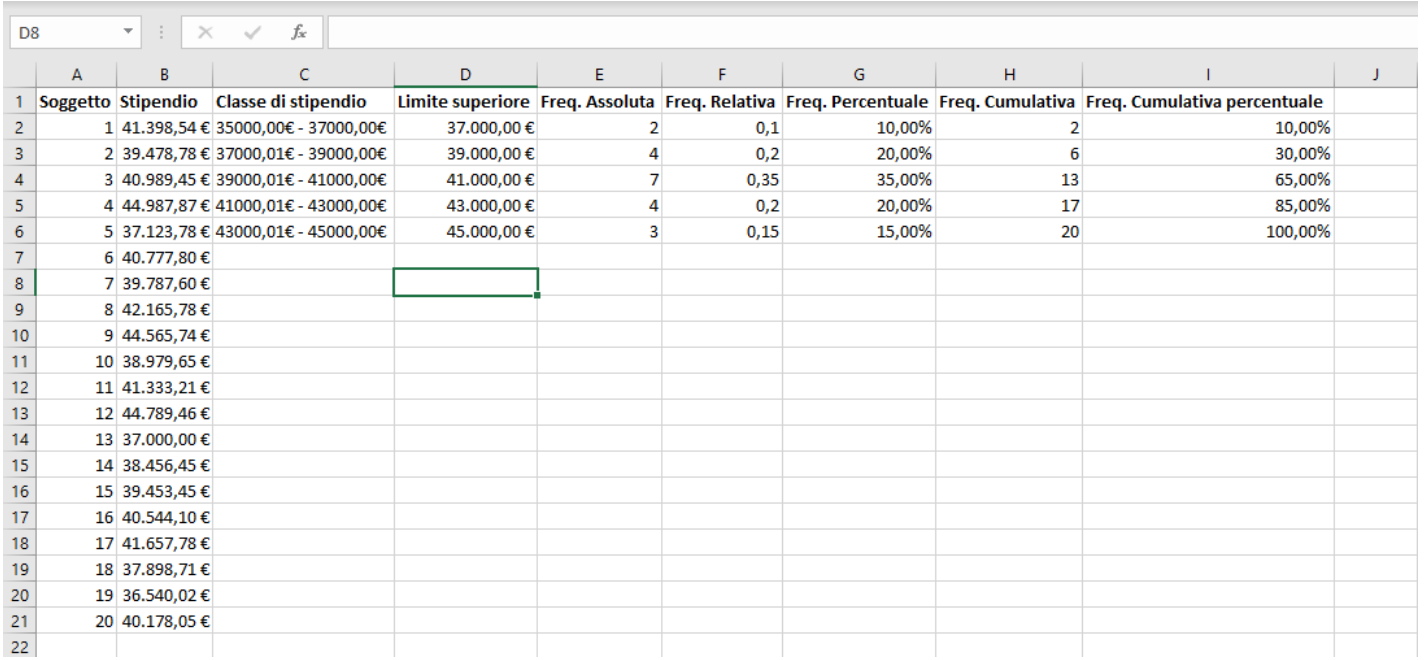


- Gli strumenti grafici maggiormente utilizzati per la rappresentazione delle **variabili quantitative** in Excel sono:
  - Il **grafico a barre** – usato per la rappresentazione della **frequenza assoluta** sull'ordinata e delle **modalità** della **variabile** in osservazione sull'ascissa
  - Il **grafico a dispersione** – usato per la rappresentazione della **frequenza cumulativa** sull'ordinata e delle **modalità** della **variabile** in osservazione sull'ascissa

Analisi di dati quantitativi in Excel – …se i dati sono continui?

- Nel contesto delle **variabili quantitative continue**, data la natura del **dato**, la **tabella di frequenza** deve essere per forza organizzata per **classi di modalità**
- Le **classi** vengono formate specificando gli **intervalli** per rappresentare i **dati**
- Come regola generale si consiglia di utilizzare tra le *5* e le *20* **classi** a seconda del fenomeno che la **variabile** rappresenta → ma non sempre vero…
- L'obiettivo è quello di utilizzare un numero sufficiente di **classi** che catturi la variabilità dei **dati** cercando di non avere **classi** con poche unità
- Inoltre, sulla base della prassi e dell'esperienza di chi con la **statistica** ci lavora quotidianamente, si consiglia di usare **classi** ad **ampiezza costante**

Analisi di dati quantitativi continui – Esempio semplice

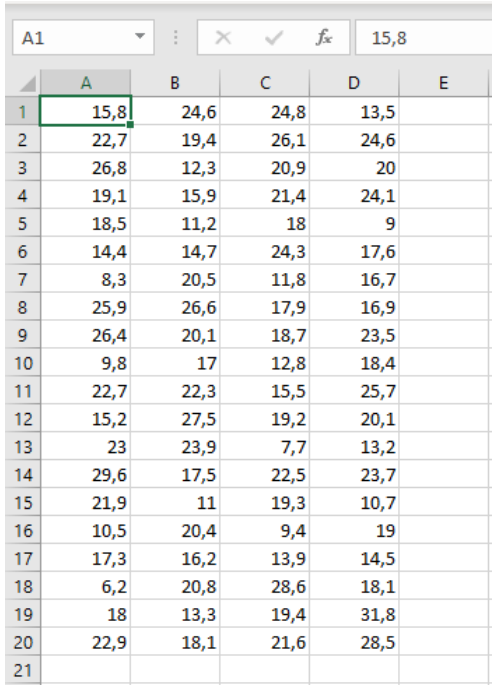


	A	B	C	D	E	F	G	H	I	J
1	Soggetto	Stipendio	Classe di stipendio	Limite superiore	Freq. Assoluta	Freq. Relativa	Freq. Percentuale	Freq. Cumulativa	Freq. Cumulativa percentuale	
2	1	41.398,54 €	35000,00€ - 37000,00€	37.000,00 €	2	0,1	10,00%	2	10,00%	
3	2	39.478,78 €	37000,01€ - 39000,00€	39.000,00 €	4	0,2	20,00%	6	30,00%	
4	3	40.989,45 €	39000,01€ - 41000,00€	41.000,00 €	7	0,35	35,00%	13	65,00%	
5	4	44.987,87 €	41000,01€ - 43000,00€	43.000,00 €	4	0,2	20,00%	17	85,00%	
6	5	37.123,78 €	43000,01€ - 45000,00€	45.000,00 €	3	0,15	15,00%	20	100,00%	
7	6	40.777,80 €								
8	7	39.787,60 €								
9	8	42.165,78 €								
10	9	44.565,74 €								
11	10	38.979,65 €								
12	11	41.333,21 €								
13	12	44.789,46 €								
14	13	37.000,00 €								
15	14	38.456,45 €								
16	15	39.453,45 €								
17	16	40.544,10 €								
18	17	41.657,78 €								
19	18	37.898,71 €								
20	19	36.540,02 €								
21	20	40.178,05 €								
22										

- Supponiamo di monitorare lo stipendio annuale di 20 soggetti
- L'analisi **statistica** suggerita è sicuramente relativa a **dati quantitativi continui**
- Definiamo *5* **classi** stipendiali in base ai valori osservati in esempio (…eye-judging)

- Per il calcolo della **tabella** della **frequenza assoluta, relativa, percentuale, cumulativa** e **cumulativa percentuale** procediamo ad utilizzare la **funzione** *FREQUENZA* esattamente allo stesso modo dei **dati quantitativi discreti**
- Ricordate!!! È importante applicare la **funzione** sui limiti superiori delle **classi**

Analisi di dati quantitativi continui – Esempio complesso



	A	B	C	D	E
1	15,8	24,6	24,8	13,5	
2	22,7	19,4	26,1	24,6	
3	26,8	12,3	20,9	20	
4	19,1	15,9	21,4	24,1	
5	18,5	11,2	18	9	
6	14,4	14,7	24,3	17,6	
7	8,3	20,5	11,8	16,7	
8	25,9	26,6	17,9	16,9	
9	26,4	20,1	18,7	23,5	
10	9,8	17	12,8	18,4	
11	22,7	22,3	15,5	25,7	
12	15,2	27,5	19,2	20,1	
13	23	23,9	7,7	13,2	
14	29,6	17,5	22,5	23,7	
15	21,9	11	19,3	10,7	
16	10,5	20,4	9,4	19	
17	17,3	16,2	13,9	14,5	
18	6,2	20,8	28,6	18,1	
19	18	13,3	19,4	31,8	
20	22,9	18,1	21,6	28,5	
21					

- Non sempre i **dati** di un'analisi sono organizzati in modo semplice da interpretare
- Nell'esempio della slide (da M. Garetto – Laboratorio di Statistica con Excel) sono indicate le misure dell'emissione giornaliera di gas inquinanti da un impianto
- Tutto quello che vediamo sono una serie di **numeri** rappresentati con un decimale
- …e quindi???

- Qui non è immediato capire quante **classi** utilizzare per rappresentare la **variabile quantitativa continua**, ma non è nemmeno immediato capire quanti **dati** ci sono a disposizione, quale **variazione** hanno, ecc.
- Si impone un'analisi **statistica** più complessa rispetto al caso precedente

Analisi di dati quantitativi continui – Funzioni utili

- Quando non si conosce il **numero di dati** a disposizione per l'analisi si può utilizzare la funzione *CONTA.NUMERI(intervallo_celle)*
  - L'argomento di questa **funzione** è l'**intervallo di celle** su cui cercare quante di esse rappresentano **numeri**. Simile alla **funzione** *CONTA.SE*… right?
- Il **numero di dati** *(n)* da analizzare può essere usato in una regola empirica che ci consente di calcolare il numero ottimale (*k*) di **classi** per rappresentare i **dati**:
  - $k = 1 + 3{,}322 \cdot Log_{10}(n)$
  - Questa regola restituisce un **numero** con i decimali. Siccome il **numero di classi** in cui suddividere i **dati** deve essere un **numero intero,** si può usare la funzione *INT* per arrotondare per difetto il risultato della formula precedente → *=INT(1+3,322*LOG10(80))*
- Una volta trovato il **numero ottimale di classi** è necessario definirne la loro **ampiezza** attraverso il **campo di variazione** dei **dati** (già! Ma cos'è???)
  - Il **campo di variazione** è la più semplice **misura di variabilità** dei **dati** che si calcola come la differenza fra il **valore massimo** e il **valore minimo** degli stessi
  - In Excel si calcola come *=MAX(intervallo_celle)-MIN(intervallo_celle)*

Analisi di dati quantitativi continui – Funzioni utili

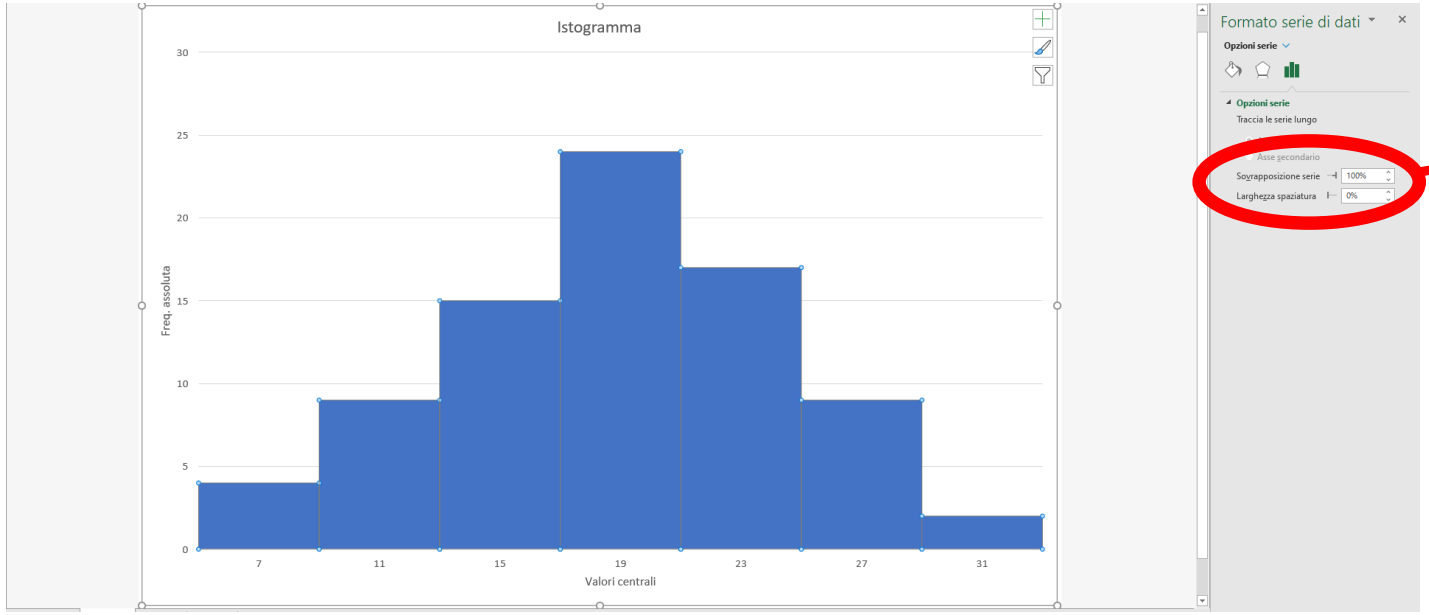
- Il **campo di variazione** diviso per il **numero di classi** in cui suddividere i **dati** rappresenta l'**ampiezza delle classi**
- **L'ampiezza delle classi** tuttavia è un **numero** che deve essere arrotondato per eccesso all'intero usando la **funzione** *ARROTONDA.ECCESSO(num;peso)*
  - Questa **funzione** prende come parametri il **numero da arrotondare** (*num*) e quale multiplo intero del parametro (*peso*) si vuole usare per l'arrotondamento
  - Ad esempio se la cella *C4* contiene il valore 24,65 *=ARROTONDA.ECCESSO(C4;1)* darà come risultato *25*, mentre *=ARROTONDA.ECCESSO(C4;10)* sarà uguale a *30*
- A questo punto vanno creati gli **estremi destri** (limite superiore) di ogni **classe**
  - Si sceglie un **valore** che rappresenti l'**estremo destro** della prima **classe**, mentre i successivi sono semplici incrementi di quest'ultimo in base all'**ampiezza** calcolata
  - Ad esempio se l'**ampiezza delle classi** è *4* e il **primo estremo destro** scelto è *9*, le **classi** successive avranno **come estremi destri** *13, 17, 21,* e così via…
- Controllate sempre che le **classi** scelte comprendano tutti i **dati**!!!!!!

…e adesso FREQUENZA di nuovo come se piovesse…

F	G	H	I	J	K	L	M
	N dati	Classi	Campo di variazione	Ampiezza classi			
	80	7	25,6	4			
	Classi	Estremi destri	Freq. Assoluta	Freq. Relativa	Freq. Percentuale	Valori centrali	
	1	9	4	0,05	5,00%	7	
	2	13	9	0,1125	11,25%	11	
	3	17	15	0,1875	18,75%	15	
	4	21	24	0,3	30,00%	19	
	5	25	17	0,2125	21,25%	23	
	6	29	9	0,1125	11,25%	27	
	7	33	2	0,025	2,50%	31	

- Ora che abbiamo tutte le informazioni che ci servono possiamo costruire la **tabella delle frequenze assolute, relative,** e **percentuali** per i nostri **dati quantitativi continui** con la ben nota **funzione** *FREQUENZA*
- Non capisco perché nella colonna *L* dell'esempio ci sia scritto *«Valori centrali»…*

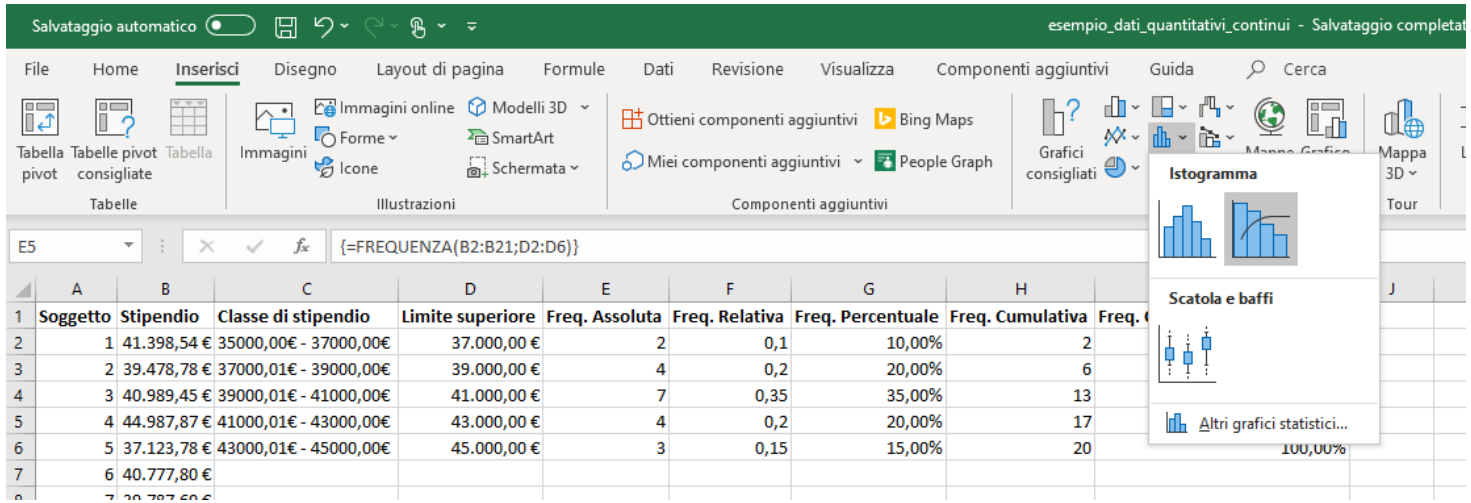
Usare gli istogrammi per plottare i dati – Procedura



- Cliccare con il *tasto destro* del mouse su una delle barre e scegliere *Formato serie di dati…*
- Impostare *Larghezza spaziatura* al valore 0%

- Per fare un **istogramma** in Excel si può procedere allo stesso modo per la creazione di un **diagramma a barre**, con l'accortezza di allargare le barre in modo da rendere gli **intervalli continui** come mostrato in figura
- L'importante è che siano definiti sull'ascissa i **valori centrali** di ogni **classe** scelta e calcolati come differenza fra **l'estremo destro** di ogni **classe** e la metà della loro rispettiva **ampiezza**

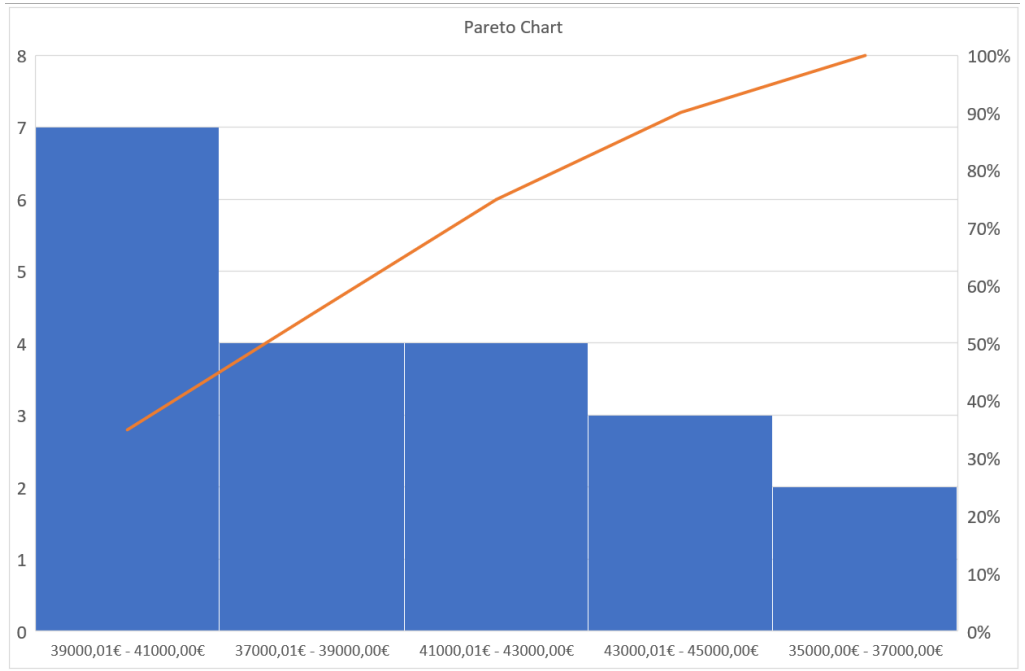
Usare le Pareto charts per plottare i dati – Procedura

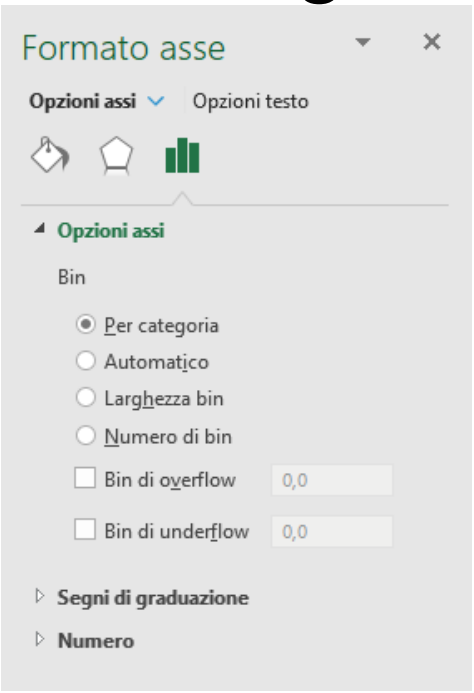


	A	B	C	D	E	F	G	H
1	Soggetto	Stipendio	Classe di stipendio	Limite superiore	Freq. Assoluta	Freq. Relativa	Freq. Percentuale	Freq. Cumulativa
2	1	41.398,54 €	35000,00€ - 37000,00€	37.000,00 €	2	0,1	10,00%	2
3	2	39.478,78 €	37000,01€ - 39000,00€	39.000,00 €	4	0,2	20,00%	6
4	3	40.989,45 €	39000,01€ - 41000,00€	41.000,00 €	7	0,35	35,00%	13
5	4	44.987,87 €	41000,01€ - 43000,00€	43.000,00 €	4	0,2	20,00%	17
6	5	37.123,78 €	43000,01€ - 45000,00€	45.000,00 €	3	0,15	15,00%	20
7	6	40.777,80 €						

- Riprendiamo l'esempio visto precedentemente circa il monitoraggio dello stipendio annuale di 20 individui
- Supponiamo di voler plottare un **grafico** che ci dica non solo la **distribuzione di frequenza** delle classi stipendiali (**istogramma**), ma che percentualmente ci dica quali **classi** pesano cumulativamente di più
- Questo tipo di **grafico** prende il nome di **Pareto chart** e per attivarlo basta scegliere il set di **dati** da plottare e successivamente *Inserisci grafico*

Usare le Pareto charts per plottare i dati – Significato



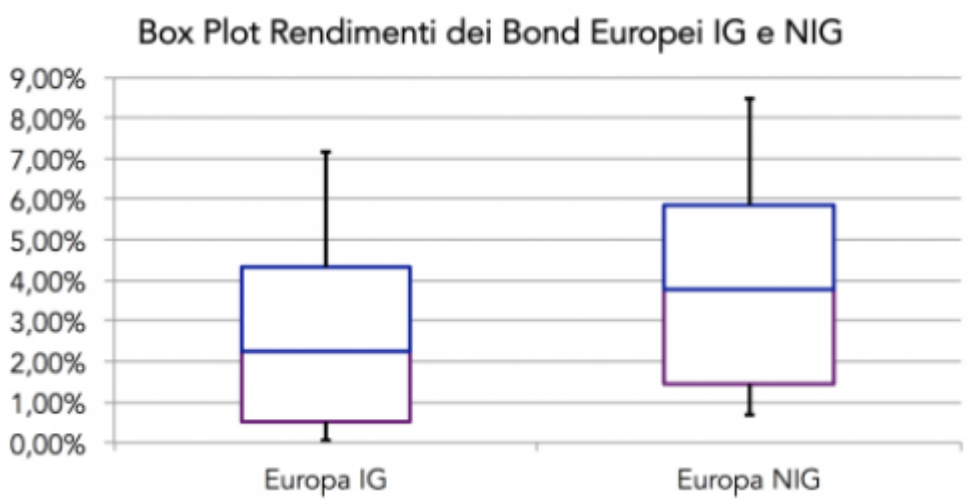


- Questo **grafico** ci dice che circa l'80% delle **frequenze** (osservate la **Pareto line** in arancione) è concentrato nelle prime 3 classi stipendiali
- Potrebbe essere necessario adattare l'ascissa del **Pareto chart** alle categorie dei **dati** usati per il **grafico**. Per fare questo cliccate con il *pulsante destro* del mouse sull'ascissa e dal menu contestuale selezione *Bin per categoria*

Usare i Boxplots – Cenni

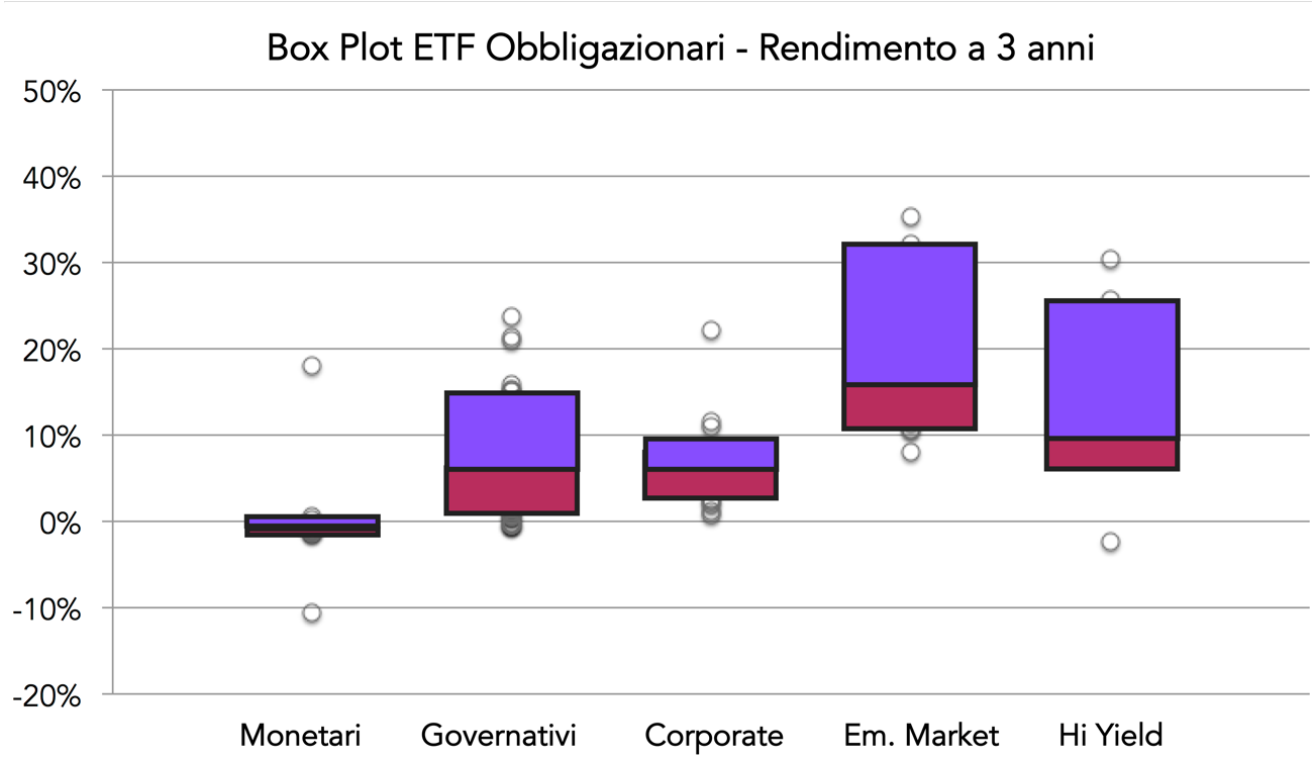
- Un grafico a scatola e baffi (box and whiskers plot abbreviato in boxplot) mostra la distribuzione dei dati in quartili, evidenziando la mediana e i valori anomali
- I rettangoli ("scatole") possono presentare linee che si estendono in verticale denominate "baffi"
- Queste linee indicano la variabilità all'esterno del quartile superiore e inferiore e qualsiasi punto all'esterno di tali linee o baffi è considerato un valore anomalo
- I grafici a scatola e baffi trovano comunemente applicazione nelle analisi statistiche. Ad esempio, è possibile usare un grafico di questo tipo per confrontare i rendimenti di diversi portafogli di investimento oppure i profitti di diverse aziende relativamente agli stessi prodotti

Usare i Boxplots – Esempi



Fonte: https://www.finanzaoperativa.com/il-mio-nome-e-bond-corporate-bond/

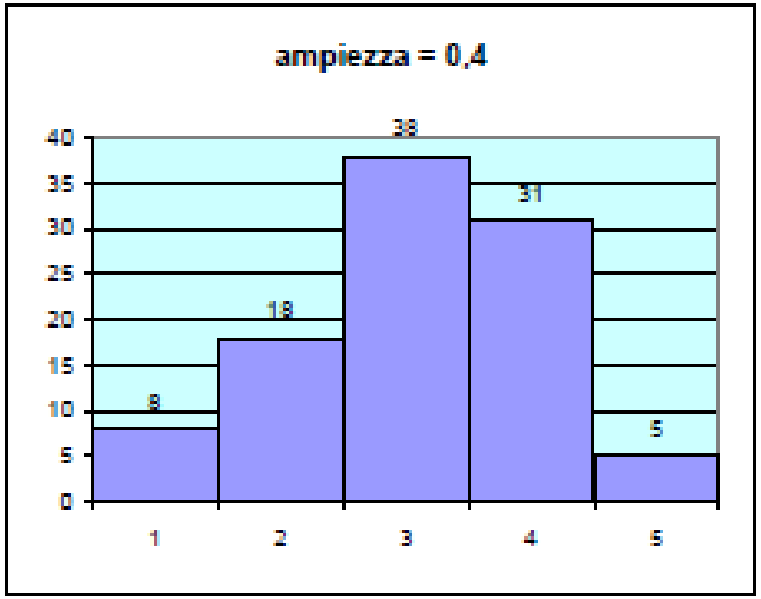
Fonte: https://www.diaman.it/blog/entry/siamo-seri.html

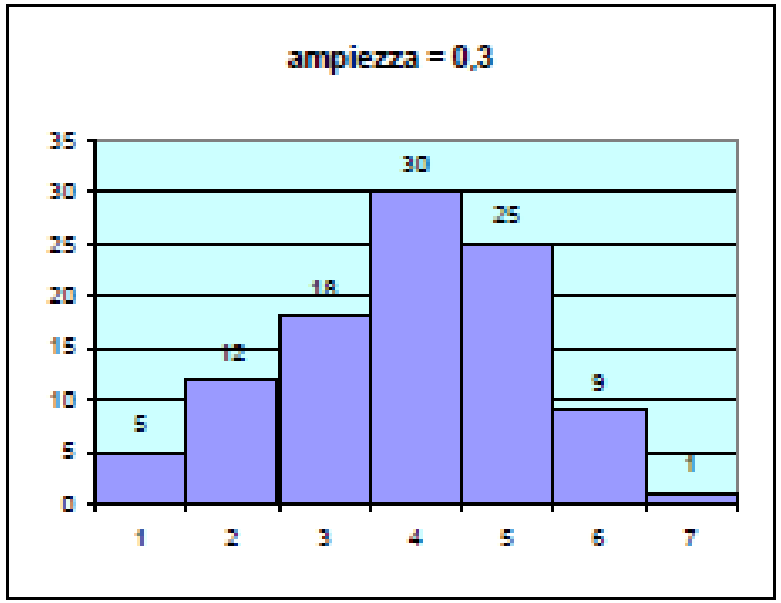


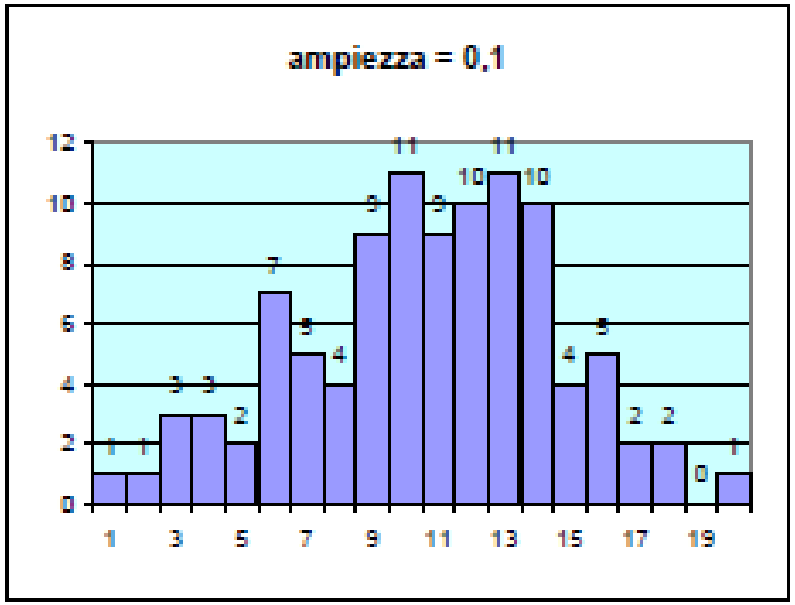
Osservazione su FREQUENZA – …ho sbagliato classi

- Come abbiamo visto la scelta delle **classi** (**ampiezza** ed **estremi destri**) è un punto fondamentale. Se si sceglie male c'è il rischio che qualche **osservazione** della **variabile** in esame non venga conteggiata
- In questo caso la **funzione** *FREQUENZA* ci può aiutare
- Se nell'esempio precedente avessimo scelto **classi di ampiezza** *3,5* avremmo alcuni valori non conteggiati (valori maggiori di *30* nell'ultima **classe**)
- Per ovviare al problema, invece di applicare la **funzione** *FREQUENZA* come al solito, si allunga l'**intervallo di celle** che dovranno contenere le **frequenze assolute** di un'unità
- In questo caso la **cella** aggiuntiva conterà il numero di **osservazioni** maggiori dell'ultimo **estremo destro** → Indicatore fondamentale per cambiare **classi**!

Confrontare istogrammi con ampiezze diverse







- A volte l'occhio umano è davvero la soluzione migliore per capire se la scelta delle **classi** è corretta
- Osservando l'**istogramma** ottenuto dopo la suddivisione in **classi** si dovrebbe evincere una buona omogeneizzazione degli **intervalli** (niente **intervalli** vuoti ma nemmeno troppo grandi…)

Analisi univariata dei dati – Misure numeriche

- Abbiamo visto la potenza di Excel nell'analisi dei **dati** mediante **tabelle di frequenza** e rappresentazioni **grafiche**
- Tuttavia per le **variabili quantitative** abbiamo anche a disposizione una serie di strumenti di sintesi ancora più potenti: le **misure numeriche**
- Le **misure numeriche** per dati **univariati** (relativi cioè ad una sola **variabile**) si possono dividere in tre categorie:
  - **Misure di tendenza centrale**
  - **Misure di variabilità**
  - **Misure di forma**

- In questo le illustreremo tutte da un punto di vista concettuale (lasciando al Prof. Bonnini nel corso di Statistica tutti i dettagli) e successivamente ne analizzeremo l'implementazione con Excel (libreria di **funzioni statistiche**)

Misure di tendenza centrale

- Le **misure di tendenza centrale** (o di **centralità**) sono **indicatori** di sintesi di una **distribuzione di frequenza**
- La **centralità** può essere definita in diversi modi, a cui corrispondono differenti **indicatori** associati
- I più utilizzati sono:
  - **La media**
  - **La mediana**
  - **La moda**
- A questa famiglia di **indicatori** possono essere associati anche i **percentili** e i **quartili**, che identificano **misure di tendenza** diverse da quella **centrale**, ma utili per una sintesi della **distribuzione di frequenza**

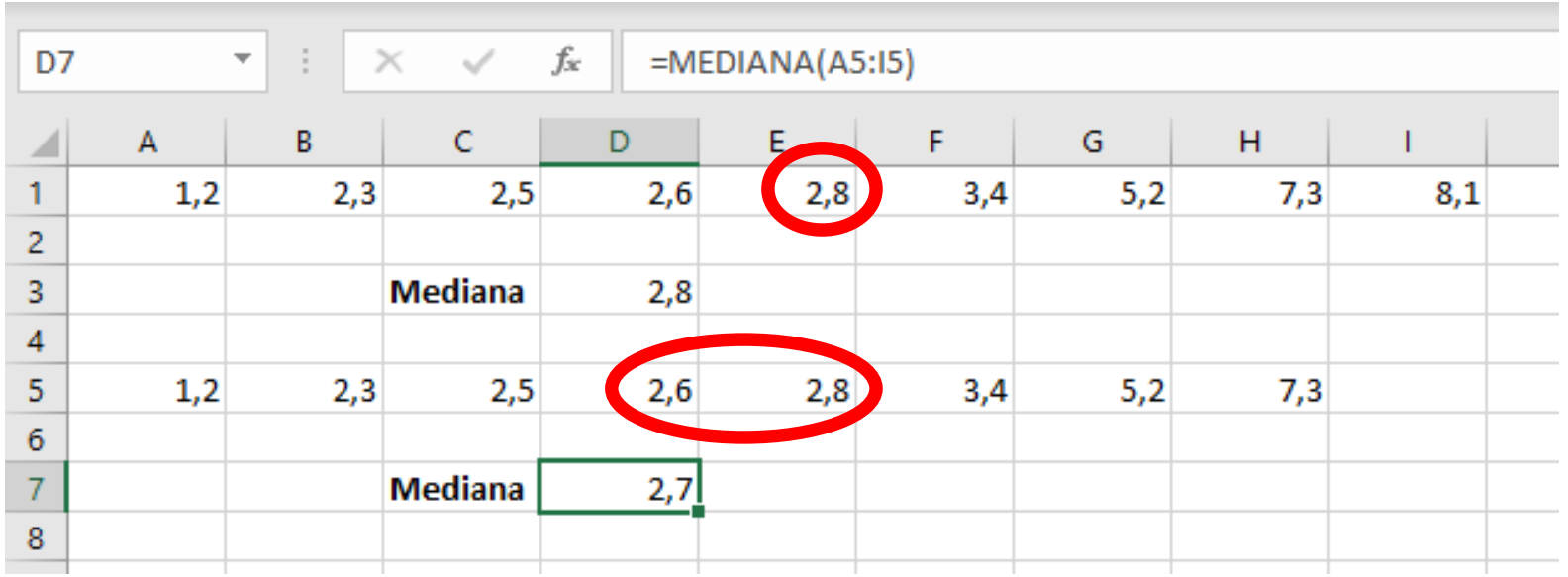
La media – l'indicatore più facile da capire…

- La **media** è la più importante e la più nota **misura di tendenza centrale**
- Essa rappresenta il baricentro della **distribuzione** dei valori di una **variabile quantitativa**, ossia il valore collocato «in mezzo» alla **distribuzione di frequenza** e attorno al quale tendono a concentrarsi le **osservazioni**
- La **misura di media** ($\bar{x}$) si calcola come la somma dei valori delle **osservazioni** (x_i) diviso il numero complessivo delle **osservazioni** stesse (n)
- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Per calcolare la **media** di un **intervallo di celle** in Excel si utilizza la **funzione** *MEDIA(intervallo_celle)* dove l'argomento *intervallo_celle* rappresenta i **riferimenti** alle **celle** contenenti le **osservazioni** della **variabile** in oggetto

La mediana – non è proprio uguale alla media…

- In certe situazioni la **variabile** oggetto di interesse potrebbe presentare **osservazioni** con **valori estremi**, ovvero eccezionalmente grandi o piccoli rispetto ad altre **osservazioni**
- In questo caso, la **media** potrebbe fallire nella sua missione di «sintesi» numerica
- La **media** è influenzata da tutte le **osservazioni** e quindi anche dai **valori estremi** (**outliers**), che possono portare ad una sovra-/sotto-stima della centralità
- La **mediana** non soffre di questa limitazione in presenza di **outliers**
- Essa rappresenta il valore di quella **osservazione** della **variabile** che occupa la posizione centrale quando le **osservazioni** sono disposte in ordine crescente
- Per calcolare la **mediana** in Excel si usa la **funzione** *MEDIANA(intervallo_celle)* dove l'argomento *intervallo_celle* rappresenta i **riferimenti alle celle** contenenti le **osservazioni** della **variabile** in oggetto

Una precisazione sulla mediana



- Quando il numero di **osservazioni** è dispari, la **mediana** corrisponde a quell'unico valore che divide esattamente in due metà la serie ordinata delle **osservazioni**
- Quando invece il numero di **osservazioni** è pari, non vi è un unico valore centrale da utilizzare
- La prassi impone di identificare la **mediana** come la **media** dei due valori collocati in **posizione centrale**

La moda – NO! Non mi riferisco a quella da indossare! O si?

- La **moda** è un'ulteriore **misura di tendenza centrale** che corrisponde al valore dell'**osservazione** che si presenta con **frequenza** più elevata nella **distribuzione**
- Diversamente dalla **media** e dalla **mediana**, la **moda** non è necessariamente una **misura** unica
- Se i valori più frequenti sono due allora si parla di **distribuzione bimodale** delle **frequenze**, se sono di più si parla di **distribuzione multimodale**
- Per calcolare la **moda** in Excel si usa la funzione *MODA.SNGL(intervallo_celle)* dove l'argomento *intervallo_celle* rappresenta i **riferimenti alle celle** contenenti le **osservazioni** della **variabile** in oggetto
- Se la **distribuzione** dei dati è caratterizzata da valori tutti a **frequenza** unitaria, la funzione *MODA.SNGL()* restituisce l'errore *#N/D* (non definibile)
- Inoltre se la **distribuzione** fosse **multimodale** *MODA.SNGL()* non è adeguata

Percentili e quartili

- Per arricchire il contenuto informativo, oltre alle **misure di tendenza centrale** può essere utile arricchire la conoscenza di valori strategici della **distribuzione** dei **dati**
- Possono essere di aiuto i **percentili**, ovvero **misure di posizione non centrali**
- Se i **dati** sono disposti in ordine crescente, si definisce **k-esimo percentile** come quel valore al di sotto del quale si trova almeno il *k%* delle **osservazioni** e al di sopra del quale si trova almeno il *(100-k)%* delle medesime
- I **quartili** sono particolari casi di **percentili** che dividono l'insieme dei **dati** in quattro parti uguali (*25-esimo, 50-esimo o mediana, 75-esimo, 100-esimo*)
- In Excel i **percentili** e i **quartili** si calcolano con la **funzione** *INC.PERCENTILE(intervallo_celle;percentile)* dove l'argomento *intervallo_celle* rappresenta i **riferimenti** alle **celle** contenenti le **osservazioni** della **variabile** in oggetto e l'argomento *percentile* rappresenta il **k-esimo percentile** espresso in centesimi (numero da *0* a *1*)

Misure di variabilità

- Le **misure di tendenza centrale** riducono la complessità dimensionale e informativa dei **dati**. Pochi numeri rappresentano anche miliardi di **dati**
- Tuttavia usare solo queste **misure** ha delle controindicazioni. Se due **distribuzioni di dati** molto diverse tra loro avessero ad esempio la stessa **tendenza centrale**…
- In generale, meno le **osservazioni** di una **distribuzione** sono concentrare intorno alla loro **tendenza centrale**, minore è la capacità delle **misure** di quest'ultima di fornire una sintesi
- Per fornire quindi un quadro di sintesi appropriato della **distribuzione** delle **osservazioni** di una **variabile** è necessario tenere conto della **variabilità dei dati**
- Questo si ottiene mediante le seguenti **misure di variabilità**:
  - **Campo di variazione o range (già visto)**
  - **Range interquartilico**
  - **Varianza**
  - **Deviazione standard**
  - **Coefficiente di variazione**

Range interquartilico – Più robusto del semplice range

- Nonostante il **campo di variazione (range)** abbia il pregio di essere un **indicatore** semplice da calcolare (max-min), basandosi solo su due **osservazioni** soffre del problemi degli **outliers**
- Una **misura di variabilità** più robusta agli **outliers** è il **range interquartilico**, dato dalla differenza tra **il terzo e il primo quartile**
- Tale **misura di variabilità** è di fatto quel **campo di variazione** che contiene il *50%* delle **osservazioni** più vicine al **valore centrale**
- Per calcolare questa **misura** in Excel, ad esempio su un intervallo di celle *A6:B6*, si usa una **formula** che include la **funzione** *INC.PERCENTILE*
- *=INC.PERCENTILE(A6:B6;0.75)-INC.PERCENTILE(A6:B6;0.25)*

Varianza – …un caposaldo della statistica

- La **varianza** è una **misura di variabilità** che sfrutta tutte le informazioni contenute nei **dati** (tutte le **osservazioni**)
- Essa raccoglie il contributo informativo che deriva dalla differenza (o **scarto**) tra il valore di ciascuna **osservazione** e la **media** della **distribuzione**
- Per definizione, la **misura** corrisponde alla **media** degli **scarti** elevati al quadrato, la quale si annulla quando tutte le **osservazioni** hanno lo stesso valore e diventa tanto più elevata quanto più le stesse sono diverse tra loro (distanti dalla **media**)
- $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ oppure calcolabile come $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- È espressa nell'**unità di misura** della **variabile** al quadrato

Varianza – Perché due formule???

- Nella circostanza in cui i **dati** da analizzare rappresentino **un'intera popolazione**, ovvero siano l'insieme completo di tutte le unità di interesse, si utilizza la **formula** in cui si divide per il numero totale di **osservazioni** (*n*)
- Se invece i **dati** da analizzare rappresentano un **campione**, ossia un sottoinsieme selezionato della **popolazione**, bisogna utilizzare la **formula** in cui si divide per *n-1*
- Il motivo di questa diversità lo capirete nel corso di Statistica. Per il momento sappiate che la prima **formula** si chiama **varianza della popolazione**, mentre la seconda si chiama **varianza campionaria**
- In Excel, dato un **intervallo di celle** in esame, la **varianza della popolazione** si calcola con la funzione *VAR.P(intervallo_celle)* mentre la varianza campionaria con *VAR.C(intervallo_celle)*

Deviazione standard

- L'interpretazione del valore numerico della **varianza** è espresso nell'**unità di misura** della **variabile** al quadrato
- Se stessimo analizzando la variazione degli stipendi in €, la **varianza** avrebbe come **unità di misura** $€^2$ → non molto significativo…
- Per ovviare a questo problema di interpretazione si usa la **deviazione standard** o **scarto quadratico medio**
- Essa è definita come la radice quadrata con segno positivo della **varianza** e con la stessa **unità di misura** dei **dati** osservati. In particolare, la deviazione standard indica di quante unità di misura i dati si discostano dalla media
- Come per la **varianza**, in Excel si calcola la **deviazione standard della popolazione** con la **funzione** *DEV.ST.P(intervallo_celle)* e quella **campionaria** con *DEV.ST.C(intervallo_celle)*

Coefficiente di variazione

- La **varianza** e la **deviazione standard** sono **misure di variabilità assoluta** perché si riferiscono all'**unità di misura** della **variabile** di interesse
- Qualora si fosse interessati a confrontare in termini di **variabilità** due o più **distribuzioni** espresse in **unità di misura** differenti o dotate di intensità **medie** differenti, serve una **misura di variabilità relativa**
- La principale utilizzata a questo scopo è il **coefficiente di variazione**
- Esso è definito come il rapporto fra la **deviazione standard** e la **media**. In questo modo si eliminano gli effetti delle **unità di misura** e si genera un numero puro idoneo al confronto
- Per calcolare il **coefficiente di variazione** in Excel su un intervallo di celle (ad esempio *A1:B6*) si usa la **formula** *=DEV.ST.P(A1:B6)/MEDIA(A1:B6)* oppure *=DEV.ST.C(A1:B6)/MEDIA(A1:B6)* a seconda che si usi la deviazione standard della popolazione o quella campionaria

Media e varianza per dati raggruppati – Che fare?

- Prendiamo i dati dell'esempio nella slide 17 relativi ad una variabile quantitativa continua. Per semplicità di analisi i dati erano stati raggruppati in classi
- Ma se i dati sono raggruppati posso usare le formule viste per media e varianza?
- Purtroppo no, è necessario modificare le formule come segue
- $\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$
- $s^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{n}$ oppure $s^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{n-1}$
- In queste formule *k* rappresenta il numero di **classi** usate per il raggruppamento dei **dati**, m_i il valore centrale della i-esima **classe**, e f_i la **frequenza assoluta** della i-esima **classe**

Usare media e varianza nei processi di decisione – Esempio

	A	B	C	D	E	F	G	H
1	**Titolo**	**Score in borsa del rendimento (1-100)**						
2	Apple	73	76	77	85	88	90	
3	Microsoft	74	74	78	84	88	91	
4	IBM	72	77	79	82	84	95	
5								
6								
7	**Media di Apple**	81,5						
8	**Media di Microsoft**	81,5						
9	**Media di IBM**	81,5						
10								
11	**Dev. Standard di Apple**	7,0639932						
12	**Dev. Standard di Apple**	7,2594766						
13	**Dev. Standard di Apple**	7,8166489						
14								

- Prendiamo l'esempio di tre titoli quotati in borsa con un rendimento da 1 a 100 e valutati in un certo arco temporale
- Se dovessimo scegliere su quale titolo investire prenderemmo quello a **valor medio** più alto (rendimento maggiore), ma in realtà in questo esempio sono tutti con la stessa **media**
- Usando la **deviazione standard** come ulteriore metrica scelgo quello con volatilità minore (**deviazione standard** più bassa) e quindi maggiore sicurezza di rendimento

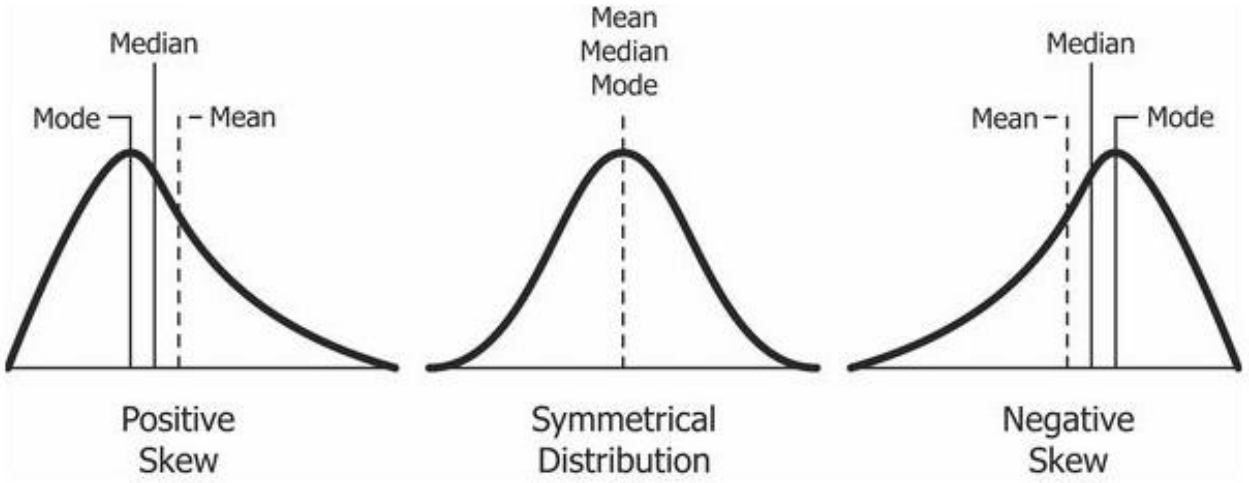
Misure di forma – Indice di Asimmetria

- L'**asimmetria** (o **skewness**) è una **misura di forma** di una **distribuzione** che può essere complementare all'uso **grafico** dell'**istogramma**
- In una **distribuzione** perfettamente **simmetrica** (**media** e **mediana** coincidono) l'**asimmetria** è pari a *0* (l'**istogramma** è perfettamente **simmetrico** e a campana)
- Se la **distribuzione** tende ad avere la **media** maggiore della **mediana** allora sarà **asimmetrica** a destra e l'indice di **asimmetria** sarà positivo
- Viceversa, l'indice diventa negativo per **asimmetria** a sinistra
- La **formula** per il suo calcolo è piuttosto complicata e quindi non sarà riportata qui
- In Excel l'asimmetria si calcola con la **funzione** *ASIMMETRIA(intervallo_celle)*

Misure di forma – Curtosi

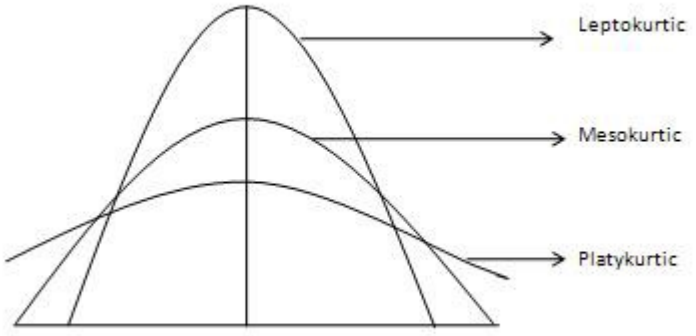
- L'indice di **curtosi** è una **misura** ulteriore **di forma** che consente di valutare il grado di appiattimento della **distribuzione** di una **variabile** attorno al suo **valore centrale**
- Se la **distribuzione** è «piatta» in gergo si dice che è **platicurtica**, ovvero l'insieme dei valori più piccoli e quello dei valori più grandi hanno **frequenza** non trascurabile. In questo caso la **curtosi** assume un valore negativo
- In caso contrario, quando una **distribuzione** è appuntita si parla di **distribuzione leptocurtica**, ed in questa situazione la **curtosi** è positiva
- Anche la **misura della curtosi** è particolarmente complicata e non oggetto di questo corso, ma se ne avete bisogno in Excel potete usare la **funzione** *CURTOSI(intervallo_celle)*

Misure di forma – Esempi (dal sito codeburst.io)

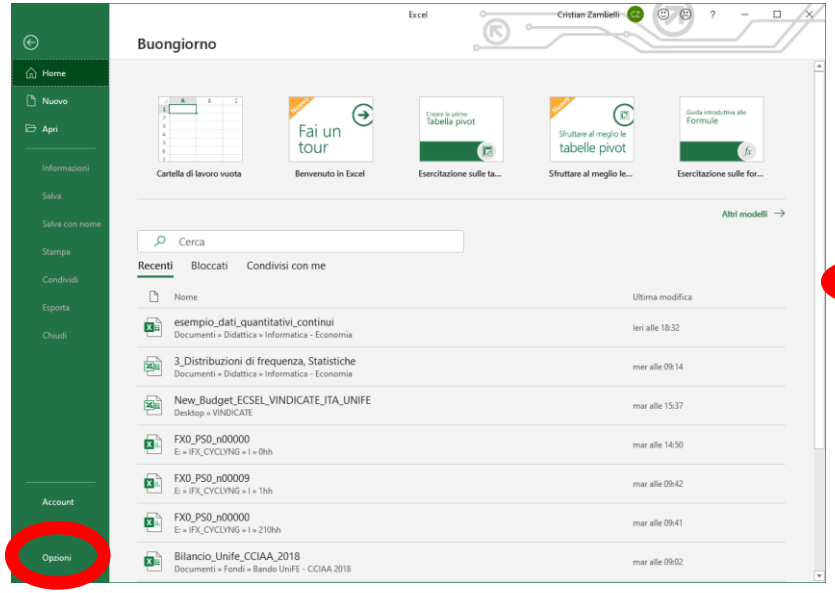


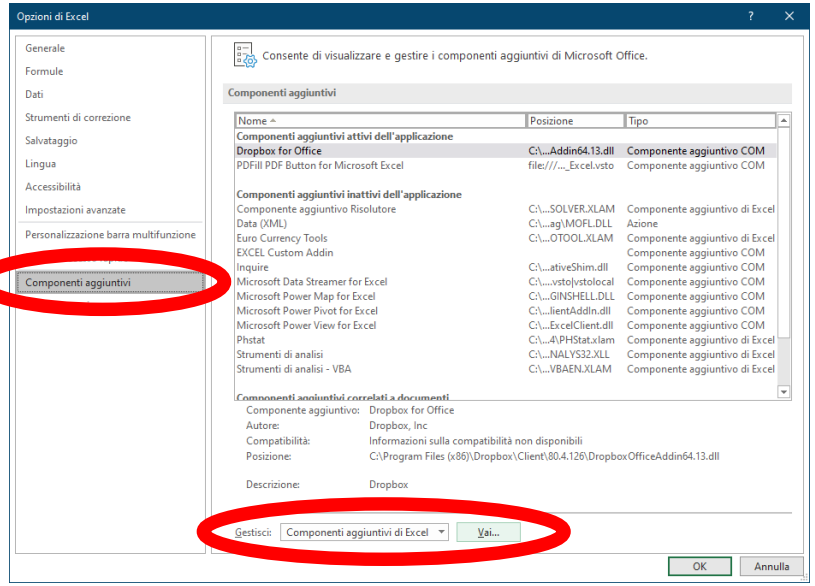
Esempi di indice di asimmetria

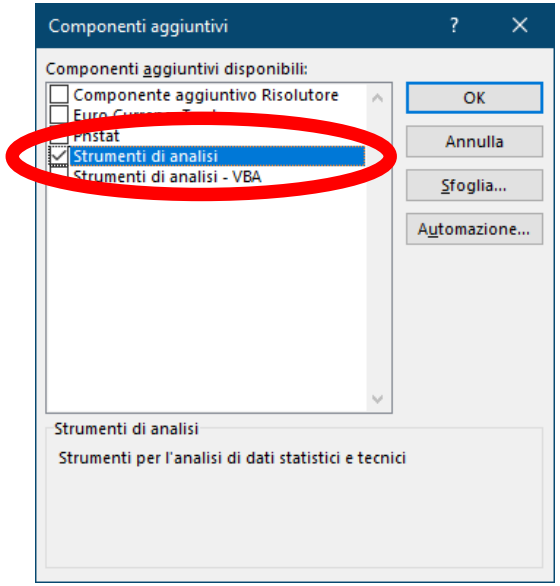
Esempi di indice di curtosi



Strumenti di analisi di Excel – Statistica descrittiva



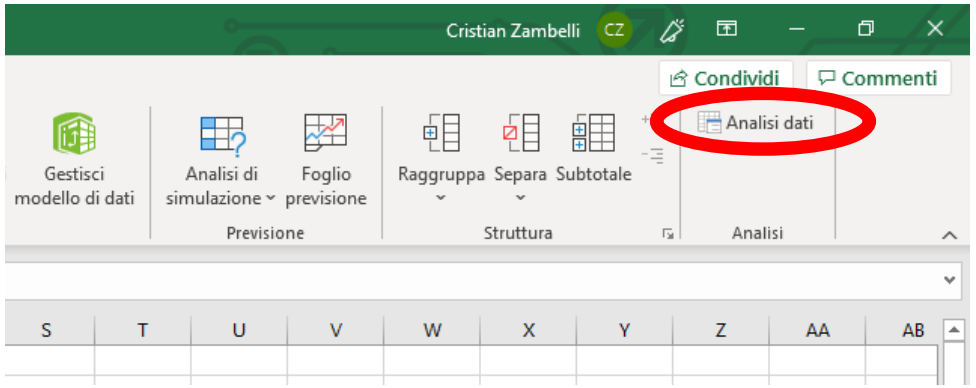


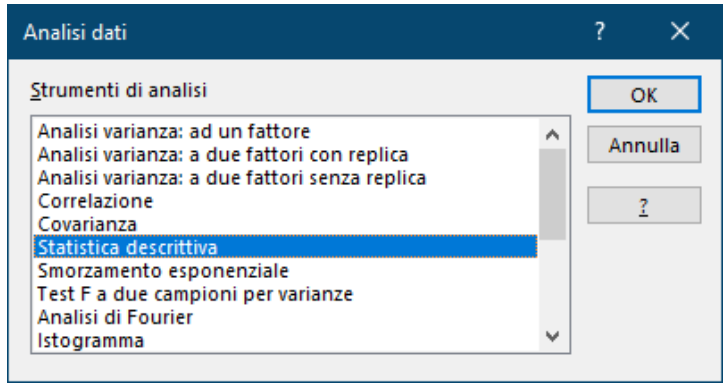


- Excel fornisce una serie di strumenti di analisi per la **statistica descrittiva** senza dover ricorrere manualmente a tutte le **formule** che abbiamo visto
- Qualora il gruppo *Analisi* non fosse presente nella scheda *Dati* della **barra multifunzione** è necessario installarlo
- Dalla scheda *File* selezionate *Opzioni*, *Componenti aggiuntivi* e cliccare su *Vai…*

Strumenti di analisi di Excel – Statistica descrittiva

	A	B	C	D
1	**Negozio**	**Fatturato**		
2	1	450000		
3	2	365000		
4	3	560000		
5	4	331000		
6	5	335500		
7	6	392500		
8	7	345000		
9	8	554000		
10	9	348000		
11	10	335550		
12	11	373560		
13	12	338900		
14	13	335500		
15	14	445000		
16	15	432150		
17				

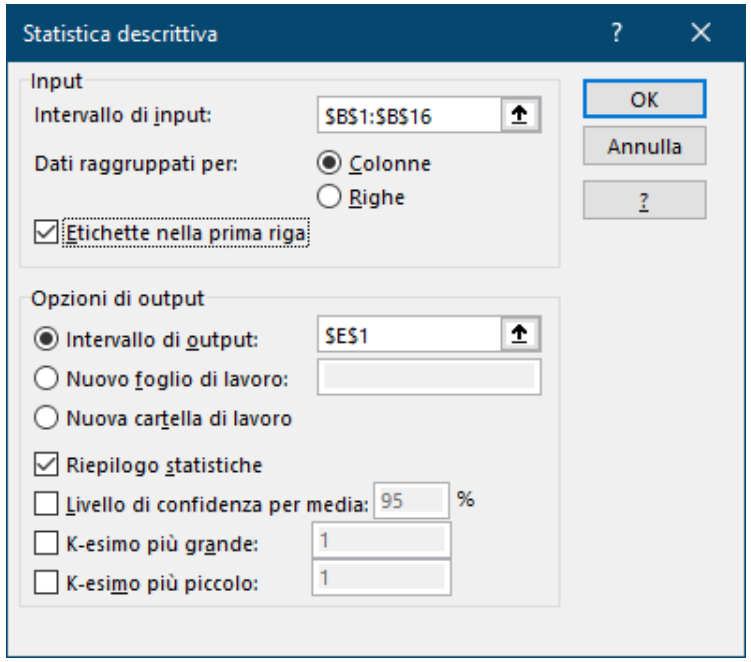




- Supponiamo di voler monitorare con un'analisi **statistica** l'andamento del fatturato in € di un anno di esercizio di 15 negozi appartenenti alla stessa catena
- A tal proposito attiviamo lo strumento *Statistica descrittiva* cliccando su *Analisi dati* nella scheda *Dati* della **barra multifunzione**
- Dalla finestra di dialogo scegliamo *Statistica descrittiva* cliccando poi su *OK*

Strumenti di analisi di Excel – Statistica descrittiva

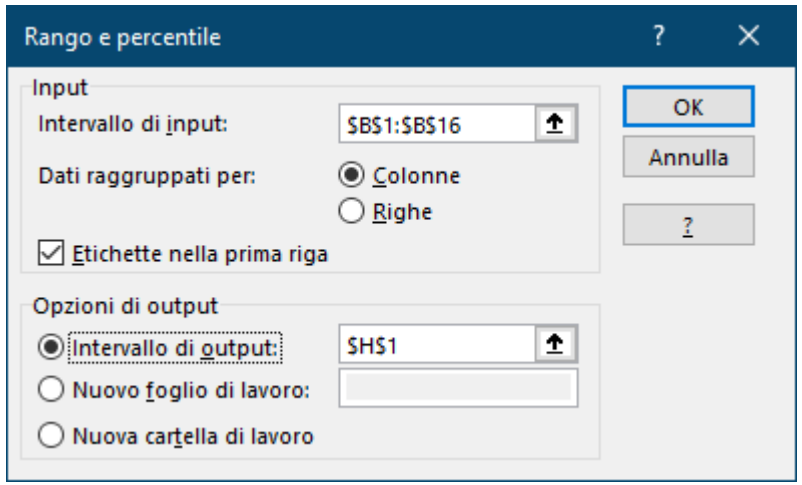
- Inseriamo come *Intervallo di Input* la **colonna** in cui si trovano i dati di fatturato e selezioniamo la casella *Etichette nella prima riga*



	A	B	C	D	E	F	G
1	**Negozio**	**Fatturato**			*Fatturato*		
2	1	450000					
3	2	365000			Media	396110,6667	
4	3	560000			Errore standard	19919,5471	
5	4	331000			Mediana	365000	
6	5	335500			Moda	335500	
7	6	392500			Deviazione standard	77148,07418	
8	7	345000			Varianza campionaria	5951825350	
9	8	554000			Curtosi	0,701665418	
10	9	348000			Asimmetria	1,299698067	
11	10	335550			Intervallo	229000	
12	11	373560			Minimo	331000	
13	12	338900			Massimo	560000	
14	13	335500			Somma	5941660	
15	14	445000			Conteggio	15	
16	15	432150					

- Come *Intervallo di output* ci scegliamo una **cella** dove posizionare il risultato dell'analisi **statistica** e selezioniamo *Riepilogo statistiche*
- Cliccando su *OK* apparirà una **tabella** riassuntiva con la maggior parte delle **misure di centralità, variabilità**, e **forma**, più altri **dati** utili per definire il comportamento della **popolazione di dati** che abbiamo chiesto di analizzare

Strumenti di analisi di Excel – Rango e percentile

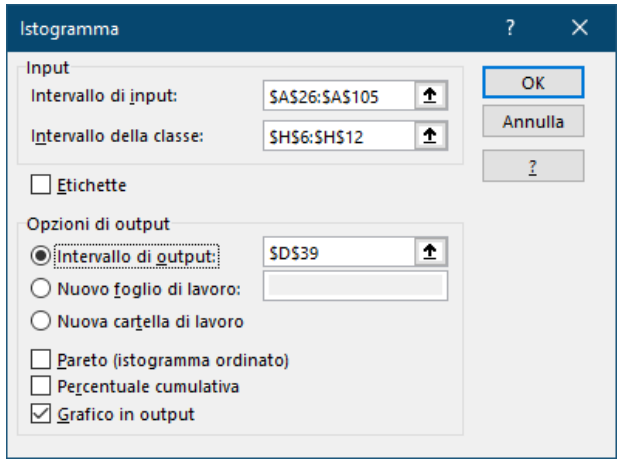


H	I	J	K
Dato	*Fatturato*	*Rango*	*Percento*
3	560000	1	100,00%
8	554000	2	92,80%
1	450000	3	85,70%
14	445000	4	78,50%
15	432150	5	71,40%
6	392500	6	64,20%
11	373560	7	57,10%
2	365000	8	50,00%
9	348000	9	42,80%
7	345000	10	35,70%
12	338900	11	28,50%
10	335550	12	21,40%
5	335500	13	7,10%
13	335500	13	7,10%
4	331000	15	0,00%

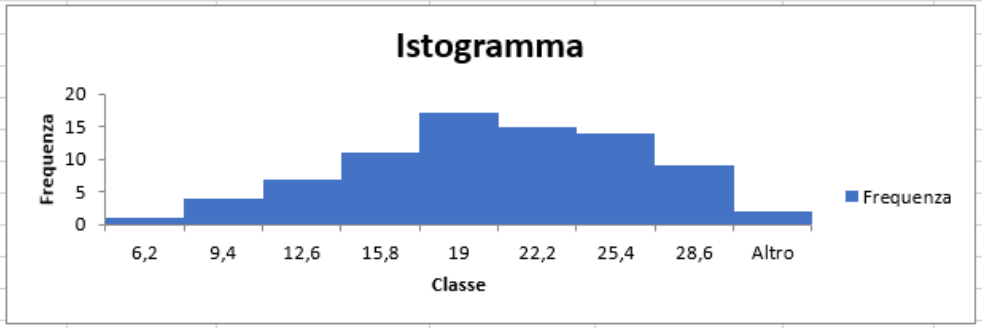
- Per esempio, da questi **dati** si può notare che l'**osservazione** *432150* corrispondente al fatturato del *15-esimo* negozio occupa il *71,40-esimo* **percentile** e che la **mediana** (il *50-esimo* **percentile**) è pari a *365000*

- Lo strumento di analisi *Rango e percentile*, applicato alle **osservazioni** della **variabile** di interesse, produce un prospetto che indica per ciascuna **osservazione**, la sua posizione occupata nella serie ordinata per **rango** e il **percentile** di riferimento
- L'inserimento e la scelta dei **dati** da computare avviene allo stesso modo di come descritto nella slide precedente

Strumenti di analisi di Excel – Istogramma



Classe	*Frequenza*
6,2	1
9,4	4
12,6	7
15,8	11
19	17
22,2	15
25,4	14
28,6	9
Altro	2



- Lo strumento di analisi *Istogramma* consente, dopo avere definito un **intervallo** di input dei **dati**, di creare un **istogramma** sia in forma **testuale** che **grafica**
- Se non viene specificato nessun **intervallo** delle **classi**, Excel calcola automaticamente i **valori centrali** delle **classi** rilevate, indicando con *Altro* tutti i valori maggiori dell'**estremo destro** dell'ultima **classe**
- Di default Excel mantiene la visualizzazione delle barre separate. Sta all'utente avvicinarle come già indicato in questo corso

Autovalutazione – Esempi ed esercizi

- Per esercitarvi con i concetti visti fino ad ora e come ausilio alla preparazione per l'esame, vi suggerisco di cimentarvi con gli esempi e gli esercizi che vi propongo nel seguente file:
  - **Analisi_statistica.xlsx** (qui imparerete ad usare le principali tecniche di analisi di dati qualitativi e quantitativi oltre ai più comuni descrittori usati nell'analisi univariata)

- Troverete tutto il materiale sul sito del corso di Informatica come indicato nelle slides di Introduzione al corso