Statistica della Formazione

Slides 3

A.A. 2020-2021

Docente: ANNA LINA SARRA

Modulo 1: elementi di statistica descrittiva

- **Le misure di tendenza centrale**

"A statistician can have his head in an oven and his feet in ice, and he will say that on the average he feels fine."

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.**

Quando il carattere è quantitativo e le modalità sono **raggruppate in classi**, si parla di **classe modale** con riferimento alla classe avente la **densità di frequenza più elevata**.

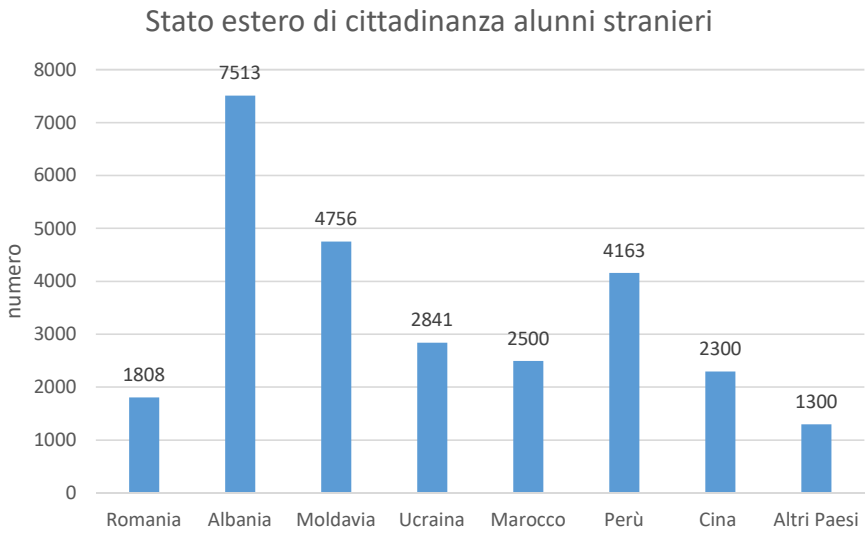
Si possono avere distribuzioni *unimodali, bimodali, multimodali e zeromodali*

Moda esempio (1)

Carattere qualitativo sconnesso

Stato estero di cittadinanza alunni stranieri

Cittadinanza	numero
Romania	1808
Albania	7513
Moldavia	4756
Ucraina	2841
Marocco	2500
Perù	4163
Cina	2300
Altri Paesi	1300
Totale	**27181**



Distribuzione unimodale

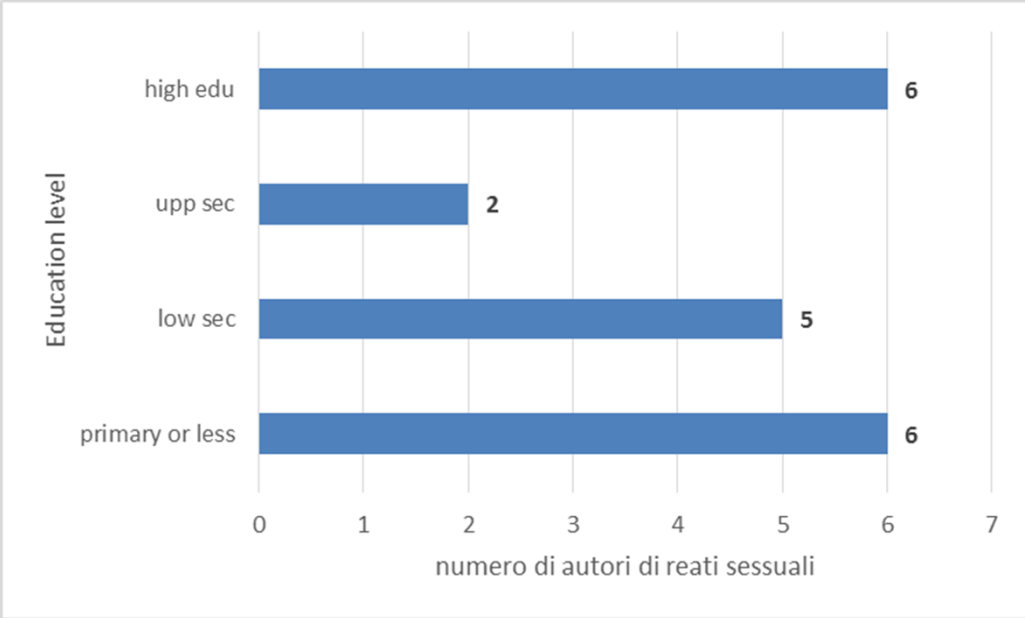
Moda: Albania

Frequenza massima: *7513*

Moda esempio (2)

Carattere qualitativo ordinabile

Education level	**frequenza assoluta**
primary or less	6
low sec	5
upp sec	2
high edu	6
totale	19



<u>Distribuzione bimodale</u>

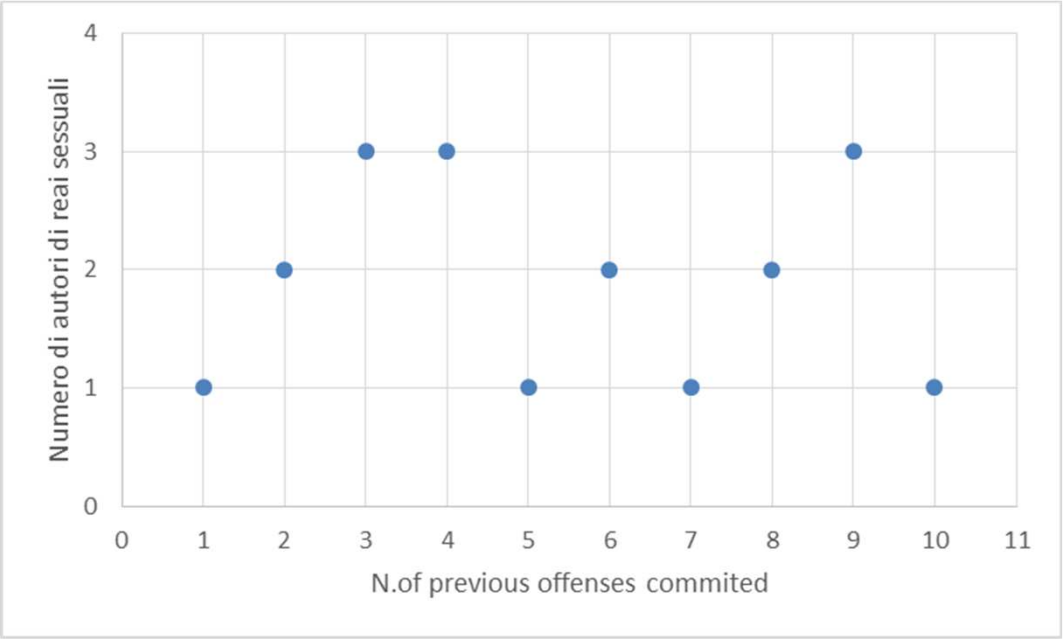
Mode: primary or less; high education

Frequenza massima: *6*

Moda esempio (3)

Carattere quantitativo discreto

N.of previous offenses commited	**frequenza assoluta**
1	1
2	2
3	3
4	3
5	1
6	2
7	1
8	2
9	3
10	1
totale	19



<u>Distribuzione trimodale</u>

Moda: 3; 4; 9

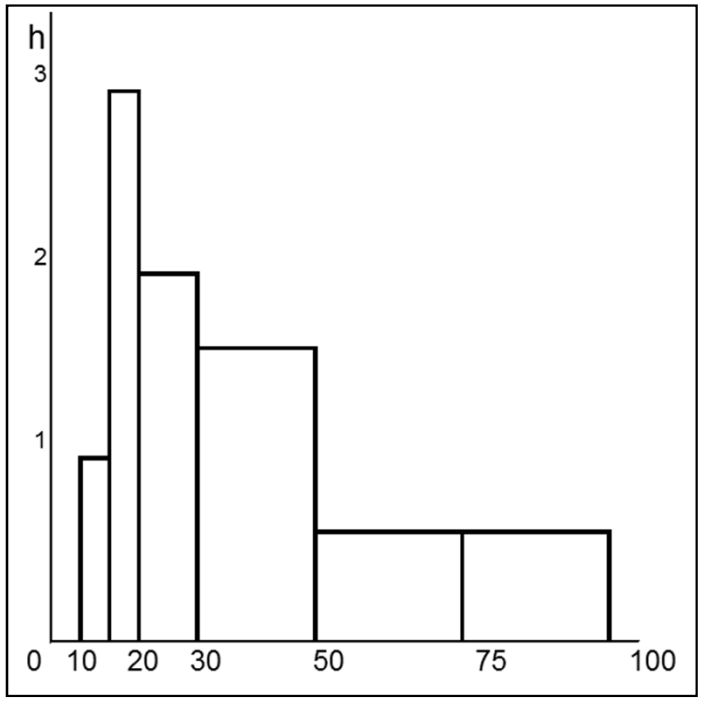
Frequenza massima: *3*

Carattere quantitativo continuo in classi: Classe modale

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** ← Densità Massima
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**		

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

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



Per determinare la moda è necessario calcolare la densità di frequenza h_i, data dal rapporto fra frequenza assoluta n_i e ampiezza di classe d_i.

La classe modale è la classe con maggiore densità di frequenza

Distribuzione unimodale

Classe modale: 15|--20

Densità di frequenza massima: *3*

Mediana

La **mediana** è quella modalità che suddivide ogni **distribuzione ordinata** in due distribuzioni aventi ciascuna una numerosità (o una quantità) che è il 50% della numerosità (o della quantità) della distribuzione totale.

Si può calcolare per i caratteri qualitativi ordinabili e quantitativi

Sia $x_1, x_2, \ldots, x_N$, una distribuzione statistica disaggregata.

Sia $y_1, y_2, \ldots, y_N$, con $y_1 \leq y_2 \leq \ldots, \leq y_N$, la corrispondente distribuzione dei termini ordinati.

Mediana per N dispari

➢ N dispari : la mediana è la modalità che nella distribuzione ordinata occupa il posto $\frac{N+1}{2}$

Distribuzione disaggregata dell'età di 9 sex offenders

$x_1=21$ $x_2=18$ $x_3=28$ $x_4=27$ $x_5=30$ $x_6=28$ $x_7=30$ $x_8=25$ $x_9=22$

Distribuzione ordinata

$y_1=18$ $y_2=21$ $y_3=22$ $y_4=25$ $y_5=27$ $y_6=28$ $y_7=28$ $y_8=30$ $y_9=30$

Posizione occupata dalla mediana: $\frac{N+1}{2}=\frac{9+1}{2}=5$

Mediana: $y_{\frac{N+1}{2}}=y_5=27$

Mediana per N pari

N pari : si hanno due modalità mediane che nella **distribuzione ordinata** occupano rispettivamente i posti $\frac{N}{2}$ *e* $\frac{N}{2}+1$

Distribuzione disaggregata dell'età di 10 sex offenders

x_1=21 x_2=18 x_3=28 x_4=27 x_5=30 x_6=28 x_7=30 x_8=25 x_9=22 x_{10}=28

Distribuzione ordinata

y_1=18 y_2=21 y_3=22 y_4=25 y_5=27 y_6=28 y_7=28 y_8=28 y_9=30 x_{10}=30

Posizioni occupate dalle mediane $\frac{N}{2}=\frac{10}{2}=5; \frac{N}{2}+1=5+1=6$

Mediane $y_{\frac{N}{2}}=y_5=27;\ y_{\frac{N}{2}+1}=y_6=28$

Mediana $\frac{y_{\frac{N}{2}}+y_{\frac{N}{2}+1}}{2}=\frac{27+28}{2}=27.5$

Se ***N* è pari**, **e il carattere è quantitativo** si può assumere come mediana la media aritmetica dei termini che occupano le due posizioni centrali della graduatoria dei termini ordinati, ossia le posizioni *N*/2 *e* *N*/2 + 1.

Esempio: Mediana per caratteri qualitativi ordinabili - N dispari

Distribuzione disaggregata:

Sufficiente, Pessimo, Insufficiente, Ottimo, Sufficiente, Insufficiente, Ottimo

Distribuzione ordinata:

Pessimo, Insufficiente, Insufficiente, Sufficiente, Sufficiente, Ottimo, Ottimo

Posizione occupata dalla mediana

$$\frac{N+1}{2}=\frac{7+1}{2}=4$$

Mediana:

$$y_{\frac{N+1}{2}} = y_4 = \text{Sufficiente}$$

Esempio: Mediana per caratteri qualitativi ordinabili - N pari

Distribuzione disaggregata:

Licenza media, Diploma, Diploma, Laurea, Licenza media, Licenza elementare

Distribuzione ordinata:

Licenza elementare, Licenza media, Licenza media, Diploma, Diploma, Laurea,

Posizioni occupate dalle mediana

$$\frac{\mathbf{N}}{\mathbf{2}}=\frac{6}{2}=\mathbf{3};\ \frac{\mathbf{N}}{\mathbf{2}}+1=\frac{6}{2}+1=\mathbf{4};$$

Mediane

$$\mathbf{y}_{\frac{\mathbf{N+1}}{\mathbf{2}}}=\mathbf{y}_{\mathbf{3}}=\text{Licenza media};$$

$$\mathbf{y}_{\frac{\mathbf{N}}{\mathbf{2}}+\mathbf{1}}=\mathbf{y}_{\mathbf{4}}=\text{Diploma}$$

Carattere quantitativo: La somma degli scarti in valore assoluto dei valori $x_1, x_2, \ldots, x_N$ da una costante c è minima quando c è uguale alla mediana

Me=5

1	**3**	**5**	**6**	**11**

$$\sum_{i=1}^{5} |x_i - Me| = |1-5| + |3-5| + |5-5| + |6-5| + |11-5| =$$

$$= 4 + 2 + 0 + 1 + 6 = 13$$

Se al posto di 5 (che corrisponde alla mediana dei valori) metto un qualsiasi altro valore, questa somma sarà sempre >13

Quartili

Sia $x_1, x_2, \ldots, x_N$ una distribuzione disaggregata.

Sia $y_1, y_2, \ldots, y_N$ la corrispondente distribuzione di termini ordinati, con $y_1 \leq y_2 \leq \ldots \leq y_N$.

- Il **primo quartile**, q_1, è la quantità che non è superata da un quarto (25%) dei termini ordinati della distribuzione
- Il **secondo quartile**, q_2, è la quantità che non è superata dalla metà (50%) dei termini ordinati.
- Il **terzo quartile**, q_3, è la quantità che non è superata dai tre quarti (75%) dei termini ordinati della distribuzione.

N.B.: Il secondo quartile coincide con la mediana

Decili

In termini discorsivi, i decili si possono definire come medie di posizione tali che:

Il primo decile: è la quantità che non è superata da un decimo (10%) dei termini ordinati

Il secondo decile: è la quantità che non è superata da due decimi (20%) dei termini ordinati

…

N.B.: i decili sono 9.

N.B.: Il quinto decile coincide con la mediana

Mediana e quartili: definizione operativa basata sulle frequenze percentuali cumulate

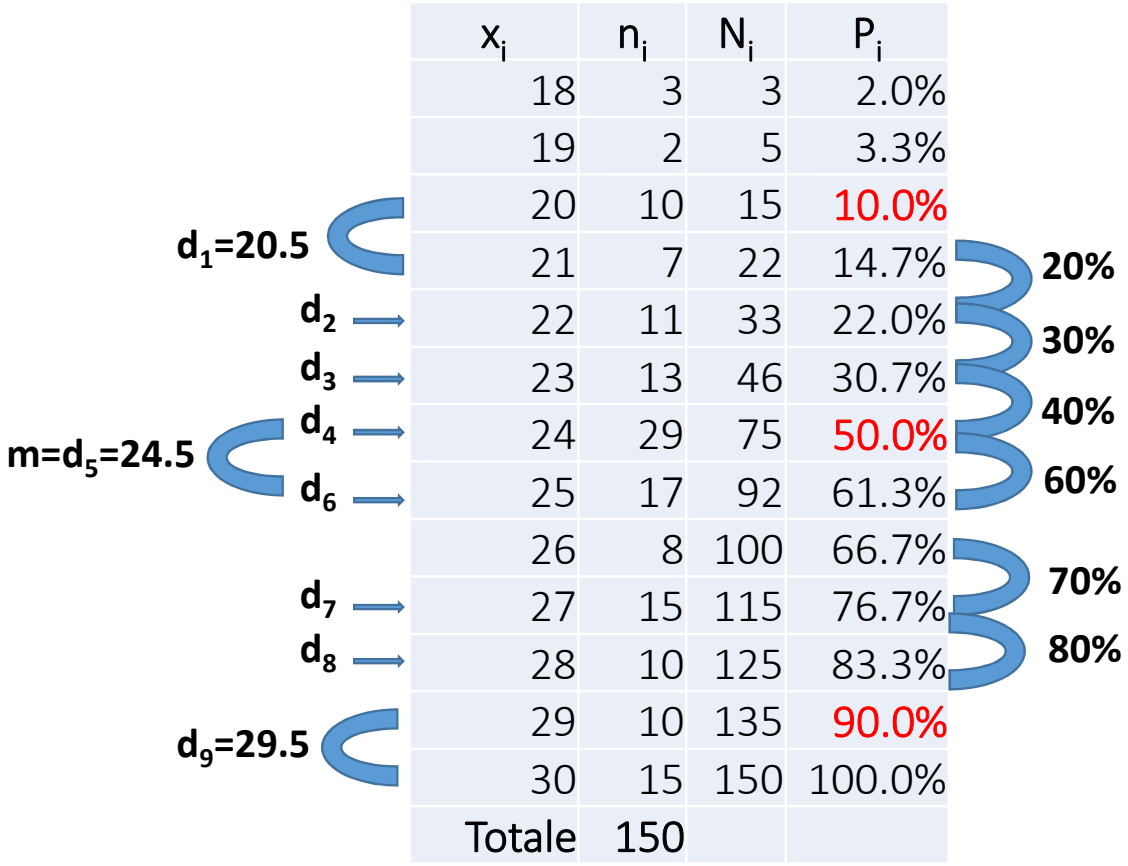
	x_i	n_i	N_i	P_i	
	41	3	3	**2.1%**	
	42	2	5	**3.6%**	
	43	6	11	**7.9%**	
	44	11	22	**15.7%**	
	45	8	30	**21.4%**	25%
q_1 →	**46**	17	47	**33.6%**	
	47	21	68	**48.6%**	50%
$m=q_2$ →	**48**	14	82	**58.6%**	
	49	17	99	**70.7%**	75%
q_3 →	**50**	15	114	**81.4%**	
	5 1	10	124	**88.6%**	
	52	10	134	**95.7%**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

Decili: definizione operativa basata sulle frequenze percentuali cumulate

	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%**	20%
d_2	45	8	30	**21.4%**	30%
d_3	46	17	47	**33.6%**	40%
d_4	47	21	68	**48.6%**	
$m=d_5$	48	14	82	**58.6%**	
$d_6;d_7$	49	17	99	**70.7%**	;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**			

Esempio

Distribuzione di frequenze dei voti di 150 studenti all'esame di statistica



x_i	n_i	N_i	P_i
18	3	3	2.0%
19	2	5	3.3%
20	10	15	10.0%
21	7	22	14.7%
22	11	33	22.0%
23	13	46	30.7%
24	29	75	50.0%
25	17	92	61.3%
26	8	100	66.7%
27	15	115	76.7%
28	10	125	83.3%
29	10	135	90.0%
30	15	150	100.0%
Totale	150		

Media aritmetica

La media aritmetica è quel valore di sintesi che sostituito alle modalità lascia inalterata

la loro somma

- Insieme alle percentuali e ai grafici, la media aritmetica è lo strumento statistico più largamente utilizzato
- La media aritmetica di una **distribuzione statistica disaggregata** si calcola come la somma dei termini x_1, x_2, …, x_N divisa per N

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

Media aritmetica per una distribuzione disaggregata: calcolo

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

Serie storica degli studenti stranieri

2009/2010	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015
673800	710263	755939	786630	803053	814208

La media annua degli studenti stranieri

$$\mu = \frac{673800+710263+755939+786630+803053+814208}{6} = 739036{,}1$$

Media aritmetica per le distribuzioni di frequenze

Modalità singole

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

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

Modalità raggruppate in classi

$$\mu = \frac{\overline{x}_1 \cdot n_1 + \overline{x}_2 \cdot n_2 + \cdots + \overline{x}_k \cdot n_k}{N} = \frac{1}{N}\sum_{i=1}^{k} \overline{x}_i \cdot n_i$$

dove $\overline{x}_i = \frac{c_{i-1} + c_i}{2}$ è il **valore centrale** della generica classe.

Media aritmetica per una distribuzione di frequenze a modalità singole: calcolo

$$\mu = \frac{1}{N}\sum_{i=1}^{k} x_i \cdot n_i$$

N.of previous offenses commited	**frequenza assoluta**	*modalità* frequenza*
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

- La media aritmetica della distribuzione è data da:

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

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

Media aritmetica per una distribuzione di frequenze a modalità raggruppate in classi: calcolo

$$\mu = \frac{1}{N}\sum_{i=1}^{k} \overline{x}_i \cdot n_i$$

Distribuzione di frequenze secondo l'età:

Classe	Valore centrale $\overline{x}_i$	n_i	$\overline{x}_i \cdot n_i$
19-21	20.0	31	620
21-24	22.5	45	1012.5
24-27	25.5	5	127.5
27-30	28.5	1	28.5
Totale		82	1788.5

- La media aritmetica della distribuzione è data da:

$$\mu = \frac{20.0 \cdot 31 + 22.5 \cdot 45 + 25.5 \cdot 5 + 28.5 \cdot 1}{82}$$

$$= \frac{1788.5}{82} = 21.8$$

La somma algebrica degli scarti dalla media aritmetica è uguale a zero

1	6	4	1

$$\sum_{i=1}^{4}(x_i-\bar{x})=(1-3)+(6-3)+(4-3)+(1-3)=$$

$$-2+3+1-2=0$$

Voto esame di matematica (x_i)	**n_i**	$x_i n_i$	(x_i-x)	$(x_i-x)ni$
18	5	90	18 - 24,76 = -6,76	-33,8
20	7	140	20 - 24,76 = -4,76	-33,32
23	8	184	23 - 24,76 = -1,76	-14,08
26	8	208	26 - 24,76 = 1,24	9,92
27	10	270	27 - 24,76 = 2,24	22,4
28	7	196	28 - 24,76 = 3,24	22,68
30	5	150	30 - 24,76 = 5,24	26,2
Totale	**50**	1238		**0**

$$\sum_{i=1}^{7}(x_i-\bar{x})n_i=0$$

La somma degli scarti dalla media aritmetica al quadrato è un minimo

1	**6**	**4**	**1**

$\mu=3$

$$\sum_{i=1}^{4}(x_i-\mu)^2=(1-3)^2+(6-3)^2+(4-3)^2+(1-3)^2=4+9+1+4=18$$

Se al posto di 3, che corrisponde alla media, metto un qualsiasi altro valore, questa somma sarà sempre >18

- **Proprietà di linearità :** Se ogni singolo termine della distribuzione, x_i , viene sottoposto alla trasformazione

$$a + bx_i,$$

con a costante qualsiasi e $b \neq 0$, la nuova media è legata a quella originaria dalla medesima trasformazione $a + b\mu$

Voti ottenuti su 4 esami	18	22	26	28	μ=23.5

Se ad ogni esame lo studente avesse ottenuto 2 punti in più la media sarebbe 25.5

	20	24	28	30	μ=25.5

Guadagno su 4 giorni	50	25	75	50	μ=50

Se ogni giorno il guadagno fosse il doppio la media sarebbe 100

	100	50	150	100	μ=100

- **Proprietà di associatività** Se un collettivo statistico di N unità viene suddiviso in L sottoinsiemi disgiunti aventi numerosità $N^{(1)}, N^{(2)}, \ldots, N^{(L)}$ e medie $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(L)}$, allora la media aritmetica del collettivo può essere così calcolata

$$\mu = \frac{1}{N}(\mu^{(1)} \cdot N^{(1)} + \mu^{(2)} \cdot N^{(2)} + \cdots + \mu^{(L)} \cdot N^{(L)}) = \frac{1}{N}\sum_{l=1}^{N} \mu^{(l)} \cdot N^{(l)}$$

Durata media in minuti delle interrogazioni: 12.8, 13.0, 13.4, 13.4 , 13.6, 13.5, 13.6, 13.7

$$\mu = \frac{12.8 + 13.0 + 13.4 + 13.4 + 13.6 + 13.5 + 13.6 + 13.7}{8} = 13.375 .$$

se suddividiamo la distribuzione data nelle due seguenti:

A. 12.8, 13.0, 13.4, 13.4 , 13.6 μ_A= 13.240
B. 13.5, 13.6, 13.7 μ_B= 13.600

la media aritmetica della distribuzione **può essere ottenuta come**

$$\mu = \frac{13.240 \cdot 5 + 13.600 \cdot 3}{8} = 13.375 .$$

Medie aritmetica ponderata

Siano x_1, x_2, …, x_k le osservazioni e w_1, w_2, …, w_k i rispettivi pesi. Allora, la **media aritmetica ponderata** di x_1, x_2, …, x_k è data dal rapporto tra la somma delle osservazioni moltiplicate per i rispettivi pesi e la somma dei pesi

$$\mu = \frac{x_1 \cdot w_1 + x_2 \cdot w_2 + \cdots + x_k \cdot w_k}{w_1 + w_2 + \cdots + w_k} = \frac{\sum_{i=1}^{k} x_i w_i}{\sum_{i=1}^{k} w_i}$$

Media aritmetica ponderata: esempio

$$\mu = \frac{\sum_{i=1}^{k} x_i w_i}{\sum_{i=1}^{k} w_i}$$

voto esame	**CFU**	*modalità* peso*
x_i	**w_i**	**$x_i * w_i$**
24	9	216
26	6	156
30	**12**	**360**
30	**6**	**180**
22	9	198
18	12	216
18	6	108
totale	60	1434

$$\mu = \frac{\sum_{i=1}^{k} x_i w_i}{\sum_{i=1}^{k} w_i} = \frac{1434}{60} = 23.9$$

Domande (1)

- La distribuzione delle altezze degli adulti in Italia è unimodale?

- Data la distribuzione disaggregata x={21,16, 21,21,23,23,17}, calcolare
  - a. x1+x3
  - b. la media della distribuzione disaggregata
  - c. la mediana della distribuzione disaggregata
  - d. la moda della distribuzione disaggregata

Domande (2)

Tasso di abbandono alla fine del primo bienno delle scuole secondarie superiori nelle regioni italiane (anno di riferimento 2012)

Calcolare i quartili della del tasso di abbandono scolastico per il periodo 2012 e interpretare i risultati.

REGIONI	%
Piemonte	6,9
Valle d'Aosta/Vallée d'Aoste	10,8
Lombardia	6,6
Trentino-Alto Adige/Südtirol	3,2
Veneto	4,3
Friuli-Venezia Giulia	4,6
Liguria	7,5
Emilia-Romagna	6,6
Toscana	7,4
Umbria	4,7
Marche	4,8
Lazio	5,8
Abruzzo	6,1
Molise	5,7
Campania	9,3
Puglia	5,3
Basilicata	5,3
Calabria	5,3
Sicilia	9,2
Sardegna	10,4
Bolzano/Bozen	2,9
Trento	3,6

** Valori %, anno di riferimento 2012*

Domande (3)

	Drug trafficking recorded by the police, 2006–12							
	Number							
	2006	2007	2008	2009	2010	2011	2012	*totale*
Germany	64,865	64,093	55,905	50,965	49,622	50,791	47,667	***383,908***
France	5,792	5,797	6,128	6,007	5,869	5,928	4,821	***40,342***
Italy	32,306	34,439	34,082	34,101	32,761	34,034	33,852	***235,575***
Totale	**102,963**	**104,329**	**96,115**	**91,073**	**88,252**	**90,753**	**86,340**	***659,825***

- Calcolare ed interpretare le medie aritmetiche per i tre Stati e per ogni anno.

Domande (4)

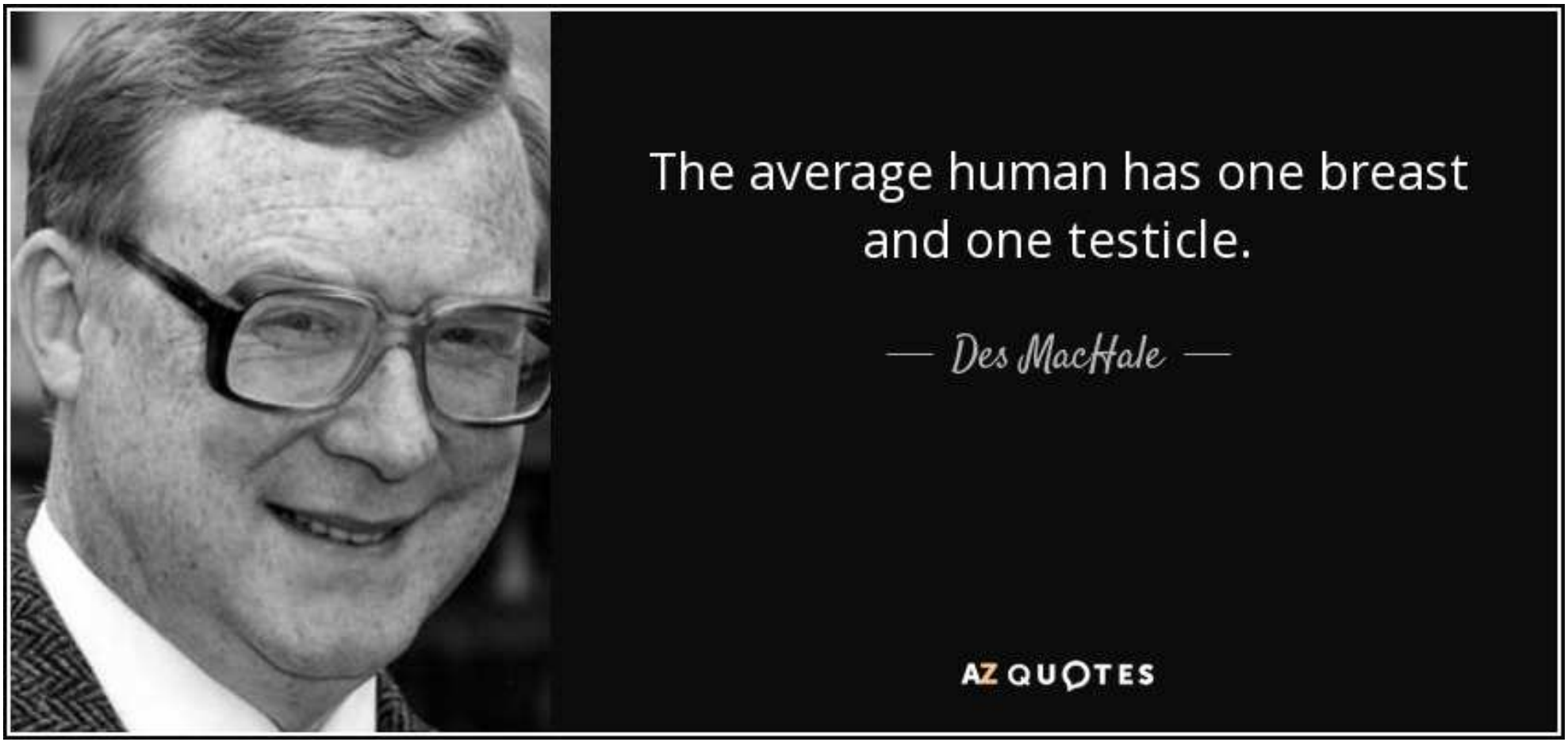
- A seguito di un controllo della guardia di finanza, sono stati rilevati casi di evasione fiscale in 10 aziende.
- Di queste, 1 azienda ha ricevuto una cartella di EQUITALIA pari a 300.000 euro e le altre 9 hanno ricevuto cartelle di 10.000 euro.
    - Qual è l'indice che consente di sintetizzare meglio l'importo delle cartelle di EQUITALIA, la media o la mediana?

Domande (5)

- I voti riportati da uno studente in 5 esami sono 18, 25, 30, 26, 27. Che voto deve prendere al prossimo esame per avere
  - una media aritmetica dei voti uguale a 26?
  - un voto mediano uguale a 26?
  - un voto modale uguale a 26?

- La media aritmetica di 5 valori è 6. La media aritmetica di 3 di questi 5 valori è 8. Qual è la media dei restanti due valori?

- La media di 121 numeri è 59. Se ogni numero è moltiplicato per 4, quale sarà la media della nuova distribuzione? E se ad ogni numero sommiamo 10?


The average human has one breast and one testicle.
— Des MacHale —
AZ QUOTES