Lezione 4
I valori centrali e di disuguaglianza

PROF. ROBERTO COSTA

SCIENZE DELL'EDUCAZIONE - STATISTICA SOCIALE (305SF)

POLO UNIVERSITARIO
FONDAZIONE
PORTOGRUARO CAMPUS

Dove eravamo rimasti

Nella scorsa lezione abbiamo visto come possiamo rappresentare una distribuzione in forma grafica.

I vari tipi di variabile richiedono modalità di rappresentazione grafica differenti.

Abbiamo anche parlato di come sintetizzare le osservazioni attraverso dei valori caratteristici (centrali e di disuguaglianza).

Parlando del calcolo dei valori centrali abbiamo che dobbiamo agire diversamente a seconda del tipo di variabili che abbiamo a disposizione.

Abbiamo visto i valori centrali analitici e di posizione.

Dove eravamo rimasti

Ritorniamo al calcolo della mediana per una distribuzione di frequenza di una variabile quantitativa in classi.

$$\mathbf{Me = l_m + (0{,}5 - F_{m-1})/(F_m - F_{m-1})\ \Delta_m}$$

Dove:

l_m è il limite inferiore della classe mediana

F_{m-1} la frequenza relativa cumulata fino alla classe precedente a quella mediana

F_m è la frequenza relativa cumulata fino alla classe mediana

Δ_m è l'ampiezza della classe mediana

Dove eravamo rimasti

$$\mathbf{Me = l_m + (0{,}5 - F_{m-1})/(F_m - F_{m-1})\ \Delta_m}$$

Significa che il valore puntuale della mediana non potrà essere inferiore a $\mathbf{l_m}$ (limite inferiore della classe mediana).

A questo assommiamo una quantità che è pari ad una certa porzione $(0{,}5 - F_{m-1})/(F_m - F_{m-1})$ dell'ampiezza della classe mediana Δ_m.

Riflettiamo sul fatto che $(0{,}5 - F_{m-1})$ è sicuramente minore di $(F_m - F_{m-1})$. Infatti sappiamo che è sicuramente maggiore di 0,5 dal momento che F_m è il limite superiore della modalità che contiene il valore mediano.

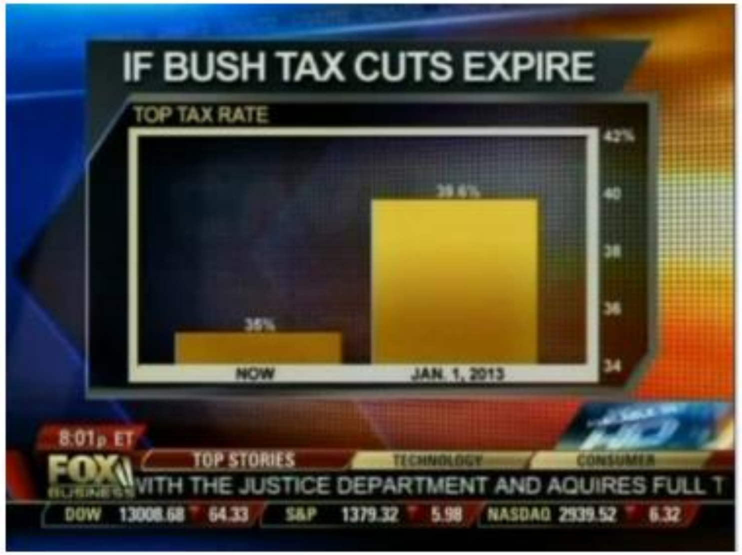
Quindi $0{,}5 < F_m$

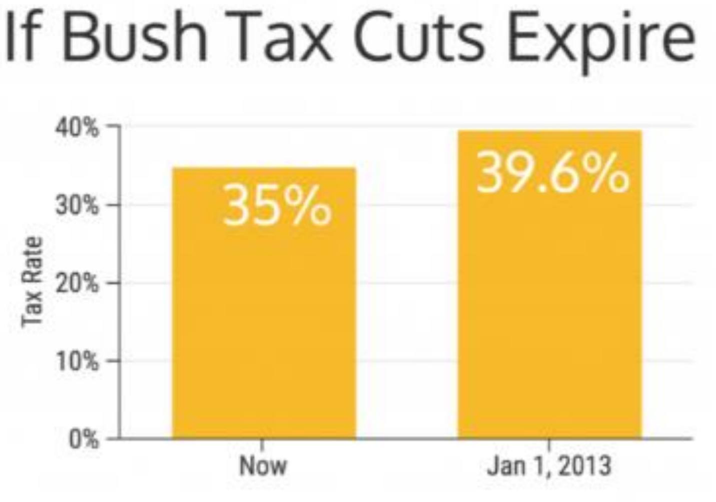
Il rapporto $(0{,}5 - F_{m-1})/(F_m - F_{m-1})$ è sicuramente > 0 e < 1.

Dove eravamo rimasti

Abbiamo anche detto che bisogna osservare con attenzione i grafici e non cadere in alcuni trucchetti.

Impariamo le regole base per costruire dei grafici che non distorcono i dati. Partiamo dal controllare gli assi X e Y!



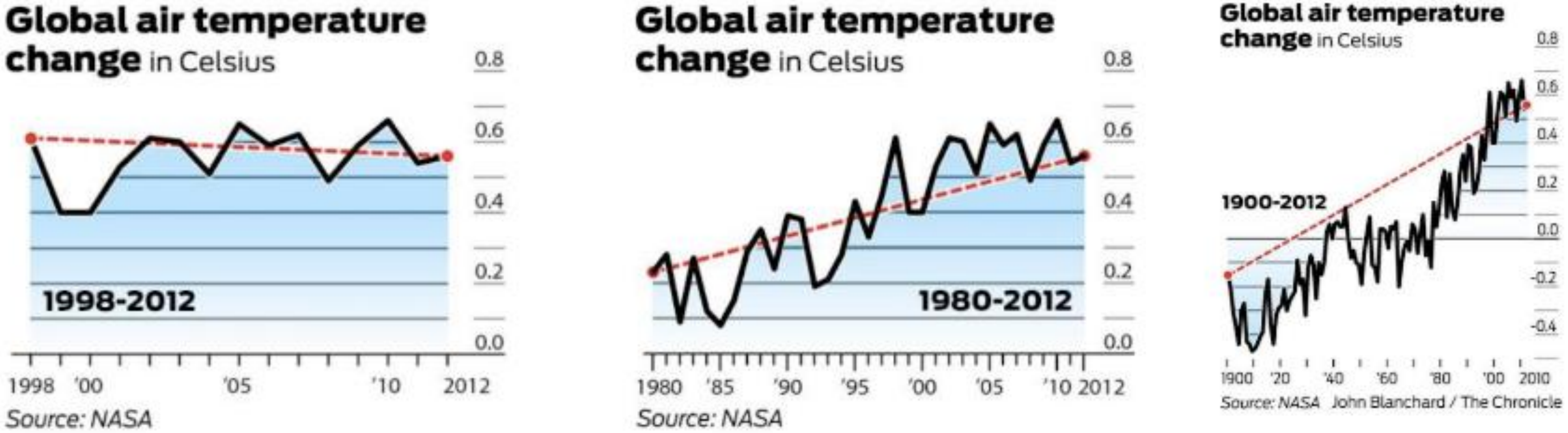


https://www.dataninja.it/2020/11/05/infografiche-che-mentono/

Dove eravamo rimasti

Abbiamo anche detto che bisogna osservare con attenzione i grafici e non cadere in alcuni trucchetti.

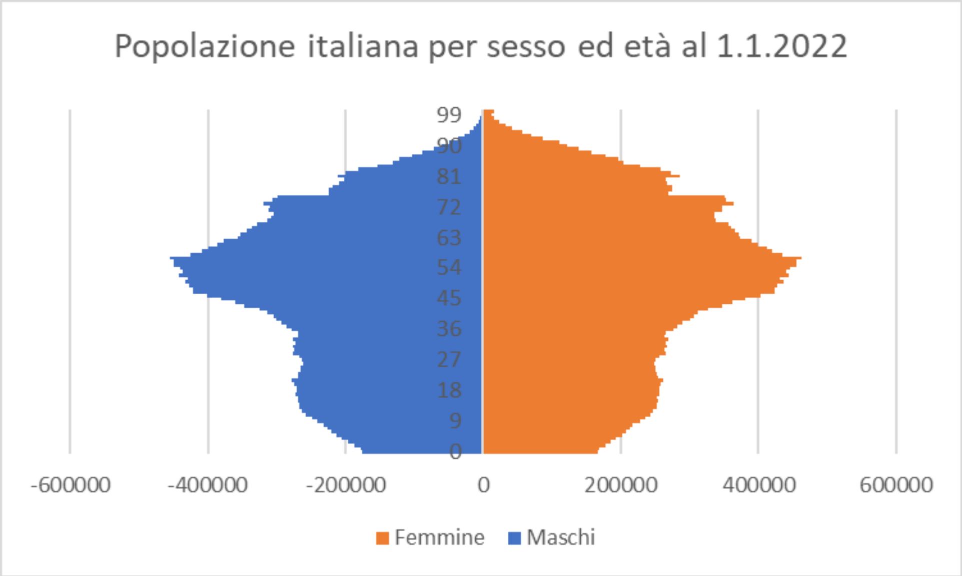
Facciamo attenzione alle serie temporali selezionate ad arte per dimostrare certe teorie! Verifichiamo sempre le fonti e le metodologie.

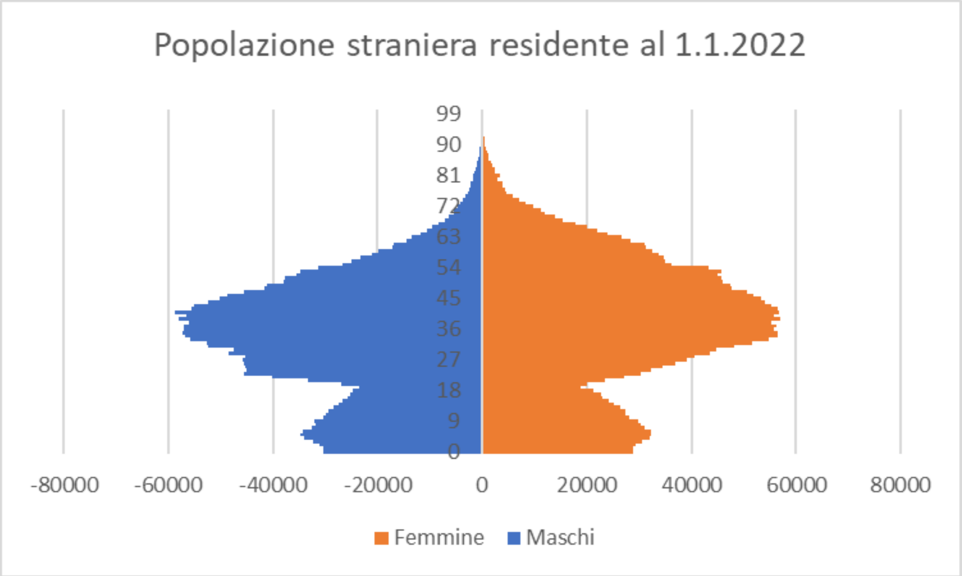


Fonte: "How climate deniers abuse statistics to mislead" di James Temple · SFGate

La piramide delle età

Vediamo come si differenziano le piramidi delle età di italiani e stranieri residenti in Italia al 1.1.2022





Facciamo chiarezza

Stati della proprietà	Procedura di operativizzazione	Definizione	Tipo di variabile	Nomenclatura adottata	Operazioni possibili
Discreti non ordinabili	Classificazione	Variabili qualitative (o categoriali, o mutabili)	Nominale	Sconessa	= o $\neq$
Discreti ordinabili	Ordinamento		Ordinale	Ordinale	= o $\neq$, > o <
Discreti enumerabili	Conteggio	Variabili quantitative (o cardinali)	Cardinale	Discreta	= o $\neq$, > o <, + - x :
Continui	Misurazione		Cardinale	Continua	= o $\neq$, > o <, + - x :

La media aritmetica

Come possiamo calcolare la **media aritmetica** quando ho una **distribuzione di frequenze**?

Possiamo fare il prodotto degli n valori moltiplicati per le frequenze e divisi per il numero totale dei casi:

$$M(x) = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

Dove k sono le classi.

Risolviamo l'esempio assieme:

Per prima cosa calcoliamo i valori centrali.

[(18 x 10) + (22 x 15) + (25 x 15) + (28 x 10)]/50

(180 + 330 + 375 + 280)/50 = 1.165/50 = 23,3

Voto	Frequenze assolute
18	10
22	15
25	15
28	10
Totale	50

La media aritmetica

Come possiamo calcolare la **media aritmetica** quando le **modalità** sono raccolte **in classi**?

Possiamo sostituire i valori reali con il valore centrale della classe.

Se abbiamo N osservazioni divise in k classi la formula sarà:

$$M(x) \approx \frac{\sum_{i=1}^{k} c_i n_i}{N}$$

Dove k sono le classi e c_i sono i valori centrali delle classi

Risolviamo l'esempio assieme:

Per prima cosa calcoliamo i valori centrali.

[(19,5 x 10) + (23 x 15) + (26 x 12) + (29 x 13)]/50

(195 + 345 + 312 + 377)/50 = 1299/50 = 24,6

Ripartizione	Frequenze assolute
18 – 21	10
22 – 24	15
25 – 27	12
28 - 30	13
Totale	50

La media aritmetica e le classi

Può capitare che in una distribuzione una classe, solitamente quella estrema, non sia limitata, ovvero sia aperta.

Pensiamo ad esempio alle classi di età: da 0 a 14 anni, da 15 a 34 anni, da 35 a 54 anni, da 55 a 74 anni, 75 anni e più.

In questo caso bisogna individuare un valore centrale adeguato della classe aperta. La valutazione è comunque soggettiva.

La media aritmetica e le classi

Numero occupati per ore settimanali lavorate (in migliaia) Anno 2020 (Fonte: Istat rilevazione sulle forze di lavoro).

Esempio 3.9 del libro

Valori centrali: 0 - 5,5 – 18 – 32,5 – 40 e ?

Possiamo assumere come limite superiore 50 ore e quindi

45,5 come valore centrale della classe aperta.

(0 x 1.694) + (5,5 X 277) + (18 x 1.171) + (40 x 5.189) + (45,5 x 2.370)

0 + 1523,5 + 21.078 + 207.560 + 107.835 = 337.999,5

337.999,5/13.280 = 25,5

Ore lavorate	Freq. Assoluta
Assenti	1.694
1-10 ore	277
11-25 ore	1.171
26-39 ore	2.478
40 ore	5.189
41 ore e più	2.370
TOTALE	13.280

La media aritmetica ponderata

In alcune situazioni bisogna dare diversa importanza ai casi che compongono una distribuzione, attribuendo a ciascuno di essi uno specifico peso che possa aumentarne o ridurne l'importanza.

In questo caso si parla di **media aritmetica ponderata**.

$$M(x) = \frac{\sum_{i=1}^{k} x_i p_i}{\sum_{i=1}^{k} x_i}$$

La media aritmetica ponderata

Siamo incaricati di stimare i visitatori di un parco giochi e per fare questo contiamo il loro numero durante due settimane.

Questi sono i risultati:

Clima	Giorni	Totale visitatori	Media visitatori
Soleggiato	7	8.750	1.250
Variabile	5	4.000	800
Pioggia	2	400	200
Totale	14	13.150	

La media aritmetica sarebbe 13.150/14 = 939 visitatori

La media aritmetica ponderata

Sappiamo però che nel periodo di apertura del parco giochi le giornate si dividono in questo modo: Soleggiato = 65%, Variabile 20%, Pioggia 15%.

Questi sono i risultati:

Clima	Media visitatori	Peso
Soleggiato	1.250	0,65
Variabile	800	0,20
Pioggia	200	0,15
Totale		1,00

La media aritmetica ponderata sarà: [(1.250 x 0,65) + (800 x 0,20) + (200 x 0,15)]/1

(812,5 + 160 + 30) =1.002,5

Proprietà della media aritmetica

- È sempre espressa nella stessa unità di misura con cui sono espressi i valori della variabile.
- È sempre compresa tra il valore minimo e il valore massimo della distribuzione dei valori della variabile.
- È il valore che, sostituito a tutti gli altri valori della distribuzione ne lascia inalterata la somma.
- La somma degli scarti dalla media aritmetica è sempre pari a 0.
- La somma degli scarti al quadrato dalla media è minore dalla somma degli scarti al quadrato da qualsiasi altro valore della distribuzione.

Trimmed mean

Abbiamo visto che la media aritmetica risente della presenza di valori anomali (troppo grandi o troppo piccoli).

I fenomeni analizzati dalla statistica sociale (redditi, consumi, ecc.) spesso presentano valori estremi troppo grandi o troppo piccoli.

Un modo per ridurre l’effetto di questi valori è calcolare la media su un insieme di valori, escludendo un porzione di quelli estremi.

Questa media si chiama ***trimmed mean***.

Trimmed mean

Riprendiamo un esempio della lezione scorsa.

Abbiamo la seguente distribuzione dei redditi:

1.670, 1.840, 2.005, 2.150, 2.280, 2.360, 2.510, 20.780

Dalla quale avevamo ricavato i seguenti valori centrali:

Mediana = 2.215 Media aritmetica = 4.449,4

Calcoliamo una *trimmed mean* del 75% dei valori, ovvero lasciamo fuori i due estremi.

1.840, 2.005, 2.150, 2.280, 2.360, 2.510

Mediana = 2.215 Media = 2.190

Criterio di Chisini

Oscar Chisini ha formalizzato una definizione generale di media ampiamente accettata, che riflette la relatività del concetto di media rispetto al particolare fenomeno in analisi.

Dato un campione $(x_1, x_2, \dots, x_n)$ di numerosità n e una funzione f, la media x rispetto a f è definita come quell'unico numero M, tale che sostituendolo a tutte le n unità, il valore della funzione rimane inalterato:

$f(x_1, x_2, \dots, x_n) = f(M, M, \dots, M)$

Le medie comunemente impiegate (aritmetica, geometrica, armonica, di potenza) sono casi particolari ottenibili tramite questa definizione, per una funzione f opportuna.

Media geometrica

La **media geometrica**, secondo il criterio di Chisini, si avvale del prodotto come operazione algebrica.

$$M_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Un esempio classico di utilizzo della media geometrica è il calcolo dell’interesse medio annuo di un investimento.

Media armonica

La **media armonica**, secondo il criterio di Chisini, si avvale della somma di quantità inverse come operazione algebrica.

$$M_{ar} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

Un esempio classico di utilizzo della media armonica è il calcolo di quantità in base al tempo come i consumi medi annui, la velocità media di un percorso a tappe, ecc..

Media quadratica

La **media quadratica**, secondo il criterio di Chisini, si avvale della somma dei quadrati come operazione algebrica. È la radice quadrata della somma dei valori osservati elevati al quadrato divisa per il numero di osservazioni.

$$M_q = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$$

I valori di disuguaglianza

I valori di tendenza centrale da soli non consentono di avere un’idea chiara di come si distribuiscono i valori in una distribuzione di frequenza.

È necessario calcolare gli **indici di disuguaglianza**.

I valori di disuguaglianza

Supponiamo di aver raccolto i voti a due appelli dell’esame della stessa materia.

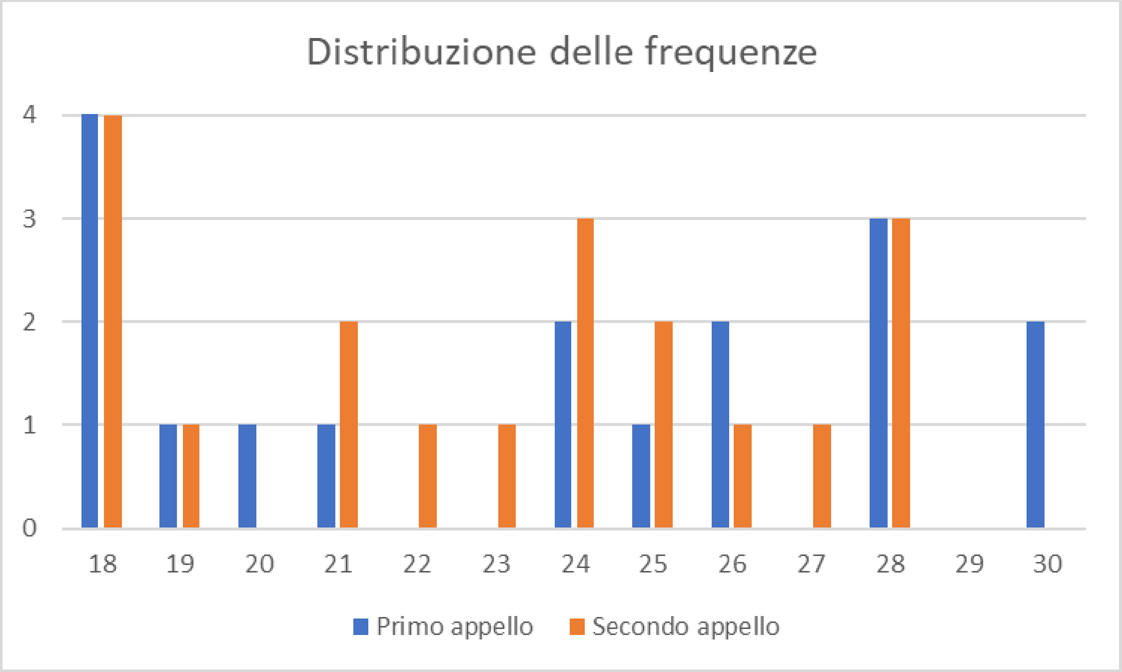
Abbiamo due differenti distribuzioni con:

Stessa media = 23

Stessa mediana= 24

Stessa moda = 18

Eppure le due distribuzioni sono diverse.



Range o escursione

Con il termine **range** o **escursione** si indica la differenza tra il valore osservato più piccolo e quello più grande.

$x_{max} – x_{min}$

E' una misura, seppur molto parziale, della dispersione di una variabile.

Range o escursione

Facciamo un semplice esempio, alla fine di un quadrimestre due studenti hanno entrambi la media del 6 in italiano. Sulla base di questo possiamo ritenere identico il rendimento dei due studenti?

Metto a confronto i voti dei due studenti:

Il primo ha preso: 5, 6, 7, 6, 6

Il secondo ha preso: 4, 9, 8, 4, 5

Calcolando il range dei voti posso fare qualche valutazione in più.

Nel primo caso il range è 7 – 5, ovvero 2

Nel secondo caso è 9 – 4, ovvero 5

Scarto medio assoluto dalla media o dalla mediana

Lo **scarto medio assoluto dalla media (o dalla mediana)** è la media degli scarti dei valori osservati rispetto alla media (o alla mediana) in valore assoluto. Si calcola per caratteri quantitativi.

$$\delta = \frac{\sum_{i=1}^{n} |x_i - m|}{n}$$

Torniamo all'esempio di prima:

$$\frac{|5-6| + |6-6| + |7-6| + |6-6| + |6-6|}{5} = \frac{2}{5}$$

$$\frac{|4-6| + |9-6| + |8-6| + |4-6| + |5-6|}{5} = \frac{10}{5}$$

Lo scarto quadratico medio

Lo scarto quadratico medio è la radice quadrata della media dei quadrati degli scarti dei valori osservati rispetto alla media:

$$\sigma = \frac{\sqrt{\Sigma(x_i - m)^2}}{N}$$

Ha pertanto la stessa unità di misura dei valori osservati.

La varianza

La varianza è lo scarto quadratico medio elevato al quadrato:

$$\sigma^2 = \Sigma(x_i - m)^2/N$$

La varianza è sempre positiva e assume valore pari a 0 solo se i valori della distribuzione sono tutti uguali e corrispondono ovviamente alla media.

Cos’è un foglio elettronico

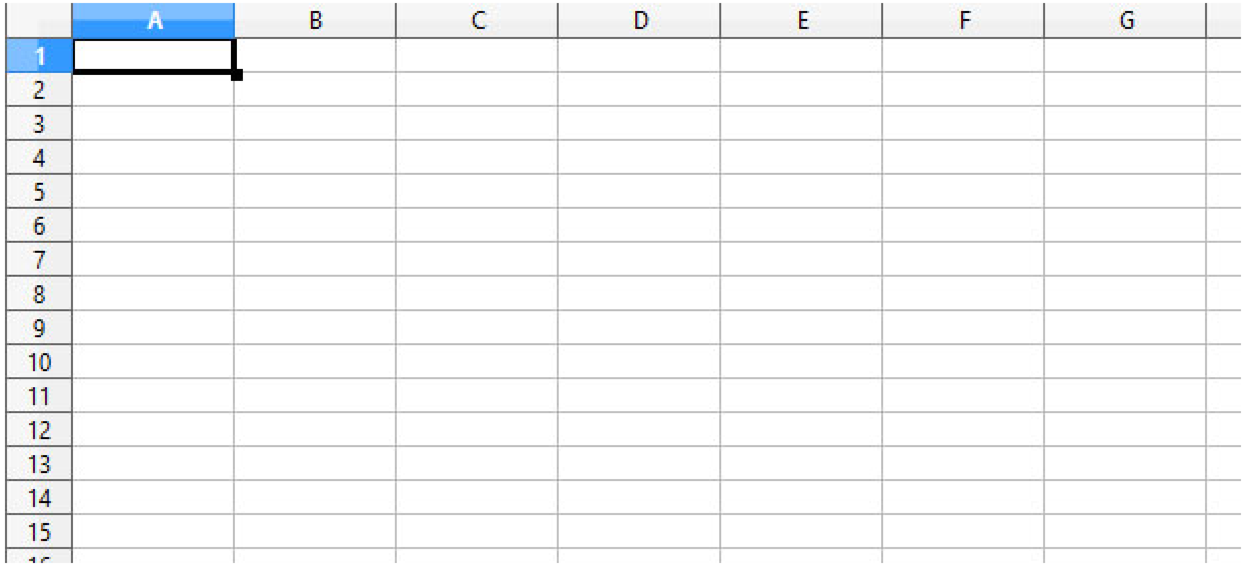
In ambito informatico si chiama foglio elettronico, o anche foglio di calcolo, un software che consente di operare in maniera efficiente su una più o meno grande mole di dati con calcoli, funzioni aritmetico-matematiche, macro e relativi grafici.

Il foglio di calcolo è costituito da fogli di lavoro, formati da celle in cui si possono inserire testi, dati, numeri o formule.

Il contenuto di una cella

Una cella può contenere un numero, del testo, oppure ci possiamo eseguire una formula e una funzione, utilizzando i valori (o le formule) di altre celle.

È uno strumento molto comodo per fare alcune operazioni statistiche elementari, costruire grafici, ecc..



Facciamo degli esercizi assieme

Esercizio 1

Questi sono i risultati degli esami di statistica sociale del giorno 26 ottobre

27

28

24

25

30

Esercizio 1

1.1 Che tipo di variabile ho davanti?

A) Qualitativa sconnessa

B) Qualitativa ordinata

C) Quantitativa discreta

D) Quantitativa continua

Possiamo calcolare dei valori centrali

1.2 Calcolate la media

1.3 Calcolate la mediana

Esercizio 1

1.1 Che tipo di variabile ho davanti?

A) Qualitativa sconnessa

B) Qualitativa ordinata

C) Quantitativa discreta

D) Quantitativa continua

Possiamo calcolare dei valori centrali

1.2 Calcolate la media **26,8**

1.3 Calcolate la mediana **27**

Esercizio 2

Ho la seguente distribuzione relativa al provincia di residenza di un gruppo di studenti.

Provincia	Studenti
Venezia	14
Pordenone	7
Trieste	5
Udine	6
Gorizia	2
Treviso	4
Totale	38

Esercizio 2

Ho la seguente distribuzione relativa al provincia di residenza di un gruppo di studenti.

2.1 Che tipo di variabile ho davanti?

A) Qualitativa sconnessa

B) Qualitativa ordinata

C) Quantitativa discreta

D) Quantitativa continua

Possiamo calcolare dei valori centrali?

2.2 Calcolate la media

2.3 Calcolate la mediana

2.4 Calcolate la moda

Provincia	Studenti
Venezia	14
Pordenone	7
Trieste	5
Udine	6
Gorizia	2
Treviso	4
Totale	38

Esercizio 2

Ho la seguente distribuzione relativa al provincia di residenza di un gruppo di studenti.

2.1 Che tipo di variabile ho davanti?

A) **Qualitativa sconnessa**

B) Qualitativa ordinata

C) Quantitativa discreta

D) Quantitativa continua

Possiamo calcolare dei valori centrali? **Solo la moda**

2.2 Calcolate la media **Impossibile**

2.3 Calcolate la mediana **Impossibile**

2.4 Calcolate la moda **Venezia**

Provincia	Studenti
Venezia	14
Pordenone	7
Trieste	5
Udine	6
Gorizia	2
Treviso	4
Totale	38