DIMA

CORSO LAUREA MAGISTRALE IN **DIGITAL MARKETING**

STATISTICA DESCRITTIVA

Prof. ssa ANNA LINA SARRA

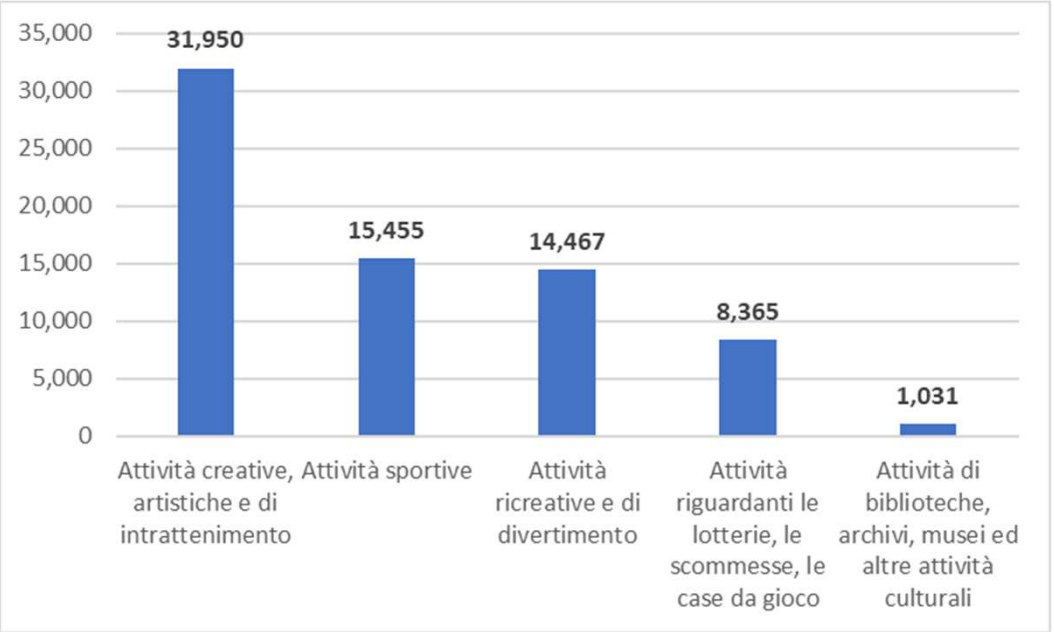
Rappresentazioni grafiche

Per ogni distribuzione statistica, è possibile individuare la rappresentazione grafica più adatta, la soluzione, cioè, che concilia **correttezza metodologica** ed **efficacia informativa**

Scala nominale: grafico a nastri

Le distribuzioni di variabili nominali (serie sconnesse) vengono generalmente rappresentate con **grafici di tipo areale**, in cui alle modalità del carattere si fanno corrispondere figure geometriche (rettangoli, quadrati, settori circolari ecc.) con **aree proporzionali** alle grandezze da rappresentare. Le figure geometriche più spesso utilizzate sono i rettangoli (**nastri orizzontali o verticali**).

ATTIVITÀ ECONOMICHE	Numero
Attività creative, artistiche e di intrattenimento	31,950
Attività sportive	15,455
Attività ricreative e di divertimento	14,467
Attività riguardanti le lotterie, le scommesse, le case da gioco	8,365
Attività di biblioteche, archivi, musei ed altre attività culturali	1,031
Totale attività artistiche, sportive, di intrattenimento e divertimento	**71,268**



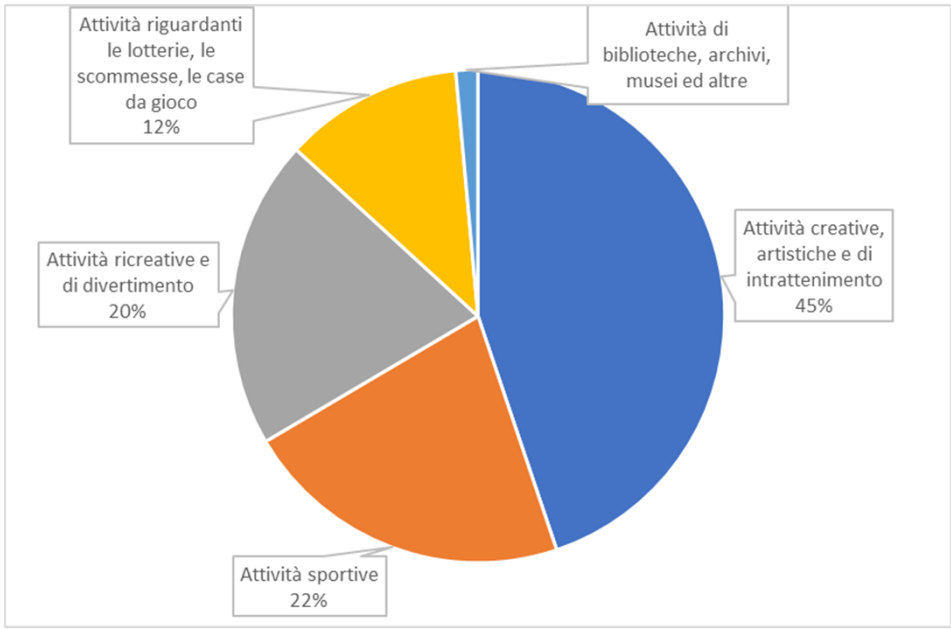
Fonte Istat

L'aspetto visivo del grafico rimane immutato se si utilizzano le frequenze relative o percentuali.

Scala nominale: grafico a settori circolari o a torta

Una rappresentazione grafica alternativa, utilizzabile quando il numero delle modalità non è elevato, è il grafico a settori circolari, in cui le frequenze o le quantità associate alle varie modalità del carattere vengono rappresentate con le aree dei settori circolari in cui è suddiviso un cerchio.

ATTIVITÀ ECONOMICHE	Numero
Attività creative, artistiche e di intrattenimento	31,950
Attività sportive	15,455
Attività ricreative e di divertimento	14,467
Attività riguardanti le lotterie, le scommesse, le case da gioco	8,365
Attività di biblioteche, archivi, musei ed altre attività culturali	1,031
Totale attività artistiche, sportive, di intrattenimento e divertimento	**71,268**

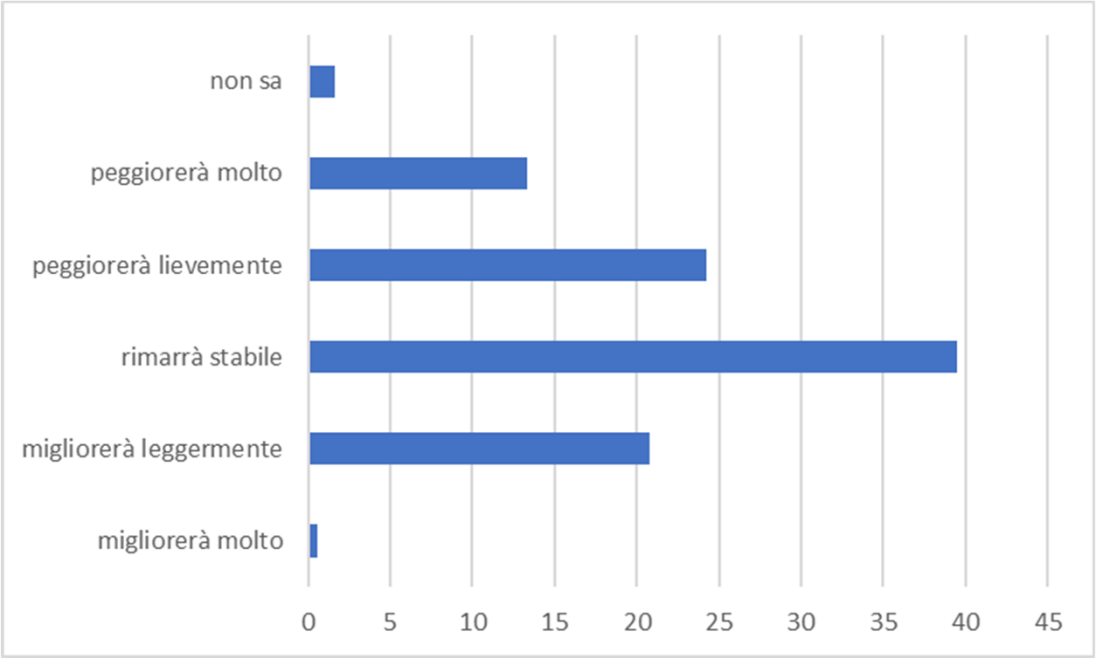


Fonte Istat

Scala ordinale: grafico a nastri con modalità ordinate

Le distribuzioni di variabili ordinali vengono generalmente rappresentate con **grafici a nastri orizzontali o verticali** in cui le modalità devono essere ordinate

Attese sulla situazione economica dell'Italia	gennaio 2023 *%*
migliorerà molto	0.6
migliorerà leggermente	20.8
rimarrà stabile	39.5
peggiorerà lievemente	24.2
peggiorerà molto	13.3
non sa	*1.6*
Totale	100

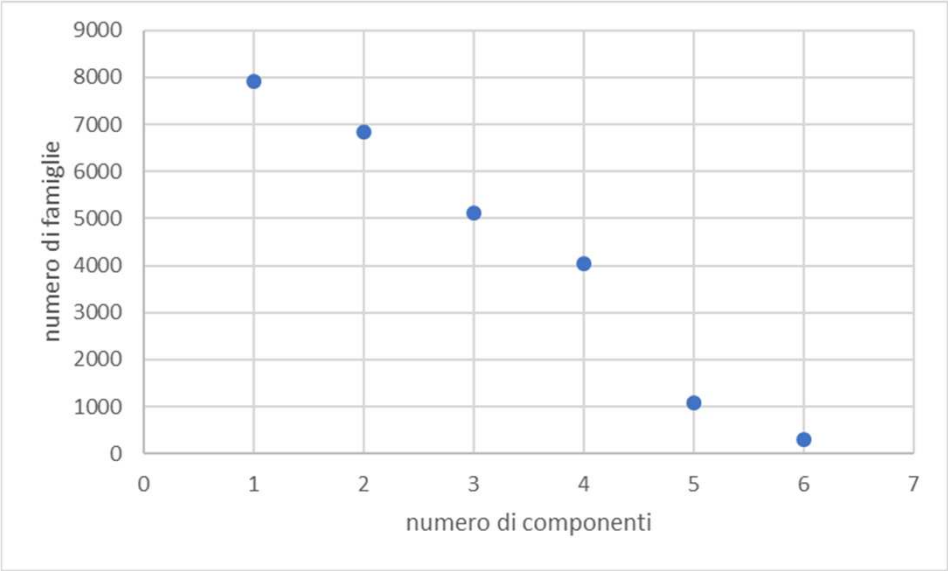


Fonte Istat

Caratteri quantitativi discreti: diagramma a punti

- La rappresentazione grafica più idonea per una distribuzione di frequenze secondo un carattere discreto è quella **cartesiana.**
- Sull'asse delle ascisse vengono poste le modalità $x_1, x_2, \ldots, x_k$, sull'asse delle ordinate le frequenze corrispondenti $n_1, n_2, \ldots, n_k$
- La rappresentazione grafica va sotto il nome di **diagramma a punti**

Numero di componenti della famiglia	Numero di famiglie (valori in migliaia)
1	7910
2	6833
3	5116
4	4051
5	1088
6	303
Totale	**25301**



Fonte Istat

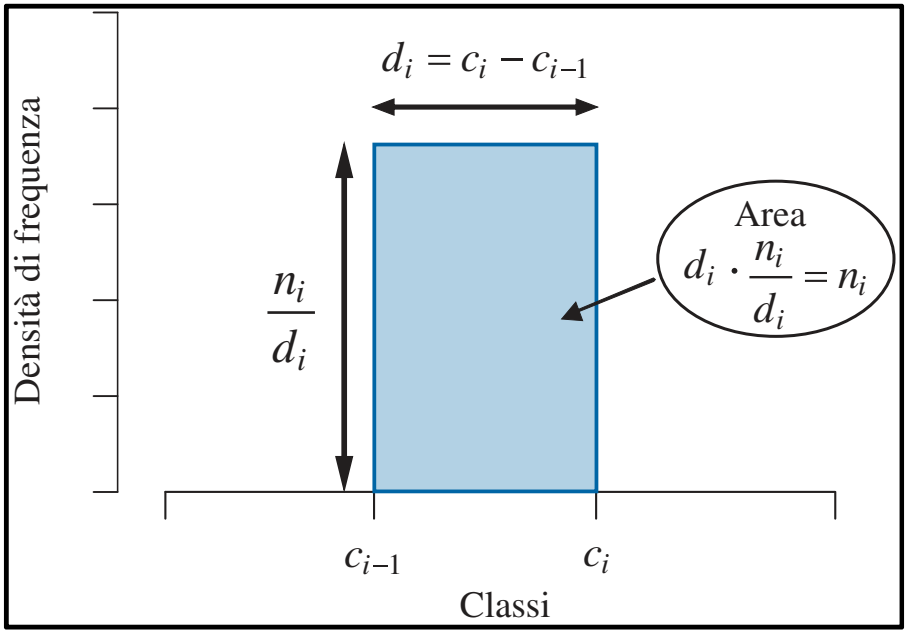
Caratteri quantitativi continui divisi in intervalli

La rappresentazione grafica più appropriata è l'**istogramma** in cui **l'area del rettangolo rappresenta la frequenza assoluta o relativa della corrispondente classe.**

L'istogramma si ottiene ponendo sull'asse delle ascisse gli estremi di classe c_0, c_1, …, c_k e disegnando per ogni classe (c_{i-1}, c_i), i = 1, 2, …, k, un **rettangolo** avente per base il segmento dell'asse delle ascisse di estremi c_{i-1} e c_i e per altezza la **densità di frequenza** n_i/d_i, dove d_i è l'ampiezza di classe.

Classe $(c_{i-1}\text{-}c_i)$	**Frequenza** (n_i)
c_0 - c_1	n_1
c_1 - c_2	n_2
⋮	⋮
c_{k-1} - c_k	n_k
Totale	N

Si chiama **distribuzione di frequenze** di un carattere *X* suddiviso in classi lo schema con cui si associa a ciascuna classe la rispettiva frequenza.

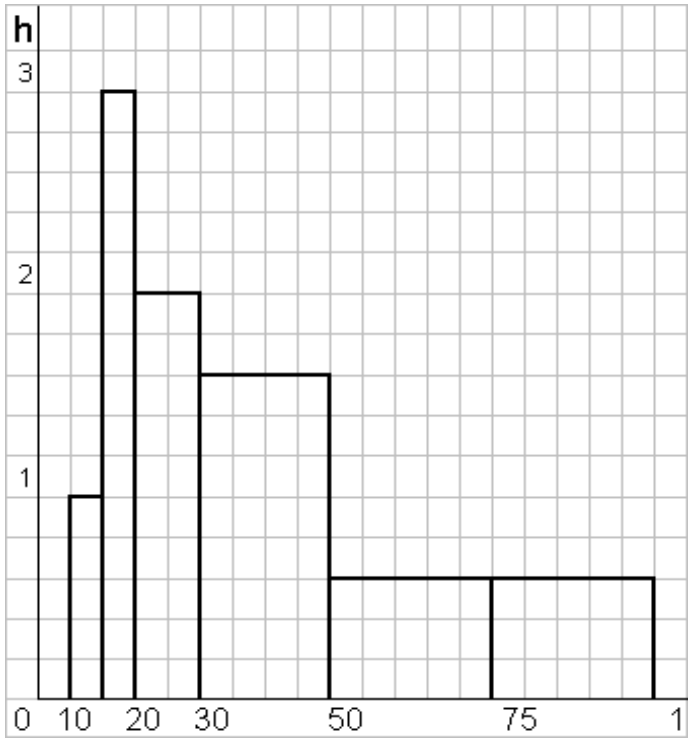


Istogramma: esempio

Classi di peso (in Kg)	Frequenza assoluta	Ampiezza di classe	Densità di frequenza
$c_{i-1} - c_i$	n_i	d_i	h_i
10 |-- 15	5	5	1
15 |-- 20	15	5	3
20 |-- 30	20	10	2
30 |-- 50	30	20	1,5
50 |-- 75	15	25	0,6
75 |-- 100	15	25	0,6
Totale	**100**		

$$d_i = c_{i-1} - c_i$$

$$h_i = \frac{n_i}{d_i}$$

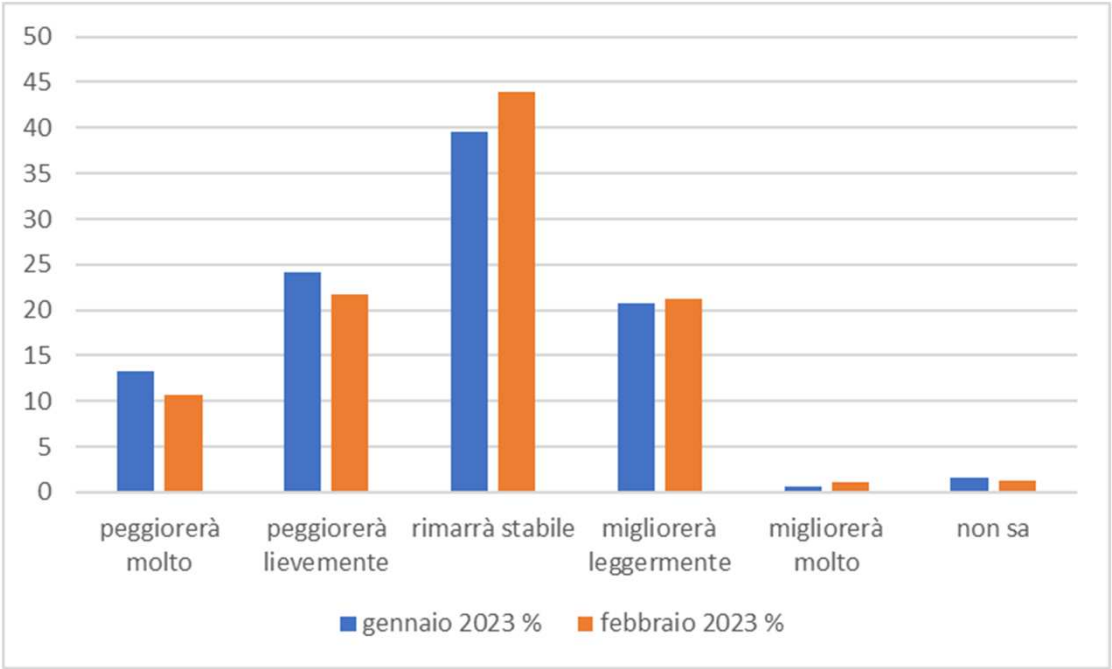


Nell'istogramma si costruiscono dei rettangoli la cui area è proporzionale alla frequenza della classe.

La base è data dall'ampiezza di classe, l'altezza dalla densità di frequenza.

Grafico su distribuzione doppia

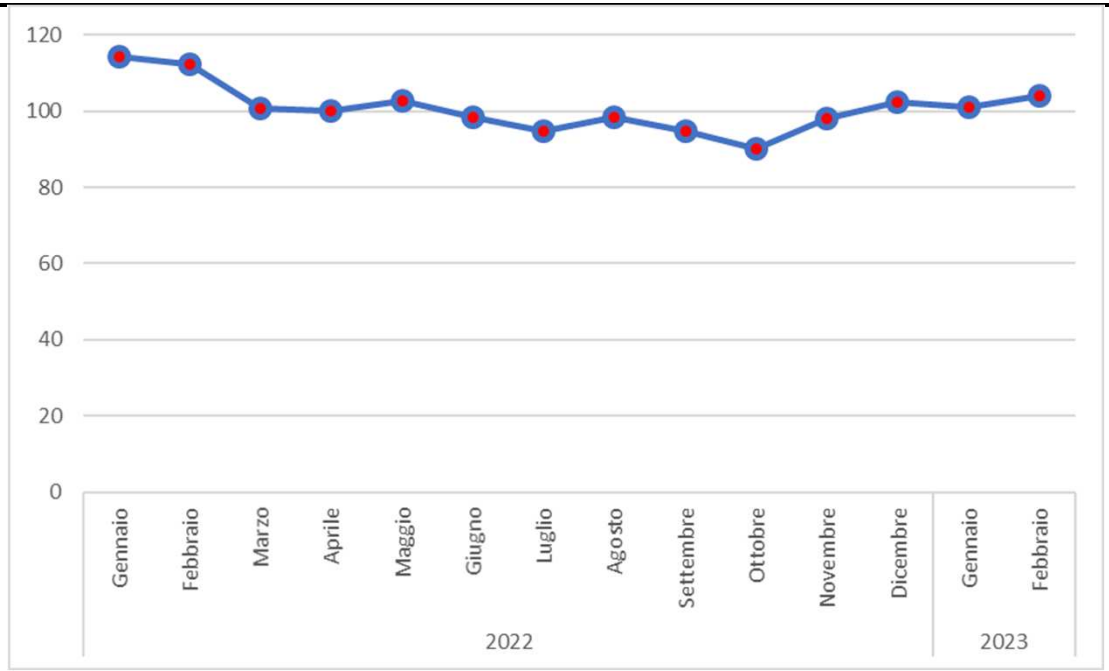
	gennaio 2023	febbraio 2023
	%	*%*
peggiorerà molto	13.3	10.7
peggiorerà lievemente	24.2	21.7
rimarrà stabile	39.5	44
migliorerà leggermente	20.8	21.3
migliorerà molto	0.6	1.1
non sa	*1.6*	*1.2*
Totale	100	100



Grafici per serie storiche

- Per la rappresentazione grafica delle serie storiche si ricorre, generalmente, ai **diagrammi cartesiani**. Si pongono sull'asse delle ascisse i tempi e su quello delle ordinate le intensità associate: i conseguenti punti del piano cartesiano vengono poi uniti con **segmenti di retta** per facilitare la percezione visiva dell'andamento del fenomeno.

	2022												2023	
	Gennaio	Febbraio	Marzo	Aprile	Maggio	Giugno	Luglio	Agosto	Settembre	Ottobre	Novembre	Dicembre	Gennaio	Febbraio
Clima di Fiducia dei consumatori	114.2	112.4	100.8	100	102.7	98.3	94.8	98.3	94.8	90.1	98.1	102.5	100.9	104



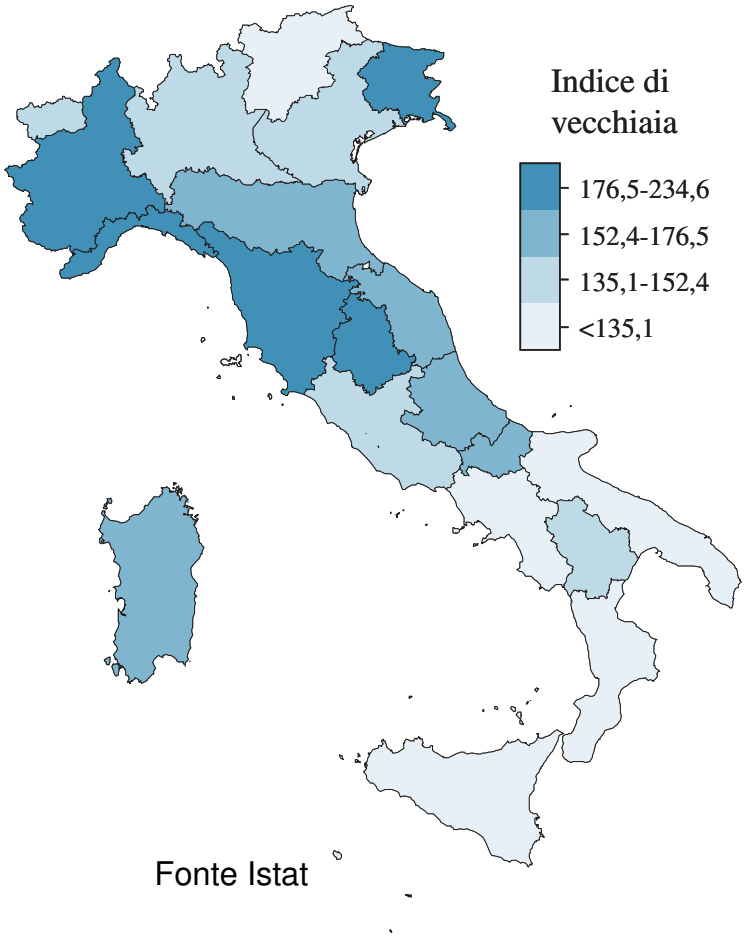
Fonte Istat

Grafici per serie territoriali

- si rappresentano spesso mediante i **cartogrammi**: le ripartizioni territoriali sono individuate sull'appropriata **cartina geografica**; le intensità corrispondenti (frequenze o quantità) vengono rappresentate tramite **colori** o **tratteggi diversi**, il cui significato è specificato in apposite **legende**

- una variante è il **cartodiagramma**: a ogni ripartizione territoriale viene associato un grafico relativo a una distribuzione o ai livelli di uno o più fenomeni

cartogramma relativo all'indice di vecchiaia per regione



Fonte Istat

Misure di tendenza centrale

Misure di tendenza centrale: Medie

- Le medie sono lo strumento con cui si **sintetizzano** i dati statistici.
- L'uso della media consente all'individuo di rappresentarsi mentalmente l'"**ordine di grandezza**" di un fenomeno, di effettuare **comparazioni** tra le manifestazioni di uno stesso fenomeno in tempi, luoghi o situazioni diverse, di comunicare ad altri tale informazione.

PROPRIETA' DI INTERNALITA': se a e b sono il minimo e il massimo dell'insieme dei numeri $x_1, x_2, \ldots, x_N$, la media è compresa tra queste due quantità: $a \leq m \leq b$

Medie che è possibile calcolare in relazione ai diversi tipi di carattere

		Indici di sintesi ed Operazioni		
		Moda	**Statistiche d'ordine** *(Mediana, Quartili, Decili, Percentili, Quantili)*	**Medie algebriche** (*media aritmetica, media armonica, media geometrica, media quadratica*)
Caratteri		=, ≠	> , <	+, -, *, /
Qualitativi	***sconnessi***	**si**	**no**	**no**
	ordinabili	**si**	**si**	**no**
Quantitativi		**si**	**si**	**si**

Moda

La **moda** di un collettivo, distribuito secondo un carattere di qualsiasi natura, è la modalità prevalente del carattere ossia quella **modalità a cui è associata la massima frequenza.**

ESEMPIO:

In quale tipologia di punti vendita ha acquistato i prodotti della nostra azienda?

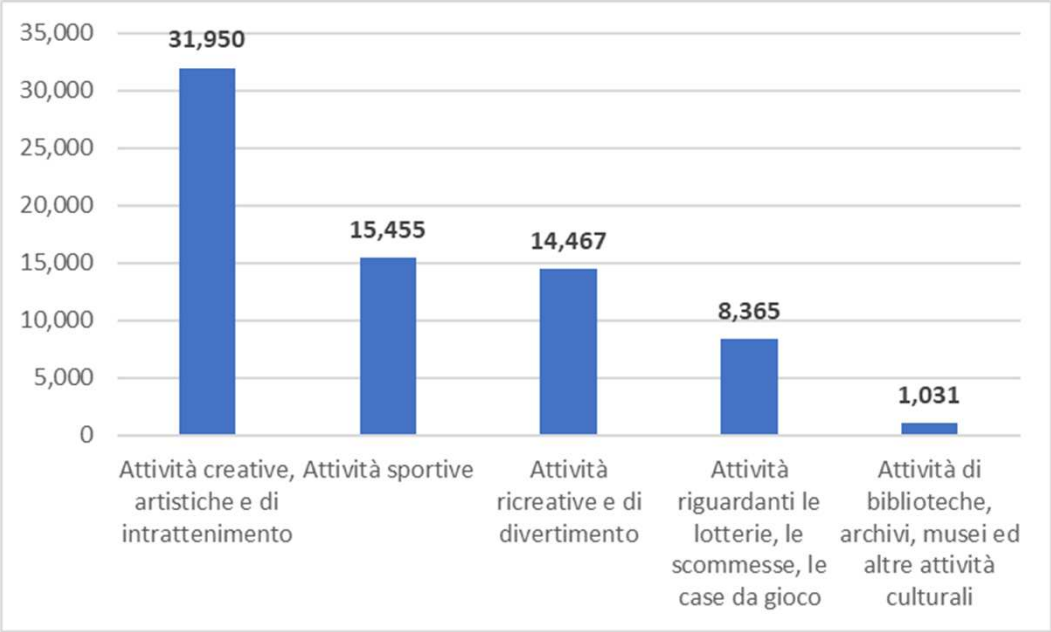
Tipologia punti vendita	ni	p (%)
On-line	454	24.4
Negozi del biologico	1227	65.9
Supermercati/Ipermercati	153	8.2
Altro	27	1.5
Totale	1861	100

Qui, la moda è la modalità "Negozi del biologico".

In un gruppo di clienti intervistati la maggior parte acquista i prodotti in negozi del biologico.

Moda: esempio

ATTIVITÀ ECONOMICHE	Numero
Attività creative, artistiche e di intrattenimento	31,950
Attività sportive	15,455
Attività ricreative e di divertimento	14,467
Attività riguardanti le lotterie, le scommesse, le case da gioco	8,365
Attività di biblioteche, archivi, musei ed altre attività culturali	1,031
Totale attività artistiche, sportive, di intrattenimento e divertimento	**71,268**



Fonte Istat

Moda: attività creative, artistiche e di intrattenimento

La Mediana

La **mediana** di una distribuzione ordinata di valori è la modalità che si trova a metà della distribuzione, in modo che il 50% dei valori della serie sia uguale o inferiore a esso e il restante 50% sia superiore.

Per calcolare la mediana **è necessario che la variabile sia quantitativa o qualitativa ordinata.**

Viene calcolata in modo diverso in base al numero di elementi della serie

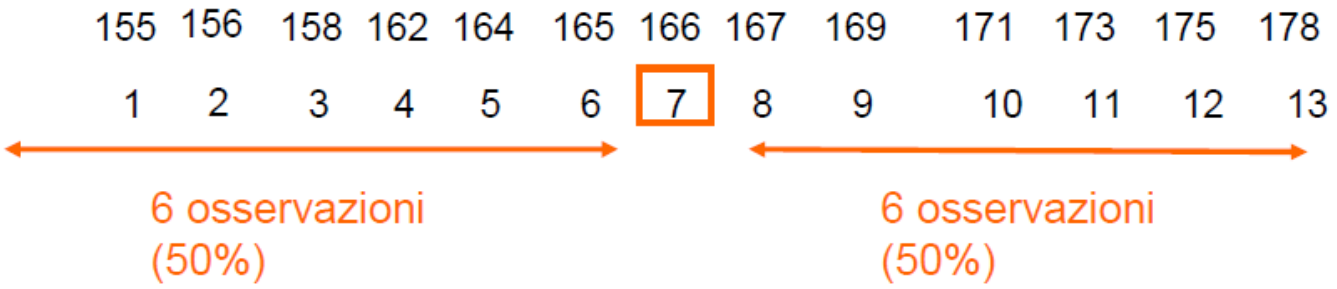
Se *n* è dispari, la mediana corrisponde alla modalità che occupa il posto $[(n+1)/2]_{mo}$ della sequenza ordinata

Se *n* è pari, la mediana è data dalla semisomma dei termini che occupano i posti $[n/2]_{mo}$ & $[(n/2) + 1]_{mo}$

La Mediana: esempio

Esempio: in un collettivo di 13 soggetti viene osservato il carattere Altezza (cm):
173 155 162 165 167 175 171 169 164 178 156 158 166

Ordiniamo in senso crescente le osservazioni, attribuendogli la posizione in graduatoria:



mediana = 166

n dispari → mediana = modalità di posto (n+1)/2

n pari → mediana = modalità intermedia fra quelle di posto n/2 e n/2+1

(ad esempio, se n=6, è la modalità centrale fra la 3° e la 4°)

Robustezza della mediana

La mediana non cambia o cambia di poco (è “robusta”) in presenza di alcuni dati molto estremi (ad es. con alcuni valori molto alti rispetto agli altri).
Vediamo per esempio che succede se nel collettivo precedente i due soggetti più alti sono ancora più alti:

La **mediana** non cambia poichè l’ordinamento delle prime n osservazioni non cambia.

Generalizzazione della mediana: quantili

La mediana separa la distribuzione in due parti, ognuna comprendente il 50% delle osservazioni.

- I quantili separano la distribuzione ad altre frazioni percentuali, ad esempio:

– Il 1 quartile (Q1) separa il primo 25% dal restante 75%
– Il 3 quartile (Q3) separa il primo 75% dal restante 25%

– Il 10 decile separa il primo 10% dal restante 90%
– Il 95°percentile è tale che solo il 5% ha un valore superiore a esso
– etc

Mediana e quartili: definizione operativa basata sulle frequenze percentuali cumulate

Carattere: Età

	x_i	n_i	N_i	P_i	
	41	3	3	**2.1%**	
	42	2	5	**3.6%**	
	43	6	11	**7.9%**	
	44	11	22	**15.7%**	
	45	8	30	**21.4%**	25%
q_1 →	**46**	17	47	**33.6%**	
	47	21	68	**48.6%**	50%
$m=q_2$ →	**48**	14	82	**58.6%**	
	49	17	99	**70.7%**	75%
q_3 →	**50**	15	114	**81.4%**	
	51	10	124	**88.6%**	
	52	10	134	**95.7%**	
	53	5	139	**99.3%**	
	54	1	140	**100.0%**	
	Totale	**140**			

Interpretazione

Primo quartile: Il 25% degli intervistati del collettivo analizzato ha un età inferiore o uguale a 46 anni.

***Mediana:** Il 50% ha un età inferiore o uguale a 48 anni.*

Terzo quartile : Il 75% ha un età inferiore o uguale a 50 anni.

Decili: definizione operativa basata sulle frequenze percentuali cumulate

Carattere: Età

	x_i	n_i	N_i	P_i	
	4 1	3	3	**2.1%**	
	42	2	5	**3.6%**	
	43	6	11	**7.9%**	10%
d_1 →	44	11	22	**15.7%**	
d_2 →	45	8	30	**21.4%**	20%
d_3 →	46	17	47	**33.6%**	30%
d_4 →	47	21	68	**48.6%**	40%
$m=d_5$ →	48	14	82	**58.6%**	
$d_6;d_7$ →	49	17	99	**70.7%**	60%;70%
d_8 →	50	15	114	**81.4%**	80%
	5 1	10	124	**88.6%**	
d_9 →	52	10	134	**95.7%**	90%
	53	5	139	**99.3%**	
	54	1	140	**100.0%**	
	Totale	**140**			

Interpretazione

Primo decile: Il 10% degli intervistati ha un età inferiore o uguale a 44 anni.

….

Nono decile : Il 90% ha un età inferiore o uguale a 52 anni.

La media aritmetica

La media aritmetica è una delle sintesi statistiche più importanti

La media è l'ammontare totale del carattere (somma di tutte le osservazioni) ripartito in parti uguali

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

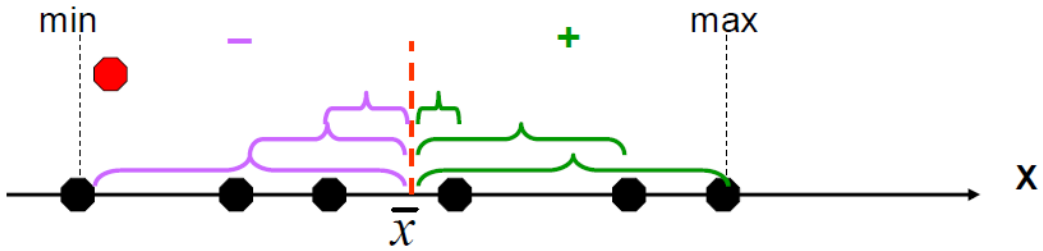
$$\mu = \frac{\sum x_i}{n} \Leftrightarrow \sum x_i = n\bar{x}$$

La media, sostituita a ciascuna osservazione, ricostituisce la somma totale delle modalità.

Voto	
26	24
18	24
28	24
72	72

Media = 72 / 3 = 24

Principali proprietà della media



La media è interna al range, ossia, è sempre compresa fra l'osservazione più bassa e quella più alta.

$$\min(x_i) \leq \bar{x} \leq \max(x_i)$$

PRINCIPIO DI INTERNALITA' DEL CHISINI

La somma degli scarti dalla media è nulla: ossia, la media si colloca "al centro" dei valori osservati, bilanciando scarti positivi e scarti negativi.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

PRIMA PROPRIETA' MEDIA ARITMETICA

Se misuriamo la distanza delle osservazioni da un valore C secondo questa misura globale, essa assume il minimo se C è la media aritmetica: ossia, la media aritmetica è il punto "globalmente meno distante" dalle osservazioni.

$$\text{dist} = \sqrt{\sum_{i=1}^{n}(x_i - C)^2}$$

SECONDA PROPRIETA' MEDIA ARITMETICA

Media aritmetica per le distribuzioni di frequenze

$$\mu = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \ldots + x_k \cdot n_k}{N} = \frac{1}{N}\sum_{i=1}^{k} x_i \cdot n_i$$

$$= x_1 \cdot f_1 + x_2 \cdot f_2 + \cdots + x_k \cdot f_k = \sum_{i=1}^{k} x_i \cdot f_i$$

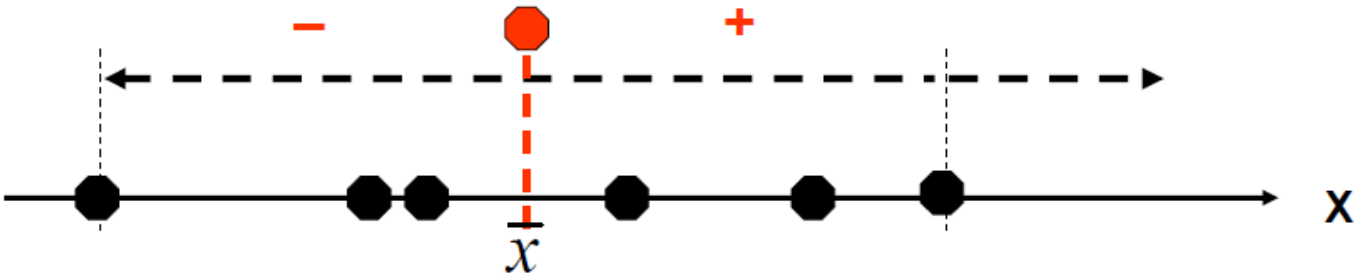
Numero di accessi al sito aziendale

x_i	n_i	$x_i * n_i$
1	1	1
2	2	4
3	3	9
4	3	12
5	1	5
6	2	12
7	1	7
8	2	16
9	3	27
10	1	10
totale	19	103

$$\mu = \frac{1 \cdot 1 + 2 \cdot 2 + \cdots + 10 \cdot 1}{140}$$

$$= \frac{103}{19} = 5.42$$

Limitazioni della media aritmetica



La media è influenzata dai valori molto alti e dai valori molto bassi .

Riprendiamo esempio ‘altezze’

173 155 162 165 167 ~~175~~ 171 169 164 ~~178~~ 156 158 166

$\mu = 166.1$

189 210

$\mu = 169.6$