

Lezione 3

Rappresentazioni grafiche e i valori centrali

PROF. ROBERTO COSTA

SCIENZE DELL'EDUCAZIONE - STATISTICA SOCIALE (305SF)

Dove eravamo rimasti

Nella scorsa lezione abbiamo parlato di unità statistica, di popolazione e di variabili statistiche.

Abbiamo approfondito le differenze tra variabili qualitative e quantitative, sottolineando come la scelta a priori del tipo di variabili determina le operazioni che possiamo fare su di esse.

Infine abbiamo visto come posso sintetizzare i microdati in tabelle di frequenza.

Abbiamo visto anche le frequenze cumulate.

Dove eravamo rimasti

Facciamo un piccolo esercizio assieme:

Persone di 14 anni e + fumatori per sigarette fumate. Anno 2022, Valori %

| Sigarette fumate | % |
|------------------|------|
| Fino a 5 | 26,4 |
| Da 6 a 10 | 37,0 |
| Da 11 a 20 | 33,4 |
| Oltre 20 | 3,2 |

Fonte Istat, Indagine Multiscopo sulle famiglie: aspetti della vita quotidiana - parte generale.

Che tipo di distribuzione di frequenza è?

Assoluta

Relativa

Percentuale

Che tipo di variabile è

Quantitativa discreta

Quantitativa continua

Quante persone fumano fino a 10 sigarette?

Quante persone fumano più di 5 sigarette?

Dove eravamo rimasti - soluzioni

Facciamo un piccolo esercizio assieme:

Persone di 14 anni e + fumatori per sigarette fumate. Anno 2022, Valori %

| Sigarette fumate | % |
|------------------|------|
| Fino a 5 | 26,4 |
| Da 6 a 10 | 37,0 |
| Da 11 a 20 | 33,4 |
| Oltre 20 | 3,2 |

Fonte Istat, Indagine Multiscopo sulle famiglie: aspetti della vita quotidiana - parte generale.

Che tipo di distribuzione di frequenza è?

Assoluta

Relativa

Percentuale

Che tipo di variabile è

Quantitativa discreta

Quantitativa continua

Quante persone fumano fino a 10 sigarette?

63,4%

Quante persone fumano più di 5 sigarette?

73,6%

La rappresentazione grafica

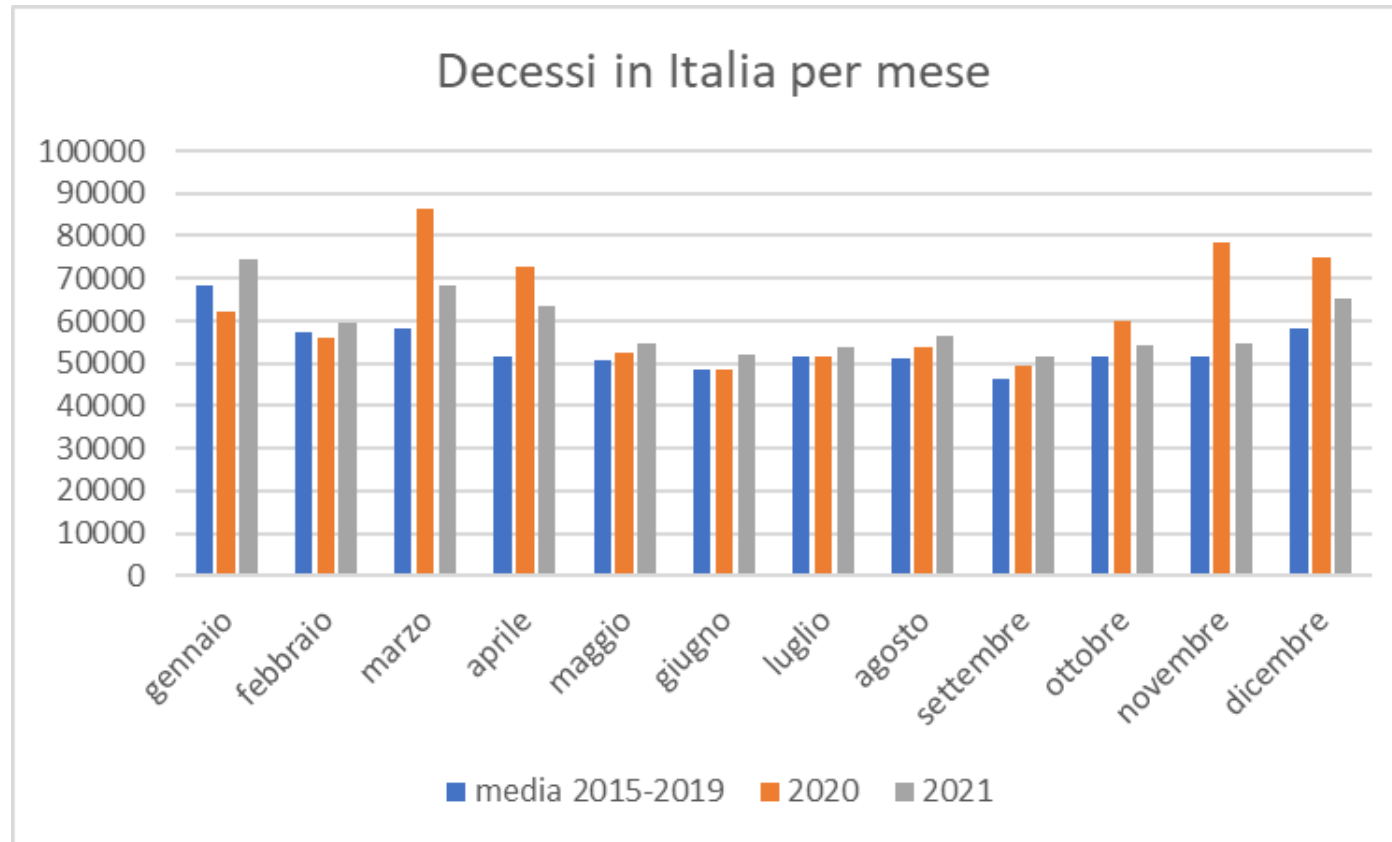
Partiamo da una tabella riferita ai decessi in diversi periodi:

| Decessi per anno e mese | | | | | | | | | | | | | |
|-------------------------|---------|----------|--------|--------|--------|--------|--------|--------|-----------|---------|----------|----------|---------|
| | gennaio | febbraio | marzo | aprile | maggio | giugno | luglio | agosto | settembre | ottobre | novembre | dicembre | TOTALE |
| media 2015-2019 | 68.324 | 57.416 | 58.267 | 51.801 | 50.724 | 48.501 | 51.811 | 51.041 | 46.548 | 51.590 | 51.462 | 58.133 | 645.620 |
| 2020 | 62.019 | 56.070 | 86.501 | 72.809 | 52.440 | 48.589 | 51.422 | 53.744 | 49.326 | 59.861 | 78.470 | 74.895 | 746.146 |
| 2021 | 74.550 | 59.389 | 68.507 | 63.434 | 54.802 | 52.201 | 53.668 | 56.594 | 51.456 | 54.463 | 54.870 | 65.101 | 709.035 |
| | | | | | | | | | | | | | |

Fonte: Istat, Tavola decessi totali regionali mensili per la media degli anni 2015-2019, per gli anni 2020-2021 e per i mesi di gennaio-agosto 2022

La rappresentazione grafica

Proviamo a tradurli in forma grafica. Lo trovate più esplicativo?



Rappresentazioni grafiche

Un grafico consente di cogliere il contenuto di una distribuzione con più facilità, rispetto ad una tabella.

Il grafico traduce le frequenze delle modalità in forme geometriche elementari.

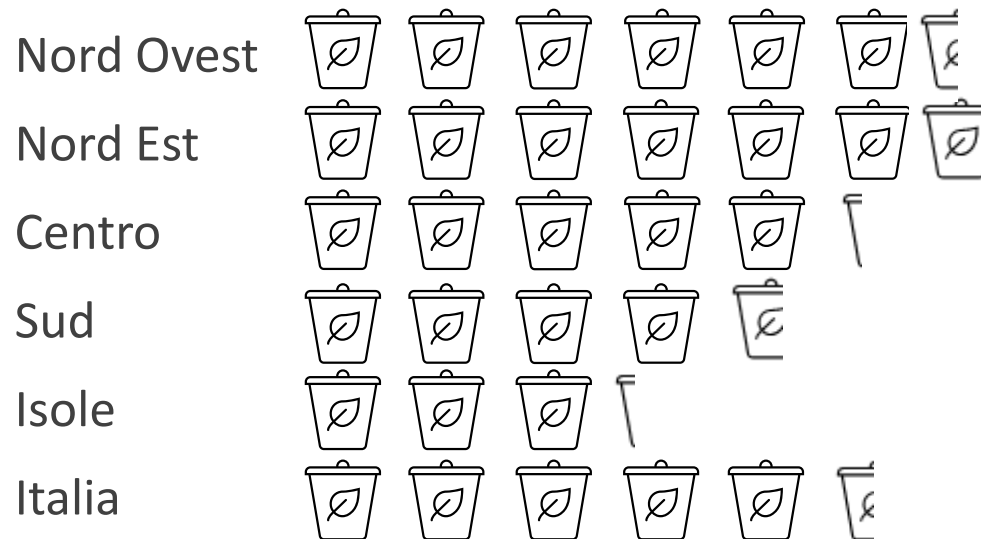
Come per le tecniche di analisi, anche i grafici variano e si differenziano a seconda del tipo di variabile che si vuole rappresentare.

Pittogrammi

Il **pittogramma** è una rappresentazione grafica che viene utilizzata di solito a fini divulgativi ed è generalmente rivolta a un pubblico di non esperto.

Consiste in figure o simboli che rappresentano la proprietà misurata e che vengono ripetuti, o hanno dimensioni proporzionali alle frequenze.

Rifiuti urbani e raccolta differenziata in Italia (2017)



| Ripartizione | Raccolta differenziata % |
|--------------|--------------------------|
| Nord Ovest | 64,5 |
| Nord Est | 68,3 |
| Centro | 51,8 |
| Sud | 47,0 |
| Isole | 31,6 |
| Italia | 55,5 |

Ortogrammi

Si utilizzano per rappresentare variabili qualitative con categorie non ordinate (mutabili sconnesse).

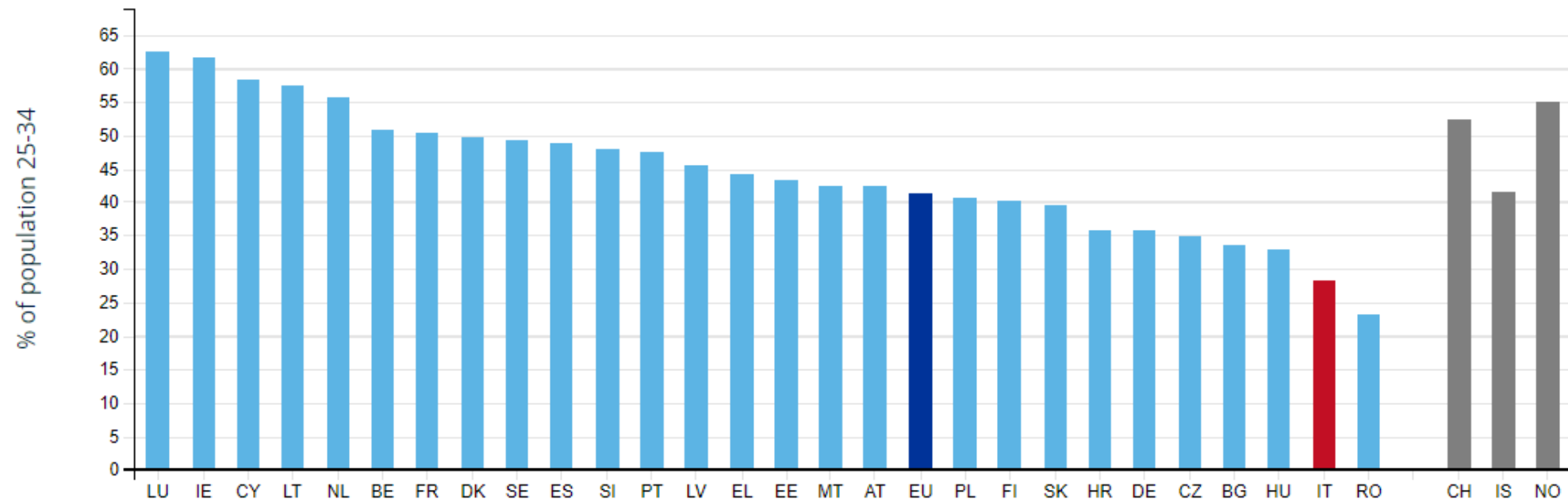
Su un asse ci sono le modalità della variabile e sull'altra le corrispondenti frequenze.

Le modalità vengono rappresentate da linee o parallelepipedi aventi base uguale ed equidistanti tra loro.

Questi possono essere disposti orizzontalmente (**diagramma a nastri**) o verticalmente (**diagramma a colonne**), lunghezza o altezza sono proporzionali alle frequenze delle modalità.

Ortogrammi

People with tertiary educational attainment, 2021
(as % of the population aged 25 to 34)

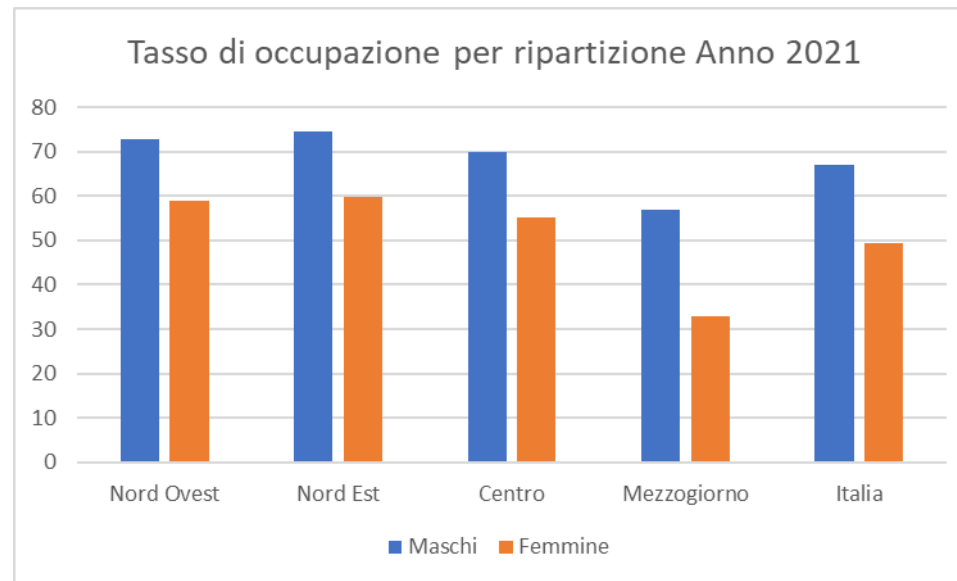


Fonte: Eurostat <https://ec.europa.eu/eurostat/cache/digpub/sdgs/index.html>

Ortogrammi

Si utilizzano gli ortogrammi per il confronto della distribuzione della frequenza di una stessa variabile in sottogruppi.

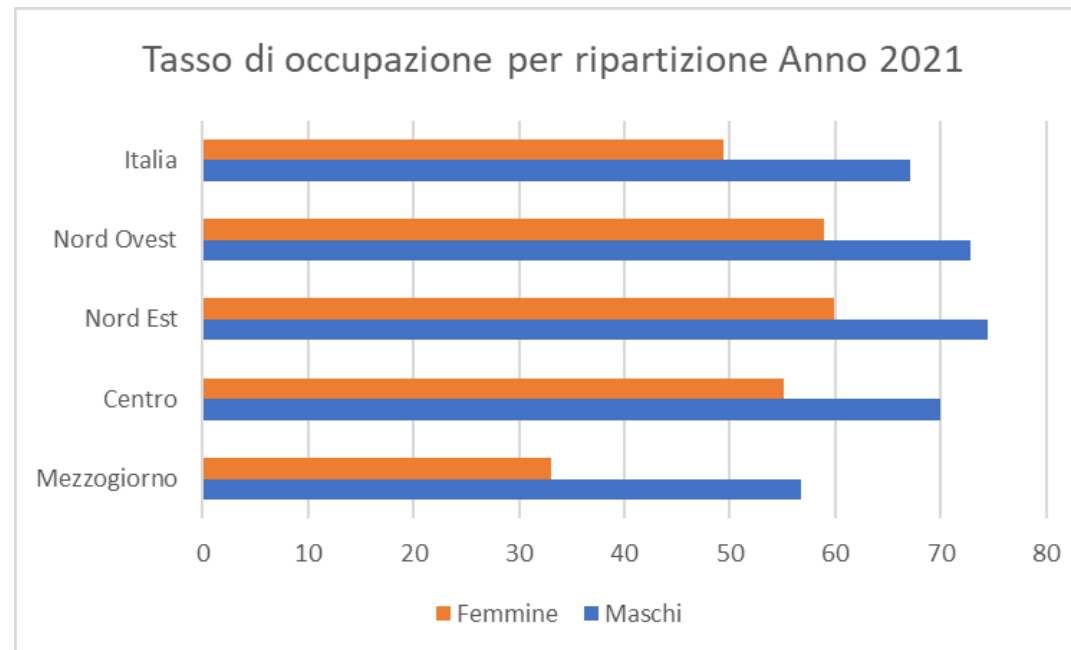
In questo caso si parla di diagramma a nastri o colonne contrapposti (o appaiati).



Fonte: Istat, Rilevazione sulle forze di lavoro

Ortogrammi

Possiamo utilizzare il diagramma a nastri appaiati.

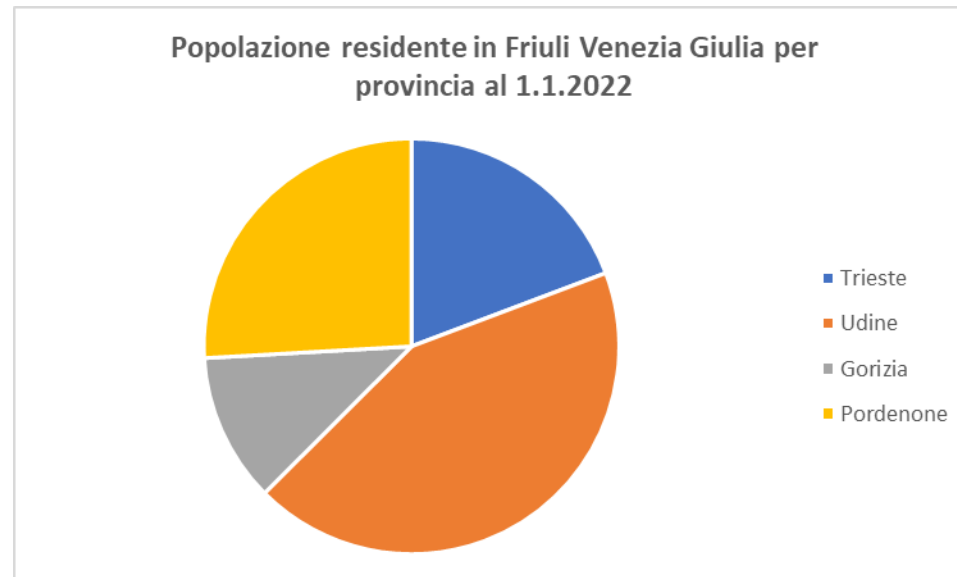


Fonte: Istat, Rilevazione sulle forze di lavoro

Aerogrammi

L'aerogramma è un grafico nel quale la distribuzione di frequenza viene rappresentata suddividendo l'area di una figura piana (solitamente un cerchio) in parti proporzionali.

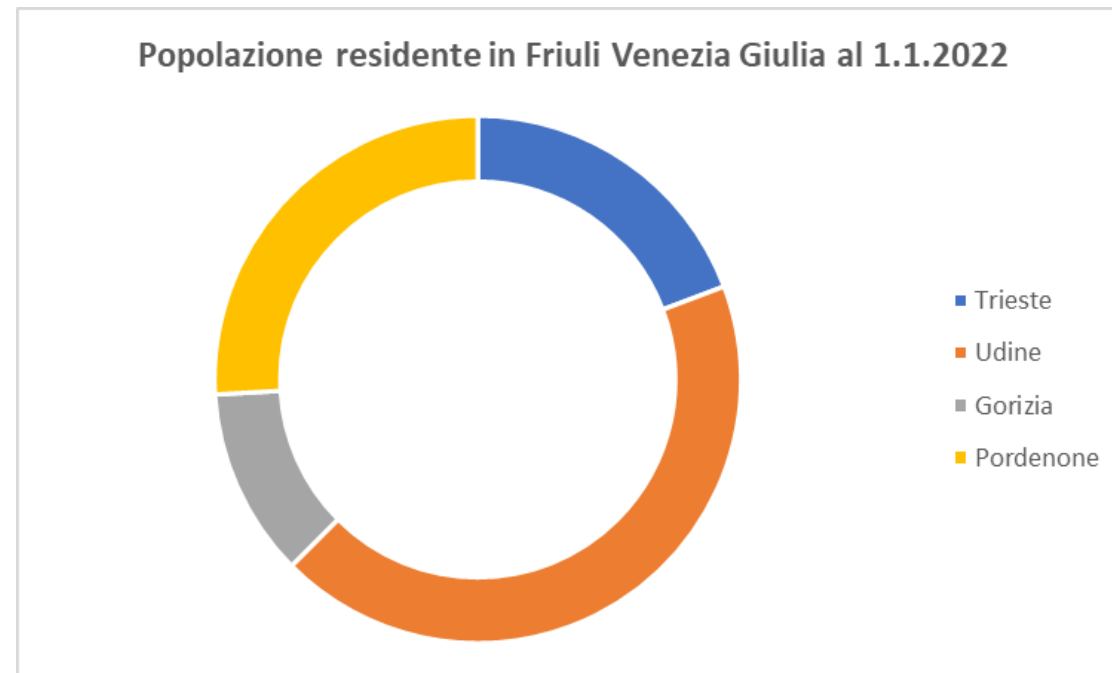
L'aerogramma più utilizzato è il **diagramma a torta** (pie chart).



Fonte: Istat, Popolazione residente comunale per sesso anno di nascita e stato civile

Aerogrammi

Un altro areogramma che viene utilizzato è il **diagramma ad anello**.

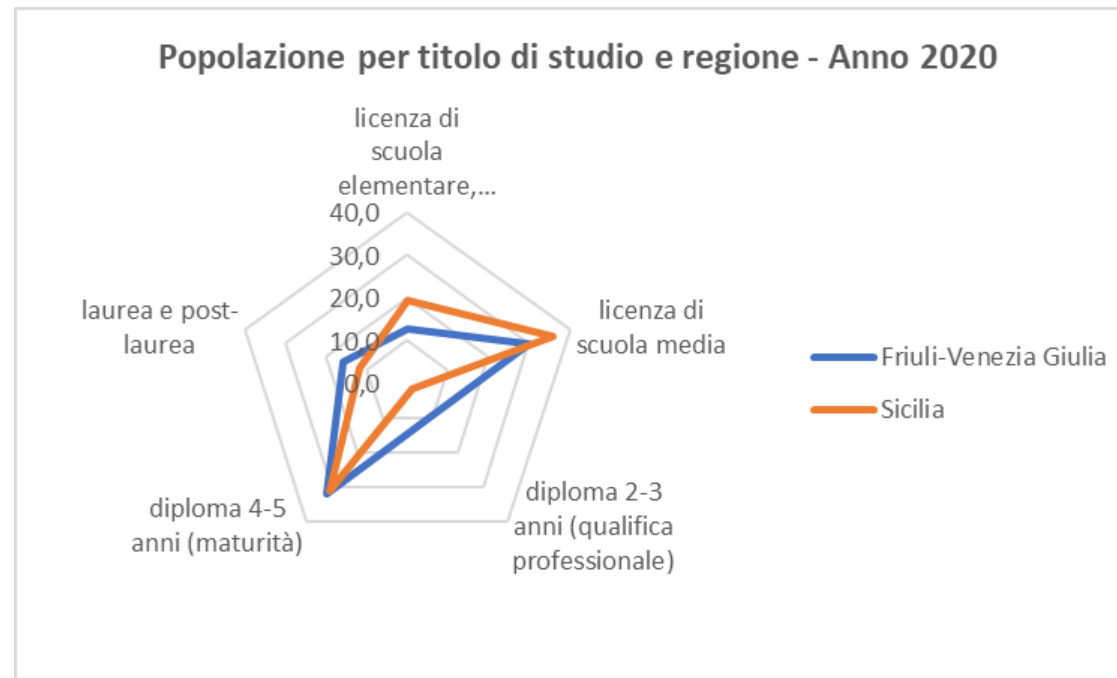


Fonte: Istat, Popolazione residente comunale per sesso anno di nascita e stato civile

Aerogrammi

Un altro areogramma che si utilizza con le variabili qualitative è il **diagramma a radar**.

E' un poligono con tanti vertici quante sono le modalità della variabile.



Fonte: Istat, Rilevazione sulle forze di lavoro

Aerogrammi

Un altro areogramma che si utilizza con le variabili qualitative ordinate è il **diagramma a barre suddivise**.



Fonte: Istat, Rilevazione sulle forze di lavoro

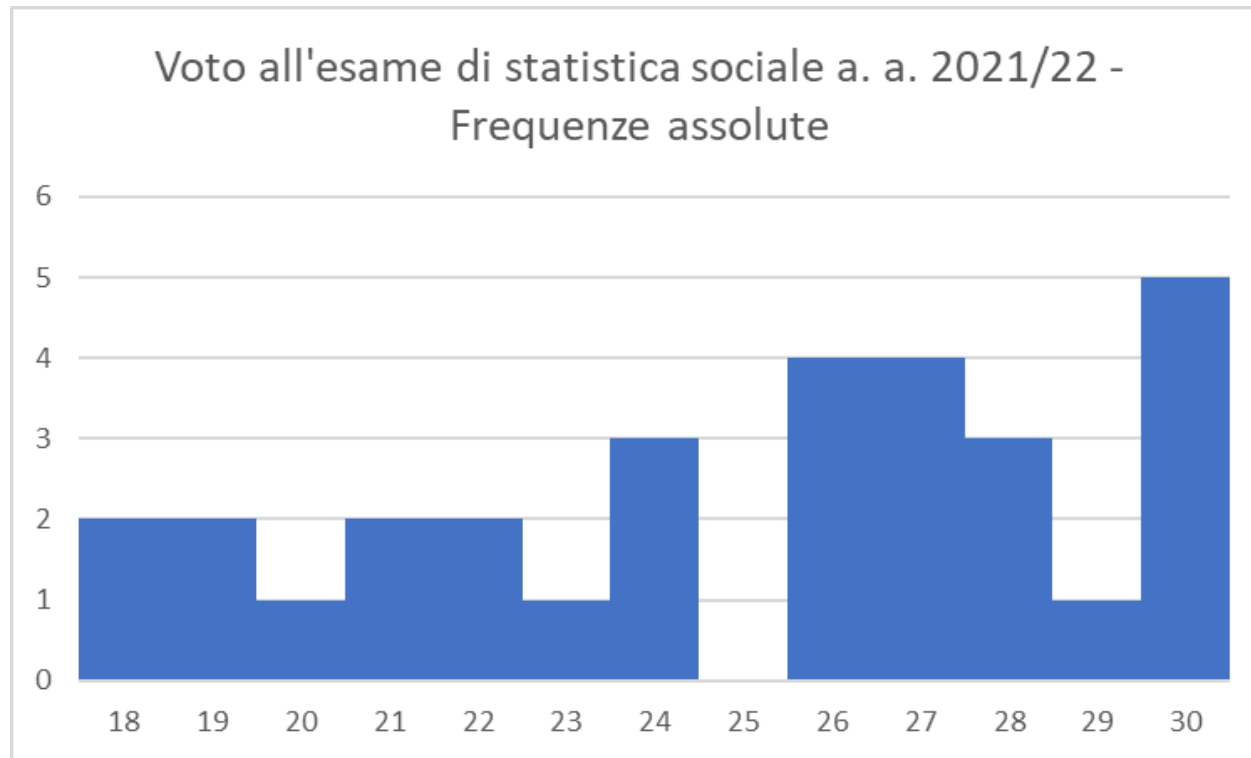
Gli istogrammi

Gli **istogrammi** vengono utilizzati per rappresentare la distribuzione di variabili quantitative discrete o continue raggruppate in classi.

Nel caso di variabili quantitative discrete, sull'asse delle ascisse troviamo le modalità della variabile ordinate in modo crescente e sull'asse delle ordinate le frequenze (assolute o percentuali).

Gli istogrammi

E' molto simile a un diagramma a colonne, con la differenza che tra le categorie non vengono lasciati spazi.



Gli istogrammi

Nel caso di variabili quantitative discrete raggruppate in classi o di variabili continue, dobbiamo tenere in considerazione l'ampiezza delle classi.

Se le classi sono di pari ampiezza si può utilizzare l'istogramma a basi uguali

Se le classi hanno ampiezze diverse non possiamo rappresentare il fenomeno con colonne di pari larghezza, ma dobbiamo calcolare la densità di frequenza (d).

La densità di frequenza si calcola dividendo la frequenza della classe e la relativa ampiezza della classe

$$D_i = \frac{\text{Frequenza assoluta}}{\text{ampiezza } i\text{-esima classe}}$$

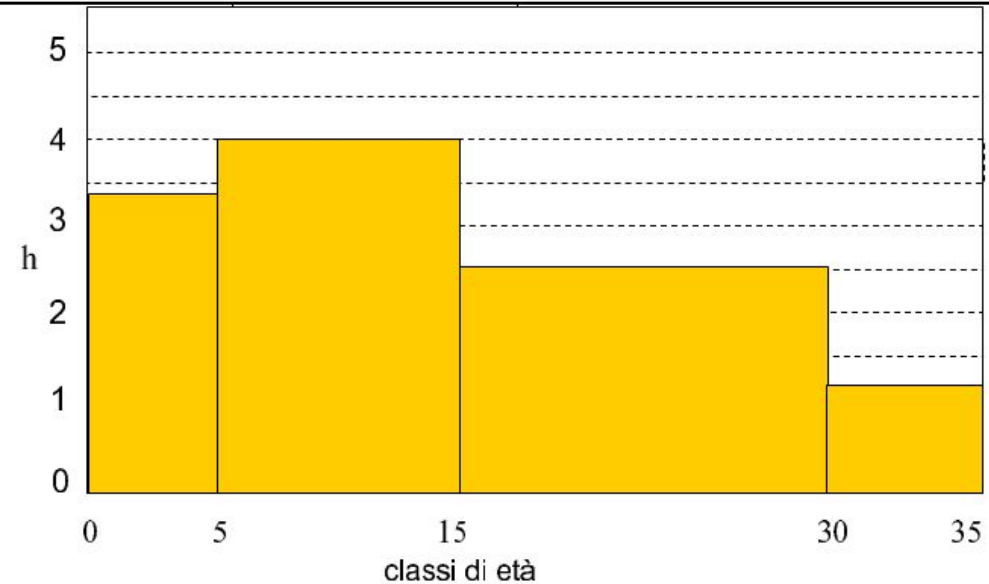
Gli istogrammi

Esempio di istogramma a basi diverse.

Dovremo riportare la densità di frequenza sull'asse delle ordinate.

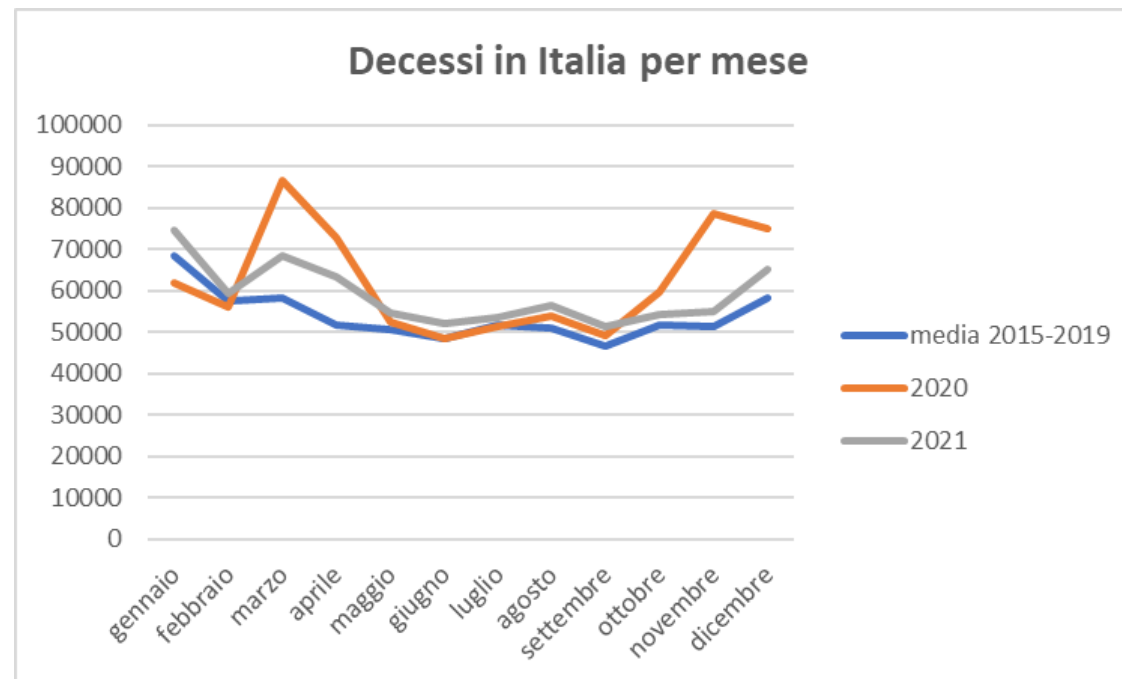
La frequenza di ciascuna classe corrisponderà all'area del rettangolo

| classi di età | amp. classe a_j | freq. % p_j | densità h_j |
|---------------|-------------------|---------------|---------------|
| 0-5 | 5 | 17,0 | 3,4 |
| 5-15 | 10 | 40,0 | 4,0 |
| 15-30 | 15 | 37,0 | 2,5 |
| 30-35 | 5 | 6,0 | 1,2 |



Le spezzate

Se immaginiamo di congiungere i punti medi dei lati superiori dei rettangoli che compongono un istogramma otteniamo una linea, che prende il nome di **spezzata**, che in alcuni casi rappresenta meglio l'andamento di una distribuzione e che consente di raffigurare nello stesso grafico più proprietà (o la stessa proprietà di più popolazioni) per metterle a confronto.



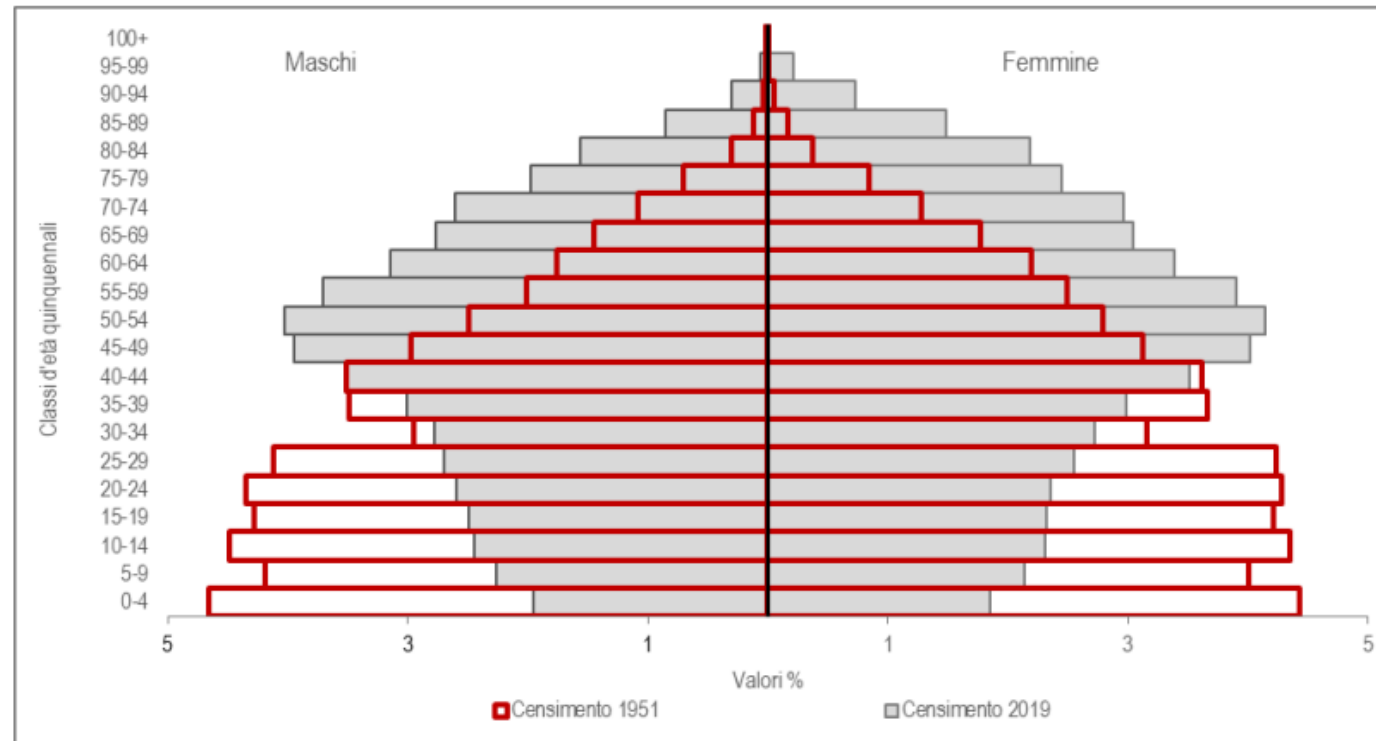
La piramide delle età

Una rappresentazione particolare è la cosiddetta piramide delle età. Si contrappongono due istogrammi che rappresentano la distribuzione della popolazione per età e genere. Otteniamo la struttura per età e sesso di una popolazione, che ci consente di osservare come si modifica una popolazione a fronte di eventi bellici, migrazioni e nascite.

La piramide delle età

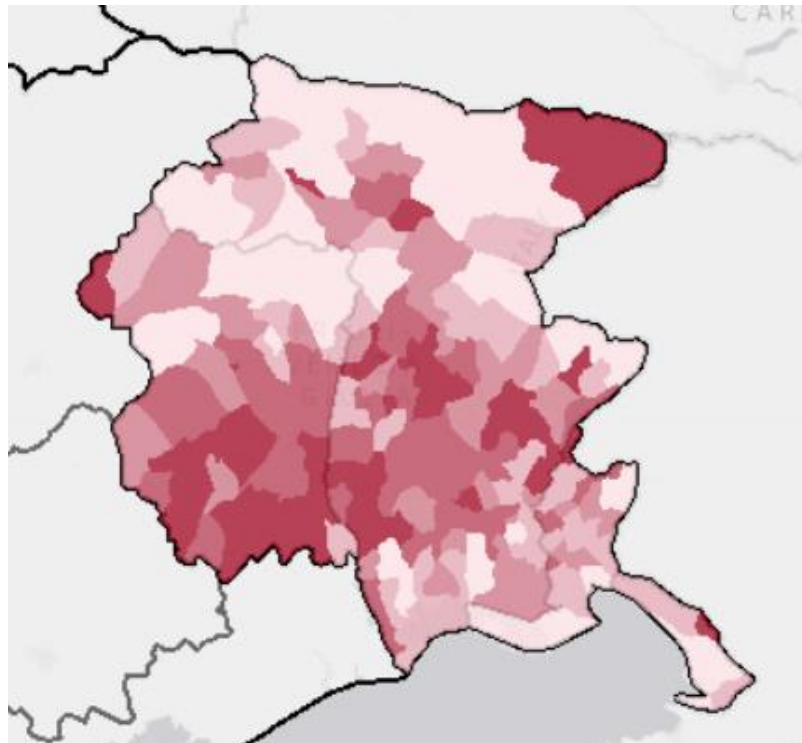
E' molto interessante vedere com'è cambiata la popolazione in Italia dal 1951 al 2019.

FIGURA 3. PIRAMIDI DELLE ETÀ E SESSO DELLA POPOLAZIONE RESIDENTE AI CENSIMENTI. Anni 1951 e 2019

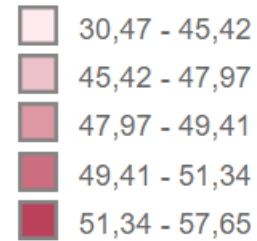


I cartogrammi

La rappresentazione grafica viene fatta sulle carte geografiche. I territori assumono colori differenti a seconda del valore che assumono le proprietà in quella partizione.



Censimento 2011: Tasso di occupazione (%) per Comuni nella regione FRIULI VENEZIA GIULIA



Sintetizzare le distribuzioni di frequenze

Fino ad ora abbiamo visto come possiamo sintetizzare una distribuzione di frequenze in una tabella o un grafico.

Nell'esempio dei voti dell'esame di statistica sociale abbiamo visto come questo da solo non ci consentiva di fare dei confronti.

Più in generale ci capiterà di ricorrere a valori essenziali della distribuzione fare confronti ad esempio nel tempo, o nello spazio.

Questi valori che definiamo **valori caratteristici** si distinguono tra:

Valori centrali, o di tendenza centrale, per riassumere la distribuzione con un unico valore caratteristico.

Valori di disuguaglianza, per evidenziare le differenze tra le distribuzioni delle modalità.

I valori centrali

Più di 200 anni fa il matematico francese Augustin Louis Cauchy diede una prima spiegazione statistica dei valori centrali, definendo valore centrale quel valore non inferiore al valore minimo e non superiore al valore massimo.

Noi usiamo continuamente i valori centrali: facciamo qualche esempio?

In generale si distinguono:

Valori centrali analitici: che vengono calcolati sui valori di una variabile quantitativa attraverso operazioni algebriche.

Valori centrali non analitici: detti anche indici di posizione, che operano sulle frequenze della distribuzione.

La moda

Si definisce moda la modalità che assume la frequenza, assoluta o relativa, maggiore.

Questo valore centrale può essere calcolato su qualsiasi tipo di variabile.

Popolazione italiana di 15 anni e più per titolo di studio - Anno 2020 valori %

| Titolo di studio | Freq. % |
|--|----------------|
| Licenza elementare o nessun titolo | 15,9 |
| Licenza di scuola media inferiore | 32,2 |
| Diploma di scuola superiore (qualifica o maturità) | 36,6 |
| Laurea e post laurea | 15,3 |

Fonte: Istat, Rilevazione sulle forze di lavoro

La moda

In questo caso troviamo la moda in una mutabile sconnessa.

Popolazione residente in Italia al 1° gennaio 2022 - Valori assoluti

| Ripartizione | Totale |
|--------------|-------------------|
| Nord-ovest | 15.848.100 |
| Nord-est | 11.561.676 |
| Centro | 11.740.836 |
| Sud | 13.451.861 |
| Isole | 6.380.649 |

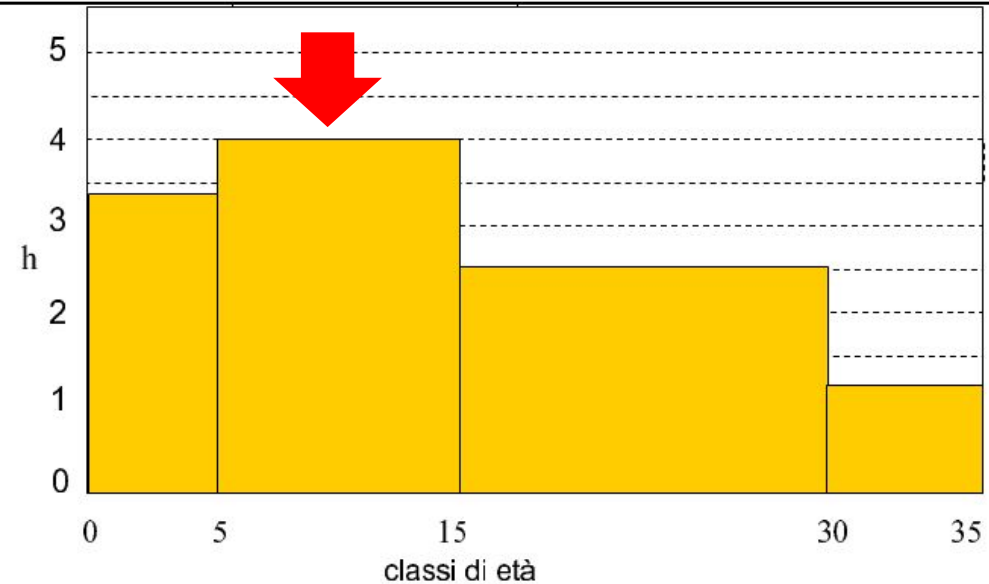
Fonte: Istat, Popolazione residente comunale per sesso anno di nascita e stato civile.

La moda

Se una distribuzione è aggregata in classe non si parlerà di moda, ma di classe modale.

Qualora le classi non fossero di ampiezza diversa, la classe modale corrisponderà alla densità di frequenza maggiore.

| classi di età | amp. classe a_j | freq. % p_j | densità h_j |
|---------------|-------------------|---------------|---------------|
| 0-5 | 5 | 17,0 | 3,4 |
| 5-15 | 10 | 40,0 | 4,0 |
| 15-30 | 15 | 37,0 | 2,5 |
| 30-35 | 5 | 6,0 | 1,2 |



La moda

La moda non sempre può essere identificata univocamente: ci possono essere due frequenze che presentano lo stesso la valore più elevato.

In questo caso si parla di distribuzione bimodale, trimodale se sono tre , ecc.

La mediana

Per le variabili qualitative con categorie ordinate si possono calcolare anche altri indici sintetici.

Possiamo infatti tenere conto non solo delle frequenze delle varie modalità, ma anche della loro posizione.

Un indice di posizione che possiamo utilizzare è la mediana, che può essere calcolata anche per le variabili quantitative.

La mediana è il valore posseduto dall'unità che si colloca nel mezzo di una distribuzione ordinata. Metà delle osservazioni hanno valori uguali o minori della mediana e metà hanno valori uguali o maggiori.

La mediana è stata introdotta dal britannico Francis Galton nel 1883.

La mediana

Come si calcola la mediana? Dobbiamo distinguere due situazioni:

Il numero di osservazioni è dispari.

Il numero di osservazioni è pari.

La prima cosa da fare è ordinare le osservazioni in modo crescente in base alla modalità.

Se il numero di osservazioni (N) è dispari la mediana è il valore che si colloca nella posizione centrale calcolata così: $\text{mediana} = (N+1)/2$

Se il numero di osservazioni è pari, invece, avremo due valori ($N/2$ e $N/2 + 1$) e la mediana, in caso di variabili quantitative sarà la semisomma di questi due valori.

Facciamo degli esempi assieme.

La mediana

Variabile qualitativa ordinata

Titolo di studio (1 = scuola primaria, 2 = scuola media, 3 = scuola superiore, 4 = università).

Ho 13 osservazioni (numero dispari):

1, 3, 2, 3, 2, 3, 4, 1, 2, 3, 4, 3, 2

Le ordino:

1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4

Calcolo la mediana

$$(N + 1)/2 = (13+1)/2 = 7$$

Al settimo valore ordinato corrisponde la mediana

La mediana

Variabile qualitativa ordinata

Titolo di studio (1 = scuola primaria, 2 = scuola media, 3 = scuola superiore, 4 = università).

Ho 14 osservazioni (numero pari):

1, 3, 2, 3, 2, 3, 4, 1, 2, 3, 4, 3, 2, 3

Le ordino:

1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4

Calcolo la mediana

$N/2 = 7$ e $N/2 + 1 = 8$

Al settimo e ottavo valore ordinato corrisponde la mediana

La mediana

Nel caso di una variabile quantitativa, in caso di un numero pari di osservazioni la mediana si otterrà facendo la semisomma dei valori delle due unità centrali.

Esempio (libro pag. 69 esempio 3.4) abbiamo il reddito di 8 famiglie:

1.670, 1.840, 2.005, 2.150, 2.280, 2.360, 2.510, 2.780

Le unità centrali sono la quarta e la quinta:

$(N/2 = 4 \text{ e } N/2 + 1 = 5)$

La mediana è la semisomma dei due valori:

$(2.150 + 2.280)/2 = 4.430/2 = 2.215$

La mediana

Se N è grande, conviene utilizzare le frequenze cumulate.

La mediana si trova nella categoria dove rientra il 50%.

Popolazione italiana di 15 anni e più per titolo di studio - Anno 2020

valori %

| Titolo di studio | Freq. % | Freq. Cum. |
|--|----------------|-------------------|
| Licenza elementare o nessun titolo | 15,9 | 15,9 |
| Licenza di scuola media inferiore | 32,2 | 48,1 |
| Diploma di scuola superiore (qualifica o maturità) | 36,6 | 84,7 |
| Laurea e post laurea | 15,3 | 100 |



Fonte: Istat, Rilevazione sulle forze di lavoro

La mediana

Se la distribuzione quantitativa è in classi, la classe che contiene la mediana è chiamata **classe mediana**.

Possiamo però individuare un valore puntuale attraverso la seguente formula:

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

Dove:

I_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

La mediana

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

Dove:

$$l_m = 26$$

$$F_{m-1} = 0,271$$

$$F_m = 0,501$$

$$\Delta_m = 14$$

$$Me = 26 + (0,5 - 0,271) / (0,501 - 0,271) \cdot 14$$

$$Me = 39,9$$

| Numero di occupati per ore settimanali lavorate | Frequenza assoluta | Frequenza relativa | Frequenza relativa cumulata |
|---|--------------------|--------------------|-----------------------------|
| 0 | 1.892 | 0,086 | 0,086 |
| 1-10 ore | 627 | 0,029 | 0,115 |
| 11-25 ore | 3.443 | 0,156 | 0,271 |
| 26-39 ore | 5.125 | 0,230 | 0,501 |
| 40 ore | 7.607 | 0,339 | 0,840 |
| 41 ore e più | 3.567 | 0,160 | 1,000 |
| Totale | 22.420 | 1,000 | |

La mediana

La mediana riesce a sintetizzare il centro della distribuzione di una variabile quantitativa anche in presenza di valori anomali o eccezionali.

Questa caratteristica viene chiamata **resistenza** o **robustezza**.

Torniamo all'esempio del reddito di 8 famiglie:

1.670, 1.840, 2.005, 2.150, 2.280, 2.360, 2.510, **2.780**

Mediana = 2.215 Media aritmetica = 2.199,4

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

Mediana = 2.215 Media aritmetica = 4.449,4

I quantili

La mediana suddivide una distribuzione ordinata in due distribuzioni che comprendono ognuna il 50% dei casi.

Questo può essere replicato costruendo un numero qualsiasi di distribuzioni parziali q , ognuna avente la q -esima parte della numerosità complessiva della distribuzione.

La modalità che si pone tra le varie distribuzioni parziali si chiama genericamente quantile.

Se $q = 2$ mediana

Se $q = 3$ terzili

Se $q = 4$ quartili

Se $q = 5$ quintili

Se $q = 10$ decili

Se $q = 100$ centili

Le medie

Le medie sono valori centrali analitici, che possono essere calcolate solo per le distribuzioni di variabili quantitative.

La media più utilizzata è la **media aritmetica**, ma esistono anche **media geometrica**, **media quadratica** e **media armonica**.

La differenza sta nell'operazione da effettuare sui valori della variabile.

Le medie

| | |
|------------------------|----------------------------|
| Addizione | -> media aritmetica |
| Moltiplicazione | -> media geometrica |
| Elevazione al quadrato | -> media quadratica |
| Somma degli inversi | -> media armonica |

La media aritmetica

La media aritmetica è la somma dei valori di una distribuzione quantitativa, divisi per il loro numero.

La formula è la seguente:

$$M(X) = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

La media aritmetica

Supponiamo di aver raccolto i risultati degli studenti che hanno sostenuto l'esame di statistica sociale nell'anno accademico 2021-22.

Questi sono i risultati:

30, 27, 22, 24, 21, 19, 26, 18, 28, 21, 24, 22, 30, 28, 18, 19, 23, 26, 29, 27, 20, 30, 27, 26, 30, 30, 26, 24, 28, 27.

La media aritmetica sarà pari alla somma dei valori diviso per N, ovvero 30.

$$M(x) = 750/30 = 25$$

| Voto 2021/22 | Frequenza assoluta | Frequenza relativa | Frequenza percentuale |
|--------------|--------------------|--------------------|-----------------------|
| 18 | 2 | 0,06667 | 6,66667 |
| 19 | 2 | 0,06667 | 6,66667 |
| 20 | 1 | 0,03333 | 3,33333 |
| 21 | 2 | 0,06667 | 6,66667 |
| 22 | 2 | 0,06667 | 6,66667 |
| 23 | 1 | 0,03333 | 3,33333 |
| 24 | 3 | 0,1 | 10 |
| 25 | 0 | 0 | 0 |
| 26 | 4 | 0,13333 | 13,33333 |
| 27 | 4 | 0,13333 | 13,33333 |
| 28 | 3 | 0,1 | 10 |
| 29 | 1 | 0,03333 | 3,33333 |
| 30 | 5 | 0,16667 | 16,66667 |
| Totale | 30 | 1 | 100 |

| Voto 2020/21 | Frequenza assoluta | Frequenza relativa | Frequenza percentuale |
|--------------|--------------------|--------------------|-----------------------|
| 18 | 2 | 0,04 | 4 |
| 19 | 4 | 0,08 | 8 |
| 20 | 2 | 0,04 | 4 |
| 21 | 3 | 0,06 | 6 |
| 22 | 3 | 0,06 | 6 |
| 23 | 5 | 0,1 | 10 |
| 24 | 4 | 0,08 | 8 |
| 25 | 3 | 0,06 | 6 |
| 26 | 6 | 0,12 | 12 |
| 27 | 4 | 0,08 | 8 |
| 28 | 5 | 0,1 | 10 |
| 29 | 3 | 0,06 | 6 |
| 30 | 6 | 0,12 | 12 |
| Totale | 50 | 1 | 100 |

Come possiamo procedere?

Ho a disposizione la distribuzione di frequenze, posso calcolare la media aritmetica in modo semplice.

Moltiplico i valori (voti) per le loro frequenze assolute. (Ad es. $18 \times 2 + 19 \times 2 + \dots + 30 \times 5$)

Successivamente sommo i prodotti che ho ottenuto.

Infine divido per il numero dei casi.

Nel 2021/2022 la media aritmetica era 25,0

Nel 2020/2021 la media aritmetica era 24,8