

Statistica Sociale

Rappresentazioni grafiche

Facilità di lettura delle caratteristiche di una distribuzione osservata di una variabile

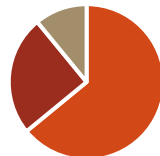
I 5 aspetti da considerare nella valutazione di un grafico:

- **Accuratezza**: precisione dei dettagli (dimensioni adeguate, legenda)
- **Semplicità**: limitare elementi grafici superflui, inserire solo ciò che aiuta la lettura del grafico
- **Chiarezza**: concentrarsi su un messaggio e rappresentare le caratteristiche salienti
- **Aspetto**: tratti, proporzioni e colori armonici
- **Struttura**: gli elementi più importanti devono avere più rilievo

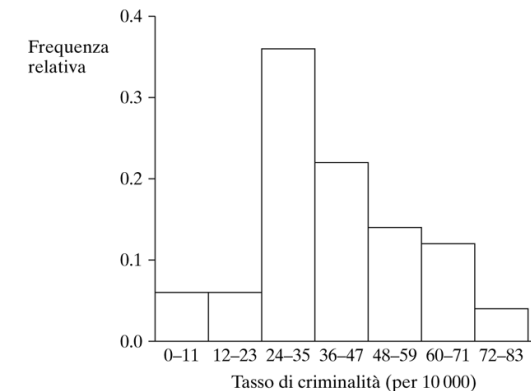
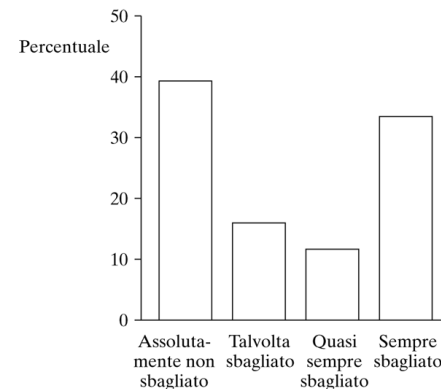
Tipi di grafici

Variabile qualitativa		Variabile quantitativa	
Nominale	Ordinale	Discreta	Continua
\neq	$\geq o \leq$	+ - * /	+ - * /
Grafico a torta	[Grafico a torta]		
[Grafico a barre]	Grafico a barre	Grafico a barre	
			Istogramma

Colore occhi



■ Marroni ■ Verdi ■ Azzurri



Grafici a torta

Le frequenze sono espresse in termini di angoli

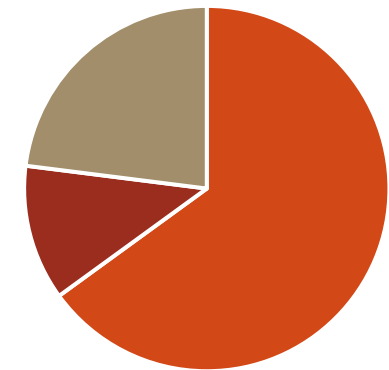
Adatto per distribuzioni con un basso numero di modalità

$$freq\ rel_{marroni} * 360 = 0,65 * 360 \approx 234$$

$$freq\ rel_{verdi} * 360 = 0,12 * 360 \approx 43$$

$$freq\ rel_{azzurri} * 360 = 0,23 * 360 \approx 83$$

Colore occhi



■ Marroni ■ Verdi ■ Azzurri

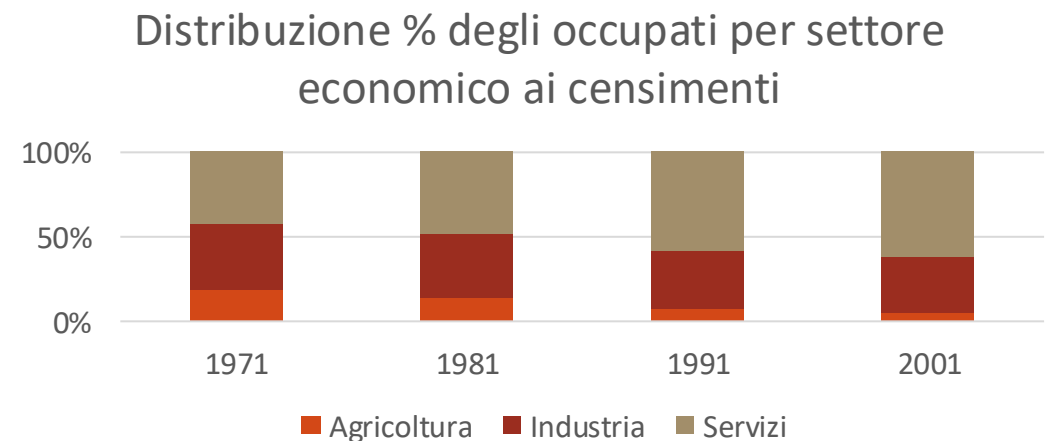
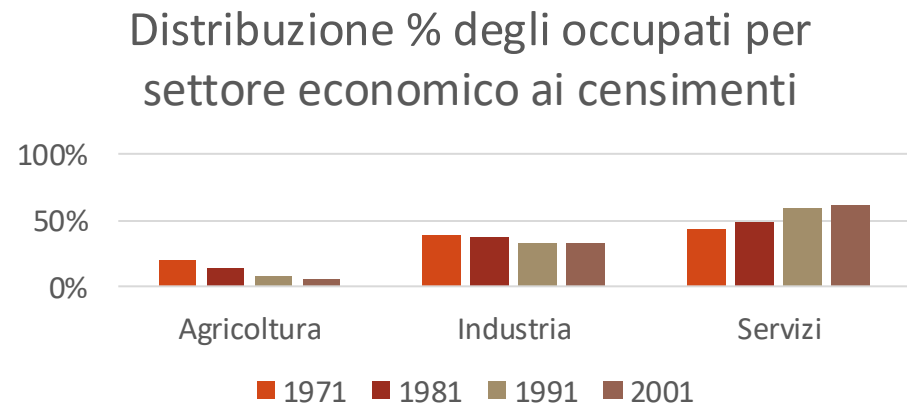
Grafici a barre

La frequenza (assoluta, relativa o percentuale) di ogni modalità viene rappresentata da una barra/rettangolo

I rettangoli hanno tutti la stessa base mentre le altezze sono proporzionali alle frequenze delle modalità

Mantiene in evidenza l'ordinamento delle modalità del carattere

Grafici a barre multipli permettono il confronto di più distribuzioni



Istogrammi (variabili quantitative suddivise in classi)

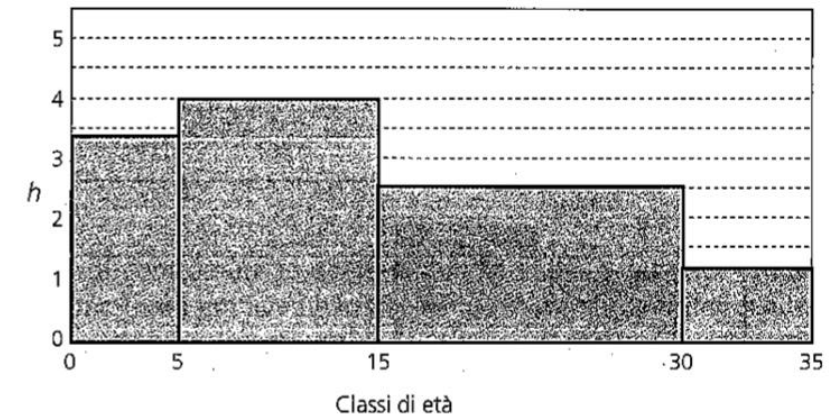
Le barre sono NON distanziate

Le basi dei rettangoli possono avere ampiezza diversa, quindi l'area del rettangolo è proporzionale alla corrispondente frequenza

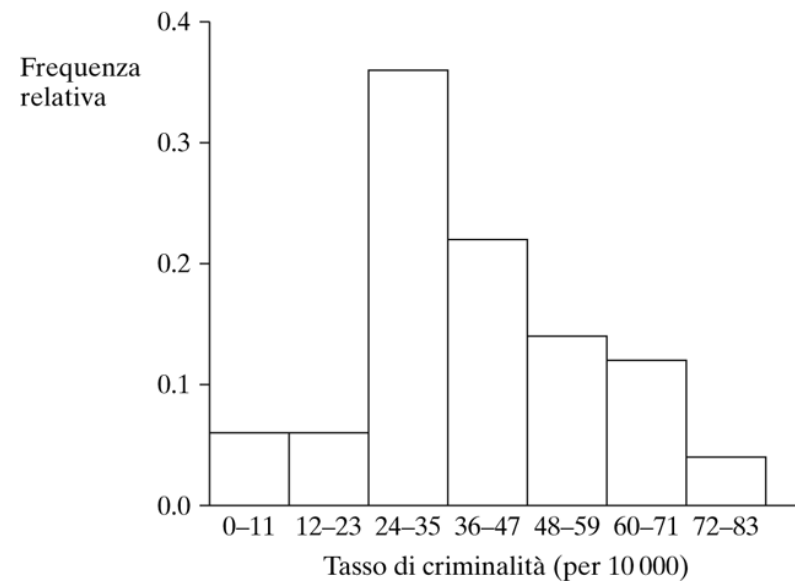
Nel caso di classi di ampiezza diversa, l'altezza del rettangolo è detta **densità** ed è proporzionale al rapporto tra frequenza e ampiezza

Attenzione all'assunzione che le modalità siano distribuite uniformemente all'interno delle classi

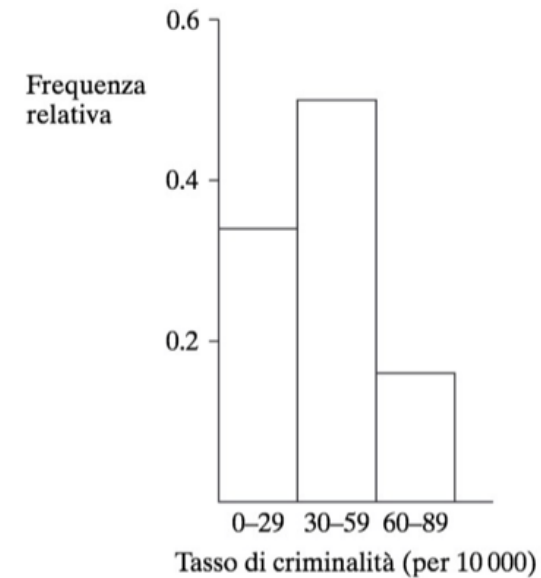
Classi di età	Ampiezza	Percentuale	Densità
[0-5)	5	17%	$3,4 = 17/5$
[5-15)	10	40%	$4 = 40/10$
[15-30)	15	37%	$2,5 = 37/15$
[30-35)	5	6%	$1,2 = 6/5$



Istogramma della distribuzione di frequenze relative del tasso di criminalità negli Stati Uniti nel 2005



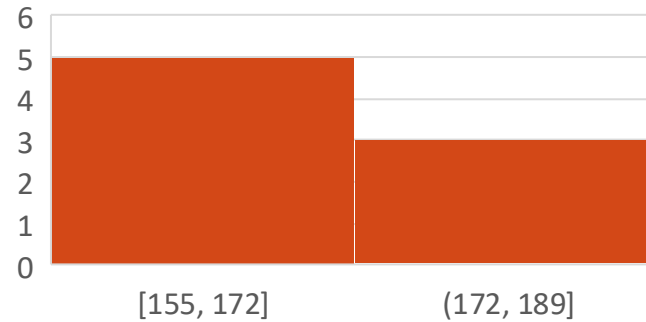
Istogramma della distribuzione di frequenze relative del tasso di criminalità negli Stati Uniti nel 2005 ottenuto utilizzando intervalli di ampiezza eccessiva



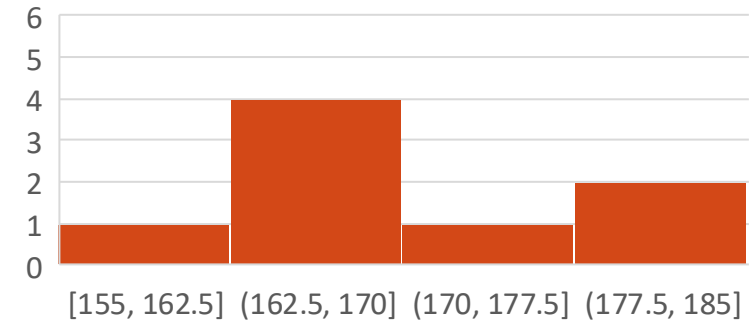
Un campione diverse rappresentazioni

Statura	Numerosità
155	1
163	1
164	1
165	1
167	1
175	1
178	1
185	1

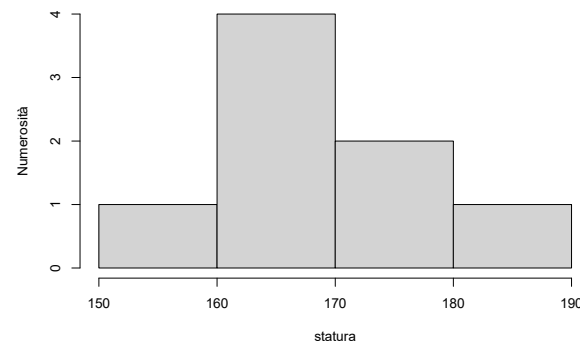
Excel default



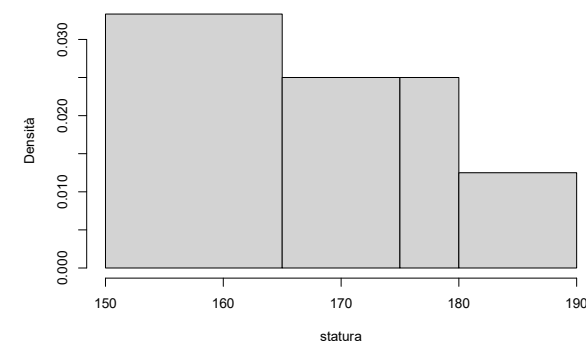
4 classi con stessa ampiezza



R default

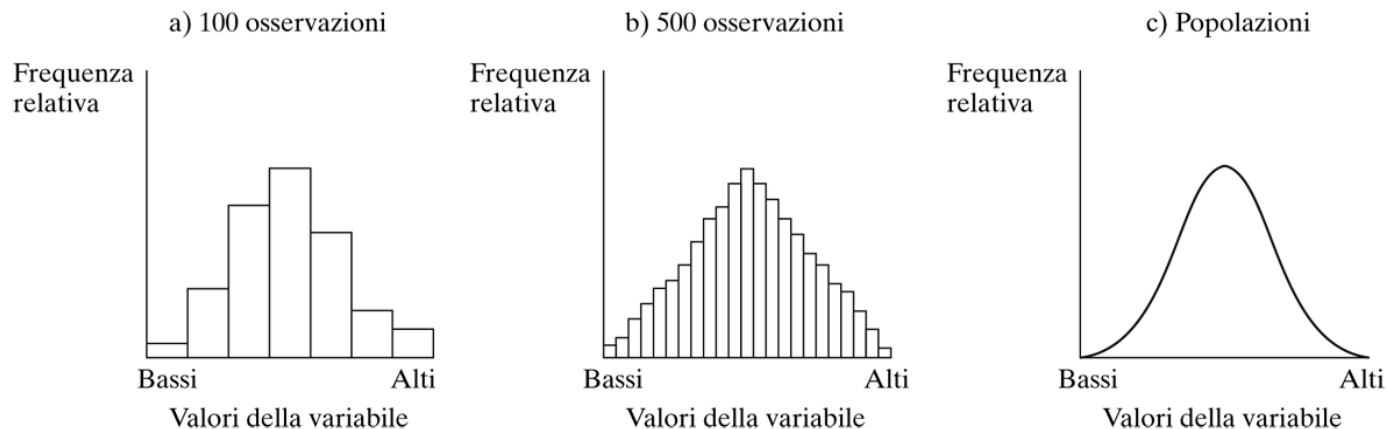


4 classi di ampiezza diversa



Distribuzioni di popolazioni e distribuzioni di dati campionari

- La costruzione di distribuzioni di frequenze o di istogrammi può essere fatta sia per i **dati di popolazione** che per quelli **campionari**.
- Nel primo caso facciamo riferimento a **distribuzioni di dati di popolazione**, nel secondo a **distribuzioni di dati campionari**.
- Si può dire che una distribuzione di dati campionari è una foto sfuocata della distribuzione dei dati di popolazione.
- A mano a mano che l'ampiezza campionaria aumenta le proporzioni campionarie in ciascun intervallo si approssimano sempre di più alle vere proporzioni della popolazione: la distribuzione dei dati campionari diventa sempre più simile alla distribuzione di popolazione.



Nel grafico c) è sovrapposta una curva liscia che rappresenta la distribuzione nella popolazione (anche se una variabile è discreta, la sua distribuzione a livello di popolazione può essere adeguatamente approssimata da una curva liscia).

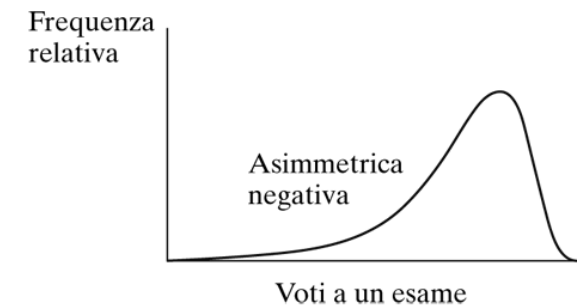
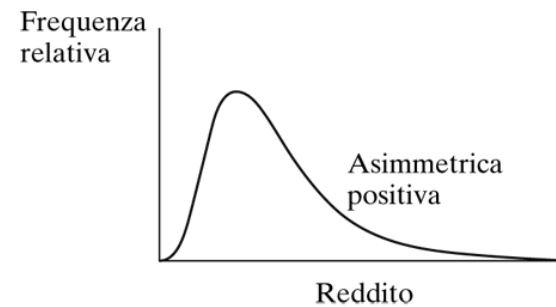
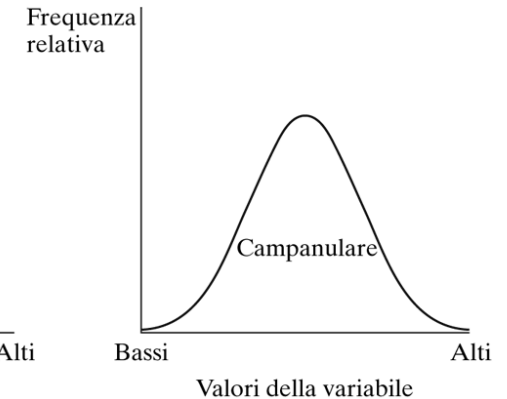
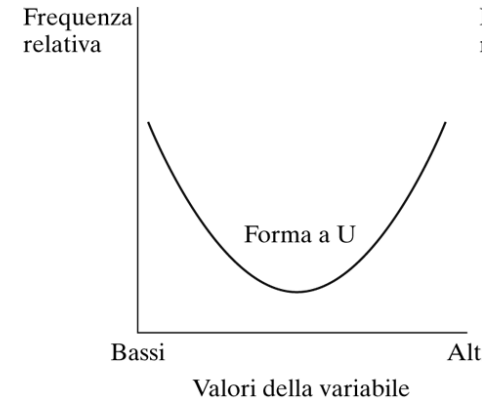
La forma della distribuzione

Entrambe le distribuzioni sono **simmetriche**: il lato della distribuzione a sinistra del valore centrale è l'immagine speculare del lato a destra dello stesso valore.

La maggior parte delle distribuzioni che vengono osservate nell'ambito della ricerca sociale **non** sono simmetriche.

Le parti della figura in corrispondenza dei valori bassi e alti sono chiamate **code della distribuzione**.

Una distribuzione viene definita **asimmetrica positiva** o **asimmetrica negativa** a seconda di quella che è la sua coda più lunga.



Esercizio

Completa la tabella e rappresenta i dati nel modo per te più opportuno

Modalità	Numerosità	Frequenza relativa	Percentuale
A	11	0,22	
B	30		
C			

$$11/n = 0,22$$

$$n = 11/0,22 = 50 \text{ da cui } C = 9$$

$$30/50 = 0,6$$

$$9/50 = 0,18$$

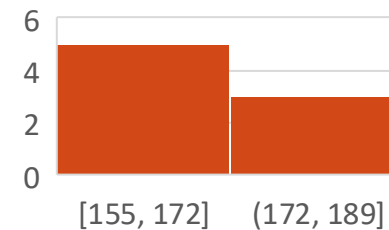
Sintesi di una distribuzione

- Distribuzione di frequenze (assolute, relative, percentuali)
- Rappresentazioni grafiche (grafico a torta, grafico a barre, istogramma)
- **Indici** per evidenziare con un solo valore alcune caratteristiche essenziali della distribuzione
 - Ad esempio gli **indici di posizione** sintetizzano con un solo numero la tendenza centrale della distribuzione

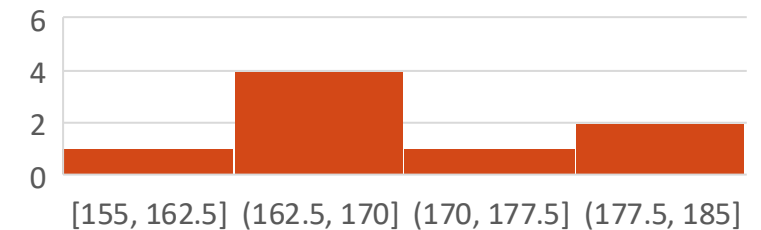
Unità	Statura
1	163
2	165
3	185
4	164
5	167
6	175
7	178
8	155

Statura	Numerosità
155	1
163	1
164	1
165	1
167	1
175	1
178	1
185	1

Statura



Statura



La media aritmetica

La **media** è la somma dei valori assunti dalle osservazioni divisa per il totale delle osservazioni

$$(155 + 163 + 164 + 165 + 167 + 175 + 178 + 185)/8 = 1352/8 = 169$$

Notazione:

- n : numerosità campionaria
- x, y, z, \dots : variabili (ad esempio statura, età, ...)
- $x_1, x_2, x_3, \dots, x_n$: osservazioni campionarie (la generica osservazione è x_i)
- \bar{x} : media della variabile x

$$\overline{statura} = (statura_1 + statura_2 + statura_3 + statura_4 + statura_5 + statura_6 + statura_7 + statura_8)/n$$

$$\overline{statura} = (statura_1 + \dots + statura_i + \dots + statura_8)/n$$

$$\overline{statura} = \sum_{i=1}^n statura_i / n$$

Equivalentemente

$$\bar{x} = (x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8)/n$$

$$\bar{x} = (x_1 + \dots + x_i + \dots + x_8)/n$$

$$\bar{x} = \sum_{i=1}^n x_i / n$$

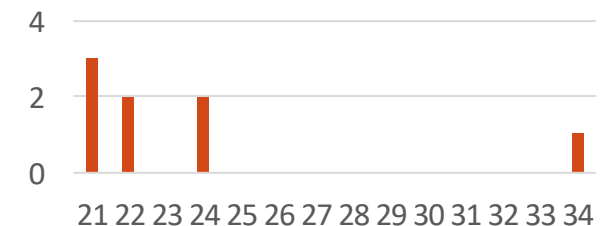
Unità	Statura
1	163
2	165
3	185
4	164
5	167
6	175
7	178
8	155

Indici sintetici: media aritmetica

- Per caratteri quantitativi
- La media è un valore sempre compreso tra il minimo ed il massimo valore osservato
- La media può essere influenzata da un'osservazione molto "lontana" (molto più piccola/molto più grande) rispetto alle altre osservazioni -> outlier. In questi casi la media può essere poco rappresentativa dei valori tipici del campione
- La media tende verso la coda più lunga della distribuzione: più forte è l'asimmetria meno rappresentativa sarà la distribuzione
- La media è il baricentro delle osservazioni: la somma delle distanze dalla media delle osservazioni più piccole è uguale alla somma delle distanze di quelle più grandi

Unità	Età
1	24
2	21
3	34
4	22
5	21
6	22
7	24
8	21

Età



Qual è l'età media del campione raccolto nella lezione precedente?

$$\begin{aligned}\overline{età} &= \sum_{i=1}^n \frac{età_i}{n} = \sum_{i=1}^8 \frac{età_i}{8} = \frac{età_1 + età_2 + \dots (+età_i) \dots + età_8}{8} = \frac{24 + 21 + \dots + 21}{8} = \\ &= \frac{189}{8} = 23,6\end{aligned}$$

La media aritmetica

- Se conosciamo la **distribuzione di frequenze** per le K modalità osservate, la media può essere calcolata come

- Se conosciamo la distribuzione di frequenze assolute:

$$\bar{x} = \sum_{k=1}^K \frac{x_k * \text{numerosità}_k}{n} = \frac{x_1 * \text{numerosità}_1 + \dots + x_k * \text{numerosità}_k}{n}$$

- Se conosciamo la distribuzione di frequenze relative:

$$\bar{x} = \sum_{k=1}^K x_k * \text{freq rel}_k = x_1 * \text{freq rel}_1 + \dots + x_k * \text{freq rel}_k$$

k	Età	Numerosità	Freq rel
1	21	3	0,375
2	22	2	0,25
3	24	2	0,25
4	34	1	0,125

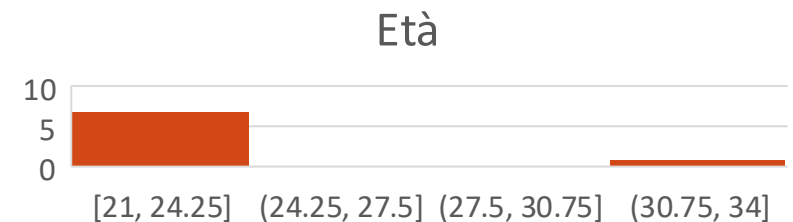
Calcoliamo di nuovo l'età media utilizzando la distribuzione di frequenze

- Con le frequenze assolute

$$\overline{\text{età}} = \sum_{k=1}^4 \frac{\text{età}_k * \text{numerosità}_k}{8} = \frac{21 * 3 + 22 * 2 + 24 * 2 + 34 * 1}{8} = 23,6$$

- Con le frequenze relative

$$\overline{\text{età}} = \sum_{k=1}^4 \text{età}_k * \text{freq rel}_k = 21 * 0,375 + 22 * 0,25 + 24 * 0,25 + 34 * 0,125 = \dots$$



La media aritmetica approssimata

Quando la variabile è **suddivisa in classi** e non conosciamo i valori osservati ma solo la classe di appartenenza, utilizziamo il **valore centrale** di ogni classe per ottenere la **media aritmetica approssimata**

Per K classi, indichiamo il valore centrale della classe con c_k , allora la media aritmetica approssimata si calcola come:

$$\bar{x} \approx \sum_{k=1}^K \frac{c_k * \text{numerosità}_k}{n} \quad \text{oppure} \quad \bar{x} \approx \sum_{k=1}^K c_k * \text{freq rel}_k$$

Età	Numerosità	Freq rel	Valore Centrale
[19-28)	7	0.7	23.5
[28-37)	2	0.2	32.5
[37-46)	1	0.1	41.5
	10	1	

Calcoliamo la media aritmetica approssimata per l'età:

$$\bar{x} \approx \sum_{k=1}^3 \frac{c_k * n_k}{10} = \frac{23.5 * 7 + 32.5 * 2 + 41.5 * 1}{10} = 23.5 * 0.7 + 32.5 * 0.2 + 41.5 * 0.1 = 27.1$$

- Se il carattere è equidistribuito all'interno della classe la media aritmetica e la media aritmetica approssimata coincidono
- Se una delle classi è non limitata (es: da [37 in su) bisogna scegliere un opportuno valore che possa rappresentare la media interna della classe.

La media aritmetica ponderata

Ad ogni osservazione si attribuisce un **peso** per aumentare/diminuire la sua importanza nel calcolo della media

Per n osservazioni, indichiamo i pesi con w_1, \dots, w_n . Allora la media aritmetica ponderata si calcola come

$$\bar{x}_w = \frac{x_1 * w_1 + x_2 * w_2 + \dots + x_i * w_i + \dots + x_N * w_N}{w_1 + w_2 + \dots + w_i + \dots + w_N} = \frac{\sum x_i w_i}{\sum w_i}$$

Calcoliamo la media aritmetica ponderata per i seguenti voti considerando come pesi i CFU di ogni insegnamento

Voto	CFU
29	9
25	3
28	6
	18

$$\bar{x}_w = \frac{29 * 9 + 25 * 3 + 28 * 6}{18} = 28$$

$$\bar{x}_w = \frac{29 * 3 + 25 * 9 + 28 * 6}{18} = 26.66 \dots$$

Voto	CFU
29	3
25	9
28	6
	18

E la media aritmetica?

$$\bar{x} = \frac{29 + 25 + 28}{3} = 27.33 \dots$$

Esempio: Media aritmetica ponderata ore di sonno/ 1

N. ore (Y)	Freq. assolute		
	Genere (X)		Tot
	F	M	
5	1	0	1
6	4	4	8
7	10	14	24
8	10	14	24
9	0	1	1
	25	33	58

Le colonne 'F' ed 'M' rappresentano le distribuzioni condizionate del carattere Y (numero di ore di sonno/notte) secondo le modalità F ed M della variabile X (sesso)

Per ogni gruppo possiamo calcolare la media **condizionata**:

Media (cond) F

$$\frac{5 \times 1 + 6 \times 4 + 7 \times 10 + 8 \times 10 + 9 \times 0}{25} = \frac{179}{25} = 7.16$$

Media (cond) M

$$\frac{5 \times 0 + 6 \times 4 + 7 \times 14 + 8 \times 14 + 9 \times 1}{33} = \frac{243}{33} = 7.364$$

La media generale è

$$\bar{y} = \frac{5 \times 1 + 6 \times 8 + 7 \times 24 + 8 \times 24 + 9 \times 1}{58} = \frac{422}{58} = 7.276$$

Media aritmetica dati raggruppati

La media aritmetica calcolata per dati raggruppati è un esempio di **media aritmetica ponderata**. Per le ore di sonno, si ha infatti

$$\bar{y} = \frac{N_f \times \bar{y}_f + N_m \times \bar{y}_m}{N_f + N_m} = \frac{1}{N}(179 + 243)$$

dove \bar{y}_f e \bar{y}_m sono le medie di gruppo (per le modalità 'F' ed 'M') e $N = N_f + N_m$ è il numero totale di osservazioni.



La media marginale è quindi la media delle medie condizionate, pesata con la numerosità dei gruppi.

Indici sintetici: mediana

- La variabile deve essere almeno ordinabile

La **mediana** è la modalità assunta dall'unità centrale del campione ordinato ovvero quell'unità che divide il campione in due parti di uguale numerosità

1. Ordinare le unità secondo le modalità del carattere
2. Se n è dispari la mediana è:

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

se n è pari la mediana è:

$$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Le parentesi tonde indicano la posizione nel campione ordinato

Esempio voti:

i valori osservati sono: $x_1 = 29, x_2 = 24, x_3 = 28$

1. Ordiniamoli: $x_{(1)} = 24, x_{(2)} = 28, x_{(3)} = 29$
2. n è dispari (è uguale a 3) quindi la mediana è

$$Me = x_{\left(\frac{3+1}{2}\right)} = x_{(2)} = 28$$

Esempio età:

i valori osservati sono: 22, 21, 21, 21, 21, 27, 31, 44, 23, 32

1. Ordiniamoli : 21, 21, 21, 21, 21, 22, 23, 27, 31, 32, 44
2. n è pari (è uguale a 10) quindi la mediana è:

$$Me = \frac{x_{\left(\frac{10}{2}\right)} + x_{\left(\frac{10}{2}+1\right)}}{2} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{22 + 23}{2} = 22.5$$

La mediana

Per una variabile qualitativa ordinale o per una variabile quantitativa suddivisa in classi si calcola la **classe mediana**

1. Ordina le classi/modalità
2. Se n è dispari la mediana è:

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

se n è pari la mediana è:

$$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Le parentesi tonde indicano la posizione nel campione ordinato

- La tabella riporta la distribuzione della variabile livello di istruzione per i cittadini statunitensi di età superiore ai 24 anni

T. studio	Freq. (in milioni)	%
Senza titolo di sc. m. sup.	30	15.9%
Sc. m. sup.	56	29.6%
Studi univ. senza titolo	38	20.1%
Laurea di I livello	14	17.4%
Laurea di II livello	32	16.9%
Master	13	6.9%
Dottorato ricerca	16	13.2%

(Fonte: *American Community Survey*, Bureau of the Census USA)

- n è uguale a 199 (30+56+...+16), quindi la posizione mediana corrisponde alla 100-esima osservazione:

$$Me = x_{\left(\frac{199+1}{2}\right)} = x_{(100)}$$

- La prima modalità «Senza titolo» include 30 osservazioni;
- Le prime due modalità insieme («Senza titolo» e «Sc.m.sup») includono 86 osservazioni (30 + 56);
- Le prime tre modalità includono 124 osservazioni quindi la mediana ricade in questa classe che è detta classe mediana:

$$Me = x_{(100)} = \text{«Studi univ. senza titolo»}$$